

A Thesis for the Degree of Ph.D. in Science

Computational pipelines for assembly, analysis, and evaluation of
genome sequences

January 2018

Graduate School of Science and Technology
Keio University

Vasanthan Jayakumar

Contents

Chapter 1 Introduction	1
1.1 DNA sequencing	1
1.1.1 First-generation sequencing (FGS)	2
1.1.2 Second-generation sequencing (SGS)	3
1.1.3 Third-generation sequencing (TGS)	4
1.2 Genome assembly	5
1.2.1 OLC approach	6
1.2.2 de Bruijn graphs	7
1.2.3 String graphs	7
1.2.4 Genome assembly in the FGS era	8
1.2.5 Genome assembly in the SGS era	8
1.2.6 Genome assembly in the TGS era	9
1.3 Factors affecting genome assembly	9
1.3.1 Sequence coverage	9
1.3.2 Repetitive sequences	9
1.3.3 Sequencing errors	10
1.3.4 Ploidy	10
1.4 Scaffolding	11
1.5 Assembly metrics: N50 and L50	11
1.6 <i>De novo</i> assembly of a plant genome	12
1.7 Evaluation of long-read assembly tools	13
Chapter 2 Construction of computational pipelines for <i>de novo</i> assembly	16
2.1 Parameters and other aspects in the pipeline	16
2.1.1 Error correction and polishing	16
2.1.2 <i>De novo</i> assembly	18
2.1.3 Scaffolding and gap-filling	18
2.1.4 Computational resources	19
2.1.5 Assembly validation	19
2.1.6 Gene prediction	19
2.1.7 Repeat prediction	20
Chapter 3 Genome sequence and analysis of the Japanese morning glory <i>Ipomoea nil</i>	21
3.1 Background	21
3.2 Results	26
3.2.1 DNA sequencing and genome assembly	26
3.2.2 Mis-assembly detection and pseudo-molecule construction	28
3.2.3 Assembly validation	32
3.2.4 Repeat analysis and identification of <i>Tpn1</i> transposons	34

3.2.5 Gene prediction and functional annotation	37
3.2.6 Genome evolution	37
3.3 Discussion	40
3.4 Methods	42
3.4.1 Plant materials and sequencing	42
3.4.2 Genome assembly	43
3.4.3 Linkage map construction and pseudo-chromosome assignment	44
3.4.4 Mis-assembly elimination and assembly validation	45
3.4.5 Repeat analysis and gene prediction	46
3.4.6 Comparative analysis	47
3.4.7 Data availability	48
Chapter 4 Comprehensive evaluation of non-hybrid genome assembly tools for third generation PacBio long-read sequence data	50
4.1 Background	50
4.2 Materials and methods	52
4.2.1 Long-read assembly pipelines	52
4.2.1.1 Hierarchical Genome Assembly Process	53
4.2.1.2 PBcR	54
4.2.1.3 Canu	54
4.2.1.4 FALCON	54
4.2.1.5 HINGE	55
4.2.1.6 Miniasm	55
4.2.1.7 SMARTdenovo	55
4.2.1.8 ABruijn	56
4.2.1.9 Wtdbg	56
4.2.1.10 Mapping, Error Correction and <i>de novo</i> Assembly Tool	56
4.2.2 Datasets for evaluation	57
4.2.3 Criteria for evaluation	58
4.3 Results	61
4.3.1 Contiguity	61
4.3.1.1 <i>Escherichia coli</i>	61
4.3.1.2 <i>Plasmodium falciparum</i>	61
4.3.1.3 <i>Caenorhabditis elegans</i>	61
4.3.1.4 <i>Ipomoea nil</i>	61
4.3.2 Completeness	62
4.3.2.1 <i>Escherichia coli</i>	62
4.3.2.2 <i>Plasmodium falciparum</i>	67
4.3.2.3 <i>Caenorhabditis elegans</i>	67
4.3.2.4 <i>Ipomoea nil</i>	67
4.3.3 Correctness	70
4.3.3.1 <i>Escherichia coli</i>	70

4.3.3.2 <i>Plasmodium falciparum</i>	70
4.3.3.3 <i>Caenorhabditis elegans</i>	70
4.3.3.4 <i>Ipomoea nil</i>	70
4.3.4 Circularity and overlapping fragmented contigs	71
4.3.5 Resource usage	74
4.3.5.1 <i>Escherichia coli</i>	74
4.3.5.2 <i>Plasmodium falciparum</i>	75
4.3.5.3 <i>Caenorhabditis elegans</i>	75
4.3.5.4 <i>Ipomoea nil</i>	75
4.3.6 Ranking	76
4.3.6.1 <i>Escherichia coli</i>	76
4.3.6.2 <i>Plasmodium falciparum</i>	76
4.3.6.3 <i>Caenorhabditis elegans</i>	76
4.3.6.4 <i>Ipomoea nil</i>	76
4.3.6.5 Mean ranking of the three eukaryotic assemblies	77
4.4 Discussion	78
Chapter 5 Conclusion and future work	82
Acknowledgements	85
References	86
Appendix A - List of publications	93
Appendix B - Supplementary of chapter 4	94

Abbreviations

DNA,	Deoxyribonucleic acid
RNA,	Ribonucleic acid
cDNA,	Complementary DNA
rDNA,	Ribosomal DNA
A,	Adenine
T,	Thymine
G,	Guanine
C,	Cytosine
bp,	Base pairs
kb,	Kilo-base pairs
Mb,	Mega-base pairs
Gb,	Giga-base pairs
FGS,	First Generation Sequencing
SGS,	Second Generation Sequencing
TGS,	Third Generation Sequencing
SMS,	Single Molecule Sequencing
HGP,	Human Genome Project
PE,	Paired-End
MP,	Mate-Pair
OLC,	Overlap-Layout-Consensus
FM,	Ferragina–Manzini
PB,	Pacific Biosciences
BLASR,	Basic Local Alignment with Successive Refinement
pbdagcon,	Pacific Biosciences Directed Acyclic Graph Consensus
ChIP,	Chromatin ImmunoPrecipitation

GAGE,	Genome Assembly Gold-standard Evaluations
NCBI,	National Center for Biotechnology Information
DDBJ,	DNA Data Bank of Japan
EMBL,	European Molecular Biology Laboratory
TTOL,	The Timescale Of Life
HGAP,	Hierarchical Genome Assembly Process
MECAT,	Mapping, Error Correction and <i>de novo</i> Assembly Tool
PBcR,	PacBio Corrected Reads
RSS,	Resident Set Size
CEG,	Core Eukaryotic Genes
CEGMA,	Core Eukaryotic Genes Mapping Approach
BUSCO,	Benchmarking Universal Single-Copy Orthologs
SRR,	Subterminal repetitive region
TIR,	Terminal Inverted Repeats
EST,	Expressed Sequence Tag
SSR,	Simple Sequence Repeat
SMRT,	Single Molecule, Real-Time
TKS,	Tokyo Kokei Standard
In-del,	Insertions-deletions
SNP,	Single Nucleotide Polymorphism
RAD-seq,	Restriction site Associated DNA Sequencing
BAC,	Bacterial Artificial Chromosome
NOR,	Nucleolar Organizer Region
LTR,	Long Terminal Repeat
TSD,	Target Site Duplication
UTR,	UnTranslated Region
BLAST,	Basic Local Alignment Search Tool

RT-PCR,	Reverse Transcriptase Polymerase Chain Reaction
LG,	Linkage Map
MCL,	Markov CLustering algorithm
MYA,	Million Years Ago
BEAST,	Bayesian Evolutionary Analysis Sampling Trees
WGD,	Whole Genome Duplication
GO,	Gene Ontology
GATK,	Genome Analysis ToolKit
BWA,	Burrows Wheeler Alignment
DP,	Depth
QD,	Quality by Depth
FS,	Fisher Strand
MQ,	Mapping Quality
BESST,	Bias Estimating Stepwise Scaffolding Tool
LOD,	Logarithm Of Odds
BLAT,	BLAST Like Alignment Tool
CDS,	Coding DNA Sequence
PAML,	Phylogenetic Analysis by Maximum Likelihood
DRA,	DDBJ sequence Read Archive
MHAP,	MinHash Alignment Process
tf-idf,	Term Frequency, Inverse Document Frequency
DDF,	Distance Difference Factor
OM,	Optical Mapping
LR,	Linked Reads
LM,	Linkage Map
RH,	Radiation Hybrid

Chapter 1

Introduction

Genome is the genetic material of an organism, which contains the instructions necessary for the proper functioning of a cell. The instructions are coded in the form of DNA constituted by four nucleotide base pairs: adenine (A), thymine (T), guanine (G), and cytosine (C). Although microscopes can be used to study the structure of chromosomes, the actual ordering of the base pairs is determined using specialized instruments called sequencers. Indeed, DNA sequencing has enabled the determination of genome sequences of numerous organisms for the reason that sequence information is essential to understand the biological functions of a cell. A genome will have both coding genes and non-coding DNA, while in fact the non-coding part makes up most of the genome. Instead of just focussing on individual genes, a focus on the genome provides an overall view of the organism's potential biological functions. For example, before the completion of the Human Genome Project (HGP), the number of genes in the human genome was estimated to be more than 100,000 (Adams et al. 1991). With the completion of the genome, the numbers came down to 30,000 to 40,000 (International Human Genome Sequencing Consortium 2001), and the most recent estimation is close to 20,000 (Ezkurdia et al. 2014). Also, genomics studies from all sorts of organisms were able to be fast-tracked, highlighting the importance of a reference genome.

1.1 DNA sequencing

The first step in sequencing a genome is to break the DNA into smaller fragments. Fragmenting a DNA is necessary, due to limitations in technology to read full-length chromosomes. The fragments, also called inserts, are independently sequenced using a sequencer, and the resulting sequence output from a sequencer are called reads, which are later computationally assembled into contiguous pieces called contigs. From Sanger of the first-generation sequencing, through second-generation sequencing, DNA

sequencing has come a very long way foraging more recently into third-generation, single molecule sequencing.

1.1.1 First-generation sequencing (FGS)

Sanger is the most famous and the most widely used sequencer from the FGS era. The first ever genome to be sequenced was that of PhiX174, a bacteriophage in 1977 (Sanger et al. 1977). The phage genome merely possessed a genome size of 5,386 bp. Even then, that particular sequencing effort was a major breakthrough in the field of genomics. It took another 18 years for the first living organism to be sequenced, which was a bacteria, *Haemophilus influenzae*, comprising of 1.8 Mb (Fleischmann et al. 1995). Thereafter, eukaryotes such as *Saccharomyces cerevisiae* (12.5 Mb) (Goffeau et al. 1996), *Caenorhabditis elegans* (100 Mb) (*C. elegans* Sequencing Consortium 1998), *Arabidopsis thaliana* (119 Mb) (Kaul et al. 2000), and *Drosophila melanogaster* (165 Mb) (Myers et al. 2000) were sequenced within a period of five years. The biggest achievement obtained using FGS is the completion of the HGP in 2001 (International Human Genome Sequencing Consortium 2001). The early history of the sequenced genomes using Sanger is represented in figure 1.1. The major caveat of FGS is the heavy consumption of money and time to complete genome projects. For example, the human genome with a size of 3.2 Gb took around 13 years and 2.7 billion dollars for completion.



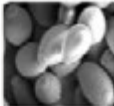




<u>Year</u>		<u>Genome size</u>	
1977		5.3 kb	<u>PhiX174</u> : First genome to be sequenced
1995		1.8 Mb	<u>H. influenzae</u> : First living organism to be sequenced
1996		12.5 Mb	<u>S. cerevisiae</u> : First eukaryotic organism to be sequenced
1998		100 Mb	<u>C. elegans</u> : First multicellular organism to be sequenced
2000		119 Mb	<u>A. thaliana</u> : First plant to be sequenced
2000		165 Mb	<u>D. melanogaster</u> : First insect to be sequenced
2001		3.2 Gb	<u>H. sapiens</u> : Completion of human genome project

Figure 1.1. Early history of the sequenced genomes. All images taken from wikipedia.

1.1.2 Second-generation sequencing (SGS)

SGS technologies came into the picture around 2000, and gaining popularity after 2004, as they featured massively parallel sequencing reactions (Barba et al. 2014). As a result, the cost and the time factors of sequencing projects drastically came down. Illumina is the most widely used sequencing technology from the SGS era and like other SGS technologies, short read lengths were a major limiting factor restricting the computational analysis of the results. Paired-end (PE) and mate-pair (MP) sequencing, as illustrated in figure 1.2, are commonly used strategies to read both the ends of longer DNA fragments to overcome the limitation of short read lengths. DNA fragment inserts of around 200 bp to 600 bp, and 2,000 bp to 40,000 bp can be handled by PE and MP sequencing strategies respectively. Because fragments of a fixed size are selected before sequencing, the insert size is approximately known and can be used to link two distant read pairs helping in increasing the contiguity of genome assembly. Although,

SGS helped assemble the genomes of a numerous organisms, the assembled genomes were mostly fragmented with long unresolved bases termed as gaps.

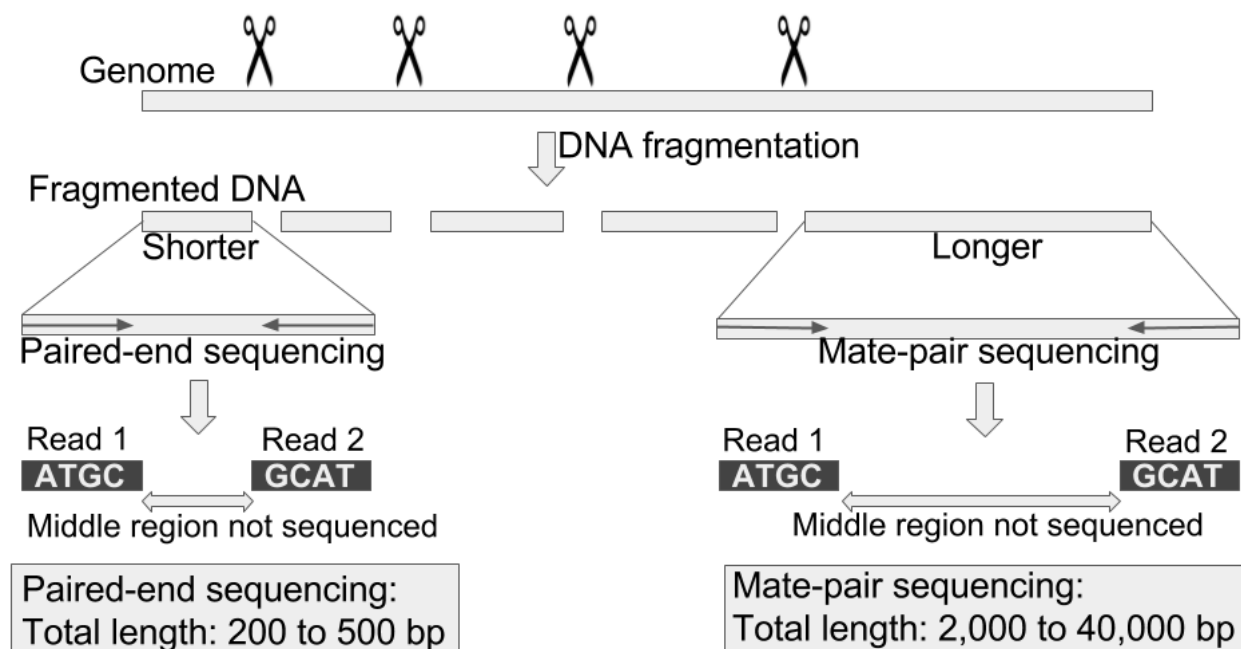


Figure 1.2. PE and MP sequencing.

1.1.3 Third-generation sequencing (TGS)

TGS technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore came into the picture more recently. TGS is also known as single-molecule sequencing owing to the longer fragments. In fact, the reads are much longer than most genomic repeats, paving way for effective assembly resolution compared to the previous generation of technologies. Sequence reads as long as 882 kb have been reported using Nanopore sequencing (Jain et al. 2017), and it might be even possible to sequence longer reads representing complete bacterial chromosomes in the near future. Unlike SGS technologies, which are sensitive to GC content and produce uneven coverage of the genome, the TGS technologies can produce even coverage of the genome (Lee et al. 2016). However, sequencing error rates, as high as 15%, are an usual phenomena in TGS reads (Lee et al. 2016). Despite the high error rates, the errors are mostly resolvable using consensus from sufficient coverage of the reads, and the assemblies achieved using TGS reads can go beyond 99% accuracy. Also, the errors from PacBio are random which makes it easier for correcting the reads because the possibility of a

random error to occur twice is minimal. However, it is common practice to use more accurate Illumina reads to correct the left-out errors from the assembled genome.

1.2 Genome assembly

The whole point of sequencing is to get longer biological information. Hence, sequencing becomes incomplete without assembling the shorter read sequences into longer contiguous sequences. Briefly, all the sequenced reads are aligned against each other to see if there is any overlap between them. Then the reads with overlaps are merged in succession to form longer contigs and the process is called *de novo* assembly (figure 1.3). Thus the assembly process is largely dependent on the overlapping regions between neighboring reads, which is ensured by the random nature of the DNA fragmentation process. The three major approaches used in *de novo* assembly are Overlap-Layout-Consensus (OLC), de Bruijn graphs, and string graphs. For a detailed discussion of the three approaches, please refer to manuscripts, Myers 2014 and Simpson and Pop 2015, which are briefly summarised below.

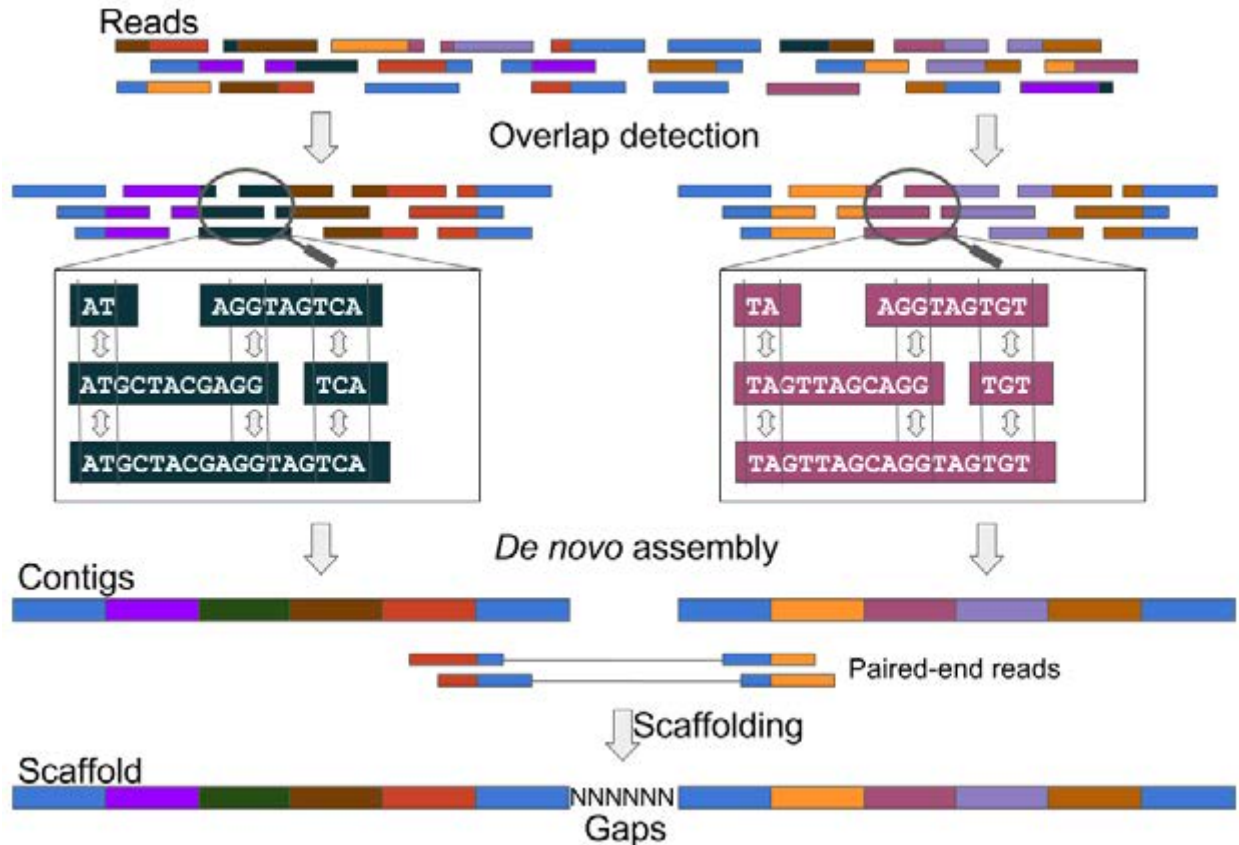


Figure 1.3. The process of *de novo* assembly and scaffolding.

1.2.1 OLC approach

As the name suggests, three steps define OLC approach: a) overlap—finding all approximate overlaps among all the reads, b) layout—using the overlap information to construct an assembly layout, and c) consensus—constructing a consensus sequence of all the reads covering a particular region. In the beginning, decreasing order of overlap lengths was used to build the overlap graphs resulting in several false positive overlap alignments. Later, when quality values began to be associated with the reads, high quality base pair overlaps were given preference when joining overlapping regions. Using such a greedy approach, a set of overlap paths are laid out. Finally, multiple sequence alignment is used to construct a consensus sequence from the layout.

1.2.2 de Bruijn graphs

To construct de Bruijn graphs, first, overlapping k-mers are derived from breaking the reads. Each k-mer is considered as a node, and the adjacent k-mers are connected by an edge to create the graph structure. Traversing the graph, visiting each edge in the graph once (Eulerian tour), will lead to an assembly solution. The most computationally time consuming step in OLC approach, overlap detection, is non-existent in de Bruijn graphs, as the overlap information is implicit in the graph structure. The graph can be constructed, while the sequences are being read by the assembler saving a vast amount of time in the order of $O(N)$ compared to $O(N^2)$ in OLC graphs, where N denotes kmers and read sequences in de Bruijn and OLC graphs respectively. As the size of the genome increases, the computational memory needed to store the graph structure also increases in the order of $O(N)$, where N becomes equivalent to the length of the genome with complete sequence coverage and absence of errors and ploidy. Recent techniques such as Bloom filters (Melsted and Pritchard 2011) does not store the actual k-mers and in the process have enabled *de novo* assembly on desktop computers.

1.2.3 String graphs

If two reads A and C ($A \rightarrow C$) are connected in a graph, and also if a third read B has connections to both A and C , such that $A \rightarrow B$ and $B \rightarrow C$, then the $A \rightarrow C$ connection is redundant. Such connections make the graph redundant and heavier and can be removed entirely and the process is called transitive reduction. Another type of redundant reads are those which are shorter and are entirely contained within an another longer read, which are also removable from the graphs, without any loss of information. By transitively reducing edges and by removing contained reads, an overlap graph can be simplified into what is known as a string graph. For this reason of simplicity, string graphs, although with a similar theoretical space complexity, are memory efficient compared to OLC graphs. Introduction of techniques such as FM-index (Simpson and Durbin 2010) have reduced the computation time for overlap identification, from $O(N^2)$ to $O(N)$ allowing string graphs to be applied for SGS read dataset as well.

All the above three approaches were already described during the FGS era. But according to the needs, shifts in the approaches were observed throughout the transitions in sequencing era.

1.2.4 Genome assembly in the FGS era

During the initial days, when the lambda bacteriophage was sequenced, a simple program was used to identify approximate overlaps between reads, but in the end, the sequences were put together by hand manually to reconstruct the genome. Celera was the first assembler to introduce string graphs, by simplifying the complex graphs produced using OLC approach. At a time, when it was still doubtful whether it was worth investing money on smaller genome projects, comparatively larger genomes such as that of *Drosophila melanogaster* were successfully assembled using Celera. Again, the biggest accomplishment of this era was the completion of the HGP, which helped accelerate various researches pertaining to human diseases and evolution. However, time taken during the overlap detection step of OLC algorithms was critical and consumed several weeks for completion. And with increase in data, the time factor only increased.

1.2.5 Genome assembly in the SGS era

SGS technologies brought more and more data into the frame and the time factor became a serious hurdle with the OLC approaches. Around this time, de Bruijn graph based approaches started gaining widespread popularity. The fact that almost no time is spent on identifying overlaps, which is the most time-consuming step in OLC approaches, made de Bruijn graphs an immediate and automatic choice in the SGS era of *de novo* assembly. Another breakthrough was the development of algorithms such as FM-index which greatly decreased the overlap detection time, and in turn making string graphs applicable to SGS read data in shorter execution times. Both de Bruijn graphs and string graphs are still a popular choice for SGS data.

1.2.6 Genome assembly in the TGS era

The high error rates of TGS meant that traditional methods can not be used for identifying overlaps among reads. Often, the reads were not alignable owing to the high error rates. Hence, approximate alignment methods were preferred for identifying overlaps and storing the information as string graphs. Instead of a single program, the assembly process was broken down to several modules, with a different program handling different aspects of the assembly in a hierarchical manner. For example, the first assembly tool for PacBio data, Hierarchical Genome Assembly Process (HGAP) used BLASR (Chaisson and Tesler 2012) for aligning the reads to identify overlaps, correct errors by consensus using pbdagcon, assemble the data using a slightly modified Celera assembler, and polish the assembled genome using quiver (Chin et al. 2013). Over time, the focus of the long-read assemblers shifted to reducing computational time leading to faster approximate methods to identify overlaps.

1.3. Factors affecting genome assembly

The aim of an assembly program is to reconstruct full-length chromosomes, however the assemblies are almost always fragmented due to practical factors. A variety of factors, which are detailed below, can affect the performance of a genome assembly.

1.3.1 Sequence coverage

The genome need not be fully sampled in every case. Although cost limitations of a project can result in reduced coverage of the genome, platform dependent limitations such as technical difficulties on AT- or GC-rich genomic regions are the major reasons for uneven coverage of the genome (Lee et al. 2016). Such regions which are not covered by sequencing will lead to gaps or fragmentation in the assembly.

1.3.2 Repetitive sequences

Highly identical stretches of nucleotides can repeat many times in a genome. Such repetitive sequences are abundant throughout the taxonomic tree. Repeats can range

from shorter to longer stretches including microsatellites, macrosatellites, centromeric repeats, transposable elements, segmental duplications and other repeats (Chaisson et al. 2015). When the reads are shorter than the repeats, the connection between the flanking regions of the repeats become ambiguous. How repeats can act as a hurdle for *de novo* assembly is illustrated in figure 1.4. Naturally in a repeat-rich genome, the ambiguous connections lead to an exponential number of assembly solutions, rather than just the correct version of the assembly (Chaisson et al. 2015).

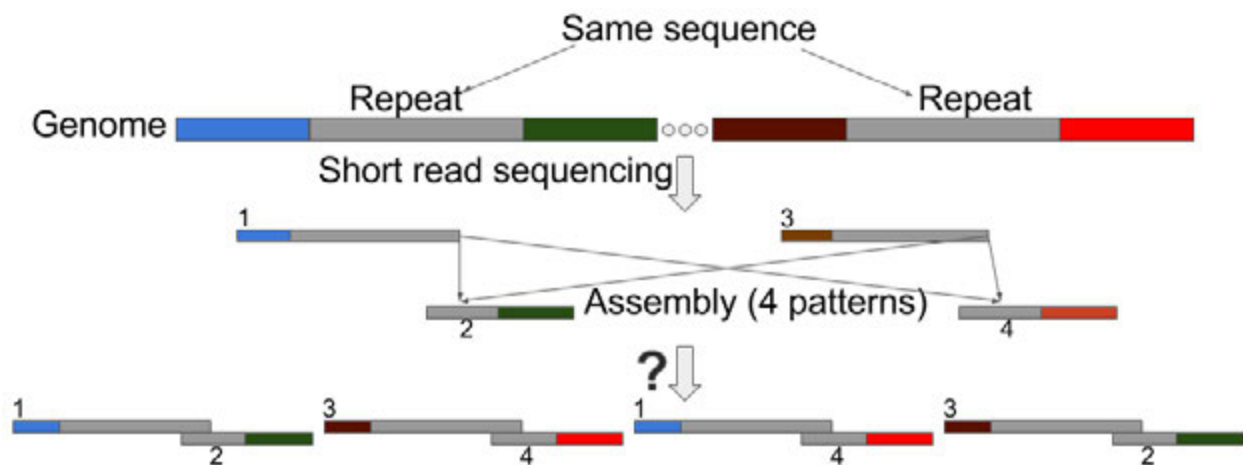


Figure 1.4. Effect of repeats in genome assembly. A repeat sequence, if present twice, can lead to four different assembly possibilities, instead of the two true possibilities.

1.3.3 Sequencing errors

SGS technologies like Illumina were highly accurate (99.99%), with a small number of systematic errors accompanying the reads. Low quality reads can also be a result of unidentified base pairs caused by defects in sequencing. Unlike SGS reads, high error rates are a standard feature of the TGS reads. Errors generally confound the overlap detection step and will lead to erroneous or extraneous paths in the assembly graph (Simpson and Pop 2015). The more the errors, the more the complexity of the assembly becomes.

1.3.4 Ploidy

An assembly program is employed with a motivation to reconstruct a haploid genome. In diploid organisms, the allelic differences act in a similar way as sequencing errors,

leading to extraneous paths in the assembly graph (Chaisson et al. 2015). The situation becomes even worse with polyploid organisms resulting in a highly complicated graph and a highly fragmented genome assembly. Allelic differences may pose as repeats, and at boundaries featuring similar and diverged sequences, and at such regions, contigs are broken without being properly assembled (Chaisson et al. 2015).

1.4. Scaffolding

A complete reconstruction of a genome is practically impossible even for bacterial-sized genomes, if only short reads are employed. In such cases, PE and MP reads can come to the rescue to achieve longer contiguity. The long-range information is inherent in the PE and MP reads, and can thus be used to connect two contigs, which contain either of the ends of the PE/MP reads. In other words, if one end of the PE read (Read1-front) is in contig A, and the other end (Read1-back) is in contig B, then both the contigs can be connected with a fixed number of Ns (unknown bases) in between the contigs. The process is called scaffolding and the inserted Ns are termed as gaps. Scaffolding is a common and an essential procedure to enhance the contiguity of SGS-based genome assemblies.

1.5 Assembly metrics: N50 and L50

One of the main goals of an assembly is to reconstruct genomes as much as long as possible. Hence, contiguity is given the main focus when evaluating an assembly. The mean or median contig lengths are useful statistics when length measures are involved. However, for genome assemblies, shorter contigs are generally more in number and may skew the distribution, which would make it difficult to get a clear picture of how good the assembly is. For this reason, two standard metrics are adopted for genome assembly known as N50 and L50. To calculate these metrics, the contigs are sorted in the decreasing order of contig lengths, and by doing so, the shorter length contigs are not considered, adding weight to only the longest contigs. After sorting, N50 is calculated as the length of the contig at 50% of the assembly length, and L50 as the

number of longest contigs until the 50% assembly mark. The N50 and L50 measures are illustrated in figure 1.5.

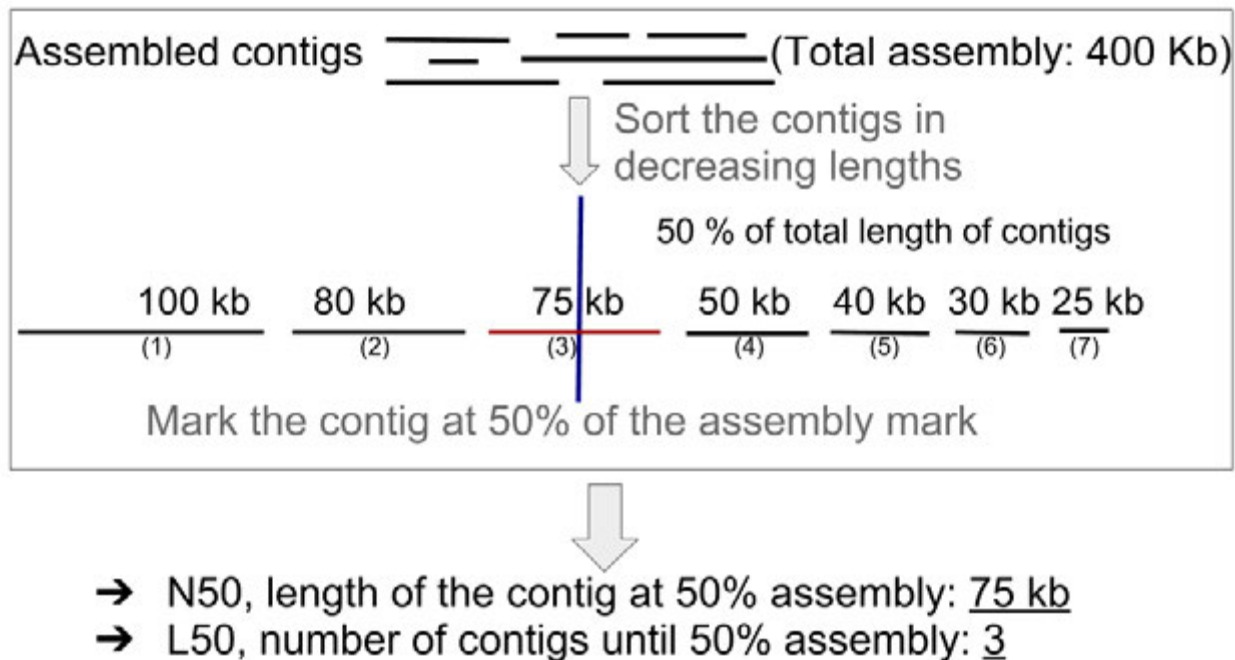


Figure 1.5. N50 and L50 assembly metrics.

1.6 *De novo* assembly of a plant genome

Plant genomes are hardest to assemble from sequenced data because of the high repetitive content and ploidy. Plant genomes are filled with transposons, which can be in the range of several kbs, with identical or nearly identical copies spread throughout the genome. Such regions deeply confound the assembly process, which can not be resolved without long-range information. The genome which was assembled as part of this thesis was that of the plant, *Ipomoea nil*, which is highly repetitive with transposons very much actively mobile. For the assembly project, initially read data (300X coverage) from Illumina was obtained. However, the quality of the data was poor and the resulting assembly was also of poor-quality, with the contig N50 not even reaching 1 kb. A push for new data helped us obtain a new set of PE (300 bp and 500 bp) and MP (3 kb, 5 kb, and 10 kb) libraries. Several attempts including hybrid approaches were tried, but were largely unsuccessful in creating an ideal assembly. Although the quality of the assembly improved from the last time, Assemblathon 2 (Bradnam et al. 2013) was published

around this time and the important lesson learned was that longer range libraries can make a drastic difference to the contiguity of assembly. Hence, longer range MP libraries (15 kb and 20 kb) were also obtained for the assembly. All the data together helped improve the assembly drastically. However, the assembly was filled with a large number of gap base pairs. In the end, the desire for a better quality of the assembly lead us to get hold of PacBio long-read data. The biggest challenge of assembling the PacBio data was the lack of resources for guidance, as compared to that of Illumina assembly. As long-read data was fairly new to the scene, assembly executions required a lot of trial and error to understand the parameters. Parameter tweaking was not just essential for obtaining a higher quality assembly, but also to make use of the computational resources effectively. After several attempts, the final attempt alone took almost a month for completion of the assembly. The numerous attempts resulted in a high quality assembly comparable to those achieved using Sanger sequencing data. As a demonstration of the quality of the genome, several insights were obtained from the genome pertaining to mutation-causing transposable elements, evolution of the Convolvulaceae family, and identification of the cause of a mutable phenotype. Without the availability of a genome, what would take several years was able to be achieved in weeks time. For instance, identification of the mutation for contracted allele had evaded researchers since 1930, however with the availability of a reference genome, the identification became possible within a couple of weeks. The study is discussed in detail in chapter 3, which also shed light on how superior the PacBio assembly is, when compared with the Illumina assembly.

1.7 Evaluation of long-read assembly tools

The assembly of *I. nil* genome was a success, however, the difficulties associated with the long-read assembly had prompted us to a study to guide researchers on assembly from the TGS reads. In the meantime, the interest began to spike in the field of long-read assembly and within a short span of time, around ten long-read assembly tools were released, prompting us to rethink what would have been the best approach for the first study. Long-read assembly is still fairly a new concept and whenever a new concept

is explored in Bioinformatics, benchmark studies are a norm. For example, evaluation studies on short read mapping tools (Hatem et al. 2013), differential ChIP-seq analysis (Steinhauser et al. 2016), RNA-seq differential expression (Zhang et al. 2014), variant calling pipelines (Hwang et al. 2015), metagenomics tools (Sczyrba et al. 2017) and numerous other studies are available to guide researchers in their respective fields. Similarly, at a time, when SGS data was used for assembling genomes, evaluation studies such as GAGE (Salzberg et al. 2012), GAGE-B (Magoc et al. 2013), Assemblathon (Earl et al. 2011), and Assemblathon 2 (Bradnam et al. 2013) were published, garnering widespread attention as a guide for assembly using SGS data. However, no such comprehensive studies had been performed on long-read datasets as of now, while more and more genomes were starting to be assembled using TGS data. A comparison of the effectiveness in the quality of the TGS data based assemblies (Lan et al. 2017; Berlin et al. 2015; Shi et al. 2016; Du et al. 2017), by comparing the N50 values with those of SGS data based assemblies (Ibarra-Laclette et al. 2013; Steinberg et al. 2014; Li et al. 2010; Schatz et al. 2014) is demonstrated in figure 1.6.

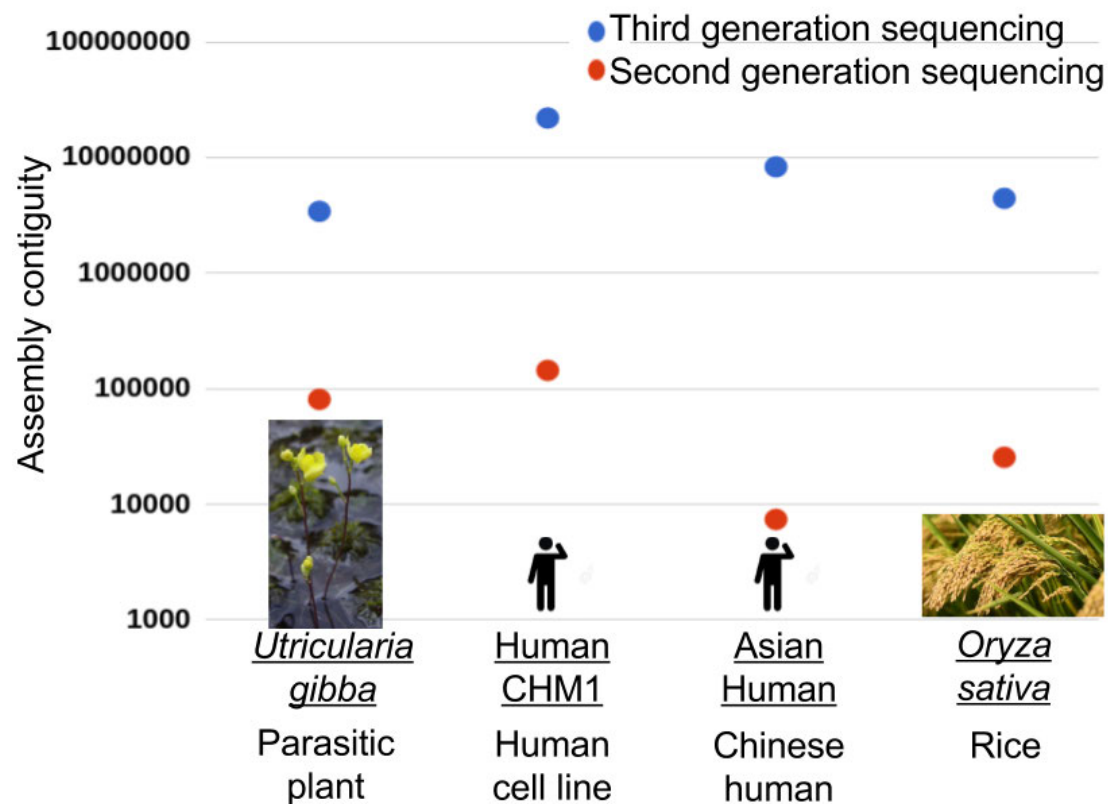


Figure 1.6. N50 values from SGS and TGS genome assemblies of different organisms

But it is hard to say which assembler would work the best. For example, PBcR is the second most used assembler in recent publications (refer chapter 4). This might guide researchers, who are new to the field, to automatically use PBcR for their assembly. However, PBcR is an outdated pipeline and is replaced by a new assembler, canu. Thus all the available long-read assemblers were put to test and therefore the theme for the second study of this thesis was set. To be as comprehensive as possible for other researchers to apply the results to their study, four organisms from very different taxonomic families were chosen such that they have huge differences in size and other features of the genome. The study was a revelation to ourselves too, as it was concluded from the study that there are better assemblers, in terms of producing lesser mis-assemblies, than what we had used for the assembly of *I. nil*. However, sufficient care was taken for the first study by extensively detecting and splitting off mis-assemblies, such that it did not have an effect on the quality of the assembly. The evaluation study was also executed with the belief that researchers might be able to choose parameters freely, when a guidance on the same is available through this study. The details of the evaluation study are discussed in chapter 4.

Chapter 2

Construction of computational pipelines for *de novo* assembly

In computational biology, a single problem is usually solved using a variety of programs employing different techniques. Some of them may share a basic layout but may differ in critical processes related to improving accuracy, speed, memory requirements etc. Hence, it is necessary to choose the right programs and right parameters for solving a computational problem. In this thesis, for the problem of long-read *de novo* assembly, there are at least ten available programs. However, *de novo* assembly using TGS reads is not straightforward, the assembly process is broken down to several modules, with a different program handling different aspects of the assembly in a hierarchical manner and thus may need a series of programs and a trial of several parameters for successful execution. For example, certain parameters such as those related to read length, in particular overlap length, can practically influence the computational speed of the program, while also influencing the contiguity and correctness of the assembly. In this thesis, we have constructed a computational pipeline that will execute an end-end analysis starting from raw read data, through *de novo* assembly, until the point of assembly validation and annotation of genomic features. The initial part of the pipeline until the end of the genome assembly is illustrated in figure 2.1.

2.1 Parameters and other aspects in the pipeline

Several parameters affecting individual parts of the pipeline are described below, along with important aspects of the pipeline.

2.1.1 Error correction and polishing

Errors are probably the most major concern in TGS technologies. Hence, they need to

be rectified either before and/or after the end of the assembly. Assemblers including Canu, FALCON, HGAP3, MECAT, PBcR, and SMARTdenovo initiate the assembly with

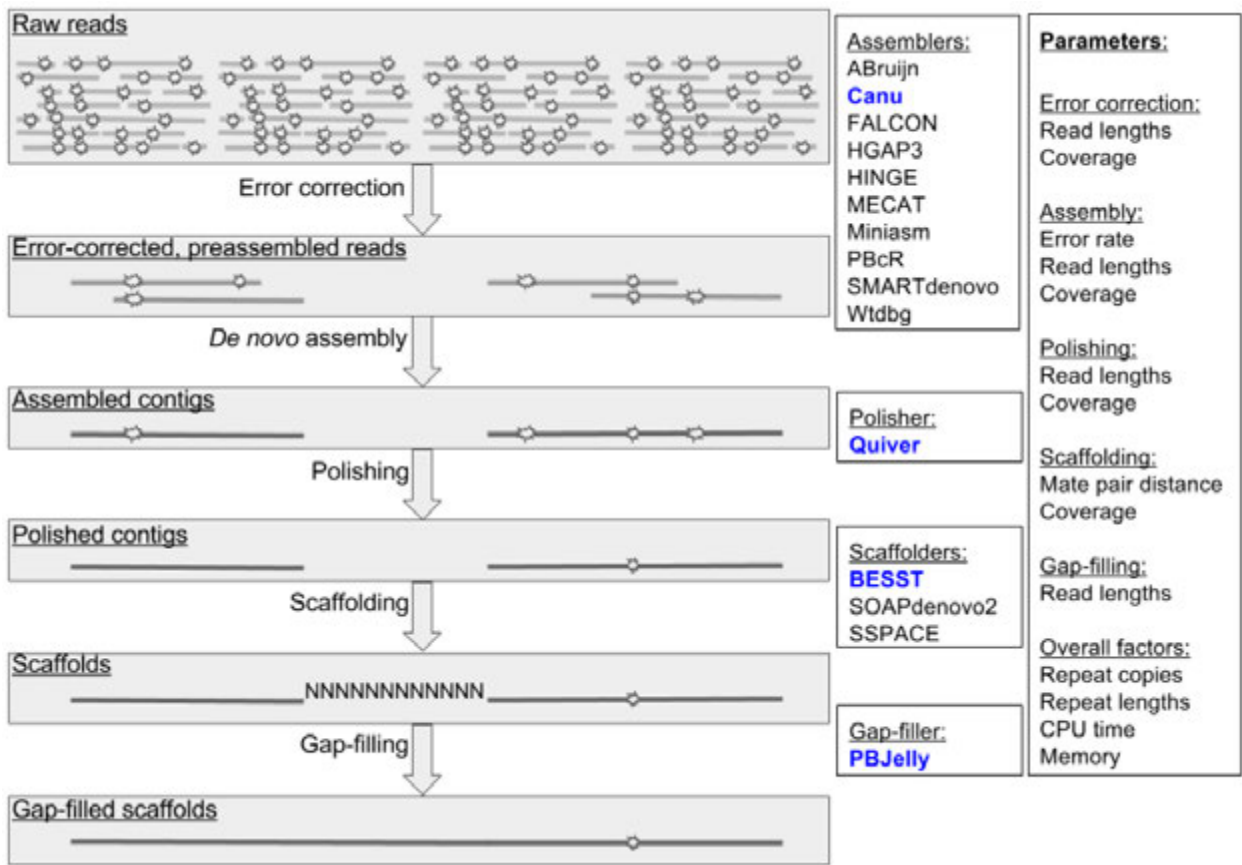


Figure 2.1. Pipeline of *de novo* assembly including parameters affecting the process.

Programs marked in blue are the programs selected as the best based on trials.

the first step being error-correction. Assemblers such as ABruijn, HINGE, miniasm, and wtdbg do not perform error correction, but have a consensus polishing step as part of the program to get rid of errors after the assembly. A few parameters, such as overlap lengths and coverage, can potentially affect the performances. Overlap lengths become crucial in differentiating an error base from actual differences stemming from repetitive regions. Because the errors in PacBio technology are random, a consensus can be derived when there is ample coverage representing genomic regions. The errors become unresolvable with lesser coverage. In fact, when the coverage is less than 20X, it is best not to assemble the data using TGS reads.

2.1.2 *De novo* assembly

Again overlap lengths and coverage play a predominant role in *de novo* assembly as well. The shorter the overlap length, the more fragmented the assembly becomes, because of extraneous connections from repetitive regions. The longer the overlap length, lesser connections are made, again fragmenting the assembly. Hence, an ideal overlap length needs to be optimized for each genome, especially in plant genomes which are rich in long transposon repeats. A step by step increase in sequence coverage and their corresponding genome assembly revealed that the contiguity of the assembly gradually increased until around 50X coverage, after which the contiguity started to plateau (Koren et al. 2017). However, after error correction, the shorter reads are thrown out and only the longest 25X–30X coverage are kept for assembly. With different coverages, the assembly quality will also differ. Another important aspect is the error rate parameter. Even after error correction, a lot of errors are still left behind in the data. Hence, based on the coverage, this parameter also needs to be adjusted accordingly to get a better resolution of the repeat specific base changes.

2.1.3 Scaffolding and gap-filling

Both PE and MP libraries are generally used for scaffolding. The libraries are added sequentially one after another starting from the shortest (300 bp) to the longest (40 kb). By doing the same, we observed that scaffolding had very little to no effect. The reason being, the PacBio reads are much more longer than the PE or even some of the MP insert sizes resulting in resolved assemblies at such locations. Most of the fragmentation was caused by longer repeats which were untenable by PacBio reads. Hence, the shorter libraries caused mis-connections and because the longest ones are added finally, conflicts arose in connections leading to no results. Hence, only 15 kb and 20 kb MP libraries were used for scaffolding, which were longer than most of the input PacBio reads. Coverage of the mate pairs is another important parameter to fine tune to obtain better results. For gap-filling, again the PacBio reads were used which largely relied on overlap lengths. Because, the PacBio reads contributed to a highly contiguous contig assembly, the scaffolding procedure managed to connect the longer

contigs resulting in a better N50.

2.1.4 Computational resources

Allocating the right parameters for computational resources is a very important aspect for long-read genome assemblies, as some of the assemblies might take more than a month with certain assemblers. If lesser resources are provided, the program will abort. Whereas, when more than sufficient resources are provided, the program may take several months to complete. So choosing the right resources for each of the process as part of the pipeline was extremely important. Some programs are memory-intensive, whereas other programs have jobs split over several computational nodes with lesser memory, however taking longer computational times. All these factors were considered for all the tested assembly tools.

2.1.5 Assembly validation

Validating assemblies was one of the easiest in terms of adjusting parameters in the pipeline. All available resources from public DNA databases, as well as, newly sequenced data, were put into use in the pipeline for validating the genome assemblies. Also, standard assembly validation tools such as CEGMA and BUSCO were used as part of the pipeline. The only major attempt was fixing the parameter for linkage maps such that all the markers are separated into exactly 15 linkage groups representing the actual chromosomes of *I.nil*.

2.1.6 Gene prediction

As there is no availability of a reference genome for the Convolvulaceae family, initially, the cDNA data from NCBI for *I. nil* was used to train gene models for gene prediction. However, the lack of sequences meant that the training was not complete. In contrast, when Tomato from Solanaceae, the sister family of Convolvulaceae, was used as a reference for training gene models, the predicted results were mostly accurate and correlated well with the available cDNA data.

2.1.7 Repeat prediction

Although, the standard repeats were well characterized using repeat prediction programs, *Tpn1* transposons, the main feature of *I. nil* was not predicted by standard programs. Hence, a separate program was written in-house to predict and catalog the transposons. The structure of the *Tpn1* transposons is used as the reference, and the sequence features are mapped using the custom program to find and catalog the *Tpn1* transposons.

Most of the tools used in the pipeline required several trial and errors, before being applied successfully. Although, the pipeline has parts specific to *I. nil* genome, the constructed pipeline can now be applied to any future related projects without any difficulties. For example, currently the common marmoset genome is being assembled and analyzed using the pipeline.

Chapter 3

Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*

Ipomoea is the largest genus in the family Convolvulaceae. *Ipomoea nil* (Japanese morning glory) has been utilized as a model plant to study the genetic basis of floricultural traits, with over 1,500 mutant lines. In the present study, we have utilized second- and third-generation sequencing platforms, and have reported a draft genome of *I. nil* with a scaffold N50 of 2.88 Mb (contig N50 of 1.87 Mb), covering 98% of the 750 Mb genome. Scaffolds covering 91.42% of the assembly are anchored to 15 pseudo-chromosomes. The draft genome has enabled the identification and cataloging of the *Tpn1* family transposons, known as the major mutagen of *I. nil*, and analyzing the dwarf gene, *CONTRACTED*, located on the genetic map published in 1956. Comparative genomics has suggested that a whole genome duplication in Convolvulaceae, distinct from the recent Solanaceae event, has occurred after the divergence of the two sister families.

3.1 Background

The genus *Ipomoea*, which includes 600–700 monophyletic species, is the largest genus in the family Convolvulaceae and is a sister group to the family Solanaceae (Austin and Huáman 1996; Stefanovic et al. 2002). These species exhibit various flower morphologies and pigmentation patterns (Clegg and Durbin 2003), and are distributed worldwide (Austin and Huáman 1996). Morning glory species, including *Ipomoea nil*, *I. purpurea*, *I. tricolor*, and *I. batatas* (sweet potato), are commercially important species. Japanese morning glory (*I. nil*), locally known as Asagao, is a climbing annual herb producing blue flowers capable of self-pollination (figure 3.1a–l). It is believed to have been introduced from China to Japan in the 8th century, and has become a traditional

floricultural plant in Japan since the 17th century. Most of Japanese elementary students grow it, as part of their school curriculum. The genetics of *I. nil* has been extensively studied for more than 100 years, and it has been a model plant for the study of photoperiodic flowering and flower coloration. A number of spontaneous mutants of *I. nil* have been identified since the early 19th century. Most of their mutations were related to floricultural traits, and several variants with combinations of mutations have been developed (figure 3.1m–aa). The unique features of *I. nil*, e.g., blue flowers and vine movements (Fukada-Tanaka et al. 2000; Kitazawa et al. 2005), have been characterized by using the cultivars carrying such mutations.

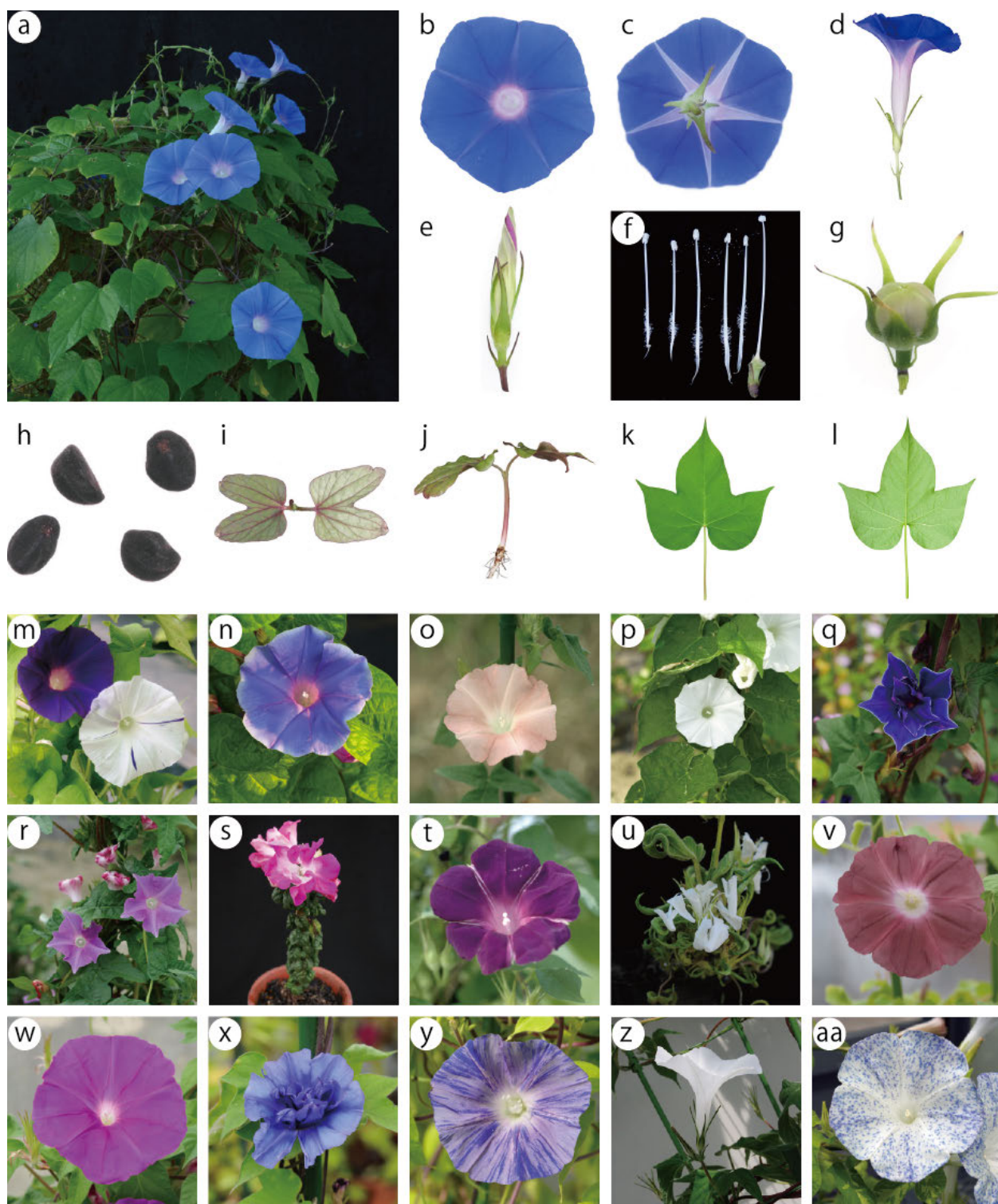


Figure 3.1. The Japanese morning glory. (a-l) The wild-type line Tokyo-kokei standard (TKS). **(a)** The individual used for whole genome sequencing. **(b)** Flower from the front. **(c)** Flower from the back. **(d)** Flower from the side. **(e)** Flower bud one day before flower opening. **(f)** Stamens (five on the left) and carpel (farthest right). **(g)** Seed pod. **(h)** Seeds. **(i)** Seedling from above. **(j)**

Side view of a seedling. **(k)** Leaf from the front. **(l)** Leaf from the back. **(m)** The Q1072 line carrying the recessive *a3-f* mutation that is the *Tpn1* insertion into the *DFR-B* gene for flower pigmentation. The mutant produces white flowers with pigmented spots or sectors (right) and sometimes produces fully pigmented flowers (left). The pigmentation patterns are caused by *Tpn1*-inducing somatic mutations and indicate that transposases TnpA and TnpD of *Tpn1* family elements are active. **(n-s)** Brassinosteroid-deficient mutants. **(n)** Q848 (*ct-1*). **(o)** Q853 (*ct-2*). **(p)** Q220 (*ct-w*). **(q)** Q708 (*s*). **(r)** Q721 (*s*). **(s)** Q837 (*ct-1*, *s*). **(t-aa)** The mutant lines carrying one of the recessive mutations that were mapped on the classic linkage map. The *cd*, *fe*, *dy*, *a3*, *mg*, *dp*, and *dk-2* mutations were assigned to classic LG1, LG2, LG3, LG4, LG5, LG6, and LG10, respectively. The recessive mutations of *c1* and *sp* were also assigned to LG3. In this study, LG3N with *dy* and LG3S containing *c1* and *sp* were found to correspond to different chromosomes. **(t)** Q557 (*cd*) showing partial transformation of floral petals into sepals. **(u)** Q459 (*fe*) showing alteration of organ polarity. **(v)** Q114 (*dy*) with dark-colored flowers. **(w)** AK62/Violet (*mg*) with reddish flowers. **(x)** Q426 (*dp*), producing double flowers. **(y)** Q531 (*dk-2*) with pale- and dull-colored flowers. **(z)** AK33 (*c1*), producing white flowers with red stems. **(aa)** AK30 (*sp*) showing speckled flowers. All *I. nil* lines are from the National BioResource Project (<http://www.shigen.nig.ac.jp/asagao/>).

More than 1,500 cultivars of *I. nil* are maintained by the Stock Center at Kyushu University as a part of the National BioResource Project. Our recent studies have revealed that many of these mutant lines have been the result of mutagenic activity by *Tpn1* family transposons (Fukada-Tanaka et al. 2000; Inagaki et al. 1994; Hoshino et al. 2009; Morita et al. 2014; Iwasaki and Nitasaka 2006; Nitasaka 2003). These transposons are class II elements and members of *En/Spm* or CACTA superfamily that can transpose via a cut-and-paste mechanism. The maize *En/Spm* elements encode two transposase genes for TnpA and TnpD, mediating transposition of *En/Spm* and its derivatives (Weil and Kunze 2002). TnpA and TnpD bind to the sub-terminal repetitive regions (SRRs) and terminal inverted repeats (TIRs) of *En/Spm*, respectively (Weil and Kunze 2002). The copy number of the *Tpn1* family was estimated to be 500–1,000, and almost 40 copies have been characterized (Fukada-Tanaka et al. 2000; Inagaki et al. 1994; Hoshino et al. 2009; Morita et al. 2014; Iwasaki and Nitasaka 2006; Nitasaka 2003; Kawasaki and Nitasaka 2004; Morita et al. 2015). All of the transposons characterized thus far are non-autonomous elements, and no elements encoding intact

transposase genes have been identified. The non-autonomous *Tpn1* family elements have a characteristic structure and are known to capture genic regions from the host genome (Kawasaki and Nitasaka 2004; Takahashi et al. 1999). Their internal sequences are substituted with the captured host sequences, whereas their terminal regions necessary for transposition are conserved. Some of the internal genic regions are transcribed; a *Tpn1* transposon integrated in the *DFR-B* gene for anthocyanin pigment biosynthesis generates chimeric transcripts consisting of both the *DFR-B* and the captured intragenic region (Takahashi et al. 1999).

I. nil has 15 pairs of chromosomes ($2n = 30$) (Yasui 1928). However, the original classical map from 1956 contained only ten linkage groups, as a result of mapping 71 genetic loci out of 219 analyzed loci to one of the ten linkage groups (Hagiwara 1956). The genetic information of *I. nil* available to date includes the linkage map (Hagiwara 1956), 62,300 expressed sequence tags (ESTs) deposited to the DDBJ/EMBL/NCBI databases, Simple Sequence Repeat (SSR) markers (Ly et al. 2012) and a recent large scale transcriptome assembly (Wei et al. 2015). The availability of a reference genome sequence would give researchers a standard with which to compare their mutant lines and would fast track genomic analysis of mutations. The genome of a closely related species of a wild sweet potato, *I. trifida*, was recently sequenced and published (Hirakawa et al. 2015), in which they reported genome sequences of two *I. trifida* lines analyzed using Illumina HiSeq platform, with average scaffold lengths of 6.6 kb (N50 = 43 kb) and 3.9 kb (N50 = 36 kb), respectively. However, the assembled scaffolds did not have chromosomal level information, and were highly fragmented.

In the present study, we report a pseudo-chromosomal level whole genome assembly of a wild-type *I. nil* line, with an estimated genome size of 750 Mb, sequenced using PacBio's Single Molecule, Real-Time Technology (SMRT) and Illumina sequencers. We have also identified two copies of *Tpn1* family transposons encoding putative TnpA and TnpD transposases, 339 other non-autonomous *Tpn1* transposon copies, as well as the most likely candidate for the dwarf gene, *CONTRACTED*, mapped on the classical genetic map.

3.2 Results

3.2.1 DNA sequencing and genome assembly

One individual plant of the wild-type line, Tokyo Kokei Standard (TKS), was used for genome sequencing. Its genome size was estimated to be approximately 750 Mb using flow cytometry. PacBio sequencing yielded 5.74 million reads (39.4 GB, 52.6× coverage and N50 of 10.3 kb), with the longest and the average read lengths being 48.1 kb and 6.8 kb respectively, whereas, sequencing using the Illumina HiSeq (table 3.1) included two short and six long insert libraries. With an initial read length of 150 bp, the short reads covered approximately 906× of the genome. The work-flow for the PacBio data assembly consisted of seven steps (figure 3.2).

Table 3.1. Statistics of raw Illumina reads.

Strategy	Insert length	# of reads (in millions)	# of bases (in Gb)	Sequence coverage	Accession number
Paired end	300 bp	602	90	123 ×	DRR013917, DRR013918
Paired end	500 bp	652	98	133 ×	DRR013919, DRR013920
Mate pair	3 kb	563	85	115 ×	DRR013921, DRR013922
Mate pair	5 kb	544	82	111 ×	DRR013923, DRR013924
Mate pair	10 kb	584	88	119 ×	DRR013925, DRR013926
Mate pair	10 kb	505	76	103 ×	DRR048755
Mate pair	15 kb	495	74	101 ×	DRR048756
Mate pair	20 kb	494	74	101 ×	DRR048757

PacBio reads

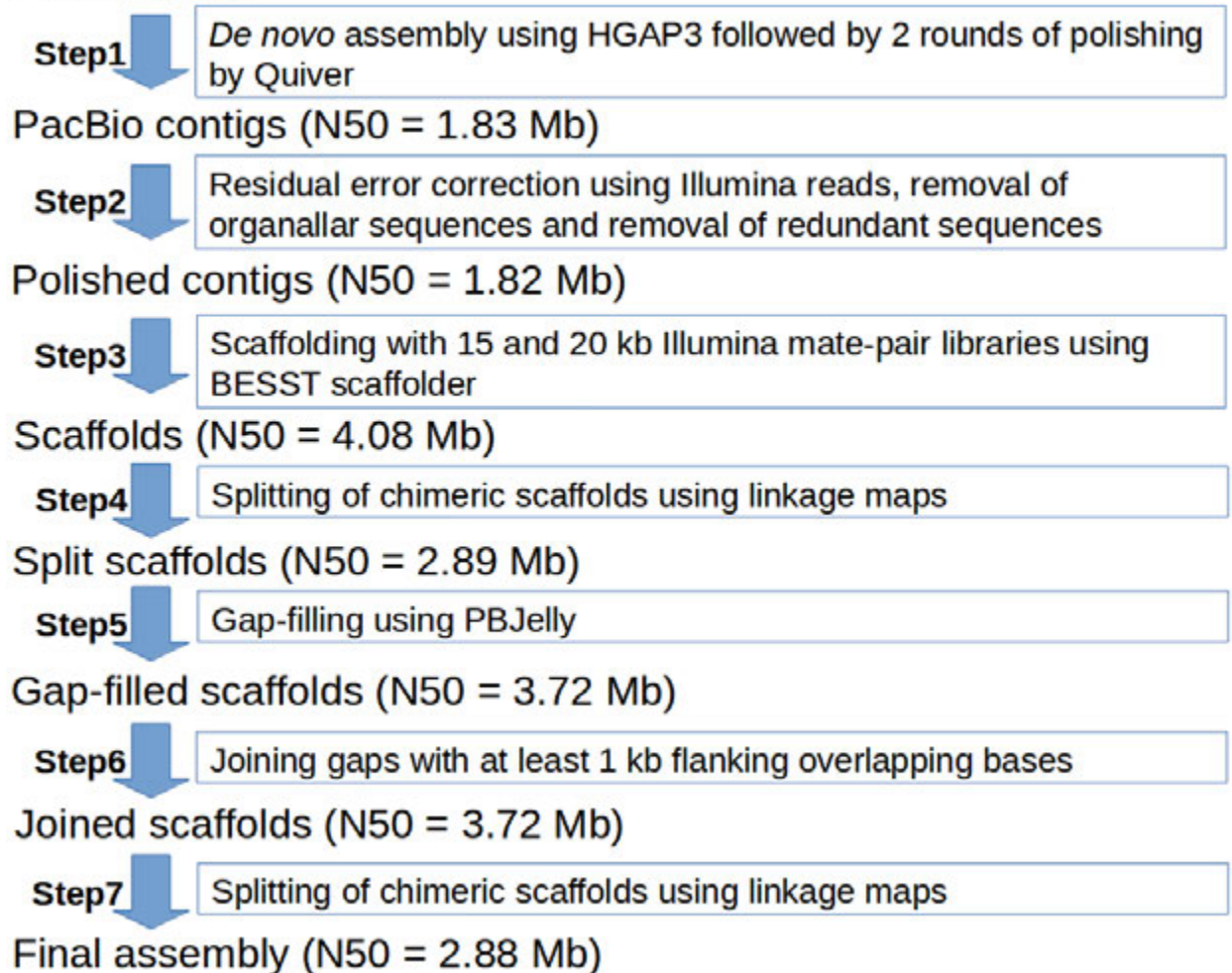


Figure 3.2. Assembly work-flow. The assembly pipeline used for assembling the *I. nil* genome utilizing PacBio and Illumina sequence reads.

Initial *de novo* assembly of the PacBio reads resulted in 736.4 Mb of genome assembly, with a contig N50 of 1.83 Mb. To remove left-over residual errors originating from PacBio sequences, the short reads from Illumina were aligned against the assembled genome to identify homozygous variants. The homozygous variants amounted to 1,532 SNPs, 20,479 deletions, and 6,549 insertions showing that the assembly had 99.99% base accuracy. The insertion-deletion (in-del) errors had outnumbered the substitution errors, similar to the results observed in PacBio-based *Vigna angularis* (Sakai et al. 2015) and *Oropetium thomaeum* (VanBuren et al. 2015) genome assemblies, and were replaced with the Illumina sequence bases. Mitochondrial and chloroplast derived

sequences were identified to be 1.15 Mb from 51 contigs and were removed. The organellar genomes were sequenced using a Sanger sequencer and assembled separately. Scaffolding using Illumina longer range mate-pair libraries and subsequent gap-filling using PacBio reads increased the N50 to 3.72 Mb. The assembly statistics at each step of the work-flow are mentioned in table 3.2. An independent assembly of Illumina reads using SOAPdenovo2 assembler (Luo et al. 2012) resulted in 1.1 Gb of genome assembly. The assembly size was reduced to 768 Mb, with a scaffold N50 of 3.5 Mb and a contig N50 of 9.5 kb, when considering only contigs and scaffolds longer than 1 kb. The assembly statistics of both the PacBio and Illumina assemblies are compared in table 3.3. The PacBio version of the assembly was chosen for downstream analysis owing to PacBio's long read lengths vastly increasing the contiguity of the assembled genome.

3.2.2 Mis-assembly detection and pseudo-molecule construction

Illumina sequencing employing the RAD-seq (Baird et al. 2008) procedure, yielded 86.1 million reads for the parent samples and 562.2 million reads for the progeny samples (read length of 150 bp). Filtering the SNP markers obtained using the STACKS (Catchen et al. 2011) pipeline resulted in 3,733 SNP markers from 176 samples. Fifteen linkage maps were constructed using the SNP and were helpful in identifying inconsistent scaffolds which were present in more than one linkage group. To eliminate the possibility of mis-assembled chimeric scaffolds, the scaffolds were split at their junction points into two separate scaffolds using the linkage maps as a reference. In the case of mis-assemblies at the contig level, each chimeric region was split into three parts such that the first and the last part would belong to two different chromosomes from the linkage map, whereas the middle part would still remain chimeric, albeit with a shorter length (figure 3.3). A first splitting procedure was employed to split 52 scaffolds, after the scaffolding phase of the assembly process. After gap-filling, another splitting procedure was used to break 29 additional scaffolds. The major achievement of the assembly procedure was that, even after splitting chimeric scaffolds, the N50 values obtained for scaffolds and contigs were still 2.88 Mb and 1.87 Mb (table 3.4) respectively, which is comparable to assemblies achieved utilizing traditional

Table 3.2. Comparison of the stepwise assemblies of PacBio data, with each step referring to the step from the assembly workflow (figure 3.2)

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
	Scaffolds						
Sequences	NA		3367	3495	3345	3345	3416
Total Length of Sequences			734061355	734055648	735418791	734768050	734803190
Gap openings			624	602	544	461	449
Gap length			327148	321441	177095	177012	211647
Longest sequence length			16099154	14441919	14449637	14449637	14449934
Shortest sequence length			638	638	638	638	638
Average sequence length			218016.44	210030.23	219856.14	219661.6	215106.32
N50			4082476	2890004	3727853	3727853	2880368
Sequences (>10 Mb)			7	2	6	6	2
Sequences (>1 Mb)			166	204	182	182	205
Sequences (>100 kb)			299	401	326	326	389
Sequences (>10 kb)			2043	2164	2120	2120	2194
Sequences (>1 kb)			3991	4097	3889	3806	3865
Sequences (>500 bp)			3991	4097	3889	3806	3865
Sequences (>100 bp)			733734207	733734207	735241696	734591038	734591543
	Contigs						
Sequences	4187	3991	3991	4097	3889	3806	3865
Total Length of Sequences	736457052	733734371	733734207	733734207	735241696	734591038	734591543
Longest sequence length	11504781	11504932	11504932	8729492	11281532	11281532	9127415
Shortest sequence length	638	638	638	638	638	638	638
Average sequence length	175891.34	183847.25	183847.21	179090.6	189056.75	193008.68	190062.49
N50	1830236	1825684	1825684	1584472	1918312	2087487	1873359
Sequences (>10 Mb)	4	4	4	0	3	3	0
Sequences (>1 Mb)	191	191	191	200	189	195	205
Sequences (>100 kb)	649	649	649	724	620	574	625
Sequences (>10 kb)	2773	2648	2648	2746	2650	2567	2629
Sequences (>1 kb)	4169	3973	3973	4075	3873	3790	3853
Sequences (>500 bp)	4187	3991	3991	4093	3885	3802	3865

Sanger sequencing data (Michael and Jackson 2013). The mapping of scaffolds to linkage maps not only aided in identifying potential mis-assemblies, but also guided the generation of pseudo-chromosomes from the available scaffolds. The pseudo-chromosomes accounted for 91.42% of the assembly (N50 of 44.78 Mb), along with unoriented scaffolds (around 25.53% of the assembly), and are represented in a circular display, with predicted genomic features along the 15 pseudo-chromosomes (figure 3.4a–f).

Table 3.3. Comparison of the Illumina and PacBio assemblies

	PacBio Assembly	Illumina Assembly
Sequences	3416	2262957
Total length of sequences	734803190	1106449450
Gap openings	449	132545
Gap lengths	211647	74798170
Longest sequence length	14449934	18182283
Average sequence length	215106.32	488.94
N50 (sequences >1 kb)	2880368	3532667
Sequences (>10 Mb)	2	3
Sequences (>1 Mb)	205	213
Sequences (>100 kb)	389	387
Sequences (>1 kb)	3404	3927
Sequences (>100 b)	3416	2262957

Mis-assemblies were not resolved in the Illumina based assembly.

Table 3.4. *I. nil* genome assembly statistics

Category	Total	N50 (Mb)	Longest (Mb)	Size (Mb)	Percentage of the assembly
Contigs*	3,865	1.87	9.12	734.6	-
Scaffolds	3,416	2.88	14.4	734.8	100
Anchored scaffolds	321	3.14	14.4	671.7	91.42
Genes	42,783	-	-	182	24.77
Repeats	-	-	-	465	63.29

*The gaps in the final version of the scaffolds were split to produce the final version of contigs.

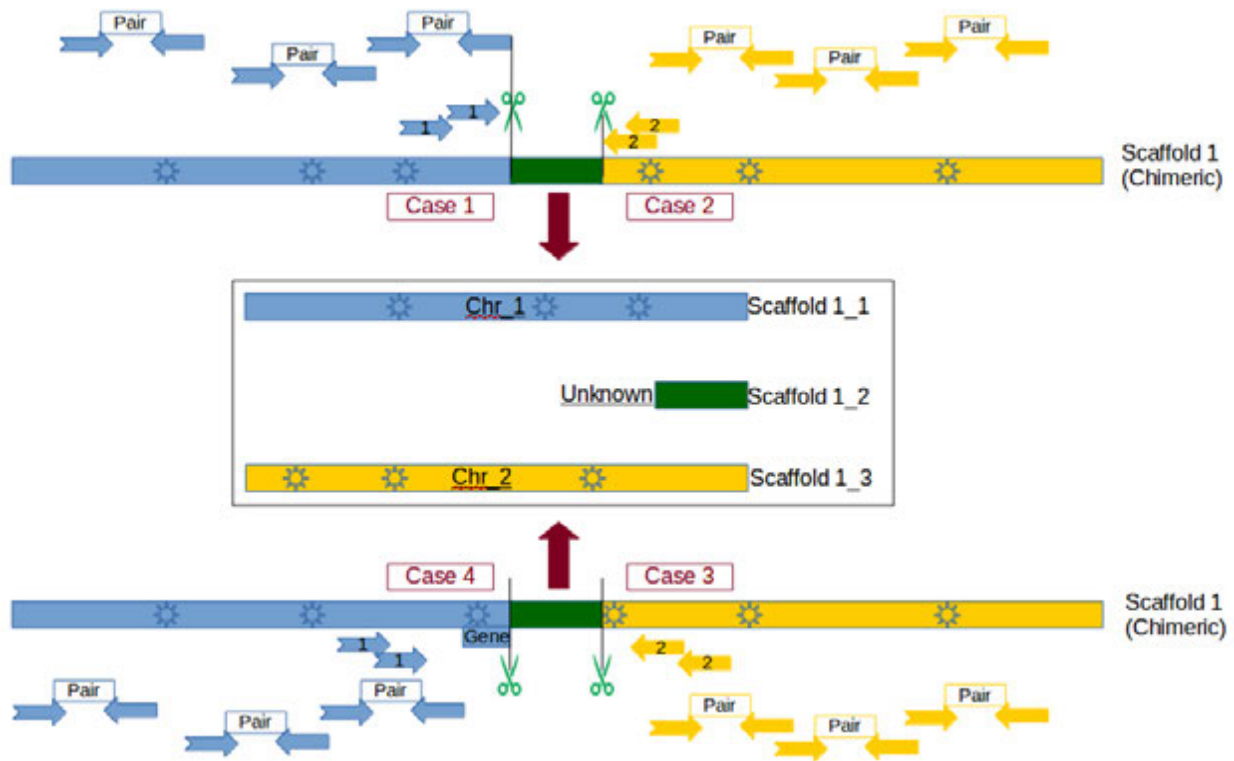


Figure 3.3. Mis-assembly Breakage Process. Case 1 and 2 depicts breakage using BAC-end pair information. In case 1, the breakpoint is at the nearest complete BAC-end pair, and in case 2, the breakpoint is at the nearest BAC-end read, whose read-pair is in a different scaffold. Also, when there is not sufficient BAC-end read information, the SNP marker from the linkage maps was used as the breakpoint (Case 3). All cases were identified using disputes in linkage maps and were split into 3 separate scaffolds. The first and last scaffolds were assigned to corresponding chromosomes from the linkage map.

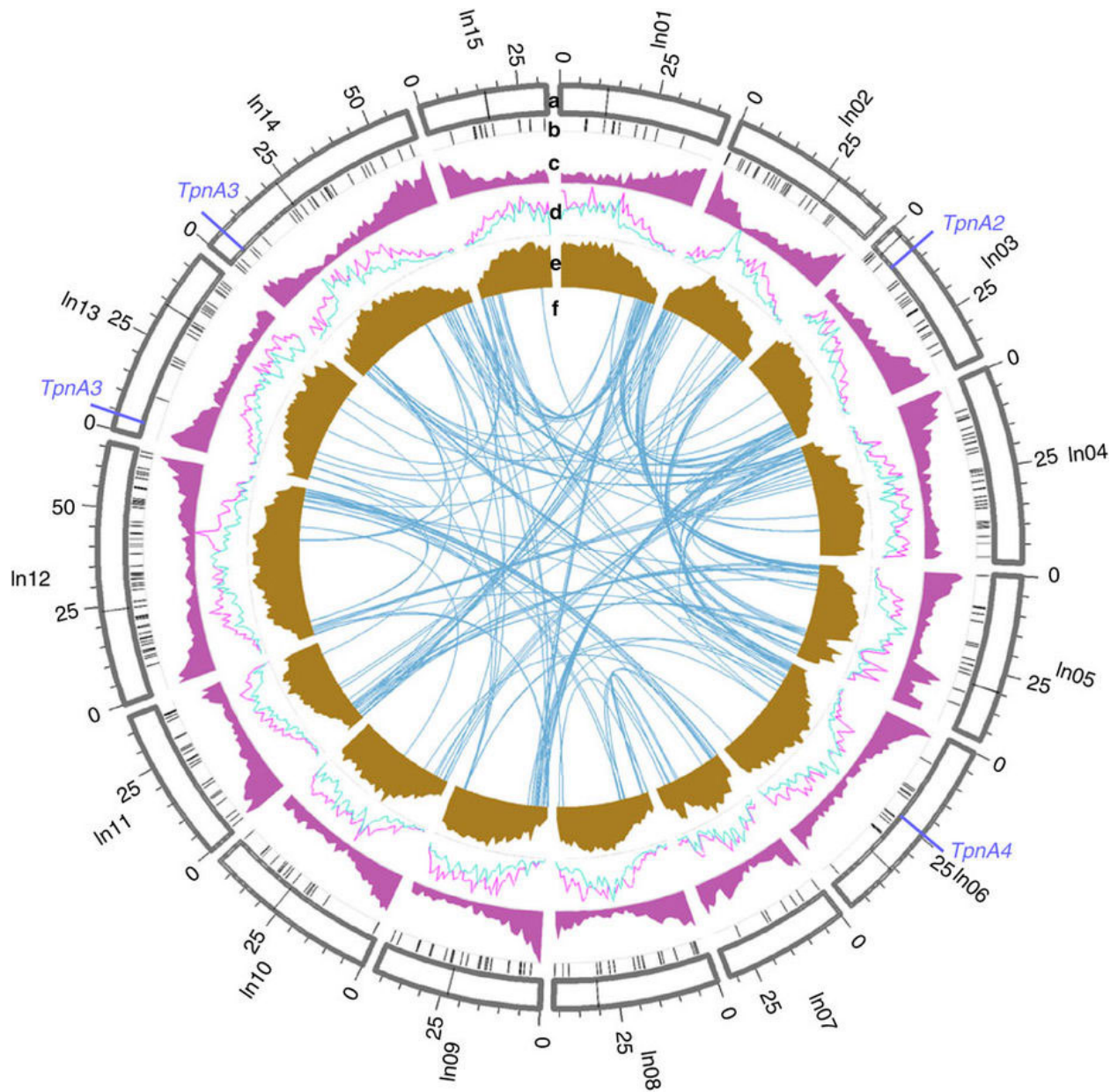


Figure 3.4. Genomic characterizations of *I. nil*. (a) Outer circle displaying the 15 pseudo-chromosomes in 1 Mb units. TpnA2–4 (blue dashes) and putative centromeric locations (black dashes) are also denoted in the outer circle. (b) Location of *Tpn1* family transposons. (c) Gene density per Mb. (d) Coverage of copia (magenta) and gypsy (turquoise) LTRs per Mb. (e) Repeat coverage per Mb. (f) Syntenic regions containing more than 10 paralogous genes.

3.2.3 Assembly validation

The Core Eukaryotic Genes Mapping Approach, or CEGMA pipeline (Parra et al. 2007)

and more recently, the BUSCO (Simão et al. 2015) pipeline have become commonly used protocols to validate the completeness of assembly projects by examination of coverage of highly conserved genes. The percentage of completeness for our assembly was 94.35% and 99.60% for completely and partially aligned core eukaryotic genes (CEGs) respectively. BUSCO analysis revealed a completeness score of around 95%. This indicated that most of the evolutionarily conserved core gene set was present in the *I. nil* assembly suggesting a high quality assembly. To further validate the assembly, the newly generated *I. nil* ESTs, BAC-end, and RNA-seq data were utilized. Comparisons against 93,691 ESTs showed that 99.11% of them were aligned, with 97.40% of the ESTs having at least 90% of their lengths covered in the alignments. Using 20,847 BAC-end read pairs, it was found that 94.92% of the reads were paired in the same scaffold with a mean insert length near the 100 kb mark, and 97.87% of the reads were paired in the same pseudo-chromosome. RNA-seq reads from six different tissues including leaf, flower, embryo, stem, root, and seed coat tissues, when aligned against the assembled sequence, showed that around 94.7% and 96% of the read pairs were aligned in the embryo sample and the remaining five samples respectively. The high quality of the assembly verified by CEGMA and BUSCO was corroborated by the ESTs and BAC-end sequences. Five whole BAC sequences (approximately 100 kb in length) were also completely covered in the scaffolds with minor in-dels. One of the BAC sequences included 12.6 kb of the *Tpn1* family transposon, *TpnA2*, suggesting that repetitive elements with high copy numbers and relatively long sequences were successfully determined. The SOAPdenovo assembly was also able to cover the five BAC sequences, but with large in-dels and an increased number of mismatches, indicating that per-base resolution was better in the assembly using PacBio reads. Telomeric repeats, centromeric repeats, and rDNA arrays were identified to further analyze the contiguity of the assembly. Thirty scaffolds, with telomeric repeat units (AAACCCT) in the range of 47.1 to 4,613.9 repeating units, were identified, of which 13 were completely covered by the tandem repeats and could not be incorporated into the linkage map. Pseudo-chromosomes 2, 6, 8, and 14 were found to have telomeric repeats at both the ends, while pseudo-chromosomes 3, 4, 5, 9, 10, 12, 13 and 15 had telomeric repeats at only one end. Although SOAPdenovo assembly captured 27

telomeric repeat sequences, the average size of the repeats was five times longer in the PacBio assembly. The ribosomal DNAs (rDNAs) in the order of 18S, 5.8S, and 25S rDNAs are found to occur in tandem arrays typically spanning several megabase pairs in regions called Nucleolar Organizer Regions (NORs) (VanBuren et al. 2015). Three scaffolds were found to contain 3 NOR units and 34 scaffolds had 2 NOR units. In total, 1,212 5S rDNA sequences were clustered in 21 scaffolds that were located away from the scaffolds carrying NORs. Centromeric repeats are known to span hundreds of kilobase pairs to several megabase pairs and are difficult to be assembled owing to their repetitive complexity. The centromeric monomer sequence was identified to be 173 bp in length. Using the monomeric sequence as a base, the longest centromeric repeat stretches were identified for each chromosome and the analysis revealed that two of the identified centromeric repeat stretches were longer than 100 kb.

3.2.4 Repeat analysis and identification of *Tpn1* transposons

Analysis using RepeatModeler showed that LTRs (long terminal repeats) comprised the largest portion of predicted repeats. The unclassified elements were mined for copia and gypsy repeats using RepBase. Copia and gypsy elements comprised 12.92% and 14.46% of the assembled genome (figure 3.4d). DNA class repeat elements represented 5.60% of the genome. Altogether, 63.29% of the genome was predicted to be repetitive (figure 3.4e). However, RepeatModeler was not able to predict *Tpn1* family transposons (figure 3.5). Hence, an in-house pipeline based on the presence of 5' and 3' TIRs as well as target site duplications (TSDs) was used to identify the *Tpn1* transposons. In total, 339 *Tpn1* transposons were identified with an average length of 7,081 bp (figure 3.4b). The smallest identified was 161 bp in length, while the longest was 40,619 bp. All the transposons had 3-bp TSDs, with the exception of one that had a

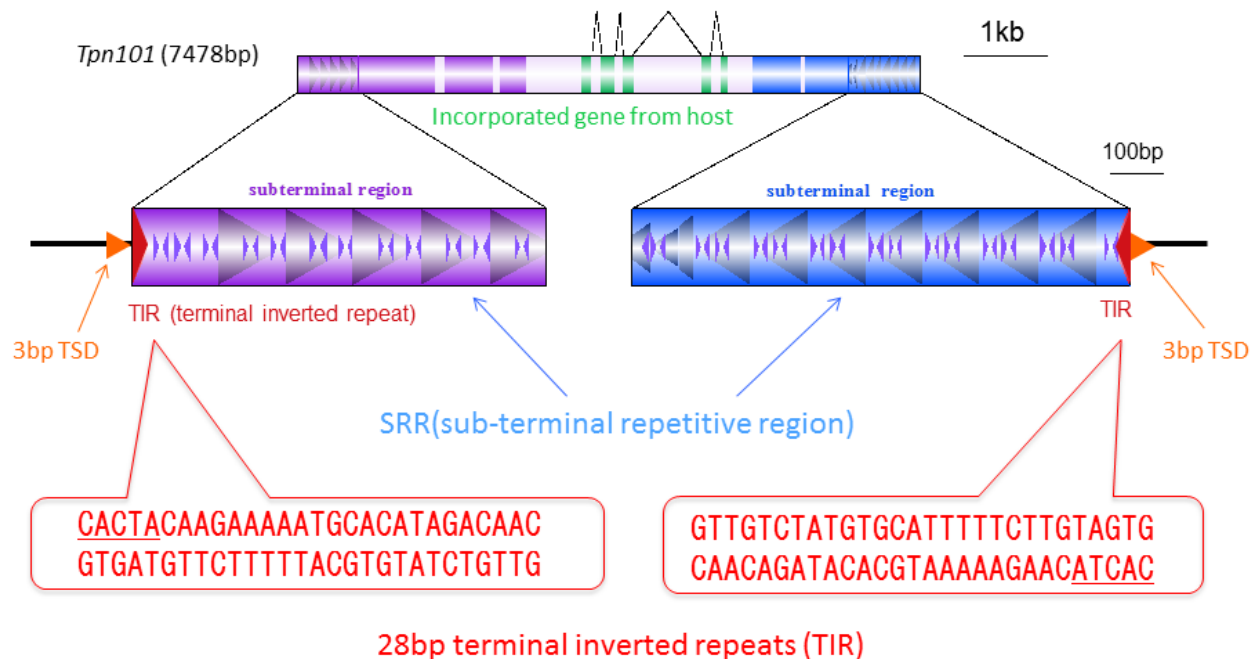


Figure 3.5. Structure of a *Tpn1* family transposon. The first and last identical 28 bp represent

Terminal Inverted Repeats (TIRs) across all the *Tpn1* transposons, flanked by typical 3-bp Transposon Site Duplications (TSDs). The TIRs are followed by sub-terminal repetitive regions (SRRs) and the region in-between can have incorporated genes from the host.

5-bp TSD. Fourteen of them had a mismatch in their TSDs. The TSDs tended to be AT rich, with at least one of A or T bp appearing in 95% of the TSDs. A nucleotide BLAST analysis revealed that most of the *Tpn1* transposons carried sub-terminal repetitive (SRR) sequences (figure 3.5) in both 5' and 3' terminal regions. Because TIR and SRR sequences are *cis*-requirements for transposition, it can be suggested that the *Tpn1* transposons are capable of transposition. However, thirty-two of the identified *Tpn1* transposons contained large rearrangements in SRR indicating that they are inactive. Twenty-nine *Tpn1* transposons were found within the 5' UTR and introns of the predicted genes, which could disrupt the function of those genes. It could be expected that the autonomous *Tpn1* family transposons carry both the TnpA and TnpD transposase coding sequences such as *En/Spm* and related autonomous transposons (Weil and Kunze 2002). A translated BLAST search against the 339 *Tpn1* family transposons, using TnpA and TnpD sequences from maize and snapdragon (Nacken et al. 1991) as queries, revealed that two transposons, named *TpnA3* and *TpnA4*, carried *TnpD* homologues, with two copies of *TpnA3* residing in the genome (figure 3.6). No

obvious *TnpA* homologues were identified in the 339 transposons. Also, no transcripts corresponding to *TnpA* and *TnpD* were found in the predicted genes or transcripts, indicating transcription of the transposase sequences was silenced in the line TKS. To identify autonomous transposons, the cDNA fragments for *TnpA* and *TnpD* homologues were isolated by a series of RT-PCRs from the line Q1072, where *Tpn1* actively transposes (figure 3.1m). A nucleotide BLAST search, against the whole genome sequence using the isolated cDNA sequences as queries, identified two transposons with *TnpA* and *TnpD* sequences, designated as *TpnA1* and *TpnA2* (figure 3.6). Of these, *TpnA2* is truncated in the genome, while the 5' terminus of *TpnA1* was not completely captured in the draft genome assembly. To characterize the entire *TpnA1* sequence, a BAC clone from TKS carrying *TpnA1* was isolated and sequenced. *TpnA1* is the putative autonomous element, because it carries apparently functional TIR and SRR sequences, in addition to the coding sequences of *TnpA* and *TnpD*. No transposons carrying *TnpA* coding sequences alone were found. In total, the genome contained two *TnpA* and five *TnpD* putative coding sequence copies (figure 3.6). The deduced amino acid sequences of the transposases were highly conserved in the genome and shared conserved domains with known transposases of *En/Spm* and snapdragon *Tam1* (Nacken et al. 1991).

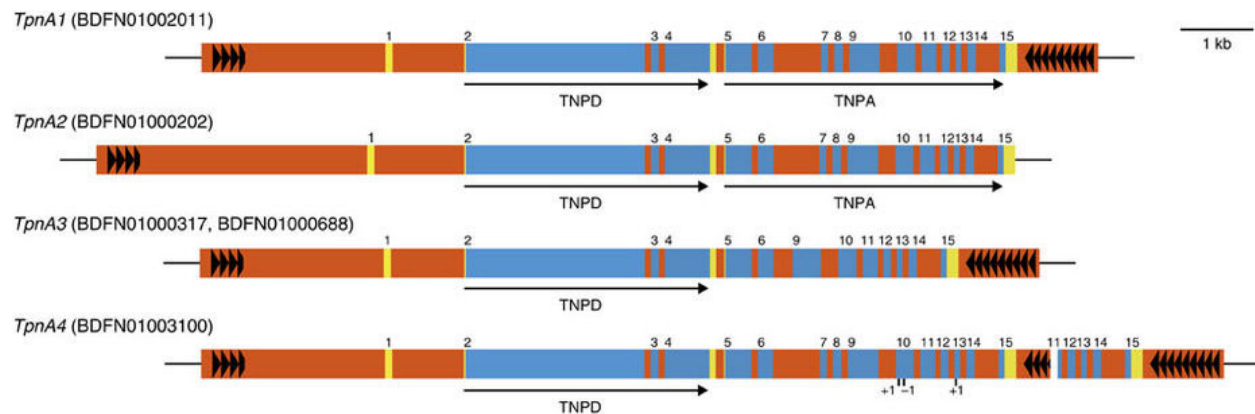


Figure 3.6. The Tpn1 family transposons encoding transposases. The orange, yellow and blue boxes indicate transposons, untranslated regions, and coding sequences respectively. The numerals above the blue boxes show exon numbers, and the arrows show the orientations of the transposase genes. The filled triangles are the 122-bp and 104-bp tandem repeats in the 5' and 3' sub-terminal regions respectively. *TpnA3* lacks exons 7 and 8, and *TpnA4* has a gap

represented by a white box, as well as three frame shift mutation indicated by the vertical bar with -1 and +1 in the exon 10 and 13.

3.2.5 Gene prediction and functional annotation

RNAseq data from leaf, flower, embryo, stem, root, and seed coat samples were used to assist in the process of gene prediction. A total of 42,783 gene models were predicted along with 45,365 transcripts, with tomato as the reference species, using Augustus (Stanke and Waack 2003). Of the transcripts, 44,916 contained a complete ORF with a start and a stop codon and 95.54% of the gene models could be assigned inside the 15 pseudo-chromosomes (figure 3.4c). Single exon genes accounted for 17.52% of the total. Two thirds of the transcripts were found to have less than or equal to 5 exons. A total of 61.99% of the gene models were annotated using the UniProt-Swiss-Prot database and in the remaining gene models, 16.93% were annotated using the UniProt-Trembl database. In addition, 61.92% of the gene models were assigned Pfam domain annotations. In total, the combined annotation procedure was able to assign annotations for 79.12% of the gene models.

3.2.6 Genome evolution

Protein sequences from rice (Ouyang et al. 2007) (monocotyledon outgroup), grape (Jaillon et al. 2007), kiwifruit (Huang et al. 2013) (from the Asterid clade), along with Solanales order members tomato (Tomato Genome Consortium 2012), potato (Potato Genome Sequencing Consortium et al. 2011), and capsicum (Kim et al. 2014) were used for gene family clustering using the OrthoMCL pipeline (Li et al. 2003) to infer phylogenetic relationships. A total of 1,353 single copy orthologs corresponding to the seven species were extracted from the clusters and were filtered to 214 single copy orthologs. Phylogenetic inference using RaxML (Stamatakis 2014) reconfirmed the phylogenetic arrangement of *I. nil*. BEAST (Bouckaert et al. 2014) estimated the divergence of *I. nil* from the other Solanales members to be around 75.25 million years ago (MYA), which was very close to the estimation from the TTOL (Hedges et al. 2015) database (figure 3.7a). Also, *I. nil* was estimated to have separated from kiwifruit

approximately 105.8 MYA. Divergence time estimates obtained for the other species also corresponded well with the estimations from TTOL database.

Syntenic analysis using MCScanX revealed that 2,275 syntenic gene blocks were found to contain 17,376 paralogous gene pairs in the assembled pseudo-chromosomes (figure 3.4f). The number of synonymous substitutions per synonymous site (Ks) of the gene pairs in the syntenic regions was plotted against the percentage of corresponding genes to infer and compare whole genome duplication (WGD) events in *I. nil*. *I. nil* and tomato were found to share 47.05% of syntenic orthologs in a 1:1 ratio, whereas, the percentage of kiwifruit orthologs in a 1:1 ratio across tomato and *I. nil* were 34.89% and 36.01% respectively. Apart from the 1:1 orthologs, both tomato and *I. nil* shared large numbers of syntenic blocks with kiwifruit, possibly because of the two recent duplication events in kiwifruit (Huang et al. 2013), which was also evident from the two Ks peaks specific to kiwifruit (figure 3.7b). A recent WGD event was estimated to have occurred in Solanaceae members, approximately 71 ± 19.4 MYA (Tomato Genome Consortium 2012; Potato Genome Sequencing Consortium et al. 2011). A Ks peak from syntenic paralogs of tomato, corresponding to the above mentioned WGD event, was found to occur after the speciation peak between tomato and *I. nil* (figure 3.7b), suggesting that the event was specific to the Solanaceae and should have occurred reasonably close, following the divergence which was estimated to be 75.25 MYA (figure 3.7a). The analysis also revealed a Ks peak specific to *I. nil* indicating that a WGD event had also occurred, independently, in the Convolvulaceae family (figure 3.7b).

Gene family clustering showed that 10,549 core gene families were shared by all four species of the Solanales members (figure 3.8). *I. nil* contained 2,242 unique gene families not shared by Solanaceae members, whereas the Solanaceae members shared 2,681 more gene families than *I. nil*. *I. nil* specific gene families had expansions of paralogs (mean value of 4.92) compared to gene families which had orthologous relationships with the other Solanales (mean value of 1.79). *I. nil* specific gene families were found to be enriched with pollination and reproductive process related gene ontology (GO) terms.

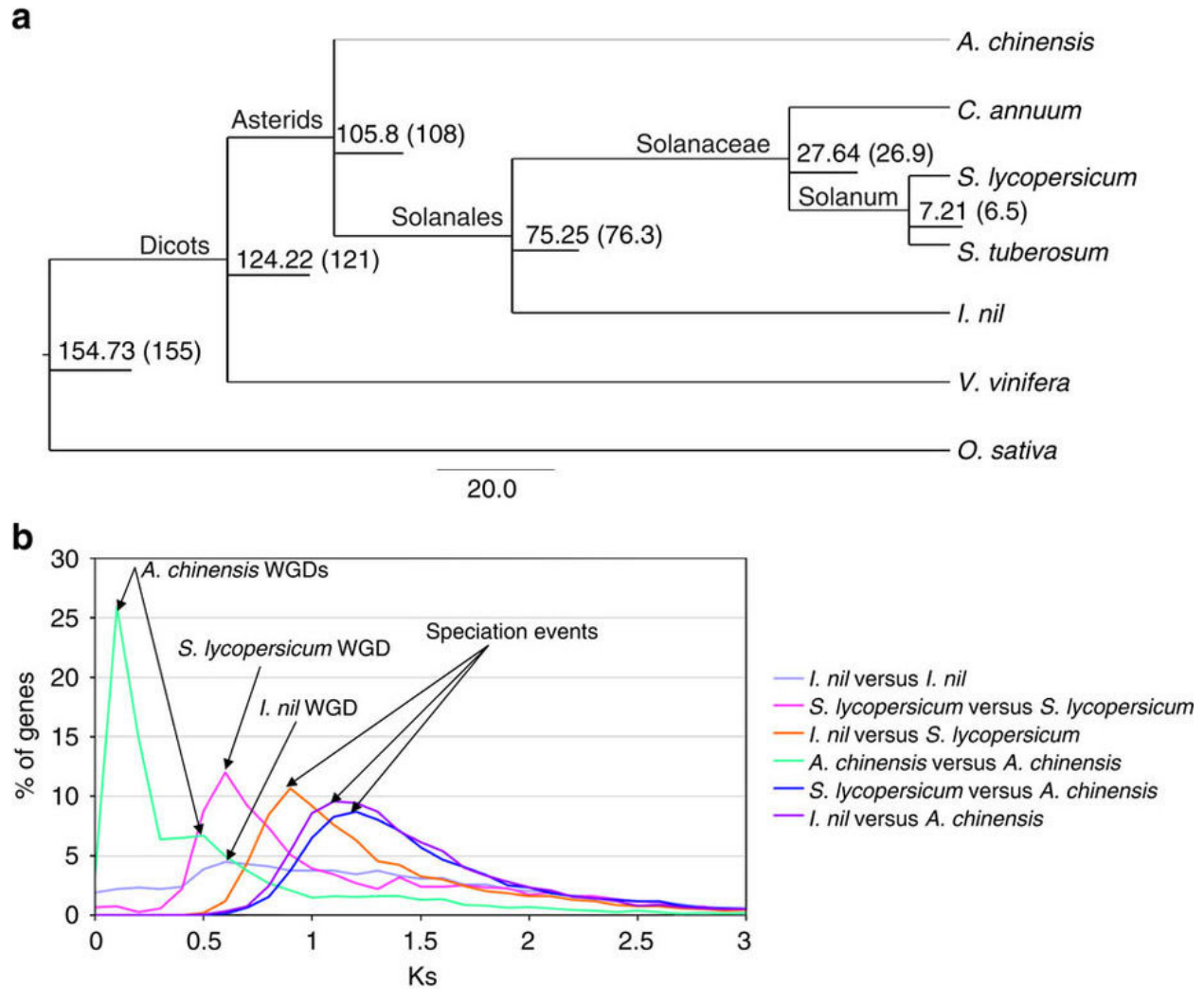


Figure 3.7. Genome evolution. (a) Divergence time estimation using BEAST. The scale bar 20.0 corresponds to Myr ago. The node labels indicate estimated divergence times in Myr ago, with estimations from TTOL in parentheses, and the branch labels indicate the clades within the branch. (b) Distribution of Ks values against the corresponding percentage of syntenic genes, comparing *I. nil* and *S. lycopersicum* against *A. chinensis*. The colours violet, magenta, orange, turquoise, blue, and purple represent the Ks values of *I. nil* versus *I. nil*, *S. lycopersicum* versus *S. lycopersicum*, *I. nil* versus *S. lycopersicum*, *A. chinensis* versus *A. chinensis*, *S. lycopersicum* versus *A. chinensis*, and *I. nil* versus *A. chinensis* respectively. Speciation events among the three species and lineage specific WGDs are highlighted with arrows.

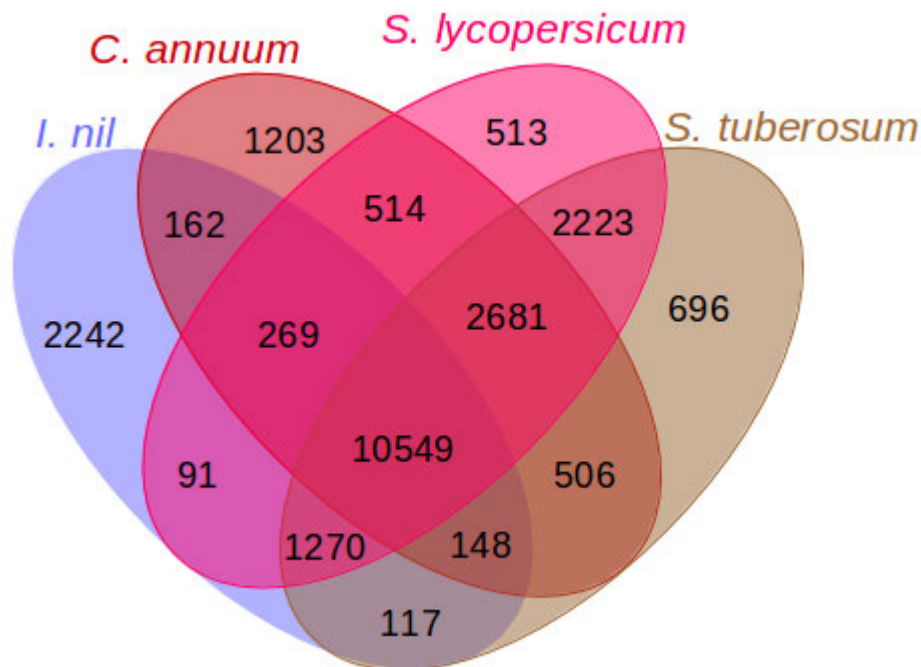


Figure 3.8. Venn diagram depicting the gene family clustering of the Solanales.

3.3 Discussion

The advent of second and third generation sequencers have fast-tracked genome assemblies of a variety of species. The current study has utilized nearly the complete potential of recent sequencing tools and has culminated in a highly contiguous genome assembly of *I. nil*. A few of the recent genome assembly projects have used PacBio data to supplement Illumina based contig assemblies, and a mild improvement in the lengths of the assembled scaffolds have been observed. However, in this study, PacBio data were used as a base to construct contig assemblies, while Illumina data were used to supplement the assembly, resulting in a marked increase in the lengths of the assemblies observed (scaffold N50 length of 2.88 Mb). The average contig N50 length for all published genomes is 50 kb (Michael and Jackson 2013), whereas *I. nil* had a contig N50 length of 1.87 Mb. The 7-kb size selected inserts of the PacBio sequence data was especially helpful in resolving *Tpn1* transposons, whose average length was approximately 7 kb, and the assembly also revealed complex repeats like telomeric repeats, rDNA clusters, and centromeric repeats. However, a better resolution of such

repeats was obtained in *Oropetium thomaeum* (VanBuren et al. 2015) genome assembly, possibly owing to the 15-kb lower end insert size selection, explaining the importance of longer read lengths in obtaining near-perfect assemblies. The potential of PacBio sequence data in long, eukaryotic genomes has been further showcased in the draft genomes of *Gorilla gorilla* (Gordon et al. 2016) (scaffold N50 of 23.1 Mb), *V. angularis* (Sakai et al. 2015) (scaffold N50 of 3.0 Mb), *O. thomaeum* (VanBuren et al. 2015) (contig N50 of 2.4 Mb) and *Lates calcarifer* (Vij et al. 2016) (scaffold N50 of 1.19 Mb). A rapid increase in PacBio sequencing for similar large-scale assemblies can be expected in the near future.

The draft genome has enhanced the understanding of the genetic basis of floricultural traits in *I. nil*. It was possible to catalog *Tpn1* family transposons along with the putative autonomous element, *TpnA1* (figure 3.6). The *Tpn1* transposons were distributed across all 15 chromosomes and one copy per 126 genes (339 copies per 42,783 genes) was observed. Most of them retain apparently functional *cis* elements, TIRs, and SSRs suggesting that they are capable of transposition. In addition, *TpnA1*, *TpnA2*, *TpnA3* and *TpnA4* also encode putative transposases (figure 3.6). These features should be the basis for *Tpn1* transposons to act as the major mutagen in the mutant cultivars of *I. nil*. The *ct* mutation is traditionally called as “*uzu*”, and the key mutation of the barley green revolution was also named after *I. nil*'s *uzu* (*contracted*) because of their common semi-dwarf phenotypes (Chono et al. 2003). It was also possible to identify the strong candidate for the *CT* gene by using the combination of the draft genome and classical linkage map, demonstrating the capability of the assembled draft genome. It can be expected that the draft genome will maximize future use of the abundant mutants and genetic knowledge of *I. nil*. Comparative analysis revealed that each of *I. nil*, tomato and kiwifruit had independent WGD events in their genomes, even though they all belonged to Asterids. One of the major reasons for the fruit-specific gene neo-functionalization in tomato is reported to be because of a large number of genes triplicated from the recent WGD event (Tomato Genome Consortium 2012). It could be assumed that the lineage specific WGDs, observed in *I. nil*, tomato and kiwifruit, could have had a major role in shaping the diverse evolution of these plant species. Being the only pseudo-chromosomal level genome assembly in Convolvulaceae, the genome

sequence, linkage map and DNA clones developed in this study will facilitate not only future studies on *I. nil* and its related species, but also aid comparative genomic studies in Solanales.

3.4 Methods

3.4.1 Plant materials and sequencing

An individual of *I. nil* Tokyo Kokei Standard (TKS) line was propagated clonally and genomic DNA isolated from the flower petals of young buds was used for whole genome sequencing. A 20 kb library (BluePippin size selection at 7 kb) for P5-C3 chemistry was constructed. Ninety SMRT cells were first sequenced on PacBio RS II system. Furthermore, sequencing libraries were prepared using the Illumina TruSeq DNA Sample Prep kit and Nextera Mate Pair Sample Prep kit. Two paired-end and six mate-pair libraries were constructed and sequenced on the Illumina HiSeq2500, with a read length of 150 bp. To validate the accuracy of the reference assembly, end sequencing of a JMHibA BAC library was carried out using the ABI 3730xl DNA Analyzer. The TKS line was also used for construction of cDNA and BAC libraries for EST sequences. The genome size was estimated using a flow cytometer. For transcriptome analysis, tissues from flowers, stems, leaves, and seed coat (maternal tissue) of the individual; embryos and roots of its self-pollinated progeny were used, and the mRNA-Seq libraries were constructed using the Illumina TruSeq mRNASeq Sample Preparation Kit (version 2) from 600 ng of total RNA, collected from each of the indicated tissues, according to the manufacturer's instructions. Sequencing was conducted as paired end reads of 101 bp on Illumina HiSeq2000. An F2 hybrid population of *I. nil* lines TKS × Africa (Q63) was used to construct a RAD-tag based linkage map. Two double-digested RAD libraries (Ly et al. 2012) were prepared, as described before (Sakaguchi et al. 2015) with slight modifications of the restriction enzymes and adapters. The restriction enzyme pairs were *NdeI/BglII* and *MseI/BglII* (New England Biolabs). The prepared libraries were sequenced on an Illumina HiSeq2500 platform as 151-bp single-end reads. Forty-three *I. nil* lines were also used to characterize the *CT* gene. The *a3-flecked* mutant, Q1072,

was used to isolate *TnpA* and *TnpD* mRNA, and an authentic *s* mutant line, Q721, was used for genetic complementation test for the *kbt* mutant, Q837.

3.4.2 Genome assembly

Before assembling the Illumina short read data set, adapters were trimmed using Cutadapt v1.2.1 (Martin 2011). Using k-mer frequencies of the short insert libraries, SOAPdenovo2's error correction module (v2.01) was used to correct errors with a low frequency cutoff of 80 kmers and a minimum trimmed read length of 50 bp. The processed reads were assembled, scaffolded and gap-filled using SOAPdenovo2 assembler v2.04 (Luo et al. 2012) with a k-mer value of 115. The work-flow (figure 3.2) of the assembly of longer PacBio reads began with contig assembly using HGAP3 pipeline (Chin et al. 2013) from SMRTanalysis v2.3.0. For HGAP3, the following parameters were used: PreAssembler Filter v1 (minimum sub-read length = 500 bp, minimum polymerase read length = 100 bp, and minimum polymerase read quality = 0.80); PreAssembler v2 (minimum seed read length = 6000 bp, number of seed read chunks = 6, alignment candidates per chunk = 10, total alignment candidates = 24, minimum coverage for correction = 6, and blasr options = "minimum read length = 200 bp, maximum score = 1000, maximum LCP length = 16, and noSplitSubReads"); AssembleUnitig v1 (genome size = 750 Mb, target coverage = 30, overlap error rate = 0.06, minimum overlap length = 40 bp, and overlapper k-mer = 14); Mapping (Maximum number of hits per read = 10, maximum divergence = 30%, minimum anchor size = 12 bp, and pbalgn options = "random number generator initializing seed =1, minimum accuracy = 0.75, minimum length = 50 bp, useQuality, and placeRepeatsRandomly"). The polymerase N50 and the sub-read N50 at the assembly phase was recorded as 12.3 kb and 10.5 kb respectively. The initial assembly was followed by two rounds of polishing by Quiver. To correct PacBio residual errors, the Illumina reads were aligned against the contigs using BWA v0.7.12 (Li and Durbin 2009). After sorting the alignments and marking duplicates using Picard tools v2.1.1 (<http://picard.sourceforge.net/>), Genome Analysis Toolkit v3.5 (McKenna et al. 2010) was used to perform local realignment around in-dels and to call variants using the module, HaplotypeCaller. Variant filtering was performed using the expression:

“DP<20.0 || QD<2.0 || FS>60.0 || MQ<40.0”. The homozygous in-dels were treated as errors, while the heterozygous in-dels were replaced with Illumina read bases in the assembled contigs using FastaAlternateReferenceMaker. MUMmer v3.23 (Kurtz et al. 2004) was used to identify and remove contigs, if more than 50% of their sequence was either mitochondrial or chloroplast sequence. Smaller contigs, which had greater than 98% sequence coverage in other contigs with at least 98% sequence identity, were also removed from the assembly. The contigs were then scaffolded with the help of 15 and 20 kb Illumina mate pair read libraries, with a minimum of 10 paired read witness links, without the default scoring option, using BESST scaffolder (Sahlin et al. 2014). A first round of splitting chimeric scaffolds was performed before gap-filling. PacBio reads were utilized to gap-fill the scaffolds using PBJelly (English et al. 2012) with the blasr options “minimum seed length = 8 bp, minimum percent Identity = 70%, report number of best alignments =1, number of candidates for best alignment = 20, maximum subread score = 500, and noSplitSubreads”. If the flanking sequences, at the gap junctions, had an overlap of more than 1 kb, those gaps were filled by joining the flanking sequences manually.

3.4.3 Linkage map construction and pseudo-chromosome assignment

The RAD-seq technique (Baird et al. 2008) was employed to sequence 2 parent samples (TKS and Africa lines) and 207 progeny samples. The Illumina short reads from the parent samples and progeny samples were aligned against the assembly using BWA v0.7.12. The reads which were not tagged as uniquely mapped, and those which did not have the requisite restriction enzyme cut site were filtered out. STACKS v1.37 (Catchen et al. 2011) was used to identify SNP and the following two criteria were used to filter markers: a) Each marker should be present in at least 80% of the samples, and b) Each sample should have at least 80% of the markers. Also, 150 bp flanking regions from either side of each SNP location was extracted from the assembly and was aligned against each other using BLAST to check for repetitive regions. Any region with an alignment length of longer than 150 bp were filtered out. Onemap (Margarido et al.

[2007](#)) was used to create linkage maps with an LOD score of 30. TMAP (Cartwright et al. 2007) was used to reorder the linkage map, along with manual inspection. The original classical map contained 10 linkage groups (LGs), although *I. nil* has 15 chromosomes (Yasui 1928). The marker genes from seven of the 10 LGs of the classical map (Hagiwara 1956) were mapped in the current RAD-based linkage maps, and the LGs were named 1 to 6 and 10 correspondingly. Because two LGs in our RAD-marker based map corresponded to LG3 in the classical map, they were accordingly assigned as LG3 and LG11 with the corresponding marker genes being *DUSKY* and *SPECKLED* respectively. This coincided with the fact that the *DUSKY* and *SPECKLED* genes were mapped on the different linkage groups in the older linkage analysis (Imai 1929). LGs 7 to 9 and 12 to 15 were numbered randomly.

3.4.4 Mis-assembly elimination and assembly validation

Before anchoring scaffolds to pseudo-chromosomes, chimeric assemblies were first resolved using linkage maps and BAC-end sequences. Contigs were first aligned against the scaffolds using the NUCmer module within MUMmer v3.23 (Kurtz et al. 2004) to identify the contig locations in the scaffolds. If a scaffold contained a stretch of linkage markers pointing to two different linkage groups with a scaffold junction (N) in between, it was considered a chimera and was split into two at the junction. If the mis-assembly occurred at the contig level, the bac-end alignments were used as a key in splitting chimeric contigs. Based on the order of the linkage maps, the scaffolds were merged using Ns as gaps to form pseudo-chromosomes. The orientations of the scaffolds were determined using the marker order, and the orientations of scaffolds with inadequate markers were ignored but included as part of the pseudo-chromosomes. The circular view of the genome was generated using Circos (Krzywinski et al. 2009). CEGMA v2.5 (Parra et al. 2007) and BUSCO (Simão et al. 2015), two commonly used genome assembly validation pipelines, were used to validate the completeness of genes in the assembly. BLAT was used to align ESTs and BAC-end paired reads against the assembly. In-house scripts were written, which calculated paired BLAT scores from both the BAC-end read pairs and picked up the best paired hits based on the combined score. RNA-seq reads were trimmed using Trimmomatic v0.33 (Bolger [et](#)

al. 2014) and TopHat v2.1.0 (Kim et al. 2013) was used to align the RNA-seq reads with default parameters. Tandem repeats finder v4.07b (Benson 1999) was used to identify tandem repeats by assigning values 1, 1, 2, 80, 5, 200, and 2000 bp to match weight, mismatch weight, indel weight, match probability, indel probability, minimum score, and maximum period size respectively. Inspection of short tandem repeats at the ends of the contigs revealed the monomer “AAACCCT” to be the telomeric repeat. Manual inspection of the tandem repeats also revealed the centromeric repeat monomer to be of approximately 173 bp in length. A tetramer centromeric repeat sequence was used to search against the whole output of tandem repeats finder using BLAST. The BLAST alignment results were screened for monomer sequences closer to 173 bp length to identify centromeric repeat candidates. Tandem centromeric repeat stretches (> 3 kb) were merged, when they were within a distance of 50 kb and the longest stretch for every chromosome was identified to approximate the putative position of the centromeres. Infernal v1.1.1 (Nawrocki and Eddy 2013) was used to identify rDNA clusters by searching against Rfam v12.0.

3.4.5 Repeat analysis and gene prediction

De novo repeat identification was done using RepeatModeler v1.0.7 which combines RECON and RepeatScout (Price et al. 2005) programs, followed by RepeatMasker v4.0.2 to achieve the final results. *Tpn1* family transposons were detected using the following approach: The TIRs of the *Tpn1* transposons (28 bp in length) were searched using BLAST; the aligned TIR coordinates were sorted by their locations; if two nearby TIRs contained the same TSDs (3 to 5 bp), they were nominated as *Tpn1* family elements. The sub-terminal repeats were also identified using BLAST to determine the orientation of the *Tpn1* elements. A translated BLAST search against the identified transposons using TnpA and TnpD sequences from maize and snapdragon as queries revealed non-autonomous *TpnA3* and *TpnA4*. To isolate autonomous *Tpn1* transposons, the cDNA fragments of *TnpA* and *TnpD* homologue were isolated from Q1072. Using the isolated cDNA sequences as query, *TpnA1* and *TpnA2* were identified by screening against the assembled scaffolds using BLAST. As the 5' terminal of *TpnA1* was not assembled completely in the genome sequence, a BAC clone from TKS

carrying the entire *TpnA1* sequence was isolated and characterized. Repeats obtained by both the above mentioned approaches were masked for gene prediction. The genes harboring *Tpn1* transposon insertions were identified using the gene and the transposon co-ordinates and were annotated using the web version of BLASTX. Gene models were predicted using Augustus v3.2.2 (Stanke and Waack 2003) with tomato as the reference species, using hints from RNA-seq alignments, and also allowing prediction of untranslated regions (UTRs). Because of the scarcity of complete CDs of *I. nil* in public databases, independently, Augustus was also used to predict gene models, after training using CEGMA predicted genes, and the procedure resulted in more than 55,000 gene models. The 189 complete CDs sequences already available in NCBI were downloaded and compared against the predicted gene models using BLAT. Tomato based gene models showed that 116 out of 189 CDs were perfectly complete, whereas CEGMA trained gene models showed that only 61 out of 189 CDs were complete and hence, the tomato based gene predictions were used for further analysis. The gene models were translated to proteins and were aligned against proteins from UniProt-Swiss-Prot and UniProt-TrEMBL databases using NCBI BLAST+ v2.2.29 (Altschul et al. 1990). Using an e-value cut-off of e-5 for annotation, alignments from the Swiss-Prot database were given preference ahead of the TrEMBL database. In other words, TrEMBL annotations were assigned for only those entries without a Swiss-Prot annotation. To extract protein domain annotations, InterProScan v5.19-58.0 (Jones et al. 2014) was used to assign Pfam domains to the gene models. GO terms were extracted from the Pfam annotations as well as UniProt annotations.

3.4.6 Comparative analysis

Protein sequences were downloaded from tomato, potato, capsicum, grape, and rice. OrthoMCL v2.0.9 (Li et al. 2003) was used to construct orthologous gene families, with an inflation parameter of 1.5. Prior to OrthoMCL, an all-vs-all BLAST was performed to find similar matches from different species, and the BLAST results were filtered with an e-value cut-off of e-5, a minimum alignment length of 50 bp, and a percentage match cut-off of 50. AgriGO (Du et al. 2010) was used for finding GO enrichments in *I. nil* specific gene families. MAFFT v7.221 (Katoh et al. 2002) was used for multiple

sequence alignments of the resultant single copy orthologs, and trimAl v1.4 (Capella-Gutiérrez et al. 2009) was used to remove poorly aligned regions and to back-translate protein alignments to CDs alignments. The alignments were filtered using the criteria that coding sequences from each of the species should have covered at least 95% of the multiple sequence alignments, thereby, reducing the gaps to less than 5% of the alignments. RAxML v8.2.4 (Stamatakis 2014) was used to build Maximum Likelihood phylogenetic trees using the GTRGAMMA model, with rice as an out-group. BEAST v2.3.1 (Bouckaert et al. 2014) was used to estimate the divergence times with the Jules Cantor substitution model, with a lognormal relaxed clock and Yule model. The chain length of MCMC analysis was 10,000,000. TimeTree (Hedges et al. 2015) is a public database containing divergence time estimates from various publications along with their own estimation. These estimates, ignoring the outliers, were used for selecting the range of lower and upper uniform calibration priors. The lower and upper calibration values, in million years, were chosen as 1.9–12.8, 15.6–41, 58.6–95.1, 93.3–128.3, 101.2–156.3, and 110–216 for the most common ancestor of the seven species belonging to *Solanum*, Solanaceae, Solanales, asterids, dicotyledons, and all plants respectively. FigTree (<http://tree.bio.ed.ac.uk/software/figtree>) was used to view the phylogenetic trees. Synteny analysis of the 15 pseudo-chromosomes against the chromosomes of other species was performed using the MCScanX toolkit (Wang et al. 2012) utilizing options such as maximum gaps = 15 genes, minimum evalue = 1e-10, and minimum match score = 50. PAML's (Yang 2007) yn00 module was used to calculate the Ks values of the orthologous and paralogous gene pairs in the syntenic regions using Nei-Gojobori method. The assembled genome was compared against the genome of *I. trifida*.

3.4.7 Data availability

All sequencing data used in this work are available from the DNA DataBank of Japan (DDBJ) Sequence Read Archive (DRA) under the accession numbers DRA001121, DRA002710, and DRA004158 for PacBio and Illumina sequencing, DRA002647 for RNA-seq, and DRA002758 for RAD-seq. The genomic assembly sequences are available from accession numbers BDFN01000001–BDFN01003416 (scaffolds), and

two organelle DNA sequences are available from accession numbers AP017303–AP017304. The EST and BAC-end sequences are available from accession numbers HY917605–HY949060 and GA933005–GA974698, respectively. Accession numbers for the *CONTRACTED* gene, its mutant alleles, and *Tpn1* family elements are LC101804–LC101815. All the above data has been released for public access, as of August 31, 2016, and the accessibility has been verified by the authors.

Chapter 4

Comprehensive evaluation of non-hybrid genome assembly tools for third generation PacBio long-read sequence data

Long reads obtained from third generation sequencing platforms can help overcome the long-standing challenge of the *de novo* assembly of sequences for the genomic analysis of non-model eukaryotic organisms. Numerous long read-aided *de novo* assemblies have been published recently, which exhibited superior quality of the assembled genomes in comparison to those achieved using earlier second-generation sequencing technologies. Evaluating assemblies is important in guiding the appropriate choice for specific research needs. In this study, we evaluated ten long-read assemblers using a variety of metrics on PacBio datasets from different taxonomic categories with considerable differences in genome size. The results allowed us to narrow down the list to a few assemblers that can be effectively applied to eukaryotic assembly projects. Moreover, we highlight how best to use limited genomic resources for effectively evaluating the genome assemblies of non-model organisms.

4.1 Background

Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) and Oxford Nanopore sequencing technologies are the two widely used third-generation, single-molecule sequencing (SMS) technologies, which can generate average read lengths of several thousand base pairs. SMRT sequencing technology suffers from high error rates reaching up to 15% (Lee et al. 2016); however, since these errors are random, high-quality error-corrected consensus sequences can be generated with sufficient coverage.

Application of SMRT sequencing to eukaryotic genomes (Hoshino et al. 2016; Korlach et al. 2017; Allen et al. 2017; Lan et al. 2017; Gordon et al. 2016; Sakai et al. 2015; Conte et al. 2017; Vij et al. 2016; Weissensteiner et al. 2017; Bickhart et al. 2017; Shi et al. 2016; Jiao et al. 2017; Pendleton et al. 2015; Du et al. 2017; VanBuren et al. 2015; Jiao et al. 2017; Steinberg et al. 2016) has already demonstrated the obvious advantages provided by long reads in *de novo* assembly, such as higher contiguity, lesser gaps, and fewer errors. The assembled contigs of recently assembled plant and animal genomes can be routinely seen to achieve an N50 of 1 Mb using SMS data. Hence a significant rise in the number of genomes sequenced using SMS technologies is imminent, raising the need for evaluation of the available long-read assemblers. Large-scale evaluation studies such as GAGE (Salzberg et al. 2012), GAGE-B (Magoc et al. 2013), Assemblathon (Earl et al. 2011), and Assemblathon 2 (Bradnam et al. 2013) have been attempted with short-read assemblers, providing conclusions that serve as a useful guide for the *de novo* assembly of a given target organism. Although such evaluations have also been attempted for SMS data, these studies were either focused on bacterial and smaller eukaryotic genomes (Sović et al. 2016; Istace et al. 2017), or were not sufficiently comprehensive to cover all of the available non-hybrid long-read assemblers (Koren et al. 2017; Vaser et al. 2017; Xiao et al. 2016), while others are already outdated because of continuous improvements in the technology (Cherukuri and Janga 2016; Liao et al. 2015). Also genome size was found to correlate with contiguity in long-read assemblies (Jiao et al. 2017), hence, diverse genome sizes can help differentiate the effect of the assemblers on each dataset. In this study, we attempted to comprehensively evaluate three important features—contiguity, completeness, and correctness (Lee et al. 2016)—of long-read assemblers (table 4.1), using SMRT data of a bacterium (*Escherichia coli*, ~5 Mb), protist (*Plasmodium falciparum*, ~23 Mb), nematode (*Caenorhabditis elegans*, ~105 Mb), and plant (*Ipomoea nil*, ~750 Mb). We also designed a pipeline (figure 4.1) for assembling the data and evaluating the results of different assemblers, which can be applied to both model organisms as well as to non-model organisms with limited genomic resources.

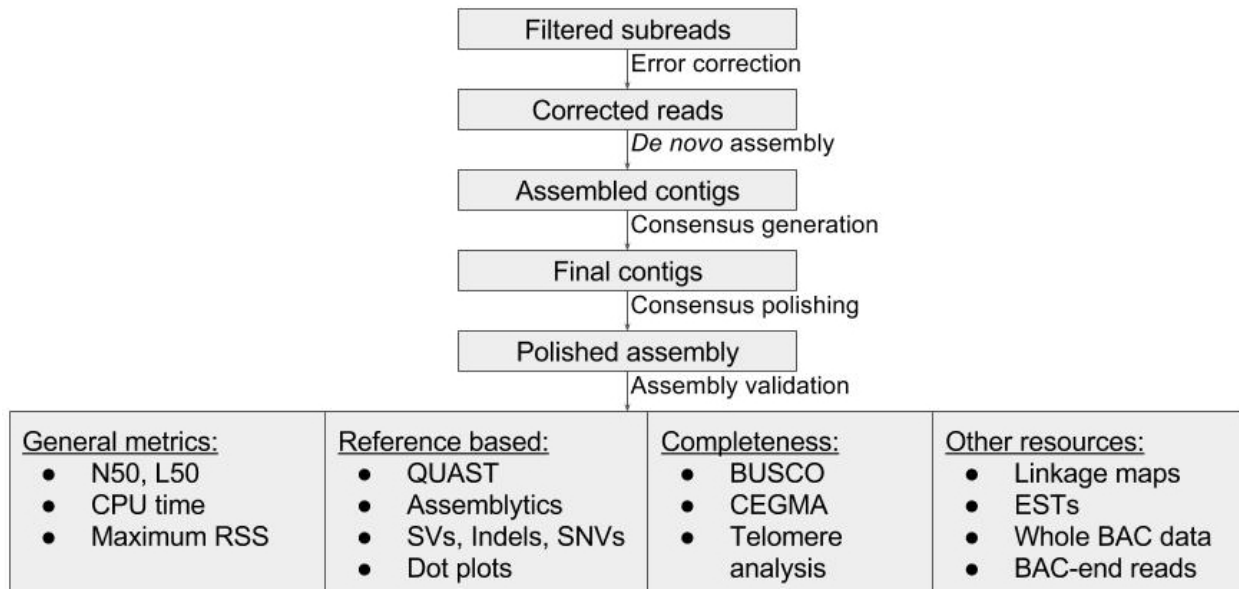


Figure 4.1. Evaluation pipeline.

4.2 Materials and methods

4.2.1 Long-read assembly pipelines

Overlap Layout Consensus (OLC) approach, de Bruijn graphs, and string graphs are the commonly used algorithms for *de novo* assembly (Myers 2014; Simpson and Pop 2015; Chen et al. 2017; Chaisson et al. 2015). The advent of SMS data introduced a new challenge in *de novo* assembly because of the high error rates. Hence, application of de Bruijn graphs was rendered unfeasible (Kamath et al. 2017), bringing back the OLC approach along with the string graphs to higher prominence. The longer the reads, the more efficient the assembly using the OLC approach, resulting in a linear increase in contiguity (Koren et al. 2012). Although second generation sequencing (SGS) reads were initially used for correcting long reads (Chin et al. 2013), most of the current long-read OLC pipelines follow a hierarchical approach (figure 4.2), exclusively using SMS data as follows: a) select a subset of longer reads as seed data; b) use shorter reads to align against the longer seed data as reference, and correct sequencing errors by consensus of the aligned reads; c) use the error-corrected reads for a draft assembly; and d) obtain a polished consensus of the draft assembly (Chin et al. 2013; Li 2016). The procedure to identify overlaps has been the key difference in most long-read

assemblers, and some of the overlap detection methods have been evaluated previously (Chu et al. 2017).

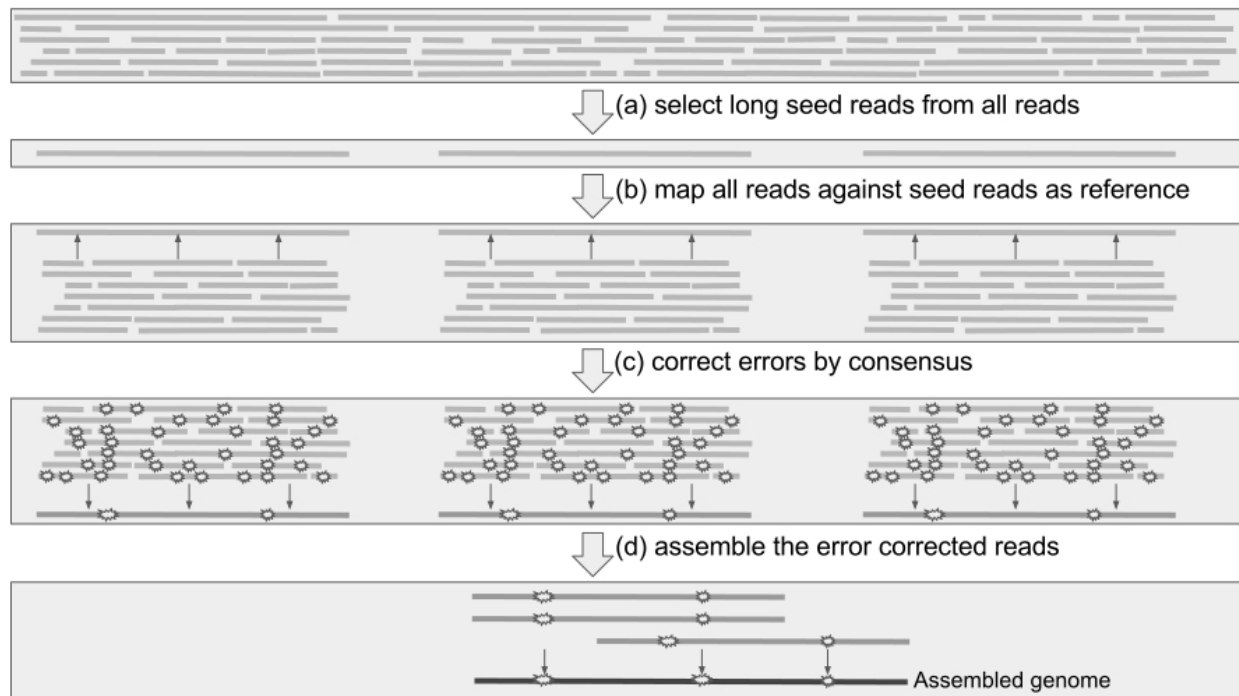


Figure 4.2. Hierarchical pipeline for OLC assembly approaches. Errors are displayed in Step C, which become reduced in number in the corrected reads. After assembly, a consensus polishing step, which is not shown in the figure, will also be performed as part of the hierarchical pipeline.

The long-read assemblers assessed in the present work are briefly summarized below.

4.2.1.1 Hierarchical Genome Assembly Process (HGAP)

HGAP (Chin et al. 2013) was one of the first hierarchical pipelines to exclusively use SMS reads for assembling a genome. Higher-quality pre-assembled reads with around 25–30× coverage are generated by aligning shorter reads against longer seed reads. The pre-assembled reads are then fed to the celera assembler (Myers et al. 2000) to obtain a draft assembly, followed by applying a consensus polishing procedure called quiver. BLASR (Chaisson and Tesler 2012) is used for aligning candidate overlaps, which are identified using an FM-index search and clustering of k-mer hits. The slower BLASR-based pipeline was replaced by FALCON in the latest version (v4). To

distinguish between HGAP v3 and v4, the version used in the present evaluation is referred to as HGAP3.

4.2.1.2 PBcR

PBcR (Berlin et al. 2015) also follows the hierarchical approach using MinHash Alignment Process (MHAP) for overlap detection. To identify k-mers shared between overlapping reads, without performing any alignments, k-mers of query reads are converted to integer fingerprints using multiple hash functions. The minimum values from the multiple hash functions are used to create a set called MinHash sketch, for each read. MHAP then calculates the Jaccard similarity index by comparing the sketches of query reads to identify overlap candidates. Like HGAP3, the assembly of the corrected reads is performed using the celera assembler.

4.2.1.3 Canu

Canu (Koren et al. 2017) is a fork of the celera assembler and improves upon the earlier PBcR pipeline into a single, comprehensive assembler. Highly repetitive k-mers, which are abundant in all the reads, can be non-informative. Hence term frequency, inverse document frequency (tf-idf), a weighting statistic was added to MinHashing, giving weightage to non-repetitive k-mers as minimum values in the MinHash sketches, and sensitivity has been demonstrated to reach up to 89% without any parameter adjustment. By retrospectively inspecting the assembly graphs and also statistically filtering out repeat-induced overlaps, the chances of mis-assemblies are reduced.

4.2.1.4 FALCON

FALCON (Chin et al. 2016) is a hierarchical, haplotype-aware genome assembly tool. The sequence data are split into blocks for comparison using daligner (Myers 2016). Daligner first compiles a list of k-mers, along with their read identifiers and read coordinates, and then sorts them lexicographically. Identical k-mers from each block are merged into a new list containing both the query identifiers and their coordinates. A second sorting procedure, accounting for the query coordinates, places neighboring matches adjacent to each other, resulting in the identification of overlap candidates. A directed string graph is created from the alignment of the overlaps, with a collapsed diploid-aware layout, while maintaining the heterozygosity information.

4.2.1.5 HINGE

HINGE (Kamath et al. 2017) is one of the few assemblers not requiring an error-correction step. DALIGNER is used for overlap detection. The key innovation of this assembler is the placement of hinges to mark repeat regions that are not spanned by longer reads. Repeats are identified using the coverage gradients of the alignments, and an in-hinge and an out-hinge are marked on the reads, which are on the boundaries of unbridged repeats. Only two reads per repeat region, which have the longest overlap within the repeat, are chosen for placing the hinges. When a repeat is spanned by a completely bridged read, the other overlapping reads are marked as poisoned and not considered for hinge placing, thereby separating bridged repeats. Hinge-aided greedy graphs are used to resolve repeat junctions before obtaining a consensus.

4.2.1.6 Miniasm

Miniasm (Li 2016) was the first long-read assembler to not employ error correction and hence is fast. Minimap is used for overlap detection, which indexes subsampled k-mers, by the principle of minimizers (Roberts et al. 2004), from all the reads in a hash table, against which the query minimizers are then compared. The matches are sorted and clustered to find the longest collinear matching chains to identify overlap candidates. An assembly graph layout is subsequently constructed from the collinear matches and output as the assembled contigs, without building any consensus. Because error-correction and consensus procedures are not executed, the error rate of the final assembly is equivalent to that of the raw reads. To circumvent this, Racon (Vaser et al. 2017), a consensus module, was shown to generate high-quality contigs within reasonable run times and is included in the present study as part of the miniasm pipeline.

4.2.1.7 SMARTdenovo

SMARTdenovo (<https://github.com/ruanjue/smartdenovo>) is another fast assembler, which can also work without error correction of the raw reads. Similar to minimap, SMARTdenovo searches subsampled query k-mers in indexed hash tables, which are then sorted and merged into collinear matches. Alignment using a dot-matrix alignment method is performed for adjacent matches, and the overlap candidates are

subsequently input to a string graph layout. The consensus module can reach an accuracy of up to 99.7%, albeit taking up much of the entire computational time.

4.2.1.8 ABruijn

A de Bruijn graph is a directed graph that is generally constructed from $k-1$ overlaps of adjacent k -mers. Rather, a set of solid strings (frequent k -mers), instead of all k -mers, is used to construct the ABruijn graphs (Lin et al. 2016), because of the high error rates in SMS reads. A fast dynamic programming approach is used to find the longest common subpaths to obtain a rough estimate of the overlaps between two reads. Overlapping read vertices are added onto the graph and the draft assembly is subsequently constructed. After aligning reads against the draft assembly, ABruijn graphs are constructed again to obtain a polished consensus assembly.

4.2.1.9 Wtdbg

Wtdbg (<https://github.com/ruanjue/wtdbg>) is another assembler that uses the framework of de Bruijn graphs. Unlike ABruijn graphs, overlapping k -mer hits are identified among the reads using a sorting approach similar to that adopted in minimap and SMARTdenovo, and the hits are used to add on and construct the fuzzy de Bruijn graphs. The resulting graphs, in comparison to ABruijn graphs, have reduced complexity and thereby consume lesser memory.

4.2.1.10 Mapping, Error Correction and *de novo* Assembly Tool

Mapping, Error Correction and *de novo* Assembly Tool (MECAT) (Xiao et al. 2016) scans for identical k -mers, in blocks of sequences among query reads, to calculate distance difference factor (DDF) between neighboring k -mer hits. When the DDF is within a specified threshold, scores are assigned to the blocks of k -mers and extended to neighboring blocks. With the scoring mechanism, a large number of irrelevant read overlap candidates are filtered out, significantly reducing the computational time before alignment. After error correction, the corrected reads are pairwise-aligned and fed into a modified canu pipeline to construct contigs.

4.2.2 Datasets for evaluation

The evaluation datasets were broadly chosen in such a way that i) data are available for public use, and ii) genomes are of diverse sizes.

Initially, the standard bacterial model organism *Escherichia coli* was chosen, and the sequence data (1 SMRT cell: approximately 140× coverage) of P6-C4 chemistry (figure 4.3A) was downloaded from the PacBio DevNet website (<https://github.com/PacificBiosciences/DevNet/wiki/Datasets>).

Plasmodium falciparum (protist) is one of the few smaller eukaryotic genomes with long-read data available. Although the genome is only approximately 23 Mb in length, it contains 14 chromosomes with a relatively high repeat content of 51.8% and a very high AT% of 80.6% (Girgis 2015). *P. falciparum* sequence data (9 SMRT cells: approximately 180× coverage) of P6-C4 chemistry (figure 4.3B) were downloaded from the National Center for Biotechnology Information's Sequence Read Archive (SRA360189) (Vembar et al. 2016).

In contrast to *P. falciparum*, *Caenorhabditis elegans* (nematode) has a genome size of approximately 105 Mb, but with only six, although much longer, chromosomes. The genome is also estimated to contain approximately 20,000 genes making it more complex when compared to those of *E. coli* and *P. falciparum*, which have only approximately 5,000 genes each. There are also relatively fewer transposons (approximately 12%), although they are sufficiently long (1–3 kb) to confound the genome assembly (Tyson et al. 2017). *C. elegans* sequence data (11 SMRT cells: approximately 45× coverage) of P6-C4 chemistry (figure 4.3C) were also downloaded from the PacBio DevNet website.

Next, we tackled the main challenge of focus for this evaluation using the genome of a non-model plant with a high repetitive content and longer repeats. For this purpose, *Ipomoea nil* (plant) data (Hoshino et al. 2016) of P5-C3 chemistry (figure 4.3D) were obtained based on our previous work (90 SMRT cells: approximately 50× coverage; DRA002710). *I. nil* has a highly repetitive (64%) genome of an estimated size of 750 Mb, with limited available genomic resources, providing a good measure for similar repetitive plant genomes. To evaluate the correctness of the *I. nil* genome assemblies, restriction site-associated DNA (RAD)-seq (DRA002758), expressed sequence tags (ESTs; HY917605–HY949060), and bacterial artificial chromosome (BAC)-end data (GA933005–GA974698) were used.

PacBio RSII was the sequencer employed in all cases. The P6-C4 chemistry, in comparison to P5-C3, has shown an increase in average read lengths and therefore the average read lengths of the *I. nil* data set are slightly shorter than those of the other data sets (figure 4.3). The reason for choosing only SMRT data for the present study is that one of the aims was to evaluate long-read assemblies without depending on SGS data, whereas the non-random errors of Nanopore data may still have to rely on more accurate Illumina data (Schmidt et al. 2017; Jain et al. 2017). All four datasets were pre-processed using HGAP3 to obtain filtered subreads for assembly. Two rounds of consensus polishing were applied to all assemblies using quiver. The jobs were executed on a node with a Intel Xeon E7-8870 processor (2.40 GHz) consisting of 160 cores and a memory of 2019.8 Gb under the operating system of RHEL v6.5. SGE was used for job management and the qacct command was used to access the maximum RSS and CPU time registered by the jobs.

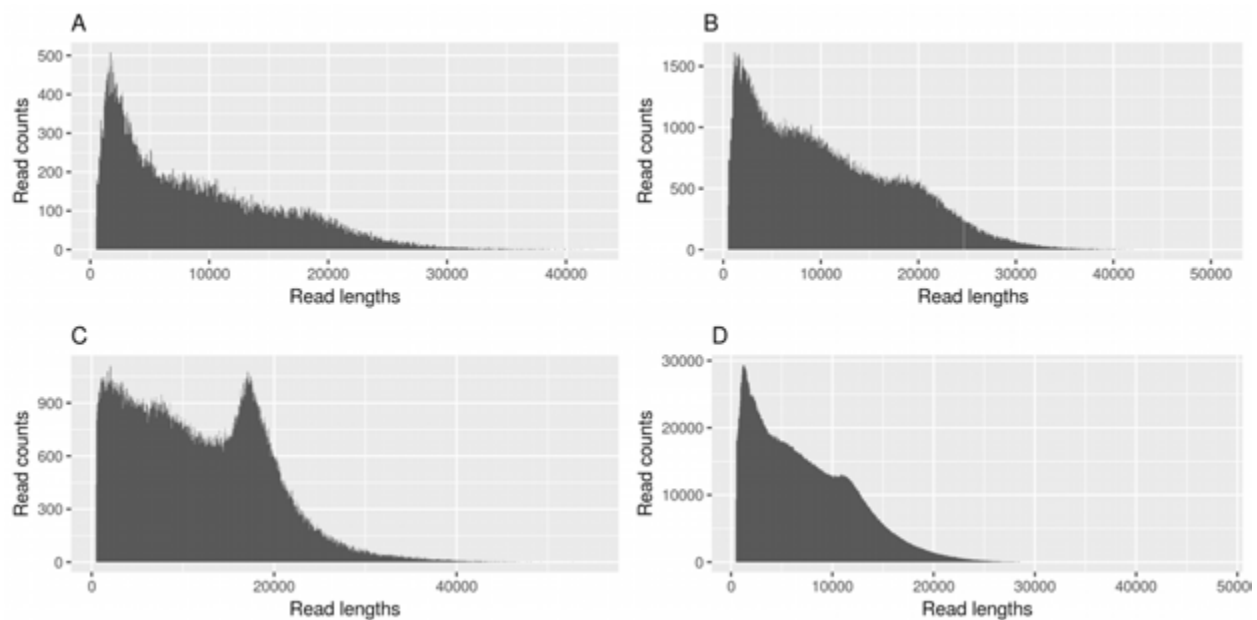


Figure 4.3. Read length distributions of A) *E. coli*, B) *P. falciparum*, C) *C. elegans*, and D) *I. nil* datasets. The binwidth used for the plotting was 50.

4.2.3 Criteria for evaluation

For assessing the assembly results, we considered various metrics (figure 4.1). Apart from N50 and L50 measures, the average contigs-to-chromosomes (ctg/chr) ratio was calculated for assessing contiguity. For gene-level completeness, BUSCO (Simão et al. 2015) and CEGMA (Parra et al. 2007) were used. In eukaryotic contigs, the terminal

regions were scanned using tandem repeats finder (Benson 1999) for the presence of telomeres. Peak computational memory in the form of Maximum resident set size (RSS) and CPU time were determined to compare computational requirements. When complete reference sequences were available, single nucleotide variations (SNVs), indels, and structural variations (SVs) were analyzed from QUAST (Gurevich et al. 2013) and Assemblytics (Nattestad and Schatz 2016) to evaluate correctness; unique SVs provided a relative measure of assembly errors. In addition, dot plots were visualized for rearrangements. The percentage of reference sequences covered by the assemblies was calculated using MUMmer (Kurtz et al. 2004) alignments.

For the non-model organism *I. nil*, linkage maps were constructed from RAD-seq (Baird et al. 2008) data using STACKS (Catchen et al. 2011), to identify mis-assembled contigs. Because the marker density of the linkage maps was low, this also provided a good measure for contiguity, as larger contigs have a better chance of being incorporated in the linkage maps. ESTs and BAC-end reads were used for assessing completeness. Longer contigs had a better chance of concordantly mapping the 100-kb insert-sized BAC-end read pairs, whereas discordant mapping rates provided an indirect measure of mis-assemblies. Whole BAC sequences, of approximately 100 kb in length, were used to assess contiguity and completeness, and also to identify SNVs and indels. *Tpn1* transposons, a unique feature of *I. nil* flowers (Hoshino et al. 2016), were also considered to assess completeness.

For *E. coli*, all the assemblers reconstructed the bacterial chromosome in one piece. Therefore only, the following metrics were taken into account for ranking the assemblers:

- Circularity resolved or unresolved
- Number of mismatches per 100 kb from QUAST, in decreasing order
- Number of SVs, in decreasing order
- Length of SVs, in decreasing order
- CPU time, in decreasing order
- Maximum RSS, in decreasing order

For *P. falciparum* and *C. elegans*, the following criteria were used for ranking the assemblers:

- Number of assembled contigs, in decreasing order
- N50 values, in increasing order
- L50 values, in decreasing order
- Number of mismatches per 100 kb from QUAST, in decreasing order
- Number of SVs, in decreasing order
- Length of SVs, in decreasing order

- Number of SVs unique to the assemblers, in decreasing order
- Mean of percentage of chromosomes covered by contigs, in increasing order
- Number of telomeres, in increasing order
- Number of complete genes from BUSCO, in increasing order
- CPU time, in decreasing order
- Maximum RSS, in decreasing order

For the non-model organism, *I. nil*, the following metrics were used for ranking:

- Number of assembled contigs, in decreasing order
- N50 values, in increasing order
- L50 values, in decreasing order
- Number of BAC-end read pairs mapped onto the same contigs, in increasing order
- Number of discordantly mapped BAC-end read pairs, in decreasing order
- Number of mapped ESTs, in increasing order
- Number of transposons, in increasing order
- Number of telomeres, in increasing order
- Number of contigs incorporated in linkage maps, in decreasing order (longer and hence fewer contigs are incorporated in the linkage maps)
- Length of contigs incorporated in linkage maps, in increasing order
- Number of mis-assembled contigs, in decreasing order
- Length of mis-assembled contigs, in decreasing order
- Average per base accuracy observed in five whole BAC sequences, in increasing order
- Number of complete genes from BUSCO, in increasing order
- CPU time, in decreasing order
- Maximum RSS, in decreasing order

The ranks for all criteria were summed up for each assembler. The summed score, in the decreasing order, was used for assigning an overall rank. Also, z-scores were calculated for all observed metrics, so that significant observations received rewards or penalties (Bradnam et al. 2013). The average of the z-scores, from all metrics, for each assembler was plotted to observe z-score based rankings, which displayed high and low scores for better and worse performances, respectively. For assemblies which failed during execution, either they were left out from the rankings or assigned arbitrary low rankings.

4.3 Results

4.3.1 Contiguity

All of the assemblers reported good contiguity (table 4.1).

4.3.1.1 *Escherichia coli*

A single contig representing the complete bacterial genome was reconstructed by all the assemblers (table 4.2).

4.3.1.2 *Plasmodium falciparum*

Fewer number of contigs (15–43 contigs), high N50 values (1.2–1.7 Mb), low L50 values (5–7), and low ctg/chr ratios (1–2.27 ratios) were generally observed in all the assemblies, representing high level of contiguity, despite the repetitive nature of the genome. MECAT, in particular, reconstructed every chromosome in one piece, whereas miniasm, SMARTdenovo, and wtdbg produced comparatively fragmented or redundant contigs (table 4.3).

4.3.1.3 *Caenorhabditis elegans*

The N50 exceeded 1 Mb in all, but the PBcR assembly. Canu had the best N50 (3.6 Mb) and L50 (11) values, while PBcR had low N50 (847 kb) and high L50 (38) values. In general, six contigs, on an average, were found to be sufficient to represent a chromosome (table 4.4).

4.3.1.4 *Ipomoea nil*

HGAP3 obtained the best contiguity (N50=1.53 Mb; L50=120) and was the only assembler to have contigs more than 10 Mb in length. Canu and FALCON shared the next best N50 (934 and 904 kb respectively) and L50 values (191), while both wtdbg and miniasm had fragmented assemblies (table 4.5).

The shorter the genome, the lesser the differences observed in contiguity among the assemblers. However, with longer genomes, the contiguity profiles progressively started to differ among the assemblers.

Table 4.1. Summarized statistics of the assemblies

Organism		# Contigs	Assembly Size (Mbp)	Longest Contig (Mbp)	N50 (Mbp)	L50	CPU time (hours)	Max RSS (GB)
<i>E. coli</i> (4.6 Mb)	Maximum	1	4.7	4.7	4.7	1	83.9	44.5
	Minimum	1	4.6	4.6	4.6	1	2.2	3.6
	Mean	1	4.7	4.7	4.7	1	19.4	15.7
<i>P. falciparum</i> (23 Mb)	Maximum	43	23.8	3.3	1.7	7	2012.6	43.9
	Minimum	15	23.1	2.1	1.3	5	20.1	4.5
	Mean	26.3	23.4	2.9	1.5	6.1	441.7	22.7
<i>C. elegans</i> (105 Mb)	Maximum	452	106.9	7.1	3.7	38	6733.8	251.7
	Minimum	68	101.9	2.7	0.8	11	13.4	10.1
	Mean	166.7	104.2	5.1	2.2	19.4	1221.4	56.9
<i>I. nil</i> (750 Mb)	Maximum	8751	752.7	11.5	1.8	1194	28504.7	331.2
	Minimum	1697	642	2.5	0.1	104	129.7	16.2
	Mean	4288	702.7	6.2	0.7	439.4	10065.8	78.2

L50 and N50 represents the number of contigs and the length of the contig, respectively, crossing 50% mark of the assembly. Higher N50 and lower L50 values indicate highly contiguous assemblies. Max RSS represents the peak memory usage of the computational node.

4.3.2 Completeness

4.3.2.1 *Escherichia coli*

In all the cases, the assembly size was slightly larger than that of the reference genome, with 99.9% BUSCO completeness (table 4.2).

Table 4.2. *E. coli* assembly statistics after Circlator and two rounds of polishing by quiver.

	ABruijn	Canu	FALCON	HGAP3	HINGE	MECAT	Miniasm	PBcR	SMARTdenovo	Wtdbg
# Contigs	1	1	1	1	1	1	1	1	1	1
Total length	4642496	4642496	4642497	4681591	4642496	4679737	4642496	4642496	4642497	4695247
# mismatches per 100 kbp	0.09	0.09	0.09	0.11	0.09	0.09	0.09	0.09	0.09	0.41
# indels per 100 kbp	0.11	0.11	0.09	0.41	0.11	0.41	0.11	0.11	0.09	3.25
GC (%)	50.79	50.79	50.79	50.74	50.79	50.75	50.79	50.79	50.79	50.79
Insertions (Count)	3	3	3	24	3	26	3	3	3	135
Insertions (Total bases)	780	780	780	801	780	803	780	780	780	944
Deletions (Count)	2	2	2	23	2	26	2	2	2	183
Deletions (Total bases)	2	2	2	25	2	28	2	2	2	244
Tandem expansions (Count)	1	1	1	1	1	1	1	1	1	1
Tandem expansions (Total bases)	181	181	181	181	181	181	181	181	181	181
Tandem contractions (Count)	1	1	1	1	1	1	1	1	1	1
Tandem contractions (Total bases)	113	113	113	113	113	113	113	113	113	113
Repeat expansions (Count)	0	0	0	0	0	0	0	0	0	0
Repeat expansions (Total bases)	0	0	0	0	0	0	0	0	0	0
Repeat contractions (Count)	1	1	1	0	1	0	1	1	1	1
Repeat contractions (Total bases)	2	2	1	0	2	0	2	2	1	171

Table 4.3. *P. falciparum* assembly statistics after two rounds of polishing by quiver.

	ABruijn	Canu	FALCON	HGAP3	MECAT	Miniasm	PBcR	SMARTdenovo	Wtdbg
# contigs	18	23	20	27	15	43	27	30	34
Total length	23358918	23113319	23170402	23756014	23282583	23223896	23116444	23580097	23633779
Longest contig	3294952	2433823	3274628	3293149	3295200	2885033	2112509	2255642	3298739
Contig to chromosome ratio	1.20	1.53	1.33	1.80	1.00	2.87	1.80	2.00	2.27
N50	1590706	1380424	1658395	1722570	1682661	1268890	1393365	1338823	1390642
N75	1322391	1212433	1314993	1345118	1420518	875865	1174904	935228	999885
L50	5	7	5	5	5	7	7	7	7
L75	9	11	9	9	9	12	12	13	11
GC (%)	19.31	19.34	19.18	19.44	19.26	19.06	19.15	19.32	19.31
# mismatches per 100 kbp	5.05	5.82	5.03	5.69	4.31	4.76	4.95	5.99	7.11
# indels per 100 kbp	32.48	32.92	31.26	31.12	30.28	33.93	31.26	36.89	44.44
Insertions (Count)	7707	7652	7702	8043	7651	9663	7587	8361	11941
Insertions (Total bases)	11630	11805	11835	12236	11718	14074	11526	12710	17036
Deletions (Count)	546	573	667	846	532	1113	530	784	2125
Deletions (Total bases)	1917	1866	1978	2351	1742	2442	1731	2763	3885
Tandem expansions (Count)	23	26	22	30	24	26	25	23	23
Tandem expansions (Total bases)	17747	19320	18897	44815	19261	23646	19289	18513	14990
Tandem contractions (Count)	7	8	7	9	8	8	7	8	7
Tandem contractions (Total bases)	9893	966	646	1144	1054	941	715	1072	1036
Repeat expansions (Count)	4	5	6	6	6	4	4	5	6
Repeat expansions (Total bases)	1139	1597	2407	2408	2407	1139	2348	1952	8035
Repeat contractions (Count)	1	1	1	2	2	2	1	2	2
Repeat contractions (Total bases)	810	810	810	925	914	899	810	916	8162
SVs unique to assemblers	389	384	583	914	311	3045	332	1176	6448
CEGMA completeness (%)	68.95	70.16	69.76	69.35	69.35	68.95	68.95	68.95	69.76
BUSCO completeness (%)	68.4	67.4	68.4	68.8	67.9	68.8	68.4	68.9	68.8

Table 4.4. *C. elegans* assembly statistics after two rounds of polishing by quiver.

	ABruijn	Canu	FALCON	HGAP3	MECAT	Miniasm	PBcR	SMARTdenovo	Wtdbg
# contigs	68	107	95	452	100	108	272	128	170
Total length	102198820	106924671	101860568	105433602	101981214	105235071	103377354	105906666	104656824
Largest contig	4905601	6865821	6911456	4285935	3700778	7083682	2667744	3707098	5325592
Contig to chromosome ratio	11.33	17.83	15.83	75.33	16.67	18.00	45.33	21.33	28.33
N50	2841666	3694474	2660163	1592494	1675869	3140582	847486	1737100	1860672
N75	1692214	2099774	1482622	770707	1035355	1910694	445476	901376	866559
L50	14	11	13	23	22	12	38	23	19
L75	26	21	26	47	42	24	80	44	42
GC (%)	35.49	35.93	35.5	35.85	35.48	36	35.79	36.07	36.06
# mismatches per 100 kbp	15.68	16.39	14.54	15.17	16.89	15.18	10.42	14.97	16.37
# indels per 100 kbp	23.11	19.22	20.17	23.71	20.09	23.84	21.25	24.71	33.47
Insertions (Count)	9003	5552	5194	6523	5127	8210	3994	10105	19057
Insertions (Total bases)	96305	92578	78466	85237	73150	99955	60654	89745	137376
Deletions (Count)	29576	23048	24046	27848	22596	28896	22960	29509	36429
Deletions (Total bases)	43829	34728	35073	41362	33203	44748	35431	42701	53885
Tandem expansions (Count)	337	337	317	297	321	329	250	314	276
Tandem expansions (Total bases)	556384	561758	480122	357164	507875	517687	273897	460610	345345
Tandem contractions (Count)	41	39	38	49	38	55	36	43	61
Tandem contractions (Total bases)	28738	20281	28188	42896	18358	64948	12434	21196	112977
Repeat expansions (Count)	73	65	65	55	65	73	42	70	70
Repeat expansions (Total bases)	186691	171802	172052	123543	172320	181680	97462	187370	156187
Repeat contractions (Count)	27	23	27	32	27	34	26	27	39
Repeat contractions (Total bases)	19847	5393	7107	18896	17750	38847	7216	13177	27248
SVs unique to assemblers	69	49	69	115	93	60	191	99	112
CEGMA completeness (%)	97.18	96.77	97.18	97.58	95.97	97.58	97.58	96.77	94.76
BUSCO completeness (%)	98.9	98.2	98.7	99.1	98.9	98.8	99.2	98.9	97.2

Table 4.5. *I. nil* assembly statistics after two rounds of polishing by quiver.

	Canu	FALCON	HGAP3	MECAT	Miniasm	PBcR	SMARTdenovo	Wtdbg
# contigs	1697	2030	5678	3365	6772	4417	3175	8751
# contigs (>= 1 Mb)	169	171	195	99	52	136	93	16
# contigs (>=5 Mb)	5	3	14	2	0	2	1	0
# contigs (>=10 Mb)	0	0	2	0	0	0	0	0
Total length	701070001	676319005	746608706	693078889	752718457	725755666	694182782	642008886
Largest contig	7370807	6459633	12514902	5654447	3041154	7746741	5220514	2501541
Contig to chromosome ratio	113.13	135.33	378.53	224.33	451.47	294.47	211.67	583.40
N50	934355	904306	1532223	443860	251632	575269	402510	126410
N75	462826	431538	651327	189611	110987	244401	189208	57292
L50	191	191	120	351	747	315	422	1194
L75	461	463	312	946	1868	804	1048	3090
GC (%)	37.08	36.98	37.7	37.04	37.34	37.08	37.08	36.67
CEGMA completeness (%)	94.76	93.55	93.95	94.76	93.95	94.35	94.76	92.34
BUSCO completeness (%)	93.8	93.5	93.7	93.9	93.7	93.9	94	92.9

4.3.2.2 *Plasmodium falciparum*

On average, the contigs covered the 14 chromosomes in the range of 95.67–99.90%. Excluding ABruijn, the apicoplast genome was assembled by all the assemblers, while the mitochondrial genome was only present in the HGAP3 assembly. Canu was able to reconstruct 23 of the 28 telomeres, whereas the PBcR and wtdbg assemblies resolved less than 10 telomeres (table 4.6). Intriguingly, Miniasm was unable to resolve even a single telomere. BUSCO analysis showed 67.4–68.9% completeness for all the assemblies, while it should be noted that the original reference sequence also yielded only 68.8% completeness.

4.3.2.3 *Caenorhabditis elegans*

At least 99% of all the chromosomes were covered by the assembled contigs on average, excluding the wtdbg assembly. Canu and HGAP3 produced 10 out of 12 telomeres, whereas wtdbg produced only a single telomere (table 4.7). All the assemblies also showed high BUSCO (97.2–99.2%) completeness ratios.

4.3.2.4 *Ipomoea nil*

Most of the assemblies fell short of the expected genome size of 750 Mb, however BUSCO reported completeness ratios in the range of 92.9–94%. Most of the assemblies mapped around 99% of the ESTs and BAC-end reads (table 4.8). PBcR (314) and HGAP3 (311) resolved the largest number of *Tpn1* transposons, followed by canu (307) and MECAT (307). MECAT (18), FALCON (16), and SMARTdenovo (16) were better at resolving telomeres (table 4.8).

Some smaller PBcR contigs were present redundantly and were covered within larger contigs with short overhangs. The high BUSCO and CEGMA ratios indicated that the gene regions were captured effectively, despite differences in the assembly sizes. The shorter, circular, and high-copy nature of the mitochondrial genomes could have possibly confounded the assemblers and were largely unassembled.

Table 4.6. Telomere composition of *P. falciparum* assemblies.

Chromosome	ABruijn	Canu	FALCON	HGAP3	MECAT	Miniasm	PBcR	SMARTdenovo	Wtdbg
1	1	2	1	1	0	0	1	2	1
2	1	1	1	2	2	0	0	0	0
3	2	2	0	1	0	0	0	2	0
4	2	1	2	2	2	0	0	1	1
5	1	2	0	1	2	0	1	1	1
6	2	1	1	2	2	0	1	2	1
7	2	2	1	2	1	0	0	2	0
8	1	2	1	1	2	0	0	2	0
9	1	2	1	0	2	0	0	1	1
10	1	1	0	0	1	0	0	2	0
11	2	2	0	2	2	0	1	2	1
12	1	1	1	1	2	0	0	2	1
13	1	2	1	1	1	0	0	1	0
14	2	2	1	2	2	0	1	2	1
Total	20	23	11	18	21	0	5	22	8

Table 4.7. Telomere composition of *C. elegans* assemblies.

Chromosome	ABruijn	Canu	FALCON	HGAP3	MECAT	Miniasm	PBcR	SMARTdenovo	Wtdbg
1	2	1	0	1	1	1	1	1	0
2	2	2	1	2	2	1	1	2	0
3	1	2	1	1	2	1	0	1	0
4	1	1	2	2	1	1	1	1	1
5	1	2	0	2	1	2	1	2	0
X	1	2	1	2	2	1	2	1	0
Total	8	10	5	10	9	7	6	8	1

Table 4.8. Mapping of Bac-end reads, ESTs, and *Tpn1* transposons against *I. nil* assemblies.

		Canu	FALCON	HGAP3	MECAT	Miniasm	PBcR	SMARTdenovo	Wtdbg
Bac-end reads	# mapped read pairs	20832	20832	20828	20832	20830	20830	20832	20830
	% of mapped read pairs	99.93	99.93	99.91	99.93	99.92	99.92	99.93	99.92
	# read pairs mapped in the same contigs	19319	18341	19679	17766	16213	17933	17492	11525
	% of read pairs mapped in the same contigs	92.67	87.98	94.40	85.22	77.77	86.02	83.91	55.28
	# discordant read pairs	851	1325	1152	981	528	967	867	217
	% of discordant read pairs	4.08	6.36	5.53	4.71	2.53	4.64	4.16	1.04
ESTs	# mapped ESTs	92864	92697	92860	92813	92826	92759	92847	91988
	% of mapped ESTs	99.12	98.94	99.11	99.06	99.08	99.01	99.10	98.18
	# mapped ESTs with >90% coverage	91984	91586	91989	91844	91898	91855	91948	90670
	% of mapped ESTs with >90% coverage	98.18	97.75	98.18	98.03	98.09	98.04	98.14	96.78
# <i>Tpn1</i> transposons		307	299	311	307	296	314	291	226
# Telomeres		8	16	13	18	14	12	16	8

4.3.3 Correctness

After two rounds of consensus polishing of the draft assemblies, the indel rates were drastically reduced.

4.3.3.1 *Escherichia coli*

Analysis using QUAST showed that all contigs had mis-assemblies. However on closer inspection using Assemblytics, the source of the mis-assemblies reported by QUAST was revealed to be because of 3 structural variations, which are likely strain-specific differences rather than mis-assemblies (table 4.2). For instance, in the ABruijn assembly, the contig length was equal to the reference length when the SVs were tallied. However, most other assemblies still had a large number of SVs (an average of 68.8 SVs compared with 9 SVs of ABruijn), even after two rounds of polishing.

4.3.3.2 *Plasmodium falciparum*

More than 5,000 SVs were shared among all the assemblies. Wtdbg (6448) produced the largest number of unique SVs, whereas ABruijn (389), canu (384), MECAT (311), and PBcR (332) performed better by producing a relatively smaller share of the unique SVs. Dot plots were used for observing rearrangements, which displayed small rearrangements only in ABruijn and wtdbg assemblies. In other cases, an approximate straight diagonal line was observed with strong congruity.

4.3.3.3 *Caenorhabditis elegans*

A total of 17,893 SVs were shared among all the assemblies. Wtdbg (30,622) produced the largest number of unique SVs, whereas canu (2,374), FALCON (3,337), MECAT (2,358), and PBcR (4,179) produced a relatively smaller share of unique SVs. A single or a couple of mis-assembled contigs were visible in the dot plots of all assemblies, barring MECAT and SMARTdenovo.

4.3.3.4 *I. nil*

Miniasm (1.2 Mb) and wtdbg (5.8 Mb) assemblies had the shortest of the mis-assembled contigs, while HGAP3 (128 Mb) showed the largest share of mis-assembled

data. HGAP3, FALCON, and MECAT had more than 100 Mb of mis-assembled contigs, whereas canu offered the best balance in incorporating longer contigs (593.3 Mb) into the linkage maps, with shorter (20.9 Mb) mis-assemblies (table 4.9). Wtdbg (1.04%) and miniasm (2.53%) had the least discordantly mapping BAC-end read pairs. Surprisingly, FALCON (6.36%) had the highest discordant mapping rate (table 4.8). When BAC sequences were completely covered by contigs, the per-base accuracy was 99.9% in four of the five BAC sequences (table 4.10), while mismatched bases were almost non-existent. Fragmented contigs were not considered for assessing per-base accuracy, as they had unresolved errors in overlapping terminal regions.

A lot of SVs were shared among all the assemblers which may be actual variations rather than assembly errors. Unlike the SMRT data, the Illumina based assembly was found to have large indels, and plenty of mismatches covering the five BAC sequences in *I. nil* (Hoshino et al. 2016). The evaluated assemblers, which are based on the overlap information of the longer reads, had benefited not just in terms of contiguity, but also in per-base accuracy for a repetitive genome like *I. nil*.

4.3.4 Circularity and overlapping fragmented contigs

With the application of Circlator (Hunt et al. 2015), it was evident that the circularity of some of the *E. coli* assemblies was clearly not resolved, and hence the presence of additional base pairs, which were subsequently trimmed out. The increased indel rates were originally concentrated on the overlapping terminal ends of the circularly unresolved contigs. As a result, the indel rates became almost identical in all the circularly resolved assemblies (table 4.2). However, Circlator was unable to resolve the circularity for HGAP3, MECAT, and wtdbg assemblies. Similarly, when the contigs were fragmented in repetitive regions, sometimes, the breakpoints happened in such a way that two nearby contigs shared considerable overlapping terminal ends. Consensus polishing did not have an impact in such overlapping regions leading to unresolved and high amount of indel errors.

Table 4.9. Linkage map based analysis of *I. nil* assemblies.

	Canu	FALCON	HGAP3	MECAT	Miniasm	PBcR	SMARTdenovo	Wtdbg
# contigs	756	761	552	1129	1466	979	1202	1813
Length of contigs	593380023	566851695	639061741	499698582	454749164	535258438	495980623	326128807
Percentage of contig size	84.64	83.81	85.60	72.10	60.41	73.75	71.45	50.80
# mis-assembled contigs	14	76	38	86	3	47	54	17
Length of mis-assembled contigs	20985814	102369939	128150514	103449124	1261012	74298762	45306443	5802006
Percentage of mis-assembled contig size	2.99	15.14	17.16	14.93	0.17	10.24	6.53	0.90

Table 4.10. Alignments of whole BAC sequences against *I. nil* assemblies.

BAC sequence	Features	Canu	FALCON	HGAP3	MECAT	Miniasm	PBcR	SMARTdenovo	Wtdbg
JMHibA010C11	Mismatches	0	0	0	0	0	0	0	0
	Query Gap openings	3	3	3	3	3	3	3	3
	Query Gap bases	3	3	3	3	3	3	3	3
	Target Gap openings	6	6	6	5	6	6	6	7
	Target Gap bases	9	9	9	8	9	9	9	11
	Perbase accuracy	99.99	99.99	99.99	99.99	99.99	99.99	99.99	99.99
JMHibA038C09	Mismatches	0	0	1	1	NA	NA	NA	NA
	Query Gap openings	5	6	6	6	NA	NA	NA	NA
	Query Gap bases	5	7	6	6	NA	NA	NA	NA
	Target Gap openings	4	3	2	2	NA	NA	NA	NA
	Target Gap bases	12	10	8	8	NA	NA	NA	NA
	Perbase accuracy	99.99	99.99	99.99	99.99	NA	NA	NA	NA
JMHibA037J13	Mismatches	0	0	0	0	0	0	0	0
	Query Gap openings	10	10	10	10	10	10	10	10

	Query Gap bases	384	382	382	376	374	374	376	379
	Target Gap openings	6	6	6	5	5	5	5	5
	Target Gap bases	9	9	9	8	8	8	8	8
	Perbase accuracy	99.63	99.64	99.64	99.64	99.64	99.64	99.64	99.64
JMHiBa001L04	Mismatches	0	0	0	0	0	0	0	0
	Query Gap openings	9	7	6	7	5	9	8	9
	Query Gap bases	17	16	11	17	8	17	16	120
	Target Gap openings	7	6	5	7	6	6	7	7
	Target Gap bases	13	10	7	13	10	11	13	115
	Perbase accuracy	99.97	99.97	99.98	99.97	99.98	99.97	99.97	99.77
JMHiBa001I06	Mismatches	NA	0	0	0	NA	0	NA	NA
	Query Gap openings	NA	4	4	4	NA	4	NA	NA
	Query Gap bases	NA	4	4	4	NA	4	NA	NA
	Target Gap openings	NA	7	8	8	NA	7	NA	NA
	Target Gap bases	NA	15	16	16	NA	16	NA	NA
	Perbase accuracy	NA	99.98	99.98	99.98	NA	99.98	NA	NA

4.3.5 Resource usage

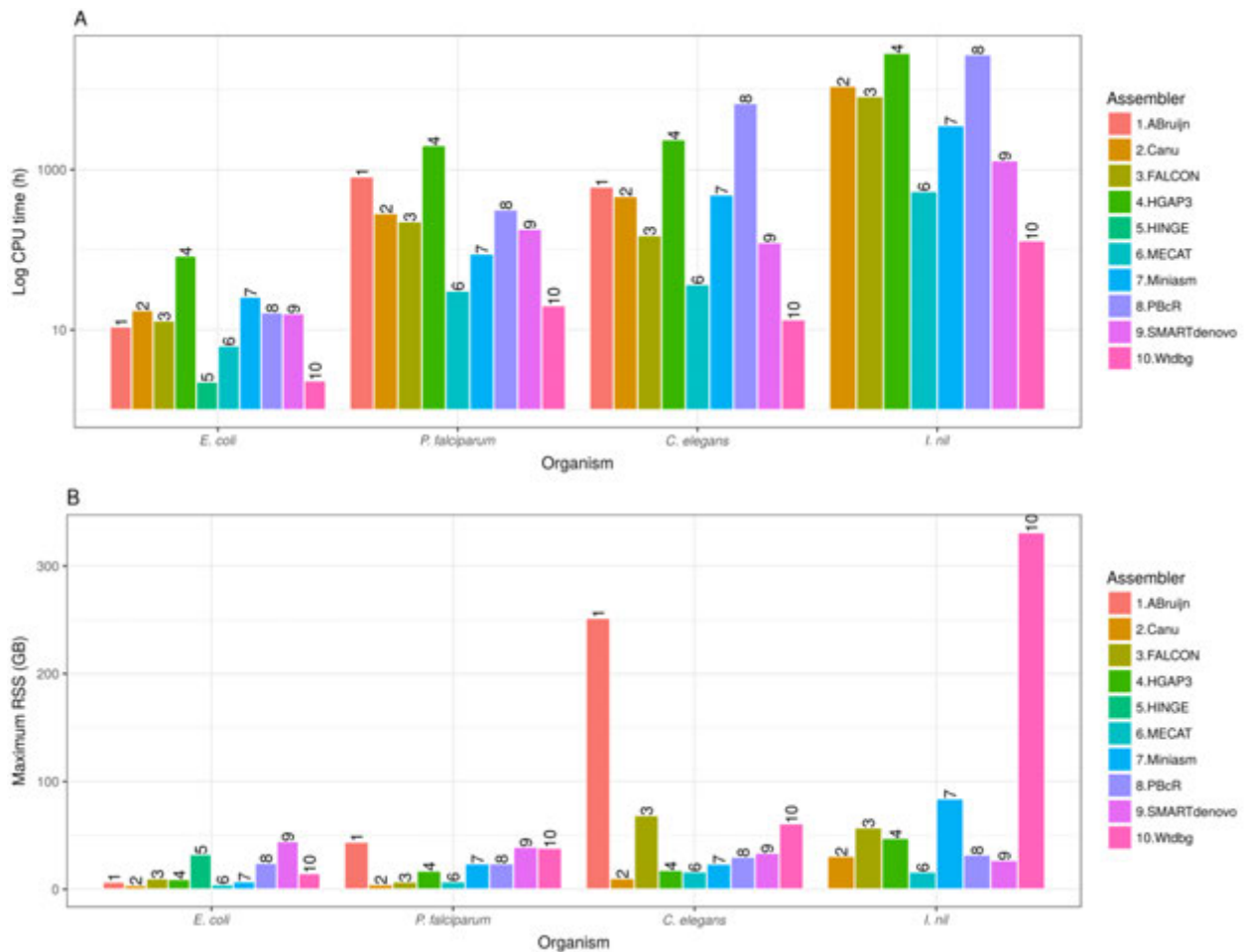


Figure 4.4. Computational resource requirements. Computational requirements are represented as (A) log CPU time and (B) maximum RSS, a measure of peak memory usage, for all assemblers.

4.3.5.1 *Escherichia coli*

HINGE and wtdbg assemblies were quickly obtained, while HGAP3 was the slowest, as expected (figure 4.4A). Miniasm was actually the fastest of all assemblers, and finished in about 16 min of CPU time; however, two rounds of RACON execution required a total of 25.81 CPU h, making this pipeline the second slowest. SMARTdenovo consumed the maximum peak memory usage, while HGAP3 consumed the least amount of memory (figure 4.4B).

4.3.5.2 *Plasmodium falciparum*

Wtdbg was the quickest assembler, closely followed by MECAT. Other assemblers generally consumed hundreds of CPU hours, with HGAP3 being almost 100-fold slower compared to the speed of wtdbg (figure 4.4A). ABruijn, SMARTdenovo, and wtdbg were memory-intensive, whereas canu, FALCON, and MECAT were memory-efficient (figure 4.4B).

4.3.5.3 *Caenorhabditis elegans*

Wtdbg followed by MECAT were the fastest in producing assemblies, while PBcR was the slowest (figure 4.4A). ABruijn consumed a huge amount of memory, while canu was the most memory-efficient, followed by MECAT and HGAP3 (figure 4.4B).

4.3.5.4 *Ipomoea nil*

Wtdbg was again the fastest assembler (129.7 CPU h). It should be noted that HGAP3 took 83.9 CPU hours even for a bacterial genome. MECAT was also fairly quick, while the celera-dependent pipelines were the slowest (figure 4.4A). Wtdbg consumed 331.15 Gb of peak memory. MECAT was the best with respect to both CPU time and peak memory usage, while canu also showed a reasonable balance in resource usage (figure 4.4B).

4.3.6 Ranking

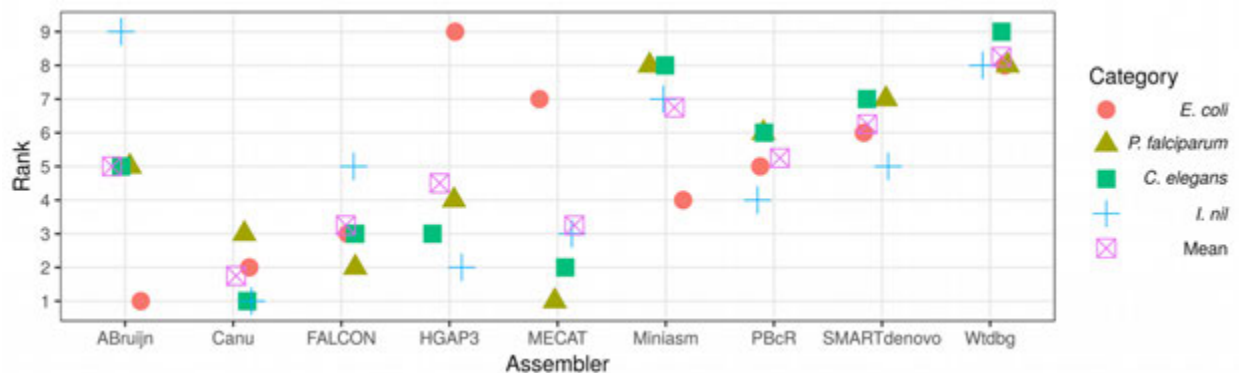


Figure 4.5. Rankings for all assemblies. The lower the rank, the better is the assembly.

4.3.6.1 *Escherichia coli*

The assemblers ABruijn, canu, and FALCON in the order were top-ranked in both the rankings (figures 4.5, 4.6A). The rankings were heavily influenced by whether the assemblies were circularly resolved or not, and hence MECAT, HGAP3, and wtdbg were pushed to the bottom of the table.

4.3.6.2 *Plasmodium falciparum*

Although HGAP3 had the highest N50 value, it was not the top-ranked assembler (figures 4.5, 4.6B). Four assemblers in the order of MECAT, FALCON, ABruijn, and canu were top-ranked according to their z-scores (figure 4.6B), corroborating that N50 should not be the sole factor in choosing an assembly. HINGE assembly was excluded from the rankings, as it resulted in a segmentation fault and therefore was not tested for the other eukaryotic datasets too.

4.3.6.3 *Caenorhabditis elegans*

Canu ranked at the top, followed by FALCON and MECAT (figure 4.6C). Although miniasm was eighth in the ranking (figure 4.5), it surprisingly ranked fourth according to the z-scores, as a result of obtaining considerably high z-scores for contiguity metrics (figure 4.6C). Without error correction, it would be difficult to distinguish duplications and repeats (Li 2016); however, the repeat-sparse nature of the *C. elegans* genome likely contributed to the better contiguity achieved by miniasm.

4.3.6.4 *Ipomoea nil*

ABruijn assembly resulted in a segmentation fault and was not considered for evaluation. The highly repetitive nature and the shorter insert size of the *I. nil* dataset prevented all of the assemblers from reaching a 1-Mb contig N50, excluding HGAP3. Nevertheless, canu ranked first, ahead of HGAP3, in either of the rankings (figure 4.5, 4.6D). If mis-assemblies were given additional penalties, the ranking of HGAP3 might come down further. For the first time, SMARTdenovo was ranked among the top five assemblers.

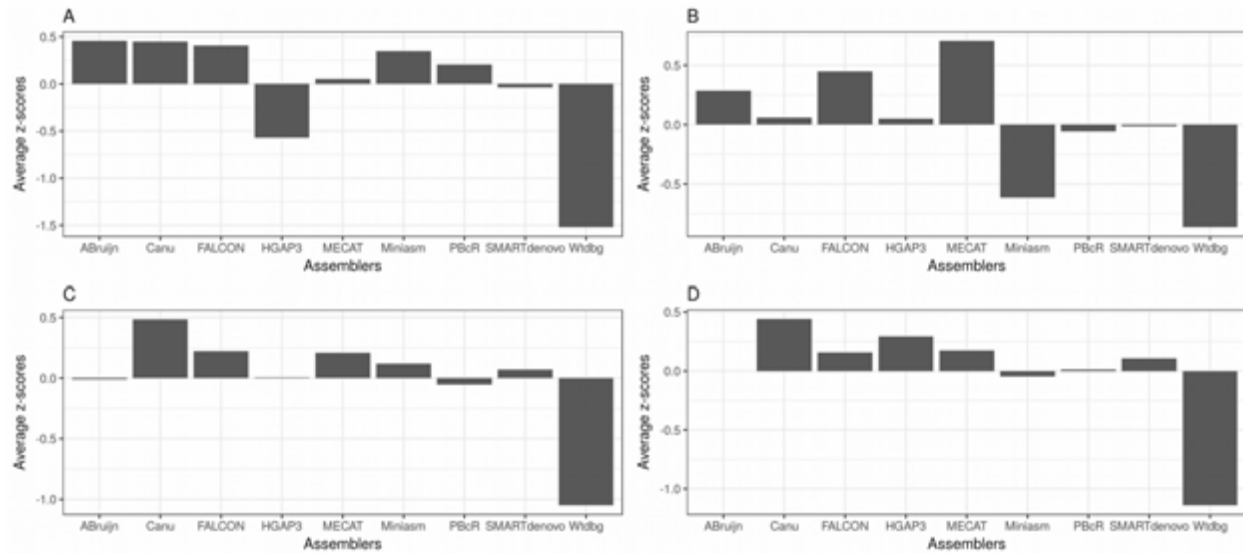


Figure 4.6. Z-score-based rankings. Average z-scores of all ranking metrics are plotted for (A) *E. coli*, (B) *P. falciparum*, (C) *C. elegans*, and (D) *I. nil*. Higher the average z-value, the better is the assembly performance. The failed ABrujn assembly is left blank for *I. nil* data set.

4.3.6.5 Mean ranking of the three eukaryotic assemblies

When the rankings of the eukaryotic assemblies were averaged (figure 4.5), canu, MECAT, FALCON, and HGAP3, in that order, were on the top of the rankings. Similarly, in the z-score based mean rankings, canu, MECAT, FALCON, and HGAP3, in that order, displayed better performances with positive mean z-scores (figure 4.7).

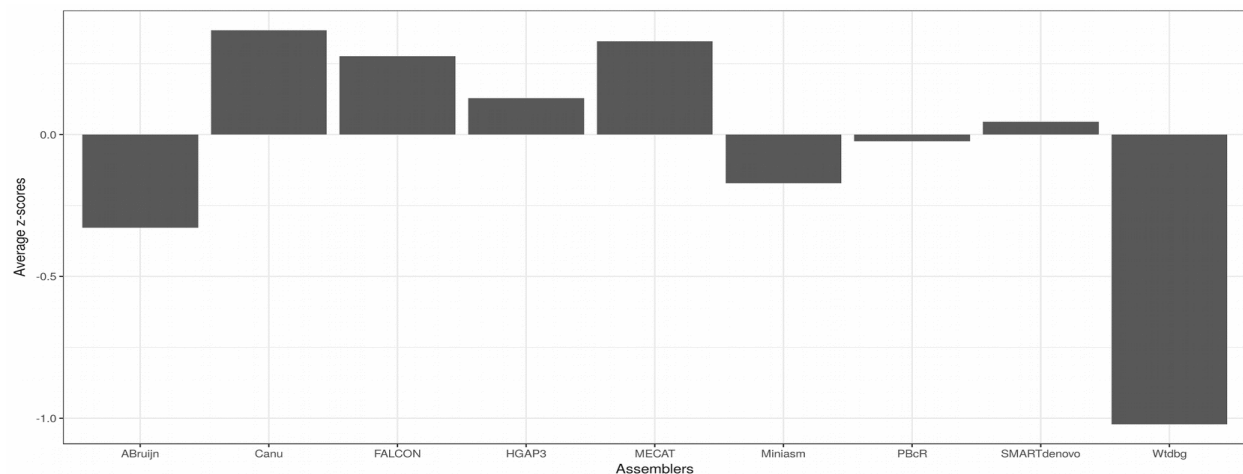


Figure 4.7. Mean z-score-based rankings. The mean scores of the individual average z-scores obtained from *E. coli*, *P. falciparum*, *C. elegans*, and *I. nil* are plotted. Higher the average z-value, the better is the assembly performance.

4.4 Discussion

De novo genome assemblies using SMRT data, when compared to earlier versions, have been shown to increase contiguity by several hundred-folds (Korlach et al. 2017; Gordon et al. 2016; Weissensteiner et al. 2017), and resolve fragmented regions into contiguous, gapless sequences (Gordon et al. 2016; Berlin et al. 2015). The average and median contig N50 values of recently assembled plant and animal genomes using long reads are 6.24 Mb and 3.60 Mb (table 4.11), respectively. In the current study, the three important features—contiguity, completeness, and correctness (Lee et al. 2016)—of long-read assemblers were evaluated.

Canu ranked the best in the average rankings of all the assemblies from all the datasets. Canu, because of its efficiency to handle repeats (Koren et al. 2017), had fewer assembly errors, sometimes trading contiguity for correctness. Indeed, it is essential to prioritize correctness rather than contiguity, which would otherwise defeat the purpose of building a reference genome for future studies.

Canu and MECAT showed the best balance in computational requirements. MECAT requires longer reads to effectively distinguish non-repetitive overlaps, and was found to underperform in the case of *I. nil*, whose transposons can be longer than the 7-kb average insert size of *I. nil* data.

FALCON, the only diploid-aware assembler, showed reasonable performance for genomes up to 100 Mb in length, similar to MECAT. The FALCON assembly was surprisingly filled with mis-assemblies for the *I. nil* data, probably because of the repeat filtering steps, leading to further loss of coverage in input data. An increase in insert sizes and coverage could yield better performance from both FALCON and MECAT.

Table 4.11. A list of recently assembled genomes using PacBio's SMRT data

Organism	Technology	Assembly tool	Contig N50/NG50	Study
<i>Taeniopygia guttata</i>	PB	FALCON	5.8 Mb	(Korlach et al. 2017)
<i>Calypte anna</i>	PB	FALCON	5.4 Mb	(Korlach et al. 2017)
<i>Drosophila serrata</i>	PB	PBcR	0.94 Mb	(Allen et al. 2017)
<i>Utricularia gibba</i>	PB	HGAP3	3.42 Mb	(Lan et al. 2017)
<i>Arabidopsis thaliana</i>	PB	PBcR	11.16 Mb	(Berlin et al. 2015)
<i>Drosophila melanogaster</i>	PB	Canu	21.31 Mb	(Koren et al. 2017)
<i>Homo sapiens</i> CHM1	PB	Canu	21.95 Mb	(Koren et al. 2017)
<i>Vitis vinifera</i>	PB	FALCON	2.39 Mb	(Chin et al. 2016)
<i>Ipomoea nil</i>	PB + Illumina + LM	HGAP3	1.87 Mb	(Hoshino et al. 2016)
<i>Vigna angularis</i>	PB + Illumina + 454	Sprai, Celera	0.8 Mb	(Sakai et al. 2015)
<i>Oreochromis niloticus</i>	PB + RH map + RAD map	Canu	3.1 Mb	(Conte et al. 2017)
<i>Gorilla gorilla</i>	PB + BAC-end + Fosmid-end	FALCON	9.56 Mb	(Gordon et al. 2016)
<i>Lates calcalifer</i>	PB + OM + LM	HGAP3	1.72 Mb	(Vij et al. 2016)
<i>Capra hircus</i>	PB + OM + HiC	PBcR	18.7 Mb	(Bickhart et al. 2017)
<i>Arabis alpina</i>	PB + OM + HiC	PBcR, FALCON	0.9 Mb	(Jiao et al. 2017)
<i>Euclidium syriacum</i>	PB + OM	PBcR, FALCON	3.3 Mb	(Jiao et al. 2017)
<i>Conringia planisiliqua</i>	PB + OM	PBcR, FALCON	3.6 Mb	(Jiao et al. 2017)
<i>Corvus corone</i>	PB + OM	FALCON	8.91 Mb	(Weissensteiner et al. 2017)
<i>Zea mays</i>	PB + OM	PBcR, FALCON	1.19 Mb	(Jiao et al. 2017)
<i>Homo sapiens</i> NA12878	PB + OM	PBcR, FALCON	1.4 Mb	(Pendleton et al. 2015)
<i>Homo sapiens</i> HX1	PB + OM	FALCON	8.3 Mb	(Shi et al. 2016)
<i>Oropetium thomaeum</i>	PB + OM	HGAP3	2.4 Mb	(VanBuren et al. 2015)
<i>Oryza sativa indica</i>	PB + Fosmids + OM + LM	PBcR	4.43 Mb	(Du et al. 2017)
<i>Homo sapiens</i> NA19240	PB + OM	FALCON	7.25 Mb	(Steinberg et al. 2016)

HGAP3 was found to be the most contiguous assembler, but with the disadvantage of extremely slow computation times. Mis-assemblies were also most abundant in the HGAP3 assemblies, possibly because of the greedier nature of celera's algorithm at the layout stage (Chin et al. 2013). In addition, as previously observed for PBcR in the rice genome assembly (Du et al. 2017), the celera-based assemblers, PBcR and HGAP3, were found to have redundant contigs.

PBcR is the second most widely used long-read assembler (table 4.11); however, it is no longer maintained, since the focus has shifted to its successor canu, which seemed to outperform PBcR in almost every analysis.

SMARTdenovo, although not the best, produced moderately good results in all metrics and would be a suitable choice for obtaining larger genome assemblies quickly.

Leaving out the consensus module, miniasm was the fastest available assembler for all genomes evaluated, excluding *I. nil*. Miniasm requires as much as 13% divergence for repeat resolution, whereas canu and FALCON require only 3% and 5% divergence, respectively (Koren et al. 2017). Hence, miniasm produced fragmented contigs for repeat-rich genomes, but obtained reasonable rankings otherwise.

HINGE may not be ideal for assembling large genomes, but would be a good choice for assembling highly repetitive bacterial genomes.

As observed in the assemblies of the slightly smaller yeast genome (Istace et al. 2017), ABruijn, despite its good contiguity, was chimeric. ABruijn failed to assemble the *I. nil* dataset; however, when the error-corrected reads of canu were used, the assembly was possible but only after consuming almost 500 Gb of maximum RSS.

Similarly, wtdbg was also memory-intensive, and both the assemblers will need high-end servers for handling larger genomes. In the case of repetitive genomes, both assemblers could collapse repeats, leading to loss of information. In particular, the wtdbg assembly was found to be more than 100 Mb short of the expected genome size in *I. nil*. Wtdbg assemblies, which always ranked last, mostly because no consensus procedure was executed, and would need additional rounds of consensus polishing to effectively compete with other assemblers. Wtdbg assemblies also had fragmented contigs.

Mitochondrial genomes were generally left unassembled. Hence it might be necessary to either extract i) reads that do not align to the assembled contigs, or ii) reads that align to an available or a closely related mitochondrial genome. The extracted reads could be used to perform an additional round of assembly, for reconstructing extra-chromosomal genomes (Vembar et al. 2016). In addition, redundancy at the ends of contigs can be a major obstacle for polishing the genome, as it might become difficult for the reads to be aligned at such regions, leaving out errors stranded in the terminal portions of the contigs. Indeed, when whole BAC sequences of *I. nil* were covered by completely spanning contigs, the error rate was approximately homogenous across all the assemblers, whereas when contigs were in overlapping fragmented pieces, the terminal overlapping regions were found to have increased error rates. The same phenomenon was observed in redundant regions from circularly

unresolved bacterial assemblies. Identifying such regions and trimming the redundant base pairs may lead to an improved overall per-base correctness.

Dot plots showed that many of the breakpoints in contig mis-assemblies originated from different locations for different assemblers. Contiguity profiles were also found to be different for FALCON and PBcR in plant genome assemblies, and a hybrid assembly utilising the different contiguity profiles was found to be highly successful (Jiao et al. 2017). Hence an alternative solution to increasing the contiguity would be to combine different assemblies by using reconciliation tools such as quickmerge (Chakraborty et al. 2016). For example, miniasm had fewer contigs and breakpoints compared to MECAT for the *C. elegans* assemblies. Using miniasm assembly as a backbone for extending the MECAT assembly may result in longer and more accurate contigs in this case.

Similar to the evaluation of short read assemblers (Salzberg et al. 2012; Magoc et al. 2013; Earl et al. 2011; Bradnam et al. 2013), the current study did not reveal a clear winner; a similar result was observed with evaluations of Nanopore sequencing data (Istace et al. 2017). That is, an optimal assembler for one dataset may not be optimal for a different dataset. Hence, it would be ideal to try out a variety of assemblers, as performed in the *Solanum pennellii* genome project (Schmidt et al. 2017), and choose the best assembly based on various evaluation strategies. Any available resources such as BAC-end data, whole BAC sequences, previously annotated gene sets, and similar resources could be effectively used for the purpose of evaluation, as demonstrated in this study.

Based on the results, we suggest that the best approach in handling larger genomes would be to generate assemblies from at least canu, FALCON, MECAT, and SMARTdenovo, and basing the final decision on the assembler according to different evaluation metrics rather than on N50 alone. When time is not a limiting factor, HGAP3 could also be used, but care should be taken in recognizing mis-assembled and redundant contigs. Recently, scaffolding techniques such as optical mapping, CHICAGO, Hi-C, and linked reads, have been applied to correct mis-assemblies (Weissensteiner et al. 2017; Bickhart et al. 2017; Shi et al. 2016; Jiao et al. 2017; Pendleton et al. 2015; Du et al. 2017; VanBuren et al. 2015; Jiao et al. 2017; Steinberg et al. 2016), which can also be used for achieving chromosome-scale assemblies.

Chapter 5

Conclusion and future work

In this dissertation, we have presented two research studies related to long-read *de novo* assembly of genomes. The first study constructs the basic layout of *de novo* assembly and analysis using long reads from PacBio technology. The result of the study is a high quality reference genome, not only for *I. nil*, but also as a representative for the whole of Convolvulaceae family. The impact of the *I. nil* reference genome was immediately witnessable. A study published soon after the publication of the assembled genome of *I. nil*, hypothesizes that severe stress events, such as mass extinctions, must have occurred at different time points in the evolutionary history and such time points were ideal for the occurrence of WGD events in plants, leading to enhanced adaptation to the modified environment (Van de Peer et al. 2017). The analysis included the results of *I. nil* WGD estimations, which fitted perfectly into their hypothesis, thus adding furthermore weight to their publication, while also serving as a validation for our estimations. Another broader impact was that the pseudo-chromosomes of *I. nil* were used as a synteny reference to create a pseudo-chromosomal map of *Ipomoea batatas* (sweet potato), a close taxonomic neighbour of *I. nil* in the Convolvulaceae family (Yang et al. 2017).

The second study improves on the assembly aspect of the first study, by evaluating the assemblies from different assemblers for various organisms. Contrary to the previous publications, which may mislead PBcR as an ideal assembly tool, the study rejects false notions and recommends the right assemblers for respective datasets. The conclusions of the study would relieve researchers from the pain of looking for the right parameters and readily apply the recommendations for their assembly projects. Another important aspect of the study is that computational resources are measured for assemblies of genomes of different complexities. Hence, a researcher can choose an assembler which will scale accordingly to their computational resources and in the process saving several weeks/months of time. Both the studies will serve as a valuable reference for *de novo* assembly and analysis of genomes for other researchers.

As for the future works stemming from the studies presented in this thesis, a couple of projects are in progress. The availability of an assembled reference genome opens up on a lot of research possibilities. Chapter 3 explained how *Tpn1* transposons can be a major mutagen in *I. nil*, which are cataloged from the genome as part of the study. Knowing the potential of the transposons, it will be intriguing if they play a larger role in mutagenic lines. To study the same, bisulfite sequencing was done for control and mutagenic plants, while the aim of the study is to observe differential methylation patterns in genes, and also in the catalogued *Tpn1* transposons across the mutagenic and control lines. By doing so, a genome-wide analysis can be performed which will pinpoint active and passive *Tpn1* transposon locations, elucidating the role of the transposons. The results can also be used as a reference for the other mutagenic lines too.

From chapter 4, recommendations for the right assemblers were chosen and applied to a different assembly project. Common marmoset's genome has already been assembled but with a lot of gaps, paving way for a lot of improvement in the quality of the genome (Marmoset Genome Sequencing and Analysis Consortium 2014; Sato et al. 2015). The common marmoset with a small body size, sharing similar physiology with humans, has garnered attention recently as a new non-human primate model organism. Hence, a high quality genome will be essential to obtain the necessary biological insights. We have obtained around 50X PacBio data for the common marmoset genome and are in the process of applying the recommended assemblers from the evaluation study to assemble the data. From the assemblies, we have identified that more than 90% of the gaps in the previous genome assembly could be filled with the results. Another suggestion from the evaluation study is that hybrid assemblies from two or more assemblies can result in a better genome assembly. In line with this, we also aim to develop a tool which would compare individual assemblies and generate a hybrid and more contiguous assembly.

Also, insights from both studies can be applied to the improvement of the *I. nil* genome. Optical mapping experiments (Iris from Bionano genomics) are currently underway for *I. nil*. Combining the idea of a hybrid assembly from the evaluation study,

along with the optical map data and the linkage map data, we can generate much more accurate and highly contiguous assemblies for the *I. nil* genome.

Acknowledgements

First of all, I would like to express my sincere thanks to Professor Yasubumi Sakakibara who has provided comprehensive support for my study as a supervisor of my doctor course. I greatly appreciate that he gave me a lot of opportunities for presenting my study in journal papers and conferences. These valuable experiences laid the foundation of my ability, and helped me make up my mind to become a professional researcher. I also appreciate his patience for supervising me to study and conduct research. It is not an exaggeration to say that he has helped me a lot to feel settled living in Japan, which greatly supported my study. I also would like to express my cordial gratitude to Assistant Professor Kengo Sato who has provided a number of technical advice for my study.

I am very grateful to my colleagues in Sakakibara Laboratory, for making a good academic environment. I would like to thank Kojiro Amano, Tatsumu Inatsuki, Wataru Shintani, Yoshimasa Aoto, Afia Hayati, Yugaku Tanaka, Sae Shirakizawa, Mariko Tsuchiya, Motoki Abe, Misato Seki, Taisuke Nishikawa, Manato Akiyama, and Genta Aoki for helping me out with personal as well as academic issues. All these people also helped me better my presentation slides from almost zero to a level which is presentable to audiences. I wish to thank MEXT scholarship program and JSPS KAKENHI Grant (Number 16H06279) for the generous financial support for my school and living expenses without which I could not have completed my doctoral program. I would also like to thank my father and mother for always supporting me throughout my career.

Finally, I would like to express my sincere gratitude to Professor Yasubumi Sakakibara, Professor Kotaro Oka, Professor Nobuhide Doi, Associate Professor Akira Funahashi, and Professor Akito Sakurai for examining and judging my doctoral dissertation.

References

- Adams, Mark D., Jenny M. Kelley, Jeannine D. Gocayne, Mark Dubnick, Mihael H. Polymeropoulos, Hong Xiao, Carl R. Merrill, Andrew Wu, Bjorn Olde, Ruben F. Moreno, Anthony R. Kerlavage, W. Richard McCombie, and J. Craig Venter. 1991. "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project." *Science* 252 (5013):1651–56.
- Allen, Scott L., Emily K. Delaney, Artyom Kopp, and Stephen F. Chenoweth. 2017. "Single-Molecule Sequencing of the *Drosophila Serrata* Genome." *G3* 7 (3):781–88.
- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3):403–10.
- Austin, Daniel F., and Zósimo Huáman. 1996. "A Synopsis of Ipomoea (Convolvulaceae) in the Americas." *Taxon* 45 (1). International Association for Plant Taxonomy (IAPT):3–38.
- Baird, Nathan A., Paul D. Etter, Tressa S. Atwood, Mark C. Currey, Anthony L. Shiver, Zachary A. Lewis, Eric U. Selker, William A. Cresko, and Eric A. Johnson. 2008. "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers." *PloS One* 3 (10):e3376.
- Barba, Marina, Henryk Czosnek, and Ahmed Hadidi. 2014. "Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology." *Viruses* 6 (1):106–36.
- Benson, Gary. 1999. "Tandem Repeats Finder: A Program to Analyze DNA Sequences." *Nucleic Acids Research* 27 (2):573–80.
- Berlin, Konstantin, Sergey Koren, Chen-Shan Chin, James P. Drake, Jane M. Landolin, and Adam M. Phillippy. 2015. "Assembling Large Genomes with Single-Molecule Sequencing and Locality-Sensitive Hashing." *Nature Biotechnology* 33 (6):623–30.
- Bickhart, Derek M., Benjamin D. Rosen, Sergey Koren, Brian L. Sayre, Alex R. Hastie, Saki Chan, Joyce Lee, et al. 2017. "Single-Molecule Sequencing and Chromatin Conformation Capture Enable *de Novo* Reference Assembly of the Domestic Goat Genome." *Nature Genetics* 49 (4):643–50.
- Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15):2114–20.
- Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, and Alexei J. Drummond. 2014. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis." *PLoS Computational Biology* 10 (4):e1003537.
- Bradnam, Keith R., Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, et al. 2013. "Assemblathon 2: Evaluating *de Novo* Methods of Genome Assembly in Three Vertebrate Species." *GigaScience* 2 (1):10.
- Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." *Bioinformatics* 25 (15):1972–73.
- Cartwright, Dustin A., Michela Troggio, Riccardo Velasco, and Alexander Gutin. 2007. "Genetic Mapping in the Presence of Genotyping Errors." *Genetics* 176 (4):2521–27.
- Catchen, Julian M., Angel Amores, Paul Hohenlohe, William Cresko, and John H. Postlethwait. 2011. "Stacks: Building and Genotyping Loci *de Novo* from Short-Read Sequences." *G3* 1 (3):171–82.
- C. elegans Sequencing Consortium. 1998. "Genome Sequence of the Nematode *C. Elegans*: A Platform for Investigating Biology." *Science* 282 (5396):2012–18.
- Chaisson, Mark J. P., Richard K. Wilson, and Evan E. Eichler. 2015. "Genetic Variation and the *de Novo* Assembly of Human Genomes." *Nature Reviews. Genetics* 16 (11):627–40.
- Chaisson, Mark J. P., and Glenn Tesler. 2012. "Mapping Single Molecule Sequencing Reads Using Basic Local Alignment with Successive Refinement (BLASR): Application and Theory." *BMC Bioinformatics* 13 (September):238.
- Chakraborty, Mahul, James G. Baldwin-Brown, Anthony D. Long, and J. J. Emerson. 2016. "Contiguous and Accurate *de Novo* Assembly of Metazoan Genomes with Modest Long Read Coverage." *Nucleic Acids Research* 44 (19):e147.
- Chen, Qingfeng, Chaowang Lan, Liang Zhao, Jianxin Wang, Baoshan Chen, and Yi-Ping Phoebe Chen. 2017. "Recent Advances in Sequence Assembly: Principles and Applications." *Briefings in Functional Genomics*, April. <https://doi.org/10.1093/bfpg/elx006>.
- Cherukuri, Yesesri, and Sarath Chandra Janga. 2016. "Benchmarking of *de Novo* Assembly Algorithms

- for Nanopore Data Reveals Optimal Performance of OLC Approaches." *BMC Genomics* 17 Suppl 7 (August):507.
- Chin, Chen-Shan, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. "Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data." *Nature Methods* 10 (6):563–69.
- Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. "Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing." *Nature Methods*, October. <https://doi.org/10.1038/nmeth.4035>.
- Chono, Makiko, Ichiro Honda, Haruko Zeniya, Koichi Yoneyama, Daisuke Saisho, Kazuyoshi Takeda, Suguru Takatsuto, Tsuguhiro Hoshino, and Yoshiaki Watanabe. 2003. "A Semidwarf Phenotype of Barley Uzu Results from a Nucleotide Substitution in the Gene Encoding a Putative Brassinosteroid Receptor." *Plant Physiology* 133 (3):1209–19.
- Chu, Justin, Hamid Mohamadi, René L. Warren, Chen Yang, and Inanç Birol. 2017. "Innovations and Challenges in Detecting Long Read Overlaps: An Evaluation of the State-of-the-Art." *Bioinformatics* 33 (8):1261–70.
- Clegg, Michael T., and Mary L. Durbin. 2003. "Tracing Floral Adaptations from Ecology to Molecules." *Nature Reviews. Genetics* 4 (3):206–15.
- Conte, Matthew A., William J. Gammerdinger, Kerry L. Bartie, David J. Penman, and Thomas D. Kocher. 2017. "A High Quality Assembly of the Nile Tilapia (*Oreochromis Niloticus*) Genome Reveals the Structure of Two Sex Determination Regions." *BMC Genomics* 18 (1):341.
- Du, Huilong, Ying Yu, Yanfei Ma, Qiang Gao, Yinghao Cao, Zhuo Chen, Bin Ma, et al. 2017. "Sequencing and *de Novo* Assembly of a near Complete Indica Rice Genome." *Nature Communications* 8 (May):15324.
- Du, Zhou, Xin Zhou, Yi Ling, Zhenhai Zhang, and Zhen Su. 2010. "agriGO: A GO Analysis Toolkit for the Agricultural Community." *Nucleic Acids Research* 38 (Web Server issue):W64–70.
- Earl, Dent, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, et al. 2011. "Assemblathon 1: A Competitive Assessment of *de Novo* Short Read Assembly Methods." *Genome Research* 21 (12):2224–41.
- English, Adam C., Stephen Richards, Yi Han, Min Wang, Vanesa Vee, Jiaxin Qu, Xiang Qin, et al. 2012. "Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology." *PloS One* 7 (11):e47768.
- Ezkurdia, Iakes, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L. Tress. 2014. "Multiple Evidence Strands Suggest That There May Be as Few as 19,000 Human Protein-Coding Genes." *Human Molecular Genetics* 23 (22):5866–78.
- Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. 1995. "Whole-Genome Random Sequencing and Assembly of *Haemophilus Influenzae* Rd." *Science* 269 (5223):496–512.
- Fukada-Tanaka, Sachiko., Yoshishige Inagaki, Toshio Yamaguchi, Norio Saito, and Shigeru Iida. 2000. "Colour-Enhancing Protein in Blue Petals." *Nature* 407 (6804):581.
- Girgis, Hani Z. 2015. "Red: An Intelligent, Rapid, Accurate Tool for Detecting Repeats *de-Novo* on the Genomic Scale." *BMC Bioinformatics* 16 (July):227.
- Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, et al. 1996. "Life with 6000 Genes." *Science* 274 (5287):546, 563–67.
- Gordon, David, John Huddleston, Mark J. P. Chaisson, Christopher M. Hill, Zev N. Kronenberg, Katherine M. Munson, Maika Malig, et al. 2016. "Long-Read Sequence Assembly of the Gorilla Genome." *Science* 352 (6281):aae0344.
- Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics* 29 (8):1072–75.
- Hagiwara, Tokio. 1956. "Genes and Chromosome Maps in the Japanese Morning Glory." *Bull. Res. Coll. Agric. Vet. Sci. Nihon Univ.* 5:34–56.
- Hatem, Ayat, Doruk Bozdağ, Amanda E. Toland, and Ümit V. Çatalyürek. 2013. "Benchmarking Short Sequence Mapping Tools." *BMC Bioinformatics* 14 (June):184.
- Hedges, S. Blair, Julie Marin, Michael Suleski, Madeline Paymer, and Sudhir Kumar. 2015. "Tree of Life Reveals Clock-like Speciation and Diversification." *Molecular Biology and Evolution* 32 (4):835–45.
- Hirakawa, Hideki, Yoshihiro Okada, Hiroaki Tabuchi, Kenta Shirasawa, Akiko Watanabe, Hisano

- Tsuruoka, Chiharu Minami, et al. 2015. "Survey of Genome Sequences in a Wild Sweet Potato, *Ipomoea Trifida* (H. B. K.) G. Don." *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 22 (2):171–79.
- Hoshino, Atsushi, Vasanthan Jayakumar, Eiji Nitasaka, Atsushi Toyoda, Hideki Noguchi, Takehiko Itoh, Tadasu Shin-I, et al. 2016. "Genome Sequence and Analysis of the Japanese Morning Glory *Ipomoea Nil*." *Nature Communications* 7 (November):13295.
- Hoshino, Atsushi, Kyeung-Il Park, and Shigeru Iida. 2009. "Identification of R Mutations Conferring White Flowers in the Japanese Morning Glory (*Ipomoea Nil*)." *Journal of Plant Research* 122 (2):215–22.
- Huang, Shengxiong, Jian Ding, Dejing Deng, Wei Tang, Honghe Sun, Dongyuan Liu, Lei Zhang, et al. 2013. "Draft Genome of the Kiwifruit *Actinidia Chinensis*." *Nature Communications* 4:2640.
- Hunt, Martin, Nishadi De Silva, Thomas D. Otto, Julian Parkhill, Jacqueline A. Keane, and Simon R. Harris. 2015. "Circlator: Automated Circularization of Genome Assemblies Using Long Sequencing Reads." *Genome Biology* 16 (December):294.
- Hwang, Sohyun, Eiru Kim, Insuk Lee, and Edward M. Marcotte. 2015. "Systematic Comparison of Variant Calling Pipelines Using Gold Standard Personal Exome Variants." *Scientific Reports* 5 (December):17875.
- Ibarra-Laclette, Enrique, Eric Lyons, Gustavo Hernández-Guzmán, Claudia Anahí Pérez-Torres, Lorenzo Carretero-Paulet, Tien-Hao Chang, Tianying Lan, et al. 2013. "Architecture and Evolution of a Minute Plant Genome." *Nature* 498 (7452):94–98.
- Imai, Yoshitaka. 1929. "Linkage Groups of the Japanese Morning Glory." *Genetics* 14 (3):223–55.
- Inagaki, Yoshishige., Yasuyo Hisatomi, Tetsuya Suzuki, Kichiji Kasahara, and Shigeru Iida. 1994. "Isolation of a Suppressor-mutator/Enhancer-like Transposable Element, Tpn1, from Japanese Morning Glory Bearing Variegated Flowers." *The Plant Cell* 6 (3):375–83.
- Istace, Benjamin, Anne Friedrich, Léo d'Agata, Sébastien Faye, Emilie Payen, Odette Beluche, Claudia Caradec, et al. 2017. "De Novo Assembly and Population Genomic Survey of Natural Yeast Isolates with the Oxford Nanopore MinION Sequencer." *GigaScience* 6 (2):1–13.
- Iwasaki, Mayumi, and Eiji Nitasaka. 2006. "The FEATHERED Gene Is Required for Polarity Establishment in Lateral Organs Especially Flowers of the Japanese Morning Glory (*Ipomoea Nil*)." *Plant Molecular Biology* 62 (6):913–25.
- Jaillon, Olivier, Jean-Marc Aury, Benjamin Noel, Alberto Policriti, Christian Clepet, Alberto Casagrande, Nathalie Choise, et al. 2007. "The Grapevine Genome Sequence Suggests Ancestral Hexaploidization in Major Angiosperm Phyla." *Nature* 449 (7161):463–67.
- Jain, Miten, Sergey Koren, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, Andrew D. Beggs, et al. 2017. "Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads." *bioRxiv*. <https://doi.org/10.1101/128835>.
- Jiao, Wen-Biao, Gonzalo Garcia Accinelli, Benjamin Hartwig, Christiane Kiefer, David Baker, Edouard Severing, Eva-Maria Willing, et al. 2017. "Improving and Correcting the Contiguity of Long-Read Genome Assemblies of Three Plant Species Using Optical Mapping and Chromosome Conformation Capture Data." *Genome Research* 27 (5):778–86.
- Jiao, Yinping, Paul Peluso, Jinghua Shi, Tiffany Liang, Michelle C. Stitzer, Bo Wang, Michael S. Campbell, et al. 2017. "Improved Maize Reference Genome with Single-Molecule Technologies." *Nature* 546 (7659):524–27.
- Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30 (9):1236–40.
- Kamath, Govinda M., Ilan Shomorony, Fei Xia, Thomas A. Courtade, and David N. Tse. 2017. "HINGE: Long-Read Assembly Achieves Optimal Repeat Resolution." *Genome Research* 27 (5):747–56.
- Katoh, Kazutaka, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. 2002. "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform." *Nucleic Acids Research* 30 (14):3059–66.
- Kaul, Samir, Hean L. Koo, Jennifer Jenkins, Michael Rizzo, Timothy Rooney, Luke J. Tallon, Tamara Feldblyum, et al. 2000. "Analysis of the Genome Sequence of the Flowering Plant *Arabidopsis Thaliana*." *Nature* 408 (6814). Nature Publishing Group:796–815.
- Kawasaki, Sayaka, and Eiji Nitasaka. 2004. "Characterization of Tpn1 Family in the Japanese Morning Glory: En/Spm-Related Transposable Elements Capturing Host Genes." *Plant & Cell Physiology* 45 (7):933–44.

- Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. "TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions." *Genome Biology* 14 (4):R36.
- Kim, Seungill, Minkyu Park, Seon-In Yeom, Yong-Min Kim, Je Min Lee, Hyun-Ah Lee, Eunyoung Seo, et al. 2014. "Genome Sequence of the Hot Pepper Provides Insights into the Evolution of Pungency in Capsicum Species." *Nature Genetics* 46 (3):270–78.
- Kitazawa, Daisuke, Yasuko Hatakeda, Motoshi Kamada, Nobuharu Fujii, Yutaka Miyazawa, Atsushi Hoshino, Shigeru Iida, et al. 2005. "Shoot Circumnutation and Winding Movements Require Gravisensing Cells." *Proceedings of the National Academy of Sciences of the United States of America* 102 (51):18742–47.
- Koren, Sergey, Michael C. Schatz, Brian P. Walenz, Jeffrey Martin, Jason T. Howard, Ganeshkumar Ganapathy, Zhong Wang, et al. 2012. "Hybrid Error Correction and *de Novo* Assembly of Single-Molecule Sequencing Reads." *Nature Biotechnology* 30 (7):693–700.
- Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation." *Genome Research* 27 (5):722–36.
- Korlach, Jonas, Gregory Gedman, Sarah Kingan, Jason Chin, Jason Howard, Lindsey Cantin, and Erich D. Jarvis. 2017. "De Novo PacBio Long-Read and Phased Avian Genome Assemblies Correct and Add to Genes Important in Neuroscience Research." *bioRxiv*. <https://doi.org/10.1101/103911>.
- Krzywinski, Martin, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. 2009. "Circos: An Information Aesthetic for Comparative Genomics." *Genome Research* 19 (9):1639–45.
- Kurtz, Stefan, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. 2004. "Versatile and Open Software for Comparing Large Genomes." *Genome Biology* 5 (2):R12.
- International Human Genome Sequencing Consortium. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822):860–921.
- Lan, Tianying, Tanya Renner, Enrique Ibarra-Laclette, Kimberly M. Farr, Tien-Hao Chang, Sergio Alan Cervantes-Pérez, Chunfang Zheng, et al. 2017. "Long-Read Sequencing Uncovers the Adaptive Topography of a Carnivorous Plant Genome." *Proceedings of the National Academy of Sciences of the United States of America* 114 (22):E4435–41.
- Lee, Hyan, James Gurtowski, Shinjae Yoo, Maria Nattestad, Shoshana Marcus, Sara Goodwin, W. Richard McCombie, and Michael Schatz. 2016. "Third-Generation Sequencing and the Future of Genomics." <https://doi.org/10.1101/048603>.
- Liao, Yu-Chieh, Shu-Hung Lin, and Hsin-Hung Lin. 2015. "Completing Bacterial Genome Assemblies: Strategy and Performance Comparisons." *Scientific Reports* 5 (March):8747.
- Li, Heng. 2016. "Minimap and Miniasm: Fast Mapping and *de Novo* Assembly for Noisy Long Sequences." *Bioinformatics* 32 (14):2103–10.
- Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (14):1754–60.
- Li, Li, Christian J. Stoeckert Jr, and David S. Roos. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Research* 13 (9):2178–89.
- Lin, Yu, Jeffrey Yuan, Mikhail Kolmogorov, Max W. Shen, and Pavel A. Pevzner. 2016. "Assembly of Long Error-Prone Reads Using de Bruijn Graphs." *bioRxiv*. <https://doi.org/10.1101/048413>.
- Li, Ruiqiang, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, et al. 2010. "De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing." *Genome Research* 20 (2):265–72.
- Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. "SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read *de Novo* Assembler." *GigaScience* 1 (1):18.
- Ly, Tong, Hiroyuki Fukuoka, Asami Otaka, Atsushi Hoshino, Shigeru Iida, Eiji Nitasaka, Nobuyoshi Watanabe, and Tsutomu Kuboyama. 2012. "Development of EST-SSR Markers of *Ipomoea Nil*." *Breeding Science* 62 (1):99–104.
- Magoc, Tanja, Stephan Pabinger, Stefan Canzar, Xinyue Liu, Qi Su, Daniela Puiu, Luke J. Tallon, and Steven L. Salzberg. 2013. "GAGE-B: An Evaluation of Genome Assemblers for Bacterial Organisms." *Bioinformatics* 29 (14):1718–25.

- Margarido, Gabriel R. A., Anete P. Souza, and Antonio A. F. Garcia. 2007. "OneMap: Software for Genetic Mapping in Outcrossing Species." *Hereditas* 144 (3):78–79.
- Marmoset Genome Sequencing and Analysis Consortium. 2014. "The Common Marmoset Genome Provides Insight into Primate Biology and Evolution." *Nature Genetics* 46 (8):850–57.
- Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBNet.journal* 17 (1):10–12.
- McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernysky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9):1297–1303.
- Melsted, Páll, and Jonathan K. Pritchard. 2011. "Efficient Counting of K-Mers in DNA Sequences Using a Bloom Filter." *BMC Bioinformatics* 12 (August):333.
- Michael, Todd P., and Scott Jackson. 2013. "The First 50 Plant Genomes." *The Plant Genome* 6 (2).
- Morita, Yasumasa, Kanako Ishiguro, Yoshikazu Tanaka, Shigeru Iida, and Atsushi Hoshino. 2015. "Spontaneous Mutations of the UDP-Glucose:flavonoid 3-O-Glucosyltransferase Gene Confers Pale- and Dull-Colored Flowers in the Japanese and Common Morning Glories." *Planta* 242 (3):575–87.
- Morita, Yasumasa, Kyoko Takagi, Masako Fukuchi-Mizutani, Kanako Ishiguro, Yoshikazu Tanaka, Eiji Nitasaka, Masayoshi Nakayama, et al. 2014. "A Chalcone Isomerase-like Protein Enhances Flavonoid Production and Flower Pigmentation." *The Plant Journal: For Cell and Molecular Biology* 78 (2):294–304.
- Myers, Eugene W., Granger G. Sutton, Art L. Delcher, Ian M. Dew, Dan P. Fasulo, Michael J. Flanigan, Saul A. Kravitz, et al. 2000. "A Whole-Genome Assembly of *Drosophila*." *Science* 287 (5461):2196–2204.
- Myers, Eugene W. 2014 "A History of DNA Sequence Assembly." It - Information Technology, ISSN (Online) 2196-7032, ISSN (Print) 1611-2776.
- Myers, Eugene W. 2016. "Efficient Local Alignment Discovery amongst Noisy Long Reads." Springer Berlin Heidelberg, 52–67. Accessed November 9, 2016.
- Nacken, Wolfgang K. F., Ralf Piotrowiak, Heinz Saedler, and Hans Sommer. 1991. "The Transposable Element Tam1 from *Antirrhinum Majus* Shows Structural Homology to the Maize Transposon En/Spm and Has No Sequence Specificity of Insertion." *Molecular & General Genetics: MGG* 228 (1-2):201–8.
- Nattestad, Maria, and Michael C. Schatz. 2016. "Assemblytics: A Web Analytics Tool for the Detection of Variants from an Assembly." *Bioinformatics* 32 (19):3021–23.
- Nawrocki, Eric P., and Sean R. Eddy. 2013. "Infernal 1.1: 100-Fold Faster RNA Homology Searches." *Bioinformatics* 29 (22):2933–35.
- Nitasaka, Eiji. 2003. "Insertion of an En/Spm-Related Transposable Element into a Floral Homeotic Gene DUPLICATED Causes a Double Flower Phenotype in the Japanese Morning Glory." *The Plant Journal: For Cell and Molecular Biology* 36 (4). Blackwell Science Ltd:522–31.
- Ouyang, Shu, Wei Zhu, John Hamilton, Haining Lin, Matthew Campbell, Kevin Childs, Françoise Thibaud-Nissen, et al. 2007. "The TIGR Rice Genome Annotation Resource: Improvements and New Features." *Nucleic Acids Research* 35 (Database issue):D883–87.
- Parra, Genis, Keith Bradnam, and Ian Korf. 2007. "CEGMA: A Pipeline to Accurately Annotate Core Genes in Eukaryotic Genomes." *Bioinformatics* 23 (9):1061–67.
- Pendleton, Matthew, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M. Stütz, et al. 2015. "Assembly and Diploid Architecture of an Individual Human Genome via Single-Molecule Technologies." *Nature Methods* 12 (8):780–86.
- Potato Genome Sequencing Consortium, Xun Xu, Shengkai Pan, Shifeng Cheng, Bo Zhang, Desheng Mu, Peixiang Ni, et al. 2011. "Genome Sequence and Analysis of the Tuber Crop Potato." *Nature* 475 (7355):189–95.
- Price, Alkes L., Neil C. Jones, and Pavel A. Pevzner. 2005. "De Novo Identification of Repeat Families in Large Genomes." *Bioinformatics* 21 Suppl 1 (June):i351–58.
- Roberts, Michael, Wayne Hayes, Brian R. Hunt, Stephen M. Mount, and James A. Yorke. 2004. "Reducing Storage Requirements for Biological Sequence Comparison." *Bioinformatics* 20 (18):3363–69.
- Sahlin, Kristoffer, Francesco Vezzi, Björn Nystedt, Joakim Lundeberg, and Lars Arvestad. 2014. "BESST--Efficient Scaffolding of Large Fragmented Assemblies." *BMC Bioinformatics* 15 (August):281.
- Sakaguchi, Shota, Takeshi Sugino, Yoshihiko Tsumura, Motomi Ito, Michael D. Crisp, David M J, Atsushi

- J. Nagano, et al. 2015. "High-Throughput Linkage Mapping of Australian White Cypress Pine (*Callitris Glaucophylla*) and Map Transferability to Related Species." *Tree Genetics & Genomes* 11 (6). Springer Berlin Heidelberg:121.
- Sakai, Hiroaki, Ken Naito, Eri Ogiso-Tanaka, Yu Takahashi, Kohtaro Iseki, Chiaki Muto, Kazuhito Satou, et al. 2015. "The Power of Single Molecule Real-Time Sequencing Technology in the *de Novo* Assembly of a Eukaryotic Genome." *Scientific Reports* 5 (November):16780.
- Salzberg, Steven L., Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen, et al. 2012. "GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms." *Genome Research* 22 (3):557–67.
- Sanger, Frederick, Gilian M. Air, Bart G. Barrell, Nigel L. Brown, Alan R. Coulson, J. C. Fiddes, C. A. Hutchison, Patrick M. Slocombe, and Mo Smith. 1977. "Nucleotide Sequence of Bacteriophage ϕ X174 DNA." *Nature* 265 (5596). Springer:687–95.
- Sato, Kengo, Yoko Kuroki, Wakako Kumita, Asao Fujiyama, Atsushi Toyoda, Jun Kawai, Atsushi Iriki, Erika Sasaki, Hideyuki Okano, and Yasubumi Sakakibara. 2015. "Resequencing of the Common Marmoset Genome Improves Genome Assemblies and Gene-Coding Sequence Analysis." *Scientific Reports* 5 (November):16894.
- Schatz, Michael C., Lyza G. Maron, Joshua C. Stein, Alejandro Hernandez Wences, James Gurtowski, Eric Biggers, Hayan Lee, et al. 2014. "Whole Genome *de Novo* Assemblies of Three Divergent Strains of Rice, *Oryza Sativa*, Document Novel Gene Space of Aus and Indica." *Genome Biology* 15 (11):506.
- Schmidt, Maximilian H-W, Alexander Vogel, Alisandra Denton, Benjamin Istace, Alexandra Wormit, Henri van de Geest, Marie E. Bolger, et al. 2017. "Reconstructing The Gigabase Plant Genome Of *Solanum Pennellii* Using Nanopore Sequencing." *bioRxiv*. <https://doi.org/10.1101/129148>.
- Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. "Critical Assessment of Metagenome Interpretation-a Benchmark of Metagenomics Software." *Nature Methods* 14 (11):1063–71.
- Shi, Lingling, Yunfei Guo, Chengliang Dong, John Huddleston, Hui Yang, Xiaolu Han, Aisi Fu, et al. 2016. "Long-Read Sequencing and *de Novo* Assembly of a Chinese Genome." *Nature Communications* 7 (June):12065.
- Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19):3210–12.
- Simpson, Jared T., and Richard Durbin. 2010. "Efficient Construction of an Assembly String Graph Using the FM-Index." *Bioinformatics* 26 (12):i367–73.
- Simpson, Jared T., and Mihai Pop. 2015. "The Theory and Practice of Genome Sequence Assembly." *Annual Review of Genomics and Human Genetics* 16 (April):153–72.
- Sović, Ivan, Krešimir Križanović, Karolj Skala, and Mile Šikić. 2016. "Evaluation of Hybrid and Non-Hybrid Methods for *de Novo* Assembly of Nanopore Reads." *Bioinformatics* 32 (17):2582–89.
- Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics* 30 (9):1312–13.
- Stanke, Mario, and Stephan Waack. 2003. "Gene Prediction with a Hidden Markov Model and a New Intron Submodel." *Bioinformatics* 19 Suppl 2 (October):ii215–25.
- Stefanovic, Sasa, Lori Krueger, and Richard G. Olmstead. 2002. "Monophyly of the Convolvulaceae and Circumscription of Their Major Lineages Based on DNA Sequences of Multiple Chloroplast Loci." *American Journal of Botany* 89 (9):1510–22.
- Steinberg, Karyn Meltz, Tina Graves-Lindsay, Valerie A. Schneider, Mark J. P. Chaisson, Chad Tomlinson, John L. Huddleston, Patrick Minx, et al. 2016. "High-Quality Assembly of an Individual of Yoruban Descent." *bioRxiv*. <https://doi.org/10.1101/067447>.
- Steinberg, Karyn Meltz, Valerie A. Schneider, Tina A. Graves-Lindsay, Robert S. Fulton, Richa Agarwala, John Huddleston, Sergey A. Shiryev, et al. 2014. "Single Haplotype Assembly of the Human Genome from a Hydatidiform Mole." *Genome Research* 24 (12):2066–76.
- Steinhauser, Sebastian, Nils Kurzawa, Roland Eils, and Carl Herrmann. 2016. "A Comprehensive Comparison of Tools for Differential ChIP-Seq Analysis." *Briefings in Bioinformatics* 17 (6):953–66.
- Takahashi, Shigekazu., Yoshishige Inagaki, Hiroyuki Satoh, Atsushi Hoshino, and Shigeru Iida. 1999. "Capture of a Genomic HMG Domain Sequence by the En/Spm-Related Transposable Element Tpn1 in the Japanese Morning Glory." *Molecular & General Genetics: MGG* 261 (3):447–51.

- Tomato Genome Consortium. 2012. "The Tomato Genome Sequence Provides Insights into Fleshy Fruit Evolution." *Nature* 485 (7400):635–41.
- Tyson, John R., Nigel J. O'Neil, Miten Jain, Hugh E. Olsen, Philip Hieter, and Terrance P. Snutch. 2017. "Whole Genome Sequencing and Assembly of a *Caenorhabditis Elegans* Genome with Complex Genomic Rearrangements Using the MinION Sequencing Device." *bioRxiv*. <https://doi.org/10.1101/099143>.
- VanBuren, Robert, Doug Bryant, Patrick P. Edger, Haibao Tang, Diane Burgess, Dinakar Challabathula, Kristi Spittle, et al. 2015. "Single-Molecule Sequencing of the Desiccation-Tolerant Grass *Oropetium Thomaeanum*." *Nature* 527 (7579):508–11.
- Van de Peer, Yves, Eshchar Mizrahi, and Kathleen Marchal. 2017. "The Evolutionary Significance of Polyploidy." *Nature Reviews. Genetics* 18 (7):411–24.
- Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads." *Genome Research* 27 (5):737–46.
- Vembar, Shruthi Sridhar, Matthew Seetin, Christine Lambert, Maria Nattestad, Michael C. Schatz, Primo Baybayan, Artur Scherf, and Melissa Laird Smith. 2016. "Complete Telomere-to-Telomere de Novo Assembly of the Plasmodium Falciparum Genome through Long-Read (>11 Kb), Single Molecule, Real-Time Sequencing." *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 23 (4):339–51.
- Vij, Shubha, Heiner Kuhl, Inna S. Kuznetsova, Aleksey Komissarov, Andrey A. Yurchenko, Peter Van Heusden, Siddharth Singh, et al. 2016. "Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-Layered Scaffolding." *PLoS Genetics* 12 (4):e1005954.
- Wang, Yupeng, Haibao Tang, Jeremy D. Debarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-Ho Lee, et al. 2012. "MCScanX: A Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity." *Nucleic Acids Research* 40 (7):e49.
- Wei, Changhe, Xiang Tao, Ming Li, Bin He, Lang Yan, Xuemei Tan, and Yizheng Zhang. 2015. "De Novo Transcriptome Assembly of *Ipomoea Nil* Using Illumina Sequencing for Gene Discovery and SSR Marker Identification." *Molecular Genetics and Genomics: MGG* 290 (5):1873–84.
- Weil, Clifford F., and Reinhard Kunze. 2002. "The hAT and CACTA Superfamilies of Plant Transposons." In *Mobile DNA II*, edited by Nancy L. Craig, Alan M. Lambowitz, Robert Craigie, and Martin Gellert, 565–610. American Society of Microbiology.
- Weissensteiner, Matthias H., Andy W. C. Pang, Ignas Bunikis, Ida Höijer, Olga Vinnere-Petterson, Alexander Suh, and Jochen B. W. Wolf. 2017. "Combination of Short-Read, Long-Read, and Optical Mapping Assemblies Reveals Large-Scale Tandem Repeat Arrays with Population Genetic Implications." *Genome Research* 27 (5):697–708.
- Xiao, Chuan-Le, Ying Chen, Shang-Qian Xie, Kai-Ning Chen, Yan Wang, Feng Luo, and Zhi Xie. 2016. "MECAT: An Ultra-Fast Mapping, Error Correction and de Novo Assembly Tool for Single-Molecule Sequencing Reads." *bioRxiv*. <https://doi.org/10.1101/089250>.
- Yang, Jun, M-Hossein Moeinzadeh, Heiner Kuhl, Johannes Helmuth, Peng Xiao, Stefan Haas, Guiling Liu, et al. 2017. "Haplotype-Resolved Sweet Potato Genome Traces Back Its Hexaploidization History." *Nature Plants*, August. <https://doi.org/10.1038/s41477-017-0002-z>.
- Yang, Ziheng. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8):1586–91.
- Yasui, Kono. 1928. "Studies on *Pharbitis Nil*, CHOIS. II Chromosome Number." *植物学雑誌* 42 (502):480–85.
- Zhang, Zong Hong, Dhanisha J. Jhaveri, Vikki M. Marshall, Denis C. Bauer, Janette Edson, Ramesh K. Narayanan, Gregory J. Robinson, et al. 2014. "A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data." *PloS One* 9 (8):e103207.

Appendix A

List of publications

Journal papers

1. Hoshino, A*, Jayakumar, V*, Nitasaka, E, Toyoda, A, Noguchi, H, Itoh, T, Shin-I, T, Minakuchi, Y, Koda, Y, Nagano, A, Yasugi, M, Honjo, M, Kudoh, H, Seki, M, Kamiya, A, Shiraki, T, Carninci, P, Asamizu, E, Nishide, H, Tanaka, S, Park, K, Morita, Y, Yokoyama, K, Uchiyama, I, Tanaka, Y, Tabata, S, Shinozaki, K, Hayashizaki, Y, Kohara, Y, Suzuki, Y, Sugano, S, Fujiyama, A, Iida, S, and Sakakibara, Y. Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. *Nat. Commun.* 7, 13295 doi: 10.1038/ncomms13295, 2016.
(*These authors contributed equally to this work.)
2. Jayakumar, V, and Sakakibara, Y. Comprehensive evaluation of non-hybrid genome assembly tools for PacBio long-read sequence data. *Briefings in Bioinformatics*. bbx147, doi.org/10.1093/bib/bbx147, 2017.

Appendix B

Supplementary of chapter 4

Assembler versions

ABRuijn v1.0 (commit 4d3bd9dbefa128c7dc72417224cbc4bbd809a5ca)
Canu v1.4 (commit 06a3a714bb1befeb0682de8467c1b83b438f29ca)
FALCON v0.7 (commit 7a6ac0d8e8492c64733a997d72a9359e1275bb57)
HGAP v3 (SMRT analysis 2.3.0.5)
HINGE (commit 4d0c4809f01bcf391d026e7ad1754e0e7969aa2d)
MECAT v1.0 (commit 8675117d0647f31e6bc630662e9d97ceafd4b4a6)
Minimap (commit 1cd6ae3bc7c7a6f9e7c03c0b7a93a12647bba244)
Miniasm (commit 17d5bd12290e0e8a48a5df5afaeaef4d171aa133)
Racon (commit 0f6d4aa4787cb8278df689e9dc92ac799a839573)
PBcR (wgs v8.3rc2)
SMARTdenovo (commit 61cf13dcaed6bb561129b60eaa833fa9f976f9b1)
Wtdbg (commit 31550398a2859cffe60f603a452cda16fff60681)

Assembly

ABRuijn

ABRuijn was executed using a kmer size of 21 bp and a overlap length of 5000 bp with 2 rounds of polishing for all organisms. The coverage parameter was set to 140, 180, 45, and 50 for *E. coli*, *P. falciparum*, *C. elegans*, and *I. nil* respectively.

Canu

The grid options for memory and threads were modified to accept 10 Gb and 10 threads respectively. Default parameters were used along with -pacbio-raw option and respective genome sizes.

FALCON

The length cut-offs were chosen as 18,000 bp, 18,000 bp, 10,000 bp, and 6,000 bp for *E. coli*, *P. falciparum*, *C. elegans*, and *I. nil* respectively. The DBsplit options were given “a minimum read length of 500 bp, and a read block size of 50 Mb”, “a minimum read length of 500 bp, and a read block size of 50 Mb”, “a minimum read length of 500 bp, and a read block size of 400 Mb”, and “a minimum read length of 500 bp, and a read block size of 400 Mb”, for *E. coli*, *P. falciparum*, *C. elegans*, and *I. nil* respectively.

HGAP3

The following parameters for the modules were used: PreAssembler Filter module (minimum subread length=500 bp, minimum polymerase read length=500 bp); PreAssembler module (compute overlap length cutoff=true, number of seed read chunks=6, alignment candidates per chunk=10, total alignment candidates=24, minimum coverage for correction=6, blasr options=”noSplitSubreads, minimum subread length=500 bp, maximum score=1000, maximum LCP length=16”); AssembleUnitig module (default fragment minimum length=500 bp, coverage=30, overlap error rate=0.06, overlap minimum length=40 bp, mer size=14 bp). All the filtered sub reads were used as filtered long reads for the pre assembly process, excluding *I. nil*. For *I. nil*, the target chunks were increased to 10.

HINGE

HINGE was executed with fasta2DB, Dbsplit, HPC.daligner, Lamerge, and DASqv tools from tools associated with daligner. It was followed by filter, layout, clip, draft-path, draft, correct-head, consensus, and get_draft_path_norevcomp.py from the Hinge package.

MECAT

The programs MECAT2pw, and MECAT2cns were given default parameters with 16 threads and corrected sequences of 25X coverage were extracted. For MECAT2canu, “error rate=0.02 maximum memory=40 Gb maximum threads=16 use grid=0 -pacbio-corrected” options were given as input, along with respective genome sizes.

Miniasm and RACON

The initial minimap of raw reads were given the options such as minimizer window size =5 bp, minimum matching length =100 bp, and fraction of shared minimizers for merging two chains =0. The later steps including miniasm, minimap mapping for RACON, and the final RACON steps were given default options. RACON was executed twice for consensus generation.

PbcR

The following options were used for PBcR: “minimum read length =500, number of consensus partitions =200, overlap memory =32, overlap store memory =32000, overlap threads =8, mer overlapper threads =8, meryl threads =8, meryl memory =32000, fragment corrected concurrency =15, overlap concurrency =15, and consensus concurrency = 15”.

SMARTdenovo

Default settings were used for SMARTdenovo.

Wtdbg

Wtdbg was executed with the options such as kmer size of 21 bp, kmer subsampling fraction of 1.01, and also with homopolymer compression turned on. The minimum coverage of graph edges was set to 15, 10, 7, and 5 for *E. coli*, *P. falciparum*, *C. elegans*, and *I. nil* respectively. The accuracy obtained from the consensus procedure

recommended in the wtdbg github page, did not yield better accuracy and hence, the draft assembly was directly used for consensus polishing using quiver.

Canu, HINGE, SMARTdenovo, miniasm, and MECAT were given either default or recommended options from the developer's site. Some of the programs worked well with default parameters, whereas other programs required trial and errors to obtain better results. The jobs were executed on a node with a Intel Xeon E7-8870 processor (2.40 GHz) consisting of 160 cores and a memory of 2019.8 Gb under the operating system of RHEL v6.5. SGE was used for job management and the qacct command was used to access the maximum RSS and CPU time registered by the jobs.

Consensus polishing

After initial assembly, two rounds of quiver polishing was applied to all assemblies to improve the quality of the assembly and to reduce errors. Quiver from SMRT analysis 2.3.0.5 was executed with the following parameters: P_Filter module (minimum sub read length=500 bp, read score=0.60, minimum polymerase read length=500); P_Mapping module (maximum hits=10, maximum divergence=30%, minimum anchor size=12 bp, placeRepeatsRandomly=true, palign_options="random number generator initializing seed =1, minimum accuracy=0.80, and minimum read length=500 bp).

Evaluations

Quast v4.4-dev (commit 9c91befca0dc1b483550059f6541f68f0f63c5c8) was used to evaluate the contiguity and mismatch statistics of the assemblies. Nucmer from MUMmer v3.23 was executed for similarity search. Assemblytics was used to analyze indels and to create dot plots. Circlator v1.5.0 was used to resolve circularity with canu. CEGMA v2.5 and BUSCO v2.0.1 (commit 89aa1ab2527f03a87a214ca90a504ad236582a11) were used to assess completeness of core conserved genes. The 28 bp Terminal Inverted Repeats (TIRs) of the Tpn1 transposons were mapped using BLAST, which were later sorted by the contig

locations. If two nearby TIRs contained the same target site duplications (3–5 bp) and the total transposons length is less than 20 kb, they were nominated as Tpn1 transposons. Tandem Repeats Finder v4.07b was used to find telomeric repeats at the 10 kb ends of the contigs by assigning values 1, 1, 2, 80, 5, 200, and 2000 bp to match weight, mismatch weight, indel weight, match probability, indel probability, minimum score, and maximum period size respectively. BLAT v36 was used to align the ESTs, BAC, and BAC-end sequences.

RAD-seq analysis

The Illumina RAD-seq short reads from the parent samples and progeny samples were aligned against the assemblies using BWA v0.7.12. The reads which were not tagged as uniquely mapped, and those which did not have the requisite restriction enzyme cut site were filtered out. STACKS v1.37 was used to identify SNPs and the following two criteria were used to filter markers: (a) Each marker should be present in at least 80% of the samples, and (b) Each sample should have at least 80% of the markers. Also, 150 bp flanking regions from either side of each SNP location were extracted from the assembly and were aligned against each other using BLAST v2.2.29+ and regions with alignment lengths longer than 150 bp were filtered out. Onemap was used to create linkage maps with a logarithm of odds score of 30. Contigs whose markers were present in more than one linkage maps were considered as mis-assembled contigs.