Chromosome-By-Chromosome Assembly: A Scalable Method For *De Novo* Assembly

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Mohamed Awad

aus Kairo, Ägypten

Köln, 2022

Die vorliegende Arbeit wurde am Max-Planck-Institut für Pflanzenzüchtungsforschung in Köln in der Abteilung für Vergleichende Entwicklungsgenetik (Direktor Prof. Dr. Miltos Tsiantis) in der Arbeitsgruppe von Dr. Xiangchao Gan angefertigt.

MAX-PLANCK-INSTITUT
FÜR PFLANZENZÜCHTUNGSFORSCHUNG

Berichterstatter:	Prof. Dr. Miltos Tsiantis
	Prof. Dr. Achim Tresch
Prüfungsvorsitzender:	Prof. Dr. Andreas Beyer
Beisitzer:	Dr. Xiangchao Gan
Tag der mündlichen Prüfung:	02. März, 2022

Table of Contents

List of Figures:

List of Tables:

List of Abbreviations:

bp	basepairs
DBG	De Bruijn Graph
Indels	Insertion and Deletion
Kbp	Thousand-basepairs
LTR	Long Terminal Repeat
Mbp	Million-basepairs
NGS	Next Generation Sequencing
OLC	Overlap-Layout-Consensus
QV	Consensus quality score
SMRT	Single Molecule Real-Time
SNP	Single Nucleotide Polymorphism
SV	Structure Variant
TE	Transposon Element

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor **Dr. Xiangchao Gan** for giving me the opportunity to work on such an exciting project. His invaluable suggestions, impressive guidance, continuous support as well as our discussions motivated and helped me to finish my Ph.D. study and revolutionize my computational skills.

I also want to extend my special appreciation and acknowledgment to **Prof. Dr. Miltos Tsiantis** for supporting my Ph.D. project and his priceless comments and discussions, as well as his constant encouragement, inspiring me to overcome challenges hindering me throughout my Ph.D. study.

I would like to thank my TAC members, **Dr. Korbinian Schneeberger** and **Dr. Stefan Laurent**, for giving significant feedbacks and valuable advice whenever I need it.

I am so thankful to **Dr. Stephan Wagner** for his help during various administrative stages of my Ph.D. I also would like to thank all **my colleagues** in the comparative developmental genetics department (Tsiantis Dept.) for their support and encouragement.

Finally, I can't express my gratitude in words to **my parents and siblings** for their endless support and encouragement. Their trust in my abilities is my source of inspiration.

Abstract

Abstract

High-quality genome assembly has wide applications in genetics and medical studies. However, reconstructing complete genome sequences from sequencing data is a complex computational problem. Numerous tools have been developed to assemble short and long reads into longer representative sequences. However, the generated genome assemblies are often fragmented due to the repetitive nature and heterozygosity, even for studies using the most updated long-read technologies. Therefore, a computational framework which can lead to gap-free chromosome-scale assemblies is an insistent demand for modern biology studies.

In this dissertation, we introduced chromosome-by-chromosome assembly, a scalable computational framework for *de novo* genome assembly. We demonstrated its efficiency with the implementation of assembler GALA. GALA achieves chromosome-by-chromosome raw sequencing data separation through a multilayer graph algorithm which can effectively identify and resolve misassembles within preliminary assemblies, and subsequently cluster contigs from preliminary assemblies and raw reads into linkage groups. For complex genomes, extra information such as Hi-C, genetic maps and even motif analyses can be used to merge multiple linkage groups into bigger linkage groups, each representing a single chromosome. Assembly of each linkage group using existing assembly tools leads to gap-free complete genome assembly. Statistics based on the real data demonstrated that the strategy of chromosome-by-chromosome assembly can significantly simplify the complexity of assembly graph for most existing assembly tools, and achieve highly accurate gap-free chromosome-scale assembly.

Firstly, we tested GALA on heterogeneous third-generation sequencing datasets with different depths to demonstrate its advantage. Our method showed outstanding performance in low-depth circumstances over the current *de novo* assembly pipelines. In addition, GALA successfully produced T2T assembly for *C. elegans* and seven human chromosomes. Furthermore, GALA assembled complete gap-free chromosome-arm pseudomolecules for *A. thaliana* and four human chromosomes. Interestingly, our method overcomes the technology barriers, facilitating straightforward assembly of genomes with heterogeneous datasets and algorithms, generating high-quality *de novo* assemblies.

Secondly, we exploited GALA's ability to handle heterogeneous data to achieve the gap-free chromosome-scale assembly of *Cardamine hirsuta*, *C. oligosperma* and *C. resedifolia,* close relatives of the model plant *Arabidopsis thaliana*. Impressively, GALA

obtained a gap-free T2T *de novo* assembly of two *Cardamine hirsuta* strains, Azores and Oxford reference strain, and the *C. oligosperma* genome. GALA also successfully assembled five T2T *C. resedifolia* chromosomes and three chromosomes with a single centromeric gap. Additionally, we conducted a comparative genomic study between the assembled genomes to examine the collinearity and prominent structural variants among them.

Finally, we applied the strategy of chromosome-by-chromosome assembly to metagenome, a more challenging scenario where multiple haplotypes were sequenced at different depths and mixed together. We developed MRDA to facilitate metagenomic data separation for chromosome-by-chromosome *de novo* assembly. MRDA was implemented through a triple-layer graph, following a *reference-guided* data separation strategy to classify the preliminary contigs and impose the chromosome-by-chromosome assembly to achieve multiple-haplotype assembly of the circular microbial molecule. Our method achieved outstanding performance in terms of contiguity and the number of recovered circular chromosomes compared to the current *de novo* assembly pipelines.

Overall, we introduced a computational framework for chromosome-by-chromosome assembly. Based on this framework, we implemented two multilayer graph algorithms for gap-free chromosome-scale assembly of heterogeneous sequencing data. Our algorithms show very promising performances in the state-of-art *de novo* assembly.

Chapter1

Introduction

1. Introduction

1.1. DNA discovery and sequencing technologies

DNA (deoxyribonucleic acid) was discovered by Miescher in (1871), and the mystery on its structure was solved by Watson and Crick (1953). DNA is a double helix molecule comprised of a chain of four simpler biochemical components called nucleotides (A) adenine, (G) guanine, (T) thymine, and (C) cytosine (Watson and Crick 1953). As a carrier of heritability information and the first level of the central dogma of molecular biology, the determination of the nucleotide sequence order in the DNA molecule is a key goal of biological science. The sequencing process is the procedure of determining the sequence of nucleotides in a nucleic acid molecule.

1.1.1 First generation sequencing technologies

In 1975 Sanger sequencing technology was developed to be the first established experimental technique to determine the order of nucleotides in genome sequences based on chain termination during DNA synthesis (Sanger and Coulson 1975). Two years later, the sequencing of the ΦX174 bacteriophage genome heralded a new era of the first-generation sequencing technology (Sanger et al. 1977). In the same year, the Maxam–Gilbert method or DNA chemical sequencing method was established as a second DNA sequencing solution based on the nucleobase-specific partial chemical modification of DNA (Maxam and Gilbert 1977).

In 1987 Applied Biosystems, Inc. announced the first semi-automated sequencer based on the Sanger sequencing technique after alternating the radiolabelling by florescent labelling. The semi-automated sequencing technology reduced the labor and manual sources of error and caused a colossal breakthrough in DNA sequencing technology (Smith et al. 1986; Hood et al. 1987). Several genome sequencing projects were established depending on this technology, including the first sequenced bacterial genome, *Haemophilus influenzae* (Fleischmann et al. 1995), *Caenorhabditis elegans (Consortium 1998), Drosophila melanogaster* (Adams et al. 2000; Myers et al. 2000), *Arabidopsis thaliana* (Arabidopsis Genome 2000) and human (Lander et al. 2001; International Human Genome Sequencing 2004).

1.1.2 Next-generation sequencing technologies

Despite the high accuracy of Sanger sequencing, it is expensive, time-consuming and labor-intensive. Many Next Generation Sequencing (NGS) technologies emerged to overcome

Sanger sequencing limitations and provide high-throughput sequencing data. The Roche 454 sequencer, a pyrosequencing technology-based sequencer introduced in 2005, was the first commercial sequencer of this generation (Margulies et al. 2005). Solexa or Genome Analyzer (GA) sequencing emerged in 2007 based on a sequencing by synthesis approach (Bentley et al. 2008). Later, Illumina developed many sequencers based on the same approach, including the Hiseq series and Miseq. In late 2007 life technologies entered the NGS market with SOLiD sequencing platform based on sequencing by ligation technology (Shendure et al. 2005; Valouev et al. 2008). Then, life technologies implemented semiconductor sequencing technology on the Ion Torrent sequencer released in 2010 (Rothberg et al. 2011).

NGS established the era of fully automated sequencers that generate high-throughput data at lower cost and with an error rate comparable to the first-generation sequencing technologies (Rothberg et al. 2011). The affordability of NGS offers unprecedented opportunities in multiple aspects of daily activities and biological fields. These activities ranged from the rapid identification of pathogens and microorganisms in different tissue and environmental samples to disease treatments and diagnosis in personalized medicine (Heikamp and Pui 2018; Wilson et al. 2019). In addition, hundreds of live organism's genomes were sequenced and implemented in various omics studies (Hodzic et al. 2017).

1.1.3 Third-generation technologies

Although NGS provides an excellent solution for sequencing studies, the read length in almost all NGS platforms is shorter than Sanger sequencing. The third-generation sequencing technology revolution arose from the single-molecule sequencing approach (Eid et al. 2009; Derrington et al. 2010). In 2011 Pacific Bioscience announced The Pacbio RS sequencing platform based on Single Molecule Real-Time (SMRT) approach. Pacbio platform can provide reads with an average length of two Kbp and up to 23 Kbp (Rasko et al. 2011). In 2014 Oxford Nanopore released the portable USB Nanopore MinIon sequencer based on Nanopore sequencing methodology (Mikheyev and Tin 2014). Theoretically, the Nanopore sequencer does not have a limitation for read length, but practically, the read length can reach 200 Kbp with N50 around 30 Kbp (Michael et al. 2018). Recently, several reports stated ~ 1 Mbp reads in Nanopore datasets (Jain et al. 2018; Rao et al. 2021).

The MinIon USB sequencer allows direct sample sequencing on the field and facilitates genome studies significantly. Unfortunately, the error rate of this long-read technology is very high compared to Sanger sequencing and NGS. The error rate is around 15 % in Pacbio

platforms and reaches 35 % in Nanopore sequencers (Ferrarini et al. 2013; Laver et al. 2015). This high error rate encourages researchers to develop hybrid error-correction algorithms using NGS data-based and self-correction algorithms (Das et al. 2019; Morisse et al. 2021). Recently, Pacbio announced the Sequel II system to provide high-quality long reads generated by calling consensus from subreads produced by several enzymes pass around a circularized template. These reads are called high fidelity (Hifi) reads with an average length of 10-25 Kbp and an error rate comparable to NGS (Hon et al. 2020).

1.2. Genome assembly general workflow

Sequencing technologies witnessed a massive development over the past decades. Unfortunately, none of these sequencing technologies can decode the whole DNA molecule as a single fragment with a straightforward sequence except for very small molecules, e.g., plasmids and virus genomes. Instead, the sequencing platforms generate an incredible amount of data as short reads/fragments ranging in length between 70 bp in NGS and 200 Kbp in Nanopore. Therefore, sophisticated computational algorithms are required to handle and process sequencing data comprehensively. **Genome assembly** is a computational biology approach that aims to reconstruct the closest representation of the actual genome from the fragmented reads. The genome assembly process includes two main stages, assembly and scaffolding (**Fig. 1.1**). The two stages can be implemented in a *de novo* manner or by aligning the reads to an existing reference genome in *the reference-guided* assembly approach.

1.2.1 *De novo* assembly algorithms

Genome *de novo* assembly is a crucial computational biology problem that aims to assemble the DNA sequence of a target genome by leveraging the overlaps between the reads generated from the sequencing technologies (Shafin et al. 2020). This process builds up a set of more extended contiguous fragments of the target genomes called contigs. For the first-generation sequencing data, simple greedy assembly algorithms were developed using a repeated operation of merging reads with the best overlap (Pop 2009). The high throughput technologies lead to a magnificent advance in genome assembly algorithms. Various assemblers were developed using graph algorithms and can be categorized in general into two

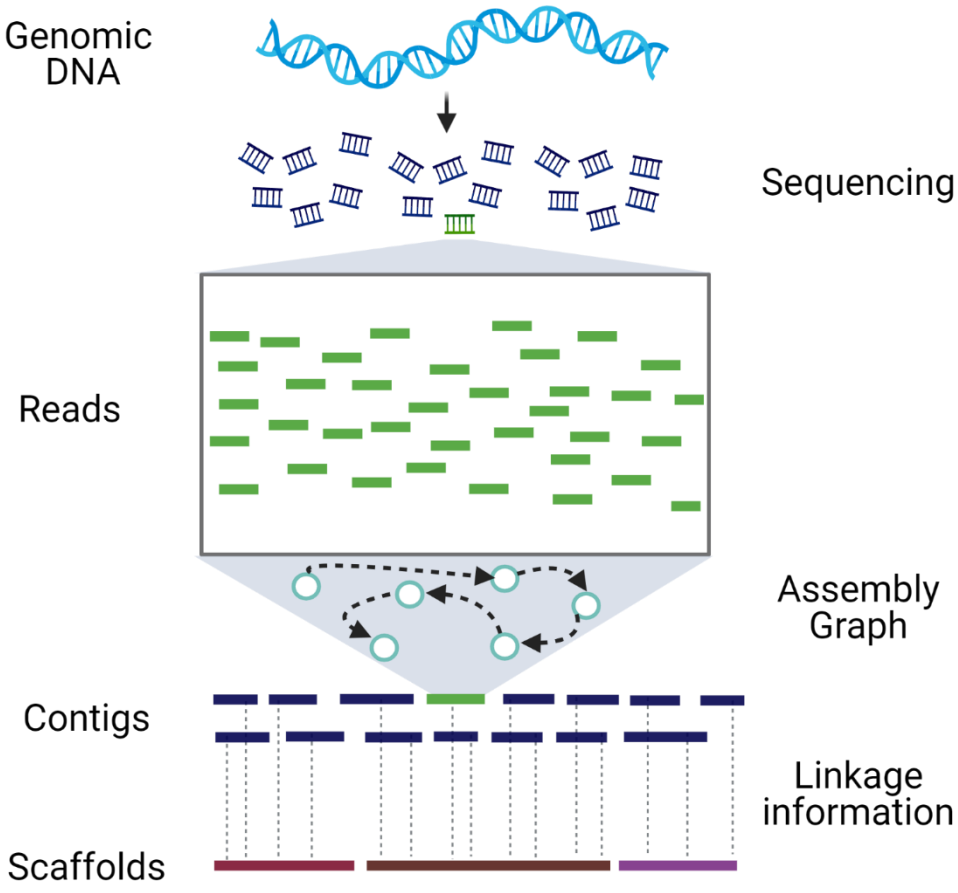


Figure 1.1: Genome assembly pipeline

categories by the used graph algorithm: Overlap-Layout-Consensus (OLC) and de Bruijn graph (DBG) (Li et al. 2012). Though there were few exceptions such as SSAKE (Warren et al. 2007) and JR-Assembler (Chu et al. 2013), which are based on the greedy algorithm.

1.2.1.1 Overlap-Layout-Consensus (OLC) algorithm

The OLC algorithm has three phases: in the first phase, it delivers a pairwise alignment between all objects in a set of reads (R). Next, all the reads inside (R) are encoded as nodes/vertices in an overlap graph, and the alignment information creates weighted directed edges representing the overlaps between the graph vertices. In this phase, the graph is simplified by resolving the branched nodes and transitively-infeasible edges to construct a contigs layout from explicit and continuous paths. Finally, in the third phase, the algorithm carries out a multiple sequence alignment to polish the sequence of each assembled contig (Li et al. 2012) (**Fig. 1.2**). Several NGS assembly tools were developed based on the OLC graph, e.g., Celera (Myers et al. 2000), Arachne (Batzoglou et al. 2002), Phrap (de la Bastide and McCombie 2007) and Newbler (Margulies et al. 2005). Furthermore, many third-generation sequencing

assemblers are also based on OLC graph, e.g., Canu (Koren et al. 2017), Mecat (Xiao et al. 2017), Miniasm (Li 2016), and Falcon (Chin et al. 2016).

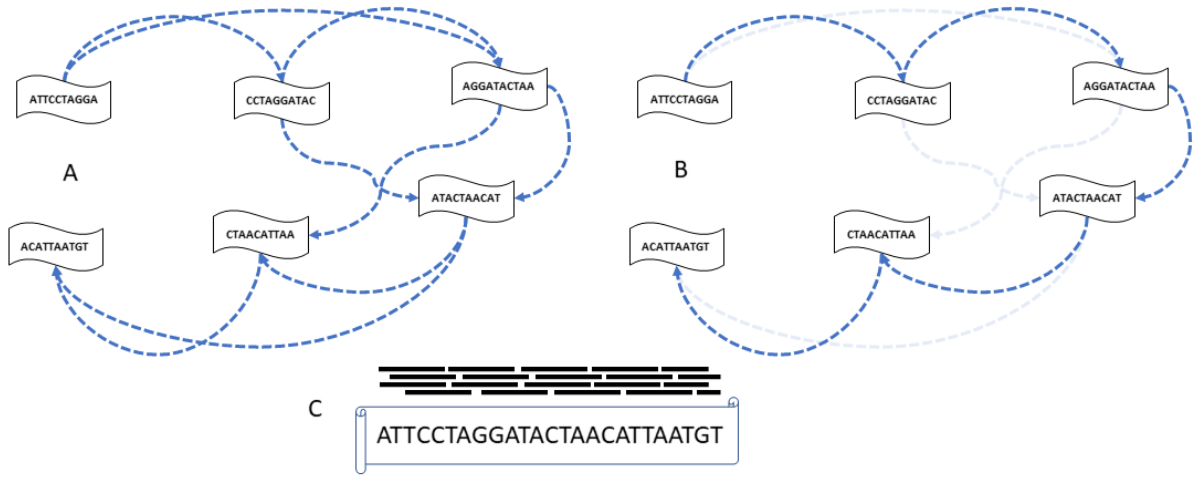


Figure 1.2: Overlap-Layout-Consensus (OLC) graph. A) overlap graph construct phase: assuming a dataset of six reads representing the graph's nodes and the weighted edges representing the overlaps between nodes. B) Layout phase: constructing the contig layout after removing the transitive edges. C) Consensus phase: carrying out the contig consensus sequence derived from multiple sequence alignment.

1.2.1.2 De Bruijn Graph (DBG)

Despite the advantage of utilizing the overlapping information from the entire read, OLC consums massive computational and storage resources, especially with large and complex genomes. Therefore, the de Bruijn graph (DBG) was introduced to reduce the computational cost and significantly enhance assembly outcomes. The DBG approach starts with the fragmentation process for each read in a set of reads (R) into substrings of a defined length *k* called *k-mers.* Next, it creates a directed graph of a set of unique *k*-1 *-mers* encoded as nodes and graph's edges encoded by *k-mers* with identical (k-1) suffix-prefix overlap. Finally, it identifies the assembled contigs by resolving the unambiguous graph paths (Li et al. 2012) (Fig. 1.3).

The de Bruijn graph fits NGS data well, leading to the emergence of numerous assemblers, e.g., Velvet (Zerbino and Birney 2008), SOAPdenovo2 (Luo et al. 2012), ABySS (Simpson et al. 2009), and AllPath-LG (Gnerre et al. 2011). Unfortunately, DBG is very sensitive to the error rate. Thus, the high error rate of third-generation sequencing hinders the application of DBG for long reads. The ABruijn assembler was the first third-generation sequencing genome assembler developed to assess the DBG influences on long reads assembly.

It was successfully applied to assemble small genomes, e.g., *E. coli and C. elegans* (Lin et al. 2016). Later, assemblers based on modified versions of DBG, e.g., Flye (repeated graph) (Kolmogorov et al. 2019) and Wtdbg2 (fuzzy Bruijn graph) (Ruan and Li 2020), were released and successfully applied to assemble complex genomes.

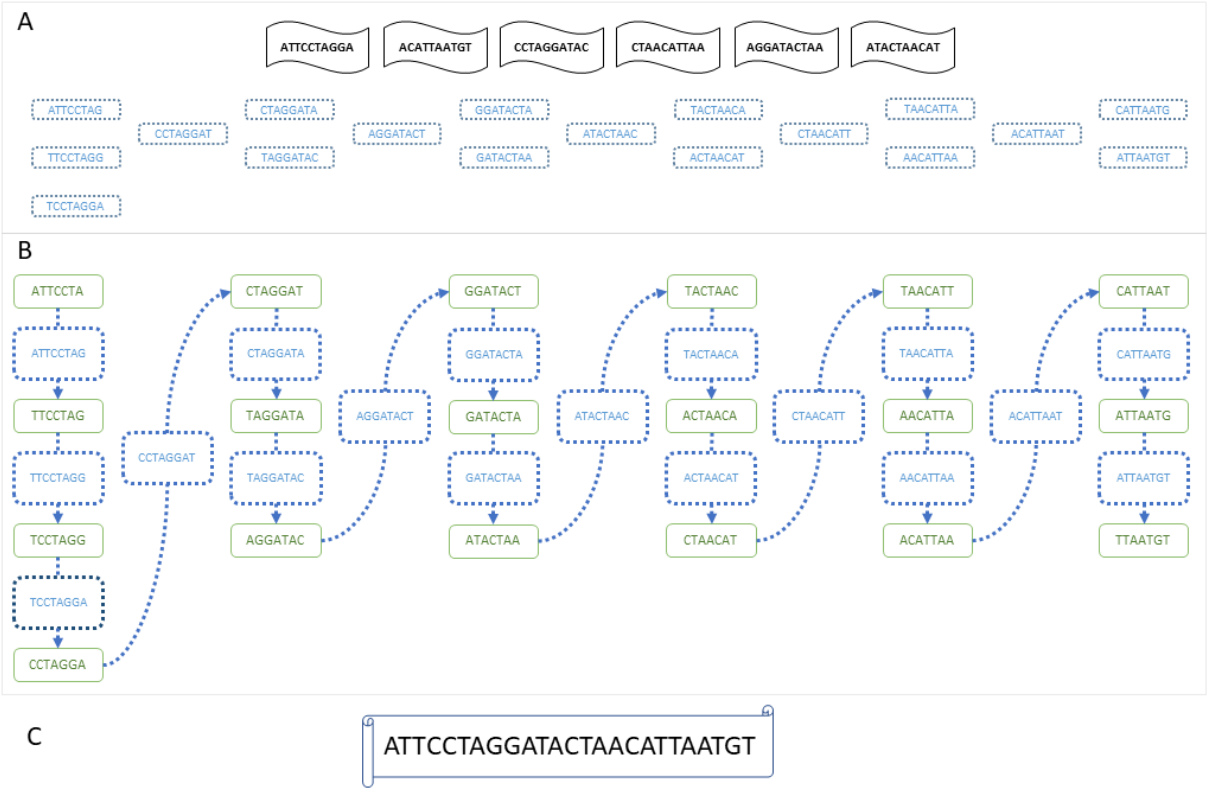


Figure 1.3: De Bruijn Graph (DBG). A) assuming a dataset of six reads partitioned into *k-mers* of length *k*=8. B) graph construction from a set of unique 7-*mers* as nodes (green rectangles) and suffix-prefix overlapped *k-mers* as edges (blue rectangles). C) the assembled contig.

1.2.2 Scaffolding approaches

The *de novo* assembly in the first stage generates a messy representation of the target genome as a large number of contigs. Unfortunately, many comparative genomics studies and population genetics investigations need highly continuous assemblies or chromosome-level assemblies. Therefore, various techniques and methods were developed to address the scaffolding stage (Rice and Green 2019). **Genome scaffolding** is a computational biology procedure that employs linkage information to infer the contig's order and orientation in the target genome. Handling the synteny information from a reference genome or a close relative species is a common scaffolding approach used to develop various scaffolding tools, e.g., Ragoo (Alonge et al. 2019), MeDuSa (Bosi et al. 2015) and CSAR (Chen et al. 2018). However, the structure variants and chromosomal rearrangements between the target genome and the

reference genome may lead to scaffolding errors (Kolmogorov et al. 2014; Bosi et al. 2015; Ghurye and Pop 2019). Consequently, *de novo* scaffolding approaches are preferable to achieve chromosome-level pseudomolecules. According to the source of linkage information, the *de novo* scaffolding strategies can be classified into mapping-based and sequencing-based strategies.

The mapping-based scaffolding approaches employ the linkage information from genomic landmarks detected by mapping techniques. The mapping markers can be identified using recombination information (Genetic Linkage map) (Nossa et al. 2014), chromosomes radiation breaks (Radiation Hybrid map) (Raudsepp et al. 2008), fluorescently labeled probes (Fluorescence In Situ Hybridization map - FISH) (Raudsepp et al. 2008) and the most recent technique, optical mapping. The optical map generates a fluorescently labeled order restriction map for the target genome. In 2014 BioNano Genomics announced the Irys system as a long-range optical mapping platform. In addition, Bionano Genomics developed IrysView to carry out scaffolded molecules by aligning the assembled contigs to the assembled optical maps (Xiao et al. 2015). Optical maps are used to achieve a chromosome-level assembly of many genomes, including *Sorghum bicolor* (Deschamps et al. 2018), *Arabis alpina* (Jiao et al. 2017) and the human genome (Seo et al. 2016).

The affordability of sequencing technology makes sequencing-based scaffolding ubiquitous in genome assembly projects. This scaffolding approach exploits the linkage information from both ends of a small genomic fragment (Paired-end) or long genomic fragment inserted in a Bacterial Artificial Chromosome BAC (mate-end) (Adams et al. 2000; International Human Genome Sequencing 2004). Moreover, various tools emerged to use low coverage long reads to scaffold the short reads assemblies, e.g., LRscaf (Qin et al. 2019), SLR (Luo et al. 2019) and LINKS (Warren et al. 2015). The high-throughput chromosome conformation capture (Hi-C) sequencing turned up in 2009 to solve chromosomes three-dimensional 3D architecture (Lieberman-Aiden et al. 2009). Currently, Hi-C is one of the most common methods used for genome scaffolding. Numerous tools were developed to utilize Hi-C in the genome scaffolding, e.g., LACHESIS (Burton et al. 2013), Juicer (Durand et al. 2016), and SALSA2 (Ghurye et al. 2019). Additionally, many genomes achieved chromosome-level assembly through Hi-C scaffolding, e.g., *Cerasus humilis* (Wang et al. 2020b), *Prunus avium* (Wang et al. 2020a), *Miscanthus lutarioriparius* (Miao et al. 2021), *Solanum melongena* (Wei et al. 2020), goose (Li et al. 2020), and human (Garg et al. 2021). 10XGenomics introduced the

linked-reads sequencing technology as a potential source of linkage information, but in 2020 the company abandoned the genomic sequencing services.

1.3. Assembly assessment

A lot of genome assembly tools have been developed over the last decades. Unfortunately, assembly software always generates certain types of errors during the process and has its advantages and disadvantages. Consequently, all genome assembly studies need to perform a comprehensive assessment to evaluate the accuracy of the assembly outcomes (Salzberg et al. 2012; Thrash et al. 2020). In the presence of a high-quality reference genome, the validation of new assemblies is often performed by comparing them to the reference genome. The quast tool was developed for this purpose (Gurevich et al. 2013). Thus, while many projects were established to assemble genomes of new species, it is become very important to conclude a reference-free validation method. The current quality assessment scheme evaluates three vital accuracy dimensions; contiguity, completeness and correctness (Thrash et al. 2020).

First, the contiguity evaluation measures the expected and observed number and length of contigs/scaffolds. The assembly will achieve a high contiguity score if the number and size of contigs in the draft assembly are closer to the expected number of chromosomes in the target genome (Thrash et al. 2020). For example, the human genome assembly achieves the perfect contiguity with 23 contigs/scaffolds from a female sample; each represents a complete chromosome. Unfortunately, the majority of assembly projects do not achieve this contiguity. Hence, various matrices are used to estimate the assembly contiguity, including the number of assembled contigs/scaffolds, the longest and shortest contig (Jayakumar and Sakakibara 2019). The N50 is a vital metric in contiguity assessment, representing the length of the contig that comprises 50% of the assembly size when summing it to the contigs with greater length. The L50 is another contiguity metric, representing the smallest number of contigs that comprises 50% of the assembly size (the descending order of the N50 contig). Generally, in the case of a target genome with the same or close chromosomes length, perfect assembly N50 and L50 will be the length and order of the median chromosome.

Second, the completeness matrices evaluate the existing and missing genomic architecture in the assembled draft (Thrash et al. 2020). Reasonably, comparing the expected and the observed genome size is an initial assessment. Regularly, the expected genome size could be estimated from flow cytometry or *K-mer* analyses (Pflug et al. 2020). However, there

are two advanced approaches to estimate assembly completeness; gene content evaluation and *K-mer* content calculation. Benchmark Universal Single Copy Ortholog (BUSCO) is one of two tools developed to determine the number of assembled ortholog genes in the draft assembly (Seppey et al. 2019); Core Eukaryotic Gene Mapping Approach (GEGMA) is the second tool (Parra et al. 2007). The *K-mer* content analysis was recently introduced as a reference-free method for completeness validation (Mapleson et al. 2017). KAT and Merqury were developed to compare the number of unique *K-mers* in a short-read dataset and an assembled genome (Rhie et al. 2020).

Finally, the correctness evaluation measures the accuracy of each nucleotide in the assembly. Indeed, it is very challenging to assess the assembly correctness in the absence of a golden standard reference genome. Nevertheless, two methods are proposed to cover this dimension, including coding sequence frameshift errors analysis (Rhie et al. 2021) and the ratio of variants between the draft assembly versus BAC library or short-read dataset (Vollger et al. 2020). Also, evaluating the accuracy of orientation and location of the contigs in the final assembly is another scope of correctness assessment (Thrash et al. 2020).

1.4. Challenges of Chromosome-scale assembly

Sequencing errors are the first factor that undermines the genome assembly even in the NGS data. Theoretically, for error-free reads from a simple genome, the assembly graph generates perfect assembly for all chromosomes (Ekim et al. 2021). The sequencing errors cause false nodes and edges in the assembly graph magnify the graph complexity and impede the contiguity (Pevzner et al. 2001). In addition, the increasing level of genome complexity hampers chromosome-level assembly. The long repeats, tandem repeats, transposons, duplications, heterozygosity and polyploidy, are vital elements that significantly influence graph complexity and assembly contiguity (Pevzner et al. 2001). Furthermore, producing linkage information for scaffolding is time cost and labor-consuming, thwarting the chromosome-level assembly (Alonge et al. 2019).

In addition to the contiguity disruption and gapped assembly, the factors mentioned above motivate two other forms of assembly errors. First, misassembled contigs are chimeric pseudomolecules constructed from reads sequenced from non-adjacent genomic regions. This error usually originated from duplicated, similar or homologous repetitive sequences in distant regions (Phillippy et al. 2008). The chimeric contigs undermine the scaffolding stage and give incorrect information that negatively affects several biological studies, e.g., Pan genomes and

population genetics. The second form of assembly errors emerges in repetitive regions by merging reads derived from distinct repeat copies into fewer copies (Collapsing) or disjoin reads originated from the exact repeat to several copies (Expanding). The collapsed and expanded contigs adversely influence the downstream analysis, e.g., gene prediction and gene ontology (Phillippy et al. 2008).

1.5. Aims and contributions of the thesis

The main goal of this thesis is to develop a framework that enables chromosome-by-chromosome assembly and is able to handle a combination of heterogeneous information from long-read technologies and other sources of scaffolding information to obtain high-quality gap-free chromosome-scale assemblies. We implemented an algorithm, GALA, using this framework and demonstrated its advantages. We benchmarked our algorithm using different types of data from different organisms.

In Chapter 2, we describe a computational framework to exploit the chromosome-by-chromosome assembly. We demonstrate GALA, a gap-free chromosome-scale assembly algorithm. GALA uses different preliminary assemblers to build a multilayer graph to detect and correct chimeric contigs accurately. Then GALA clusters the error-free contigs into multiple linkage groups, each representing a single chromosome/scaffold. The experimental results showed that GALA successfully overcomes barriers between sequencing technologies. GALA implementation achieves gap-free chromosome-scale assembly of *C.elegans*, seven chromosomes of the human genome and gap-free chromosome-arm-scale assembly of *A. thaliana* genome.

In Chapter 3, we take advantage of GALA, incorporating heterogeneous data to assemble the second draft of the *Cardamine hirsuta* genome and its two relative species, C. *oligosperma* and *C. resedifolia.* We successfully assembled the genome of *C. hirsuta* reference strain (Oxford), closing thousands of gaps in the published draft and resolving inter-chromosomal discordances. At the same time, we assembled *C. resedifolia* and C. *oligosperma* genomes to gap-free chromosome-scale assembly with only three gaps in the centromeric regions of the *C. resedifolia* genome. Finally, we conducted a comparative study to demonstrate the karyotype differences between the assembled genomes.

In Chapter 4, we describe MRDA, a metagenome *reference-guided* data separation and *de novo* assembly module. We applied the chromosome-by-chromosome assembly concept to

assemble circular and complete bacterial genomes from long-read metagenome datasets. First, MRDA builds a triple-layer graph to bin the contigs into taxa-specific linkage groups using a *reference-guided* approach. Then it follows a chromosome-by-chromosome *de novo* assembly approach to generate circular bacterial genomes. Eventually, MRDA showed better results over the standard *de novo* assembly tools.

References

Adams MD Celniker SE Holt RA Evans CA Gocayne JD Amanatides PG Scherer SE Li PW Hoskins RA Galle RF et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**: 2185-2195.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363-376.

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* **20**: 224.

Arabidopsis Genome I. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796-815.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**: 663-675 e619.

Awad M, Gan X. 2020. GALA: gap-free chromosome-scale assembly with long reads. *bioRxiv*.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.

Barchi L, Rabanus-Wallace MT, Prohens J, Toppino L, Padmarasu S, Portis E, Rotino GL, Stein N, Lanteri S, Giuliano G. 2021. Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. *Plant J* **107**: 579-596.

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

Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, Genete M, Berrabah W, Chevre AM, Delourme R et al. 2018. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants* **4**: 879-887.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* **53**: 474-485.

Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvornicic M, Soldo JP, Koh JY, Tong C et al. 2019. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol* **37**: 937-944.

Bickhart DM, Liu GE. 2014. The challenges and importance of structural variation detection in livestock. *Front Genet* **5**: 37.

Bishara A, Moss EL, Kolmogorov M, Parada AE, Weng Z, Sidow A, Dekas AE, Batzoglou S, Bhatt AS. 2018. High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat Biotechnol* doi:10.1038/nbt.4266.

Boock JT, Freedman AJE, Tompsett GA, Muse SK, Allen AJ, Jackson LA, Castro-Dominguez B, Timko MT, Prather KLJ, Thompson JR. 2019. Engineered microbial biofuel production and recovery under supercritical carbon dioxide. *Nat Commun* **10**: 587.

Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lio P, Crescenzi P, Fani R, Fondi M. 2015. MeDuSa: a multi-draft based scaffolder. *Bioinformatics* **31**: 2443-2451.

BSong B SQ, Wang H, Pei H, Gan X and Wang F. 2019. Complement Genome Annotation Lift Over Using a Weighted Sequence Alignment Strategy. *Front Genet* **10**.

Buisine N, Quesneville H, Colot V. 2008. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics* **91**: 467-475.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**: 1119-1125.

Bzikadze AV, Pevzner PA. 2020. Automated assembly of centromeres from ultra-long error-prone reads. *Nat Biotechnol* **38**: 1309-1316.

Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**: e4958.

Cao MD, Nguyen SH, Ganesamoorthy D, Elliott AG, Cooper MA, Coin LJ. 2017. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat Commun* **8**: 14515.

Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, Behrenfeld MJ, Boetius A, Boyd PW, Classen AT et al. 2019. Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* **17**: 569-586.

Cepeda V, Liu B, Almeida M, Hill CM, Koren S, Treangen TJ, Pop M. 2017. MetaCompass: Reference-guided Assembly of Metagenomes. *bioRxiv*.

Chang C, Bowman JL, Meyerowitz EM. 2016. Field Guide to Plant Model Systems. *Cell* **167**: 325-339.

Chen KT, Liu CL, Huang SH, Shen HT, Shieh YK, Chiu HT, Lu CL. 2018. CSAR: a contig scaffolding tool using algebraic rearrangements. *Bioinformatics* **34**: 109-111.

Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF. 2020. Accurate and complete genomes from metagenomes. *Genome Res* **30**: 315-333.

Chen X, Tompa M. 2010. Comparative assessment of methods for aligning multiple genome sequences. *Nat Biotechnol* **28**: 567-572.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**: 170-175.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563-569.

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**: 1050-1054.

Chu TC, Lu CH, Liu T, Lee GC, Li WH, Shih AC. 2013. Assembler for de novo assembly of large genomes. *Proc Natl Acad Sci U S A* **110**: E3417-3424.

Consortium CeS. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**: 2012-2018.

Das AK, Goswami S, Lee K, Park SJ. 2019. A hybrid and scalable error correction algorithm for indel and substitution errors of long reads. *BMC Genomics* **20**: 948.

de la Bastide M, McCombie WR. 2007. Assembling genomic DNA sequences with PHRAP. *Curr Protoc Bioinformatics* **Chapter 11**: Unit11 14.

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol* **10**: e1003998.

Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, Niederweis M, Gundlach JH. 2010. Nanopore DNA sequencing with MspA. *Proc Natl Acad Sci U S A* **107**: 16060-16065.

Desai A, Marwah VS, Yadav A, Jha V, Dhaygude K, Bangar U, Kulkarni V, Jere A. 2013. Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PLoS One* **8**: e60204.

Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H. 2018. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat Commun* **9**: 4844.

Dolezel J, Vrana J, Safar J, Bartos J, Kubalakova M, Simkova H. 2012. Chromosomes in the flow to simplify genome analysis. *Funct Integr Genomics* **12**: 397-416.

Domanska D, Kanduri C, Simovski B, Sandve GK. 2018. Mind the gaps: overlooking inaccessible regions confounds statistical testing in genome analysis. *BMC Bioinformatics* **19**: 481.

Driscoll CB, Otten TG, Brown NM, Dreher TW. 2017. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Stand Genomic Sci* **12**: 9.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**: 95-98.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133-138.

Ekblom R, Wolf JB. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl* **7**: 1026-1042.

Ekim B, Berger B, Chikhi R. 2021. Minimizer-space de Bruijn graphs. doi:10.1101/2021.06.09.447586 %J bioRxiv: 2021.2006.2009.447586.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**: 18.

Ellison CE, Cao W. 2020. Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of Drosophila melanogaster. *Nucleic Acids Res* **48**: 290-303.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**: e47768.

Ferrarini M, Moretto M, Ward JA, Surbanovski N, Stevanovic V, Giongo L, Viola R, Cavalieri D, Velasco R, Cestaro A et al. 2013. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics* **14**: 670.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496-512.

Gan X, Hay A, Kwantes M, Haberer G, Hallab A, Ioio RD, Hofhuis H, Pieper B, Cartolano M, Neumann U et al. 2016. The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. *Nat Plants* **2**: 16167.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT et al. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**: 419-423.

Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J et al. 2021. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol* **39**: 309-312.

Ghurye J, Pop M. 2019. Modern technologies and algorithms for scaffolding assembled genomes. *PLoS Comput Biol* **15**: e1006994.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol* **15**: e1007273.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**: 1513-1518.

Goel M, Sun H, Jiao WB, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**: 277.

Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Korasidis N, Gavryushkin A, Carlson JM, Beerenwinkel N, Ludington WB. 2018. Microbiome interactions shape host fitness. *Proc Natl Acad Sci U S A* **115**: E11951-E11960.

Gupta S, Mortensen MS, Schjorring S, Trivedi U, Vestergaard G, Stokholm J, Bisgaard H, Krogfelt KA, Sorensen SJ. 2019. Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. *Commun Biol* **2**: 291.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072-1075.

Guyomar C, Delage W, Legeai F, Mougel C, Simon J-C, Lemaitre C. 2018. Reference guided genome assembly in metagenomic samples %+ Institut de Génétique, Environnement et Protection des Plantes (IGEPP) %+ Scalable, Optimized and Parallel Algorithms for Genomics (GenScale). In *RECOMB 2018 - 22nd International Conference on Research in Computational Molecular Biology*, pp. 1 %8 2018-2004-2021, Paris, France.

Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK, Jr., Maiti R, Chan AP, Yu C, Farzad M, Wu D et al. 2005. Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol* **3**: 7.

Hay AS, Pieper B, Cooke E, Mandakova T, Cartolano M, Tattersall AD, Ioio RD, McGowan SJ, Barkoulas M, Galinha C et al. 2014. Cardamine hirsuta: a versatile genetic system for comparative studies. *Plant J* **78**: 1-15.

Heikamp EB, Pui CH. 2018. Next-Generation Evaluation and Treatment of Pediatric Acute Lymphoblastic Leukemia. *J Pediatr* **203**: 14-24 e12.

Hodzic J, Gurbeta L, Omanovic-Miklicanin E, Badnjevic A. 2017. Overview of Next-generation Sequencing Platforms Used in Published Draft Plant Genomes in Light of Genotypization of Immortelle Plant (Helichrysium Arenarium). *Med Arch* **71**: 288-292.

Holusova K, Vrana J, Safar J, Simkova H, Balcarkova B, Frenkel Z, Darrier B, Paux E, Cattonaro F, Berges H et al. 2017. Physical Map of the Short Arm of Bread Wheat Chromosome 3D. *Plant Genome* **10**.

Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC et al. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* **7**: 399.

Hood LE, Hunkapiller MW, Smith LM. 1987. Automated DNA sequencing and analysis of the human genome. *Genomics* **1**: 201-212.

Hou S, Wolinska KW, Hacquard S. 2021. Microbiota-root-shoot-environment axis and stress tolerance in plants. *Curr Opin Plant Biol* **62**: 102028.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* **43**: 476-481.

Huang C, Ying H, Yang X, Gao Y, Li T, Wu B, Ren M, Zhang Z, Ding J, Gao J et al. 2021. The Cardamine enshiensis genome reveals whole genome duplication and insight into selenium hyperaccumulation and tolerance. *Cell Discov* **7**: 62.

Human Microbiome Project C. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207-214.

International Human Genome Sequencing C. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931-945.

International Wheat Genome Sequencing C. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* **345**: 1251788.

Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J, Jones SJM et al. 2018. Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* **19**: 393.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338-345.

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A et al. 2020. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**: 284-289.

Jayakumar V, Sakakibara Y. 2019. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief Bioinform* **20**: 866-876.

Jiao WB, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, Willing EM, Piednoel M, Woetzel S, Madrid-Herrero E 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 Res* **27**: 778-786.

Jiao WB, Schneeberger K. 2020. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun* **11**: 989.

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL et al. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**: 1103-1110.

Kolmogorov M, Raney B, Paten B, Pham S. 2014. Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* **30**: i302-309.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**: 540-546.

Koornneef M, Meinke D. 2010. The development of Arabidopsis as a model plant. *Plant J* **61**: 909-921.

Koren S, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2019. Reply to 'Errors in long-read assemblies can critically affect protein prediction'. *Nat Biotechnol* **37**: 127-128.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-736.

Kuderna LFK, Solis-Moruno M, Batlle-Maso L, Julia E, Lizano E, Anglada R, Ramirez E, Bote A, Tormo M, Marques-Bonet T et al. 2019. Flow Sorting Enrichment and Nanopore Sequencing of Chromosome 1 From a Chinese Individual. *Front Genet* **10**: 1315.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* **40**: D1202-1210.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, Tan Y, Li X, Lai Q, Han L et al. 2020. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *Gigascience* **9**.

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R et al. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814-821.

Lapidus AL, Korobeynikov AI. 2021. Metagenomic Data Assembly - The Way of Decoding Unknown Microorganisms. *Front Microbiol* **12**: 613791.

Latorre-Perez A, Villalba-Bermell P, Pascual J, Vilanova C. 2020. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. *Sci Rep* **10**: 13588.

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* **3**: 1-8.

Li C, Lin F, An D, Wang W, Huang R. 2017. Genome Sequencing and Assembly by Long Reads in Plants. *Genes (Basel)* **9**.

Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, Yamashita H, Lam TW. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**: 3-11.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987-2993.

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103-2110.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li Y, Gao G, Lin Y, Hu S, Luo Y, Wang G, Jin L, Wang Q, Wang J, Tang Q et al. 2020. Pacific Biosciences assembly with Hi-C mapping generates an improved, chromosome-level goose genome. *Gigascience* **9**.

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B et al. 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics* **11**: 25-37.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289-293.

Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. 2016. Assembly of long error-prone reads using de Bruijn graphs. *Proc Natl Acad Sci U S A* **113**: E8396-E8405.

Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, Mueller A, Schaberle TF, Hughes DE, Epstein S et al. 2015. A new antibiotic kills pathogens without detectable resistance. *Nature* **517**: 455-459.

Lischer HEL, Shimizu KK. 2017. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics* **18**: 474.

Luo J, Lyu M, Chen R, Zhang X, Luo H, Yan C. 2019. SLR: a scaffolding algorithm based on long reads and contig classification. *BMC Bioinformatics* **20**: 539.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**: 18.

Ma Z, Zhang Y, Wu L, Zhang G, Sun Z, Li Z, Jiang Y, Ke H, Chen B, Liu Z et al. 2021. High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. *Nat Genet* doi:10.1038/s41588-021-00910-2.

Mantere T, Kersten S, Hoischen A. 2019. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet* **10**: 426.

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**: 574-576.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**: 560-564.

McKim SM, Routier-Kierzkowska AL, Monniaux M, Kierzkowski D, Pieper B, Smith RS, Tsiantis M, Hay A. 2017. Seasonal Regulation of Petal Number. *Plant Physiol* **175**: 886-903.

Mende DR, Letunic I, Maistrenko OM, Schmidt TSB, Milanese A, Paoli L, Hernandez-Plaza A, Orakov AN, Forslund SK, Sunagawa S et al. 2020. proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res* **48**: D621-D625.

Miao J, Feng Q, Li Y, Zhao Q, Zhou C, Lu H, Fan D, Yan J, Lu Y, Tian Q et al. 2021. Chromosome-scale assembly and analysis of biomass crop Miscanthus lutarioriparius genome. *Nat Commun* **12**: 2458.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat Commun* **9**: 541.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA et al. 2019. Telomere-to-telomere assembly of a complete human X chromosome. doi:10.1101/735928 %J bioRxiv: 735928.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* doi:10.1038/s41586-020-2547-7.

Mikheyev AS, Tin MM. 2014. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* **14**: 1097-1102.

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

Molina-Mora JA, Campos-Sanchez R, Rodriguez C, Shi L, Garcia F. 2020. High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. *Sci Rep* **10**: 1392.

Morisse P, Marchet C, Limasset A, Lecroq T, Lefebvre A. 2021. Scalable long read self-correction and assembly polishing with multiple sequence alignment. *Sci Rep* **11**: 761.

Moss EL, Maghini DG, Bhatt AS. 2020. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* **38**: 701-707.

Muggli MD, Puglisi SJ, Ronen R, Boucher C. 2015. Misassembly detection using paired-end sequence reads and optical mapping data. *Bioinformatics* **31**: i80-88.

Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, Roest Crollius H, Salse J. 2015. Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol* **16**: 262.

Murigneux V, Rai SK, Furtado A, Bruxner TJC, Tian W, Harliwong I, Wei H, Yang B, Ye Q, Anderson E et al. 2020. Comparison of long-read methods for sequencing and assembly of a plant genome. *Gigascience* **9**.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA et al. 2000. A whole-genome assembly of Drosophila. *Science* **287**: 2196-2204.

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Lambing C, Kuo P, Yelina N, Hartwick N, Colt K et al. 2021. The genetic and epigenetic landscape of the Arabidopsis centromeres. *bioRxiv*.

Nakamura K, Iizuka R, Nishi S, Yoshida T, Hatada Y, Takaki Y, Iguchi A, Yoon DH, Sekiguchi T, Shoji S et al. 2016. Culture-independent method for identification of microbial enzyme-encoding genes by activity-based single-cell sequencing using a water-in-oil microdroplet platform. *Sci Rep* **6**: 22259.

Nissen JN, Johansen J, Allesoe RL, Sonderby CK, Armenteros JJA, Gronbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O et al. 2021. Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* **39**: 555-560.

Nossa CW, Havlak P, Yue JX, Lv J, Vincent KY, Brockmann HJ, Putnam NH. 2014. Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. *Gigascience* **3**: 9.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**: 824-834.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020a. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *bioRxiv*.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020b. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30**: 1291-1305.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**: 1061-1067.

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W et al. 2008. A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* **322**: 101-104.

Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420-1428.

Perumal S, Koh CS, Jin L, Buchwaldt M, Higgins EE, Zheng C, Sankoff D, Robinson SJ, Kagale S, Navabi ZK et al. 2020. A high-contiguity Brassica nigra genome localizes active centromeres and defines the ancestral Brassica genome. *Nat Plants* **6**: 929-941.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* **98**: 9748-9753.

Pflug JM, Holmes VR, Burrus C, Johnston JS, Maddison DR. 2020. Measuring Genome Sizes Using Read-Depth, k-mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). *G3 (Bethesda)* **10**: 3047-3060.

Phillippy AM, Schatz MC, Pop M. 2008. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol* **9**: R55.

Pop M. 2009. Genome assembly reborn: recent computational challenges. *Brief Bioinform* **10**: 354-366.

Pucker B, Holtgrawe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, Weisshaar B. 2019. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. *PLoS One* **14**: e0216233.

Qin M, Wu S, Li A, Zhao F, Feng H, Ding L, Ruan J. 2019. LRScaf: improving draft genomes using long noisy reads. *BMC Genomics* **20**: 955.

Rao G, Zhang J, Liu X, Lin C, Xin H, Xue L, Wang C. 2021. De novo assembly of a new Olea europaea genome accession using nanopore sequencing. *Hortic Res* **8**: 64.

Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D et al. 2011. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med* **365**: 709-717.

Raudsepp T, Gustafson-Seabury A, Durkin K, Wagner ML, Goh G, Seabury CM, Brinkmeyer-Langford C, Lee EJ, Agarwala R, Stallknecht-Rice E et al. 2008. A 4,103 marker integrated physical and comparative map of the horse genome. *Cytogenet Genome Res* **122**: 28-36.

Rellstab C, Zoller S, Sailer C, Tedder A, Gugerli F, Shimizu KK, Holderegger R, Widmer A, Fischer MC. 2020. Genomic signatures of convergent adaptation to Alpine environments in three Brassicaceae species. *Mol Ecol* **29**: 4350-4365.

Rhie A McCarthy SA Fedrigo O Damas J Formenti G Koren S Uliano-Silva M Chow W Fungtammasan A Kim J et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**: 737-746.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**: 245.

Rice ES, Green RE. 2019. New Approaches for Genome Assembly and Scaffolding. *Annu Rev Anim Biosci* **7**: 17-40.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348-352.

Ruan J, Li H. 2019. Fast and accurate long-read assembly with wtdbg2. doi:10.1101/530972 %J bioRxiv: 530972.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**: 155-158.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* **22**: 557-567.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**: 441-448.

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

Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N et al. 2011. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc Natl Acad Sci U S A* **108**: 10249-10254.

Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim J et al. 2016. De novo assembly and phasing of a Korean human genome. *Nature* **538**: 243-247.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**: 227-245.

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* **38**: 1044-1053.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**: 1728-1732.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**: 1117-1123.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407-410.

Slankster EE, Chase JM, Jones LA, Wendell DL. 2012. DNA-Based Genetic Markers for Rapid Cycling Brassica Rapa (Fast Plants Type) Designed for the Teaching Laboratory. *Front Plant Sci* **3**: 118.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**: 674-679.

Song B, Mott R, Gan X. 2018. Recovery of novel association loci in Arabidopsis thaliana and Drosophila melanogaster through leveraging INDELs association and integrated burden test. *PLoS Genet* **14**: e1007699.

Steinbiss S, Willhoeft U, Gremme G, Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res* **37**: 7002-7013.

Tamokou JDD, Mbaveng AT, Kuete V. 2017. Chapter 8 - Antimicrobial Activities of African Medicinal Spices and Vegetables. In *Medicinal Spices and Vegetables from Africa*, doi:https://doi.org/10.1016/B978-0-12-809286-6.00008-X (ed. V Kuete), pp. 207-237. Academic Press.

Tang H, Lyons E, Town CD. 2015. Optical mapping in plant comparative genomics. *Gigascience* **4**: 3.

Taylor TD, Noguchi H, Totoki Y, Toyoda A, Kuroki Y, Dewar K, Lloyd C, Itoh T, Takeda T, Kim DW et al. 2006. Human chromosome 11 DNA sequence and analysis including novel gene identification. *Nature* **440**: 497-500.

Thompson OA, Snoek LB, Nijveen H, Sterken MG, Volkers RJ, Brenchley R, Van't Hof A, Bevers RP, Cossins AR, Yanai I et al. 2015. Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856. *Genetics* **200**: 975-989.

Thrash A, Hoffmann F, Perkins A. 2020. Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics* **21**: 249.

Tomaszkiewicz M, Rangavittal S, Cechova M, Campos Sanchez R, Fescemyer HW, Harris R, Ye D, O'Brien PC, Chikhi R, Ryder OA et al. 2016. A time- and cost-effective strategy to sequence mammalian Y Chromosomes: an application to the de novo assembly of gorilla Y. *Genome Res* **26**: 530-540.

Tsai YC, Conlan S, Deming C, Program NCS, Segre JA, Kong HH, Korlach J, Oh J. 2016. Resolving the Complexity of Human Skin Metagenomes Using Single-Molecule Sequencing. *mBio* **7**: e01948-01915.

Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. 2018. MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res* **28**: 266-274.

Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**: 158.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**: 1051-1063.

Vaser R, Sovic I, Nagarajan N, Sikic M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737-746.

Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, Peluso P, Wenger AM, Concepcion GT, Kronenberg ZN, Munson KM et al. 2020. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann Hum Genet* **84**: 125-140.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.

Wang J, Liu W, Zhu D, Hong P, Zhang S, Xiao S, Tan Y, Chen X, Xu L, Zong X et al. 2020a. Chromosome-scale genome assembly of sweet cherry (Prunus avium L.) cv. Tieton obtained using long-read and Hi-C sequencing. *Hortic Res* **7**: 122.

Wang P, Yi S, Mu X, Zhang J, Du J. 2020b. Chromosome-Level Genome Assembly of Cerasus humilis Using PacBio and Hi-C Technologies. *Front Genet* **11**: 956.

Warren RL, Sutton GG, Jones SJ, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**: 500-501.

Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJ, Birol I. 2015. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* **4**: 35.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737-738.

Wei Q, Wang J, Wang W, Hu T, Hu H, Bao C. 2020. A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Hortic Res* **7**: 153.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27**: 757-767.

Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, Federman S, Stryke D, Briggs B, Langelier C et al. 2019. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N Engl J Med* **380**: 2327-2340.

Wotzel S, Andrello M, Albani MC, Koch MA, Coupland G, Gugerli F. 2021. Arabis alpina: a perennial model plant for ecological genomics and life-history evolution. *Mol Ecol Resour* doi:10.1111/1755-0998.13490.

Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**: 605-607.

Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, Luo F, Xie Z. 2017. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* **14**: 1072-1074.

Xiao S, Li J, Ma F, Fang L, Xu S, Chen W, Wang ZY. 2015. Rapid construction of genome map for large yellow croaker (Larimichthys crocea) by the whole-genome mapping in BioNano Genomics Irys system. *BMC Genomics* **16**: 670.

Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, Smith CL, Edgley ML, Rougvie AE, Fire AZ et al. 2019. Recompleting the Caenorhabditis elegans genome. *Genome Res* **29**: 1009-1022.

Yuan Y, Bayer PE, Scheben A, Chan CK, Edwards D. 2017. BioNanoAnalyst: a visualisation tool to assess genome assembly quality using BioNano data. *BMC Bioinformatics* **18**: 323.

Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, Jiao WB, Patel V, Velikkakam James G, Koornneef M, Ossowski S et al. 2016. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proc Natl Acad Sci U S A* **113**: E4052-4060.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821-829.

Zhang X, Goodsell J, Norgren RB, Jr. 2012. Limitations of the rhesus macaque draft genome assembly and annotation. *BMC Genomics* **13**: 206.

Zhang X, Zhang S, Zhao Q, Ming R, Tang H. 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* **5**: 833-845.

Chapter2

GALA: gap-free chromosome-scale assembly with long reads

GALA: gap-free chromosome-scale assembly with long reads

Abstract

High-quality genome assembly has wide applications in genetics and medical studies. However, it is still incredibly challenging to achieve gap-free chromosome-scale assemblies using current workflows for long-read platforms. Here we propose GALA, a computational framework for chromosome-by-chromosome *de novo* assembly implemented through a multi-layer graph that identifies mis-assemblies within preliminary assemblies and partitions the data into chromosome-scale linkage groups. The subsequent independent assembly of each linkage group generates a gap-free assembly free from the mis-assembly errors which usually hamper existing workflows. This flexible framework also allows us to integrate data from various technologies, such as Hi-C, genetic maps, a reference genome, and even motif analyses to generate gap-free chromosome-scale assemblies. We *de novo* assembled the *C. elegans* and *A. thaliana* genomes using combined Pacbio and Nanopore sequencing data from publicly available datasets. We also demonstrated the new method's applicability with a gap-free assembly of the human genome. In addition, GALA showed promising performance for Pacbio high-fidelity long reads. Thus, our method enables straightforward assembly of genomes with multiple data sources and overcomes barriers that at present restrict the application of *de novo* genome assembly technology.

2.1. Introduction

De novo genome assembly has wide applications in plant, animal, and human genetics. However, it is still very challenging for long-read platforms, such as Nanopore and Pacbio, to provide chromosome-scale sequences (Cao et al. 2017; Li et al. 2017). To date, numerous *de novo* assembly tools have been developed to obtain longer and more accurate representative sequences from raw sequencing data (Koren et al. 2017; Xiao et al. 2017; Kolmogorov et al. 2019). In most studies, however, assemblies by these tools comprise hundreds or even thousands of contigs. To produce chromosome-scale assembly, various information sources, such as Hi-C, genetic maps, or a reference genome, have been increasingly used to anchor contigs into big scaffolds (Jiao et al. 2017; Ellison and Cao 2020). As a consequence, the final genome assembly usually contains numerous gaps, and sometimes, is also plagued with mis-assemblies, as reported in (Muggli et al. 2015).

Gaps and mis-assemblies in a genome assembly can seriously undermine genomic studies. For example, a lot of sequence alignment tools have much lower performances when query sequences contain gaps (Chen and Tompa 2010; Song et al. 2018). In intraspecific genome comparisons, large gaps not only significantly increase the possibility of failure to detect long structure variants, but also produce inaccurate results of gene annotation (Bickhart and Liu 2014; BSong B 2019). Moreover, gaps and mis-assemblies have been reported to account for a large number of gene model errors in existing genome assembly studies (Zhang et al. 2012; Denton et al. 2014).

In this study, we report on GALA (**Ga**p-free **l**ong-read **a**ssembler), a scalable chromosome-by-chromosome assembly method implemented through a multi-layer computer graph. (**Fig. 2.1**). GALA separates two steps: firstly, it identifies multiple linkage groups in the genome, each representing a single chromosome (sometimes a chromosome arm) and it also describes chromosome structure with raw reads and assembled contigs from multiple *de novo* assembly tools; secondly, it assembles each linkage group by integrating results from multiple assembly tools and inference from raw reads. Moreover, our method can also exploit the information derived from Hi-C data to obtain chromosome-scale linkage groups in studies even with a complicated genome structure or those with low sequencing quality. Of note is that our method can be easily extended to incorporate other sources of information such as genetic maps or even a reference genome. Here, we show the utility of GALA by gap-free and chromosome-scale assemblies of Pacbio or Nanopore sequencing data from two publicly available datasets for which the original assembly contains large gaps and a number of unanchored scaffolds. Notably, our new method significantly outperforms existing algorithms in both datasets. Finally, we also demonstrate the application of our method to assemble a human genome with the help of a reference genome using Pacbio high-fidelity (HiFi) long reads.

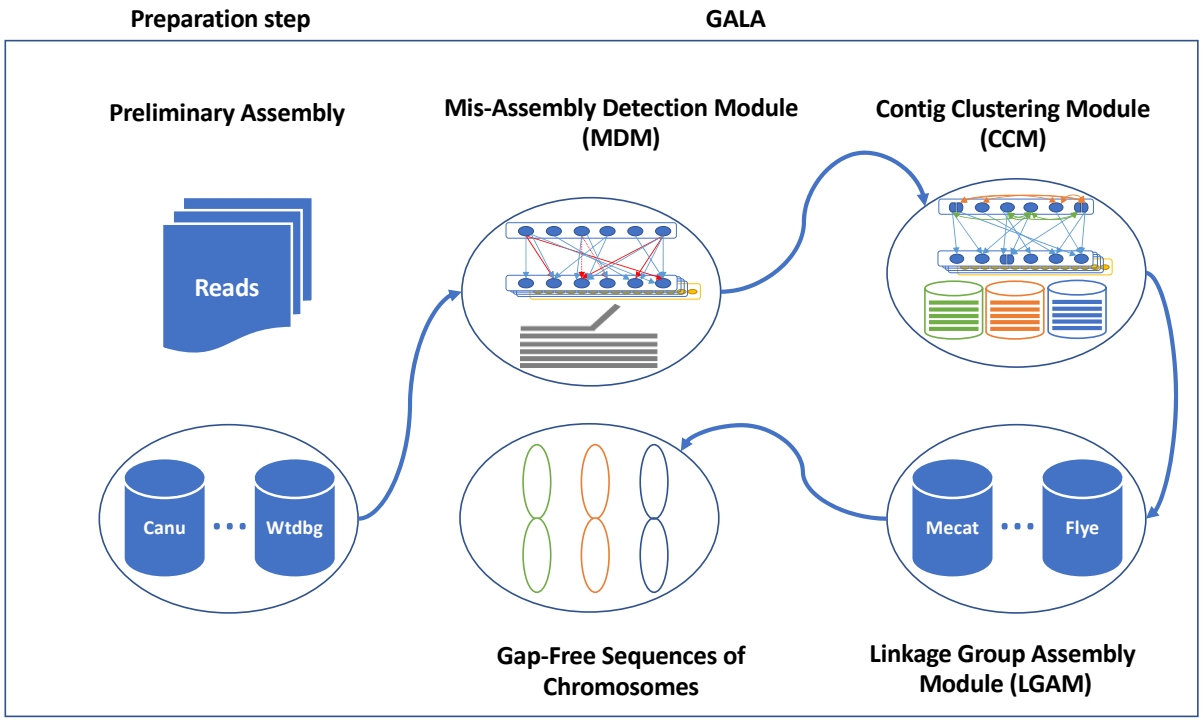


Figure 2.1. Overview of GALA. After *de novo* assembling with various tools, preliminary assemblies and raw reads are encoded into a multi-layer computer graph. Mis-assemblies are identified with MDM by browsing through the inter-layer information. The split nodes are clustered into multiple linage groups by the CCM. Each linkage group is assembled independently using LGAM to achieve the final gap-free sequences of chromosomes.

2.2. Results

2.2.1 Overview of the GALA framework

GALA exploits information from multiple *de novo* assembly tools and raw reads, as well as other information sources, such as Hi-C, genetic maps, or even a reference genome, if they exist. In GALA, various *de novo* assembly tools are selected first to create preliminary assemblies. These preliminary assemblies and raw reads are then aligned against each other. We use a multi-layer computer graph to model the GALA, with each assembly encoded as one layer, together with an extra layer representing the raw reads. Inside each layer, a contig (or a read in the raw-read layer) is encoded as a graph node. GALA browses through the reciprocal alignments and creates two types of edges. Any contradictory information between multiple assemblies or raw reads is recorded as a cross-layer edge. Inside each layer, if two nodes both partially overlap with the same node inside a different layer, a within-layer edge is created between them (**Fig. 2.2**).

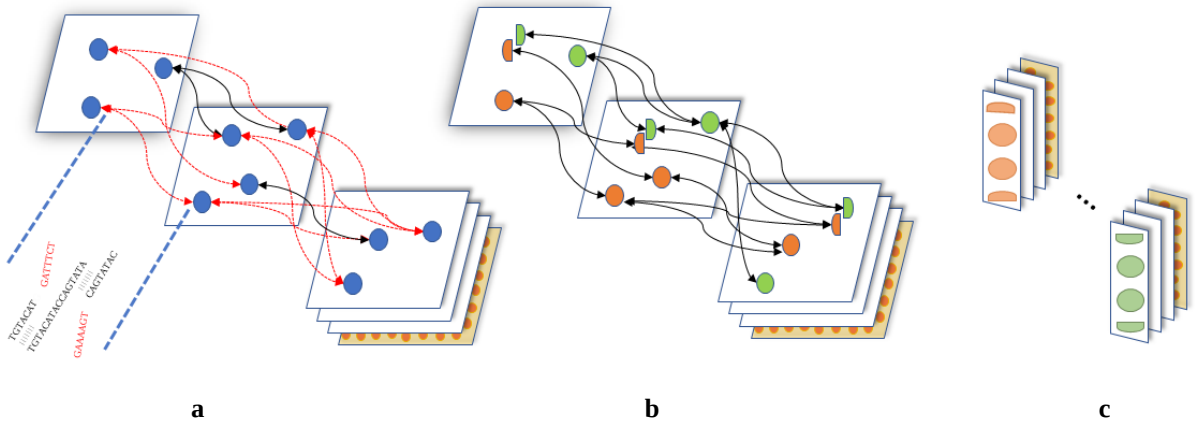


Figure 2.2. Illustration of a multi-layer computer graph in GALA. (a) The preliminary assemblies and raw reads are aligned against each other and encoded into a multi-layer graph. Conflicted alignments are encoded with edges in red. (b) The conflicted alignments are removed iteratively by splitting the nodes involved and new edges are assigned accordingly. The procedure stops only after all conflicted alignments in the system have been resolved. (c) Nodes connected by edges are clustered into linkage groups.

Depending on the sequencing quality and complexity of the genome structure, existing assembly tools usually exhibit different performances in terms of the number of misassembled contigs and N50. To prevent the spread of errors, we developed a mis-assembly detection module (MDM), which works by estimating the probability of mis-assemblies based on the contradictory cross-layer edges and splitting those nodes highly likely containing mis-assemblies to resolve the discordance in the computer graph (Methods). After removing contradictory cross-layer links, the contig-clustering module (CCM) pools the linked nodes within different layers and those inside the same layer into different linkage groups, usually each representing a chromosome (Methods). In several experiments, we identified orphan contigs. Interestingly, most of them come from external sources such as bacterial or sample contamination.

The successful partitioning of existing preliminary assemblies and raw reads into separate linkage groups allows us to essentially perform a chromosome-by-chromosome assembly. The raw reads from each linkage group are extracted and assembled with multiple assembly tools and merged together if necessary. For those tools which take corrected reads as input, we correct reads using suggested methods. Interestingly, we found that chromosome-by-chromosome assembly always provides better performance, especially for the repetitive fragments in terms of contiguity. In contrast, the improvement of read correction with chromosome-by-chromosome analysis is negligible. We also tested GALA in a fast mode,

where the consensus assembly for each chromosome is obtained by merging the assembled contigs within the linkage group without working on raw reads. However, in many cases, the fast mode generated gapped assemblies, thereby highlighting the distinct advantage of the chromosome-by-chromosome assembly strategy over existing tools.

2.2.2. *Caenorhabditis elegans* genome assembly

We used a publicly available dataset for *Caenorhabditis elegans* VC2010. The dataset was generated on the Pacbio platform with a 290X coverage along with an extra 32X coverage of Nanopore sequences (Yoshimura et al. 2019). As no current assembly tools support pooled sequencing data from Pacbio and Nanopore platforms, we used both datasets separately to generate preliminary assemblies (**Supp. Fig. 2.1**). Preliminary assemblies were generated using Canu, Flye, Mecat2/Necat, Miniasm, and Wtdbg2 (Methods). Among all our preliminary assemblies, the one produced by Pacbio-Flye showed the smallest number of contigs, with 41 contigs for 102 Mbp of overall sequences.

We applied GALA to the raw reads and the preliminary assemblies. The numbers of mis-assemblies in each preliminary assembly derived by the MDM algorithm ranged from 0 to 19. After resolving the discordances through the node-splitting operation, GALA modelled the input into 14 independent linkage groups. Seven of them contain a very small amount of sequencing data and apparently come from short continuous contigs. Among them, four contigs are from bacterial contamination or organelle DNA and two of them can be pooled into seven large linkage groups using Nanopore sequencing data. The remaining one contains a telomeric repetitive motif. We then performed telomeric motif analyses for the seven large linkage groups. Four of them contain complete chromosomes. Two groups contain the telomeric repetitive motif at one end and apparently come from two arms of the same chromosome and one group misses the telomeric repetitive motif at one end. We thus were able to merge 14 linkage groups further into six ones (**Supp. Fig. 2.2** and Methods). Of note is that the integrative assembly of each linkage group generated gap-free T2T complete sequences for all six chromosomes.

We polished our assembly using Pacbio and Illumina short reads and then compared it to the published VC2010 assembly and the N2 reference genome. Note that the VC2010 sample is derived from the N2 reference sample and their assemblies are supposed to be very close. The evaluation from Busco 3.0.0 indicated that our assembly successfully assembled two more

genes. Furthermore, the alignment of Illumina short reads against our assembly also reveals a better alignment rate as well as fewer variants (**Table 2.1** and **Supp. Fig. 2.3**).

	N2 reference genome	VC2010 assembly	GALA assembly
Assembly length	100,286,401	102,092,263	102,301,025
Number of contigs	7	7	7
Busco complete	968/982	968/982	970/982
Busco duplicated	6/982	6/982	6/982
Busco fragmented	8/982	8/982	6/982
Busco Missing	6/982	6/982	6/982
QV	36.4155	36.0716	36.2818
Mapped reads	130,604,410	130,639,345	130,652,108
Unmapped reads	4,568,540	4,533,605	4,520,842
Variants	17,385	14,839	14,169
SNPs	16,179	14,167	13,701
Deletions	412	282	124
Insertions	794	390	344
Indels	1,206	672	468

Table 2.*1*. The assembly performance evaluation of GALA with Busco scores and statistics of alignment of Illumina short reads. The Busco scores are computed using Busco V.3.0.0 with nematoda odb9 database. The QV scores are calculated using merqury reference free assessment tool.

We performed additional analyses to test the performance of our assembly using the Hi-C dataset generated by the same research group. No discordances were revealed by aligning the Hi-C data against our assembly using BWA-MEM, then detecting the discordances using Salsa (Ghurye et al. 2019). Salsa also supported the merging of two linkage groups suggested by the telomeric motif analyses in our assembly. For comparison, we also applied Salsa with Hi-C data to the best preliminary assembly from Flye with Pacbio data. This Flye/Hi-C assembly contains seven scaffolds and 14 unanchored contigs after excluding those from sample contamination. We observed 17 spanned gaps in the Flye/Hi-C assembly, with the two largest gaps being 495 Kbp and 159 Kbp (**Fig. 2.3**). Furthermore, we aligned the raw Pacbio reads to

different assemblies and examined the distribution of the depth-of-coverage across the genome (**Supp. Fig. 2.4**). Apart from being free of gaps, the GALA assembly shows comparable performance to the VC2010 assembly in terms of assembly error in repetitive regions.

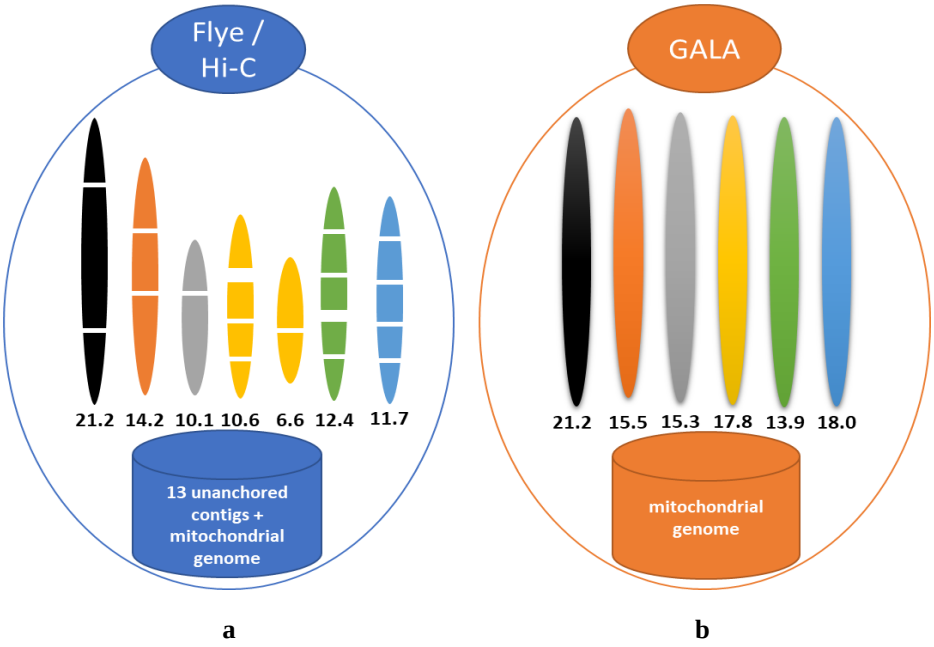


Figure 2.3. Comparison of Flye assembly with Hi-C scaffolding and GALA assembly of long reads of the *C. elegans* genome. (a) The Flye assembly with Hi-C scaffolding contains numerous gaps and 13 unanchored contigs in the assembly. (b) GALA produces gap-free assembly for each chromosome. Note this is not a fair comparison since GALA did not use Hi-C data in this assembly.

2.2.3. *Arabidopsis thaliana* genome assembly

We assembled *Arabidopsis thaliana* accession KBS-Mac-74 by combinatory analysis of two publicly available datasets using GALA: one is from Pacbio with a 58X coverage and the other is from Nanopore with a 28X coverage (Michael et al. 2018). We used both datasets separately to generate preliminary assemblies. Both raw and corrected reads by Canu and Mecat2/Necat (**Supp. Fig. 2.5**) were used as input for Canu, Flye, Mecat2/Necat, Miniasm, and Wtdbg2 assemblers (Methods).

GALA analyses on the raw reads and the preliminary assemblies highlighted a number of potential mis-assemblies for each preliminary assembly, which ranged from 1-18. GALA modelled the input into 15 independent linkage groups. Among them, one was from the mitochondrial genome, one from the chloroplast genome, and three are continuous fragments of 1.6 Kbp, 7.6 Kbp, and 18.5 Kbp. The remaining ten linkage groups represent a chromosome arm each. Previous studies have indicated that the Pacbio and Nanopore platforms seldomly

sequence through centromeric regions in *A. thaliana* (Michael et al. 2018; Pucker et al. 2019). Therefore, we only aimed to assemble each chromosome arm in this study. In this context, our algorithm assembled each linkage group into a continuous sequence. In total, we were able to identify a telomere motif in eight assembled chromosome-arm sequences. Interestingly, only two telomere motifs have been observed in the Col-0 reference genome, indicating possible missing sequences in the reference genome.

In summary, our final assembly contains ten complete chromosome-arm sequences and three unanchored contigs. We also further analysed the three unanchored contigs and all of them were mapped to the pericentromeric region in the reference genome. Thus, the successful assembly of the *A. thaliana* genome by combining Pacbio and Nanopore sequencing data indicates that GALA provides a flexible framework for integrated assembly of sequencing data from multiple sequencing platforms.

2.2.4. Human genome assembly

We next assembled a human genome using high-fidelity (HiFi) long reads generated by Pacbio using the circular consensus sequencing (CCS) mode. For simplicity, we used the published preliminary *de novo* assembly by HiCanu and the current human reference genome GRCh38.p13 as input for GALA. The raw reads and the input HiCanu preliminary assembly are partitioned by the contig-clustering module (CCM) of GALA. Here, CCM only serves as a raw-read separation tool to make it possible for subsequent chromosome-by-chromosome *de novo* assembly. Both information from the input reference genome, which could be different from the genome to be assembled, and information from the preliminary assembly, which is consistent with the genome of interest, have been used for raw-read separation. GALA revealed 23 independent linkage groups and assembled them one-by-one. Interestingly, when assembling linkage groups, we used two software, namely HiCanu and Hifiasm, and they provided significantly different assemblies in terms of the length of sequences. Taking chromosome 17 as an example, HiCanu assembled its linkage group into three contigs with a total length of 83.2 Mbp (40 Mbp, 24.7 Mbp, and 18.5 Mbp). In contrast, Hifiasm produced one single telomere-to-telomere contig of a total length of 82.1 Mb. To resolve this, we aligned the raw HiFi reads to both assemblies and examined the distribution of the depth-of-coverage. We selected the better genome assembly by taking into account the number of assembly errors as well as gaps. The comparison between our GALA assembly and the published assembly can be found in (**Fig. 2.4a** and **Supp. Fig. 2.6)**. Overall, our assembly comprised of 38 continuous

contigs, including seven telomere-to-telomere gap-free pseudomolecular sequences (3, 7, 10, 11, 16, 17, and 20), four near-complete chromosomes (5, 8, 12, and 19) each with a small telomeric fragment unanchored, and four chromosomes (4, 6 9, and 18) with gapped centromeric regions. Note that we only assembled the long arms of the five acrocentric chromosomes (13, 14, 15, 21, and 22) since the sequencing and assembly of their *p* arms are too challenging as they are almost all missing in both the reference genome and the published assembly.

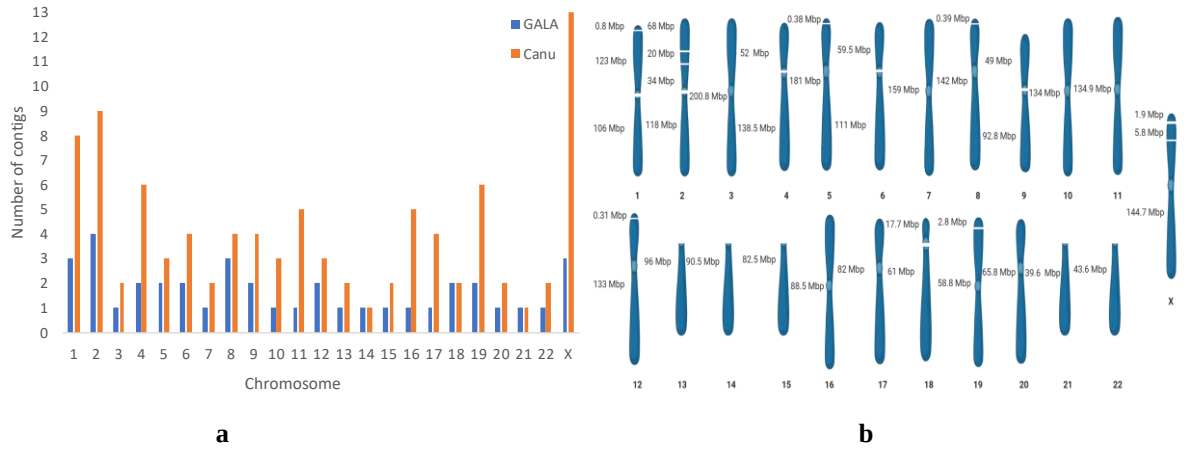


Figure 2.4. Human genome assembly by GALA. (a) Comparison of the number of contigs in assemblies by Canu and GALA. (b) A cartoon presentation of each chromosome assembled by GALA with the lengths of contigs labelled.

Our human genome assembly is depicted chromosome-by-chromosome in **Fig. 4b**. Here, two chromosomes are of key interest: chromosome 11 and chromosome X. In the reference genome GRCh38.p13 and also the published HiCanu assembly, chromosome 11 has several gaps and unanchored contigs. Interestingly, it is considered as one of the chromosomes with the highest density of genes linked with genetic diseases (Taylor et al. 2006). GALA successfully assembled this chromosome into a single contig free of gaps of a total length of 134.9 Mbp (**Supp. Fig. 2.6)**. The assembled chromosome 11 has two telomeric regions at both ends; however, one of them is missing in GRCh38.p13. The second example is chromosome X, whose assembly is regarded as highly challenging and extra effort has been devoted to this in a recent paper (Miga et al. 2020). Our assembly only contains two short gaps (about 0.75Kbp and 1.8Kbp) compared to the published one. The successful assembly of the human genome indicates that GALA can efficiently be applied to Pacbio HiFi data.

In the above assembly of CHM13 by GALA, the reference genome has been used to help to separate raw-read into linkage groups. One might wonder whether this would lead to a vulnerability that plagues traditional *reference-guided* assemblies or scaffolding. It has been reported that traditional *reference-guided* assemblies suffer from short-length assembly errors and mis-scaffolds because of reference biases and chromosomal rearrangements among different strains and cell lines, as well as errors of sequence alignment (Schneeberger et al. 2011; Ekblom and Wolf 2014; Lischer and Shimizu 2017). In addition, *reference-guided* assembly leads to missing sequences in highly divergent regions (Lischer and Shimizu 2017). Fortunately, GALA can avoid both problems. Firstly, GALA only uses the reference genome to cluster contigs from the preliminary assembly and raw reads, so the role of the reference is more like the genetic map thus insensitive to the sequence variation between the query genome and the reference. Moreover, the subsequent *de novo* assembly of linkage groups prevents assembly errors and mis-scaffolds. For example, if raw reads have been mistakenly put into the same linkage group, it leads to fragmented assembly but not errors. Secondly, GALA's linkage groups contain contigs from preliminary assembly, so unique and highly divergent regions would not miss out when aligning raw reads to linkage groups. For comparison, we performed the *reference-guided* scaffolding of the HiCanu preliminary assembly - using Ragoo (Alonge et al. 2019) and gap-filled it using PBJelly (English et al. 2012). Ragoo scaffolded ~ 12 Mbp of centromeric and pre-centromeric sequences of chromosome 9 to chromosome 4 (**Supp. Fig. 2.7**) with big gaps. In contrast, GALA clustered and assembled the reads from highly similar centromeric regions and constructed two continuous contigs in the two regions. On the other hand, the unique sequences and divergence between the used reference and the query genome lead to fragmented chromosomes if the user assembles the reads aligned to the reference genome directly.

2.2.5. Effect of the sequencing depth on the performance of GALA

We next investigated how the performance of GALA changes depending on the sequencing depth. We subsampled the original *C. elegans* Pacbio sequencing data using software Fastq-sample to 20X, 30X, 40X, 50X, 60X, 70X, 80X, 90X, 100X, and 150X coverage, together with Hi-C data, and performed *de novo* assembly independently. Preliminary assemblies were generated using Canu, Flye, Mecat2, Miniasm, and Wtdbg2 with raw and corrected reads. A detailed comparison between the resulting assemblies can be found in (**Fig. 2.5** and **Supp. Table 2.1**). This study revealed two interesting findings. Firstly, the gap-free *de novo* assembly is not a suitable option when the data coverage is less than 40X due to the

limitation of current *de novo* assembly tools. As a consequence, GALA switches to gapped assembly for this scenario. Secondly, without Hi-C for scaffolding, Flye and GALA reach the performance curve plateau at 60X and 40X coverage, respectively, regarding the number of scaffolds and N50 of their assemblies. When Hi-C data are applied, the performance curve plateau starts from 40X for Flye and GALA (**Fig. 2.5a, b**). The higher coverage leads to better assembly for Flye with or without Hi-C data by lowering down the number of big gaps and mis-assemblies; however, no notable effects on N50 and the number of scaffolds are observed (**Fig. 2.5c**). Thus, the higher coverage of data has no notable effect on GALA assembly in general.

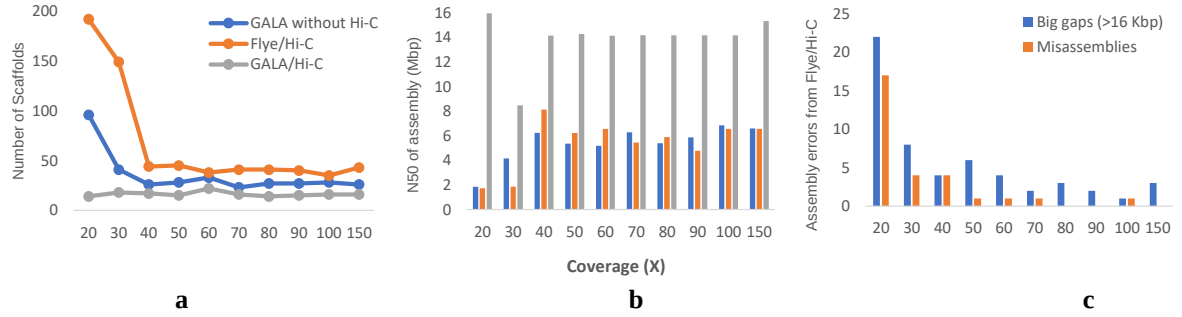


Figure 2.5. The assembly performances of GALA and Flye with Pacbio sequencing data at various coverages. Three assembly procedures have been tested: GALA without Hi-C data, Flye/Hi-C, and GALA/Hi-C. The assembly performances are evaluated in terms of (a) the number of scaffolds, (b) N50, and (c) the number of big gaps (>16Kbp) and mis-assemblies. For simplicity, only the number of gaps and mis-assemblies for Flye/Hi-C have been shown, as only one mis-assembly has been identified in the assembly by GALA using 30X coverage sequencing data without the application of Hi-C data.

The performance of GALA, as well as almost all assembly software tools, changes significantly with raw read length and sequencing error. Note that the above analyses are based on the Pacbio sequencing data generated with Pacbio RSII. Consequently, the lengths of the raw reads are notably smaller and sequencing error is significantly higher than the current Pacbio Sequel II. In practice, the sequencing length distribution often varies significantly between different sequencing platforms, genome centers, and sample preparation. Therefore, it is difficult to set a straightforward threshold value for the minimum coverage of data for GALA assembly. As a rule of thumb, GALA can produce gap-free assembly from 25X coverage of Pacbio Sequel II data or Nanopore MinION data if N50 of the raw data is larger than 20 Kbp. For Pacbio HiFi, 20X coverage works well for GALA due to its low sequencing error rate.

2.2.6. Effect of chromosome-by-chromosome assembly on the assembly graph

We investigated why GALA achieved complete assembly while existing assembly software tools had failed. We postulated that the chromosome-by-chromosome assembly strategy had played a role, and thus, we compared our assembly of *C. elegans* to that from Miniasm. This comparison revealed a much simpler computer graph in the chromosome-by-chromosome assembly. In terms of the number of overlaps between reads (graph edges) in the assembly of *C. elegans*, the whole genome assembly generated 190,936,281 edges, whereas the chromosome-by-chromosome assembly only generated 138,678,842 edges (27.37% less). A comparison between the whole genome and the chromosome-by-chromosome assembly is depicted in **Fig. 2.6**.

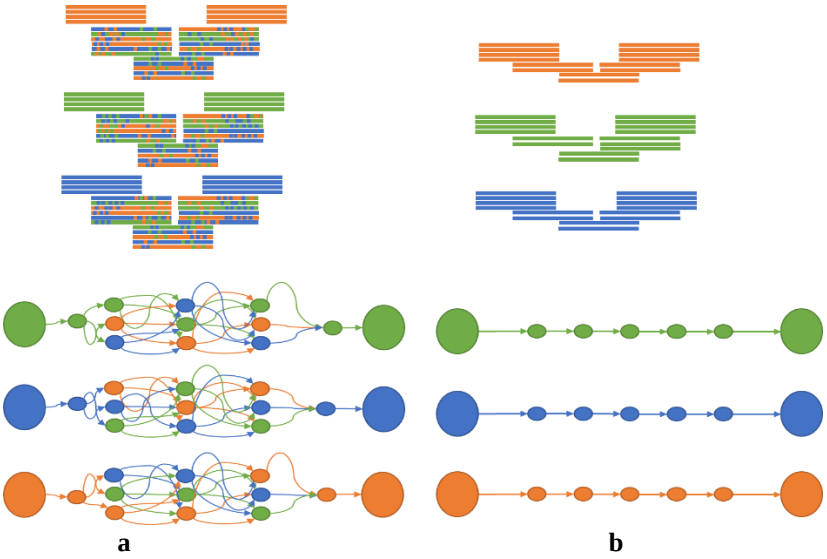


Figure 2.6. Comparison of the overlap graphs used by Miniasm during assembly of a region in the *C. elegans* genomes when the chromosome-by-chromosome strategy is applied or not. (a) In the whole genome assembly mode, the overlap graph used by Miniasm contains numerous edges and extra effort is needed to collapse edges. (b) The chromosome-by-chromosome assembly allows a linear overlap graph to be derived by Miniasm in the same region.

The advantage of chromosome-by-chromosome assembly is more obvious in the regions which contain highly similar sequences, but still have unique markers, e.g., regions with ancient transposons (**Fig. 2.6**). In addition, the regions which contain repetitive sequences, but are expanded by long reads, usually allow for a complete assembly by overlap graph-based algorithms, such as Canu or Mecat. However, such assembly is too challenging for *de Bruijn*

graph-based algorithms like Wtdbg2. In both scenarios, the GALA method can obtain superior results.

2.3 Discussion

Here, we have presented GALA, a scalable chromosome-by-chromosome assembly method implemented through a multi-layer computer graph. Compared to existing state-of-art assembly workflows and computational tools, GALA improved the contiguity and completeness of genome assembly and also has considerable advantages in a variety of settings. Furthermore, our new method is highly modular. In detail, the mis-assembly detection module (MDM) should be applicable for error correction regardless of the specific algorithm used for assembly and the contig-clustering module (CCM) can be widely applied for generating consensus assembly from multiple sequences. Although we have focused on *de novo* assembly in this paper, the modules in GALA should also work equally well in other applications.

In this study, we generated chromosome-scale gap-free assemblies in most of our experiments. We notice that the GALA assembly is usually smaller than preliminary assemblies. This is due to the fact that contigs usually contain duplicated sequences around the break point (**Supp. Fig. 2.8**). In certain circumstances, we failed to assemble challenging regions such as centromeres in *A. thaliana* and also certain regions in the human genome. This failure is mainly due to the absence of raw sequencing data in these regions (**Supp. Fig. 2.9**), and thus, also occurred in most of the other commonly used computational tools (Arabidopsis Genome 2000; Zapata et al. 2016; Pucker et al. 2019; Jiao and Schneeberger 2020). The strength of GALA comes from the multi-layer computer graph model, which is highly flexible in incorporating heterogenous information. As clearly demonstrated in the assembly of the *C. elegans* and *A. thaliana* genomes, combinatory analyses of Pacbio and Nanopore sequencing data were achieved.

The performance of our new GALA method also reflects the advantage of chromosome-by-chromosome assembly. Notably, the concept of chromosome-by-chromosome assembly was successfully tested on genome assembly in wheat, for which expensive devices and time-consuming procedures have had to be applied (Paux et al. 2008; Holusova et al. 2017). GALA is the first method to demonstrate that this can be achieved computationally. The concept of chromosome-by-chromosome assembly can also be applied to existing computational tools to refine an existing assembly. In addition, linkage group-based assembly provides a flexible

framework for GALA to support haplotype assembly in the future. This can be achieved by updating the linkage group assembly module (LGAM) to support haplotype assembly tools.

Finally, there is still room to improve GALA's assembly quality. Specifically, GALA assembly sometimes collapses long repetitive regions **(Supp. Figs. 4** and **6)**. In this context, we compared the raw reads aligned to chromosome X of the T2T v1.0 assembly and the reads in GALA's chromosome X linkage group. Interestingly, only a single read aligned to the chromosome X of the T2T v1.0 assembly is missing from GALA's chromosome X linkage group, indicating the bottleneck of the performance of GALA is the linkage group assembly module (LGAM) which relies on existing assembly tools. Thus, a new tool that can fully exploit the chromosome structure and depth-of-coverage, similar to centroFlye (Bzikadze and Pevzner 2020) but applicable to all long repetitive fragments, would be helpful in the future.

2.4 Methods

2.4.1 Reciprocal alignment between preliminary assemblies:

Minimap2 (Li 2018) (-x asm5) was used to map preliminary assemblies against each other. The raw and corrected reads were aligned to an assembly using BWA-MEM (Li and Durbin 2009) with default parameters.

2.4.2 Mis-assembly detection module (MDM):

We built a multi-layer graph by encoding the information from various preliminary assemblies D_n. Each preliminary assembly D_x represented a layer that consists of a set of nodes C_m, each node representing an assembled contig. The starting point of the MDM was the reciprocal alignment of D_n, which produced $n * (n-1)$ mapping results. We filtered the mapping results based on four criteria: (I) mapping quality (default 20), (II) contig length (default 5 Kbp), (III) alignment block length (default 5 Kbp), and (IV) sequence identity percentage (default 70%). All parameters are tunable in GALA. A simple merging procedure was performed to merge nodes within the same layer if they satisfy these four criteria to reduce the burden on computational resources.

We then linked the nodes between different layers by retrieving the information from reciprocal alignment. Assuming that a contig in node *C* in query layer D_x, denoted as C^{D_x}, is mapped to a set of nodes in layer ($D_{1..n}$), denoted as $\{C_1^{D_1}, \dots, C_i^{D_1}, \dots C_i^{D_n}\}$, a discordance at region *M* occurs if and only if contig $C_i^{D_k} \in \{C_1^{D_1}, \dots, C_i^{D_1}, \dots C_i^{D_n}\}$ is partially mapped to C^{D_x} as

exemplified in **Fig. 2a**. Two sequences are partially mapped if they cannot be merged together but their substrings, usually from one end, can be merged together according to the above four criteria.

Let L be the length of the contig C^{D_x}, N_A be the number of contigs partially mapped to M, N_B the number of contigs with complete alignment, and N_S be the number of contigs starting or ending at M. We considered M as a genuine mis-assembled locus if:

$$N_A \geq (n/2) \quad (1)$$

$$N_B = 0 \;\&\; N_A \geq 2 \quad (2)$$

$$N_A \geq 2 \;\&\; \left(\frac{N_B}{N_A}\right) \leq 0.5 \quad (3)$$

$$N_S > 0 \;\&\; \left(\frac{N_B - N_S}{N_A}\right) \leq 0.6 \quad (4)$$

If a mis-assembly is identified, the node is split into two nodes from the region *M*. This procedure iterates until the whole graph is free of mis-assemblies.

2.4.3 Contigs clustering module (CCM):

The multi-layer computer graph output by MDM was expanded by adding into an extra layer representing the raw reads. So far, within each layer, nodes were separate from each other and no intra-layer edge existed. We first built intra-layer edges by browsing through the existing cross-layer edges. For node C^{D_x} and its linked cross-layer node $\{C_0^{D_1}, \ldots, C_i^{D_1}, \ldots C_i^{D_n}\}$, CCM starts by traversing all $\{C_0^{D_1}, \ldots, C_i^{D_1}, \ldots C_i^{D_n}\}$. An intra-layer edge was built up if more than one node in the same layer was linked to the same cross-layer node. Then, CCM pooled all connected nodes into a linkage group.

In the previous step of MDM, only contigs with a length larger than a certain threshold value, 5 Kbp at default, were encoded into our computer graph. Thus, those with smaller sizes were not used for mis-assembly detection. To avoid the situation where unique sequences could be missed out by accident, we kept them and classified them into existing linkage groups for further analysis.

If Hi-C information or a genetic map is available, extra links can be created between internal nodes. This approach would essentially lead to the merging of multiple independent linkage groups. CCM could also be performed in an iterative mode together with the linkage group assembly module (LGAM), as demonstrated in the examples below.

2.4.4 Linkage group assembly module (LGAM):

The reads within a linkage group were assembled using assembly tools, e.g., Flye, Mecat, and Miniasm. In most cases, the assembly tool can produce a gap-free chromosome-scale assembly. We noticed that when a single continuous contig cannot be achieved for a linkage group, the breakpoint usually contains a very long repetitive sequence (most of the time in centromeric regions). LGAM provides a simplified version of the overlap graph-based merging algorithm to merge two contigs if necessary. However, this procedure sometimes causes collapsing of repetitive regions.

The long repetitive regions could also confuse existing assembly tools in a similar way. When assemblies from multiple software tools are significantly different in terms of length of sequence, we suggest the user to align the raw reads to different assemblies and examine the distribution of the depth-of-coverage. The user should select the best assembly by taking into account the number of assembly errors as well as gaps.

2.4.5 *Caenorhabditis elegans* assembly:

The Pacbio dataset contains three different runs and there was a clear batch effect with the sequencing quality and the amount of data between runs. We thus tested the assembly tools with either all runs (290X in coverage) or the biggest run alone (240X in coverage). We also used the reads-correcting-and-trimming module from Canu 1.8 (Koren et al. 2017) to correct the raw reads if the assembly tools take corrected reads as input. Preliminary assemblies were generated using Canu 1.8, Mecat2/Necat (Xiao et al. 2017), Flye 2.4 (Kolmogorov et al. 2019), Miniasm 0.3-r179 (Li 2016), and Wtdbg2 (Ruan and Li 2019), from Pacbio raw and corrected reads as well as Nanopore raw reads. By comparing the summary statistics of preliminary assemblies, ten preliminary assemblies were chosen for GALA.

GALA modelled the preliminary assemblies and raw reads into 14 independent linkage groups. Seven of them were short continuous contigs and the others represented individual chromosomes or chromosome arms. Further analyses by blasting the seven short contigs in the NCBI database indicated that three of them were from *E. coli* contamination and one from the

C. elegans mitochondrial genome, and thus, were excluded from the subsequent analyses. Of the remaining three short contigs, two of them can be reliably put into the seven previously created linkage groups with the help of the assembly of Nanopore reads with Miniasm (**Supp. Fig. 2.2**).

We assembled seven linkage groups with LGAM, each into a continuous sequence. Among the seven continuous sequences and one unanchored short contig, four of them revealed the telomere repetitive motif at both terminals, indicating they are complete assemblies of single chromosomes. One chromosome-scale sequence had a telomere repetitive motif at one end, and its missing telomeric repetitive motif can be identified in the unanchored short contig, indicating they both should be merged as a single linkage group. The remaining two had a telomere repetitive motif at either side and their sizes clearly indicated they were two arms from a single chromosome. We thus pooled their linkage groups together. Finally, we re-assembled the two newly created linkage groups and were able to create complete sequences for the two chromosomes with a telomeric repetitive motif at both terminals. Further analyses indicated that the split of this single chromosome into two linkage groups in the first run was mainly due to several tandem repeats.

2.4.6 *Caenorhabditis elegans* genome assembly polishing and quality control:

For a more accurate comparison, we polished our assembly with Pacbio and Illumina sequencing data. For this purpose, we first ran racon (Vaser et al. 2017) with corrected Pacbio reads. The assembly was then polished using quiver 2.3.2 (Chin et al. 2013) with Pbmm2 1.1.0 as an aligner. Finally, we ran pilon 1.23 (Walker et al. 2014) using Illumina sequencing data to correct short errors, especially those in homomorphic regions.

We evaluated the completeness of our polished assembly with Busco 3.0.0, and compared it to the published assembly, which is also polished using the same Illumina sequencing data as well as the reference genome (**Table 2.1** and **Supp. Fig. 2.10**). We also aligned the Illumina short reads to our assembly using BWA-MEM and called the variants using BCFtools 1.9. Finally, we collected the statistics and compared them to those from the published assembly as a benchmark for the precision of the assembly.

2.4.7 *Arabidopsis thaliana* assembly:

Pacbio and Nanopore raw reads datasets were corrected using the Canu 1.8 correcting and trimming module. We found that Mecat2/Necat showed good performance for read

correction in this dataset. Therefore, we additionally generated corrected Pacbio reads with Mecat2 with the minimal length of reads parameter as 5000 or 3000 separately and corrected Nanopore reads with Necat with the minimal length of reads parameter 500. The preliminary assemblies were generated using Canu, Flye, Mecat2/Necat, Miniasm, and Wtdbg2 from corrected and raw reads in both datasets. Thirteen preliminary assemblies were chosen for GALA with default parameters. In LGAM, we used Flye, Mecat2/Necat, Miniasm, and Wtdbg2 tools.

2.4.8 Data availability:

The Pacbio, Nanopro sequencing data, and Illumina reads of *C. elegans* are available at PRJNA430756. *A.thaliana* KBS-Mac-74 Pacbio and Nanopore sequencing data are available at PRJEB21270. We downloaded the human dataset from https://obj.umiacs.umd.edu/marbl_publications/hicanu/index.html. Genome assemblies that were generated by GALA in this study are available at https://doi.org/10.5281/zenodo.4672329.

Reference:

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* **20**: 224.

Arabidopsis Genome I. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796-815.

Bickhart DM, Liu GE. 2014. The challenges and importance of structural variation detection in livestock. *Front Genet* **5**: 37.

BSong B SQ, Wang H, Pei H, Gan X and Wang F. 2019. Complement Genome Annotation Lift Over Using a Weighted Sequence Alignment Strategy. *Front Genet* **10**.

Bzikadze AV, Pevzner PA. 2020. Automated assembly of centromeres from ultra-long error-prone reads. *Nat Biotechnol* **38**: 1309-1316.

Cao MD, Nguyen SH, Ganesamoorthy D, Elliott AG, Cooper MA, Coin LJ. 2017. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat Commun* **8**: 14515.

Chen X, Tompa M. 2010. Comparative assessment of methods for aligning multiple genome sequences. *Nat Biotechnol* **28**: 567-572.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563-569.

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol* **10**: e1003998.

Ekblom R, Wolf JB. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl* **7**: 1026-1042.

Ellison CE, Cao W. 2020. Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of Drosophila melanogaster. *Nucleic Acids Res* **48**: 290-303.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**: e47768.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol* **15**: e1007273.

Holusova K, Vrana J, Safar J, Simkova H, Balcarkova B, Frenkel Z, Darrier B, Paux E, Cattonaro F, Berges H et al. 2017. Physical Map of the Short Arm of Bread Wheat Chromosome 3D. *Plant Genome* **10**.

Jiao WB, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, Willing EM, Piednoel M, Woetzel S, Madrid-Herrero E 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 Res* **27**: 778-786.

Jiao WB, Schneeberger K. 2020. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun* **11**: 989.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**: 540-546.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-736.

Li C, Lin F, An D, Wang W, Huang R. 2017. Genome Sequencing and Assembly by Long Reads in Plants. *Genes (Basel)* **9**.

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103-2110.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Lischer HEL, Shimizu KK. 2017. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics* **18**: 474.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat Commun* **9**: 541.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* doi:10.1038/s41586-020-2547-7.

Muggli MD, Puglisi SJ, Ronen R, Boucher C. 2015. Misassembly detection using paired-end sequence reads and optical mapping data. *Bioinformatics* **31**: i80-88.

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W et al. 2008. A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* **322**: 101-104.

Pucker B, Holtgrawe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, Weisshaar B. 2019. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. *PLoS One* **14**: e0216233.

Ruan J, Li H. 2019. Fast and accurate long-read assembly with wtdbg2. doi:10.1101/530972 %J bioRxiv: 530972.

Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N et al. 2011. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc Natl Acad Sci U S A* **108**: 10249-10254.

Song B, Mott R, Gan X. 2018. Recovery of novel association loci in Arabidopsis thaliana and Drosophila melanogaster through leveraging INDELs association and integrated burden test. *PLoS Genet* **14**: e1007699.

Taylor TD, Noguchi H, Totoki Y, Toyoda A, Kuroki Y, Dewar K, Lloyd C, Itoh T, Takeda T, Kim DW et al. 2006. Human chromosome 11 DNA sequence and analysis including novel gene identification. *Nature* **440**: 497-500.

Vaser R, Sovic I, Nagarajan N, Sikic M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737-746.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.

Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, Luo F, Xie Z. 2017. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* **14**: 1072-1074.

Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, Smith CL, Edgley ML, Rougvie AE, Fire AZ et al. 2019. Recompleting the Caenorhabditis elegans genome. *Genome Res* **29**: 1009-1022.

Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, Jiao WB, Patel V, Velikkakam James G, Koornneef M, Ossowski S et al. 2016. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proc Natl Acad Sci U S A* **113**: E4052-4060.

Zhang X, Goodsell J, Norgren RB, Jr. 2012. Limitations of the rhesus macaque draft genome assembly and annotation. *BMC Genomics* **13**: 206.

Chapter3

Telomere-to-telomere *de novo* assembly of Cardamine species using GALA

Telomere-to-telomere *de novo* assembly of Cardamine species using GALA

Abstract

In modern biology, comparative genomics is a vital tool to investigate the biological, developmental and genetic principles of various biosystems. Availability of high-quality reference genomes is a fundamental substance for all applications of comparative genomics. Here we take advantage of the GALA assembler in incorporating heterogeneous data to achieve gap-free chromosome-scale assembly of three Cardamine species. Each assembly was sorted into eight pseudomolecules, including all assembled sequences. As a result, we accomplished a complete gap-free assembly of *C. oligosperma* and two *C. hirsuta* strains; Azores and the Oxford reference strain. At the same time, the *C. resedifolia* genome comprises only three gaps in three centromeric regions. Furthermore, our analysis proved that the new *C. hirsuta* assembly resolved several structure discrepancies in the reference genome. We exploited comparative genomics analysis to demonstrate the synteny, collinearity and karyotype differences between the assembled genomes. Finally, we believe that these assembled genomes provide valuable genomic material to facilitate various comparative and biological studies.

3.1 Introduction

The assembly of the *Arabidopsis thaliana* genome triggered the plant genomic revolution in 2000 based on the BAC strategy and Sanger sequencing (Arabidopsis Genome 2000). The usefulness of model systems in modern biology comes from achieving the highest possible quality of reference genome assembly. Therefore, since the first version's release, many updates have been introduced to enhance the accuracy of the plant model system golden genome (Haas et al. 2005; Lamesch et al. 2012; Berardini et al. 2015). Next-generation sequencing (NGS) technologies facilitated the establishment of more than 200 plant genome assembly projects. Unfortunately, most of these assemblies are low-quality and fragmented genomes (Belser et al. 2018). The establishment of the long-read technologies provided an excellent opportunity to obtain high-quality assemblies. Moreover, the introduction of Hi-C and optical mapping techniques with the long-read technology accelerated the high-quality plant genomes construction (Tang et al. 2015; Jiao and Schneeberger 2020; Murigneux et al. 2020). Unfortunately, the dynamics of transposable elements in plant genomes underlying the assembly algorithms and force the scaffolding process to generate gaped scaffolds.

Brassicaceae is a large plant family consisting of 372 genera and more than 4000 species (Tamokou et al. 2017). Moreover, The Brassicaceae family has a particular economic and scientific significance within plant families. For example, it includes several important crops and the first emerged plant model system *Arabidopsis thaliana* (Perumal et al. 2020). Hence, revealing the genomic architecture of representative members of the Brassicaceae family is a crucial objective of comparative genomics to understand the physiological, biochemical and genetic properties of different biological systems (Murat et al. 2015; Chang et al. 2016). Therefore, from 598 released plant genomes, 43 Brassicaceae genomes were published till (27.08.2021) (www.plabipd.de). In addition to *Arabidopsis thaliana*, several Brassicaceae species were chosen as model plants to reflect the unique features of different biological systems in comparative studies (Koornneef and Meinke 2010; Slankster et al. 2012; Wotzel et al. 2021). For instance, *Cardamine hirsuta* was proposed as an experimental model plant system to investigate several developmental pathways and comparative developmental studies with its close relative *Arabidopsis thaliana* (Hay et al. 2014).

The first draft of the *C.hirsuta* genome was published in 2016 based on Illumina sequencing, BAC and genetic map to facilitate the utilization of *C.hirsuta* in comparative and developmental studies (Gan et al. 2016). The assembled genome comprises eight chromosome-scale super-scaffolds and more than 600 unanchored contigs. Unfortunately, the draft assembly also includes around 7.5Mbp of unknown bases (Ns), which introduced thousands of gaps. The genomic gaps may undermine the resolution of developmental and comparative investigations or increase the false statistical outcomes (Domanska et al. 2018; Rhie et al. 2021). Here, we *de novo* assembled two accessions of *Cardamine hirsuta* into a telomere-to-telomere assembly using a combination of heterogeneous data including Pacbio, Nanopore and Hi-C sequences. After proving superiority over the reference genome on almost all evaluation matrices, we released the second draft of the *C. hirsuta* (ox) gapless telomere-to-telomere genome. Moreover, we constructed gapless chromosome-scale assembly of two *C. hirsuta* close relative species, *Cardamine oligosperma* and *Cardamine resedifolia*, to facilitate inter-species comparative studies. Finally, we conducted a repeat-load and chromosome karyotype comparative study among the assemblies.

3.2 Results:

3.2.1 *C. hirsuta* (ox) genome assembly

We used two long-read datasets of *C. hirsuta* reference strain Oxford derived from Pacbio and Nanopore sequencing technologies. The Pacbio (RS) dataset consisted of 3,399,653 reads, containing 25,120,054,417 bases with a read N50 of 11.5 Kbp and coverage depth of 125 folds (**Table 3.1; Fig. 3.1**). The Nanopore (GridION) dataset consisted of 2,926,910 reads, containing 17,761,301,168 bases with a read N50 of 33 Kbp and coverage depth of 88 folds (**Table 3.1** and **Fig. 3.1**). To improve the quality of the raw reads, we used the self-correction module of Canu 1.6 to correct the raw reads of both datasets, retrieving 495,799 pre-assembled corrected reads, containing 7,512,894,482 bases and a coverage depth of 37 folds from the Pacbio dataset (**Table 3.1; Fig. 3.1**). On the other hand, the Nanopore database holds 161,812 pre-assembled corrected reads, containing 7,749,626,415 bases and a coverage depth of 38 folds (**Table 3.1; Fig. 3.1**).

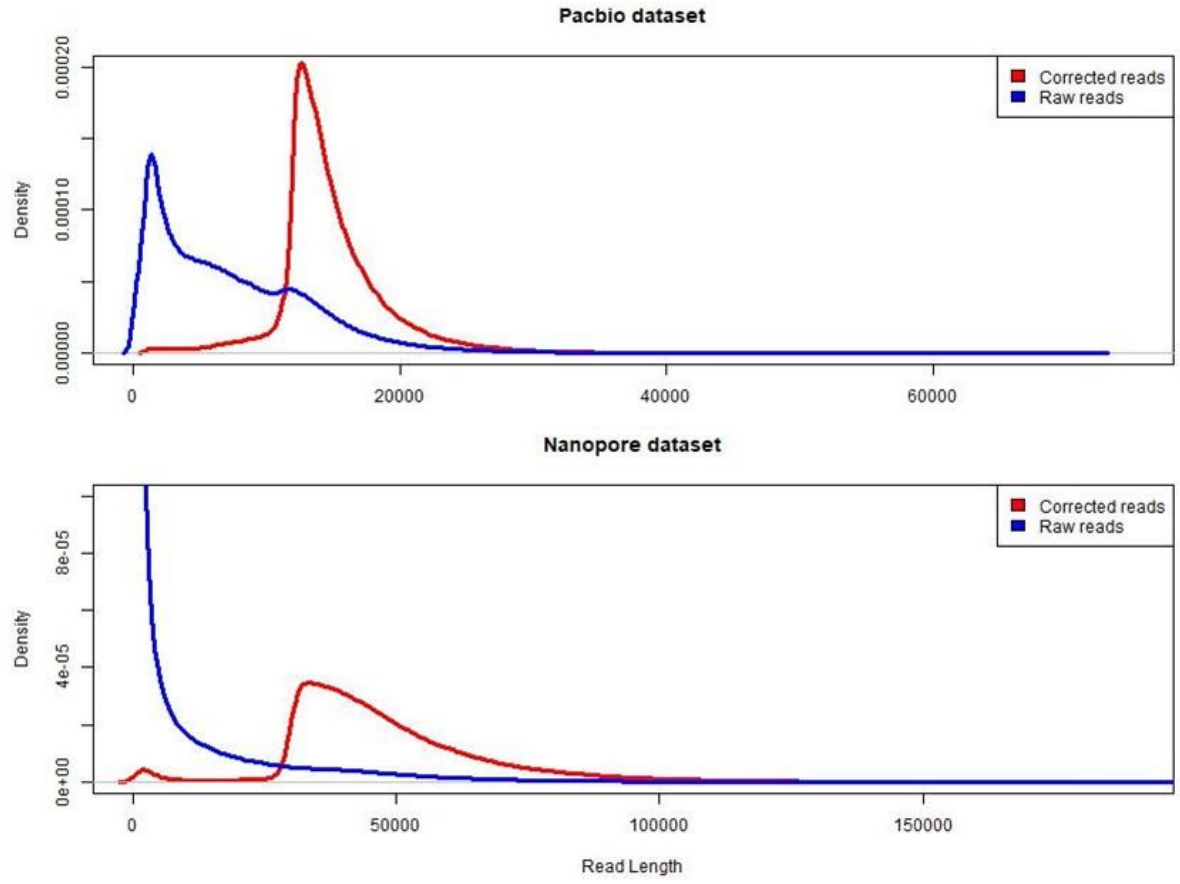


Figure 3.1: *Cardamine hirsuta* long-reads datasets length distribution.

	Pacbio		Nanopore	
	Raw	**corrected**	**Raw**	**corrected**
Total bases (Gb)	25.12	7.34	17.76	7.74
Total reads	3399653	495359	2926910	161812
Read N50 (Kb)	11.5	14.7	33	49.8
Read mean (kb)	7.3	14.8	6	47.8
Read L50	790524	201489	169734	57852
Coverage	125	36	88	38

Table 3.1: PacBio and Nanopore subreads statistics.

In total, 14 draft assemblies were constructed. First, self-corrected reads from both datasets employed Canu, Miniasm, Wtdbg2 and Flye to generate eight draft assemblies. Furthermore, raw reads carried out six draft assemblies using Flye, Miniasm and Mecat/Necat software. Finally, a polishing step hired quiver for Pacbio drafts and nanopolish for Nanopore drafts to improve contigs correctness. The overall results showed that assemblies derived from the Nanopore dataset possess higher contiguity than those obtained from the Pacbio dataset, except for the Wtdbg2 draft. Moreover, Flye and Miniasm assemblies generated from corrected reads noted significant influence on the contiguity, contrasting to those derived from raw reads except Pacbio/Miniasm assembly (**Table 3.2; Supp. Fig. 3.1**).

In more detail, from the corrected-reads Pacbio dataset, the Flye draft had the lowest number of contigs of 108 contigs. On the other hand, the Mecat software produced the longest N50 of 19.08 Mbp, the most extended contig of 26.89 Mbp, the lowest L50 of five contigs, and the longest single contig 26.89 Mbp. While, in the corrected-reads Nanopore dataset, Miniasm draft had the lowest number of contigs, 37 contigs. In comparison, the Flye draft had the longest contig of 26.98 Mbp, L50 of four contigs and N50 of 23.28 Mbp. On the other hand, Necat generated the longest N50 of 23.98 Mbp and the lowest L50 of four contigs. (**Table 3.2; Supp. Fig. 3.1**).

Dataset	Assembler	Assembly size	Contigs	N50	L50	Longest contig
Pacbio-Raw	Flye	195.30	348	3.89	13	15.90
	Mecat	198.91	111	19.08	5	26.89
	Miniasm	210.69	585	2.42	16	13.96
Pacbio-Corrected	Canu	200.36	114	10.13	6	24.57
	Flye	197.84	108	13.89	6	23.68
	Miniasm	208.46	894	1.06	36	8.78
	Wtdbg2	195.30	251	3.02	19	9.98
Nanopore-Raw	Flye	204.86	120	19.67	5	24.49
	Necat	200.88	43	23.80	4	26.82
	Miniasm	206.51	74	14.47	6	27.05
Nanopore-Corrected	Canu	199.07	38	16.10	5	26.71
	Flye	199.07	49	23.35	4	27.06
	Miniasm	199.41	37	19.02	5	24.17
	Wtdbg2	225.65	1040	2.96	15	13.82

Table 3.2: Preliminary assemblies statistics.

Assembler	Assembly size	Scaffolds	Unplaced contigs	N50	L50	Longest contig	Number of joins	Number of Gaps
HiRise/Nanopore	198.53	8	27	25.26	4	27.60	16	23
HiRise/Pacbio	198.08	8	60	25.37	4	27.49	42	42
Gala	198.54	10	0	25.77	4	27.71	-	0
Ref	198.65	10	614	22.86	5	26.17	-	26683

Table 3.3: Hi-C scaffolding, GALA draft and reference genome comparison.

3.2.2 Hi-C scaffolding

To generate a chromosome-level assembly, we carried out three Hi-C sequences libraries, which hold 164 X coverage, comprising 108.86 million reads (2x, 151 bp). The HiRise assembly pipeline was implemented on Flye drafts from Nanopore and Pacbio corrected-read datasets. The HiRise pipeline uses a modified version of SNAP mapping software to align the Hi-C dataset to the target genome, 81 % of the reads aligned to Nanopore target genome as well as 82.57% of the reads aligned to Pacbio target genome. Furthermore, 18.28 % of the reads aligned to the Nanopore draft and 17.75 % of the reads aligned to the Pacbio draft mapped on different contigs, providing a promising source of linking/scaffolding information (**Supp. Fig. 3.2**).

Next, the HiRise pipeline identified and trimmed four potential misassemblies in each draft. Running the scaffolding module on the Nanopore draft created 16 contig-joins and 35 final scaffolds, including eight pseudomolecules at the chromosome level and 27 organelles/unanchored contigs. As a result, the N50 increased to 25.26 Mbp, the L50 decreased to four contigs and the maximum scaffold length raised to 27.67 Mbp. On the other hand, the Pacbio draft scaffolding also produced eight pseudomolecules at the chromosome level from 42 contig-joins and 68 final scaffolds with N50 of 25.37 Mbp, L50 of four contigs and maximum scaffold length of 27.49 Mbp (**Table 3.3; Supp. Fig. 3.3**).

3.2.3 GALA assembly

In addition to Hi-C, we also implemented GALA on the eight corrected-read drafts and Necat/Mecat drafts as preliminary assemblies to produce gap-free chromosome-scale assembly. However, the GALA's MDM module detected various misassemblies in the preliminary drafts ranging between 0-203 (**Supp.Fig. 3.4**). Therefore, GALA split the potential misassembled contigs and applied the CCM module to misassembly-free drafts. GALA modelled the input into 13 independent linkage groups. Six groups represent six chromosomes, and two represent the organelle genomes. The Pacbio raw-reads/Flye draft helps to merge three remaining groups into a single linkage group representing the seventh chromosome. The last two linkage groups represent the arms of the last chromosome.

Nanopore Raw-reads and corrected-reads datasets were hired for the linkage group assembly module (LGAM) implementation. The LGAM employs Flye, Miniasm and Necat to

assemble each linkage group individually, producing gap-free chromosome-scale pseudomolecules for all linkage groups. Furthermore, the telomere motif analysis can identify the telomere repeats and the orientation of the two assembled chromosome arms. Therefore, we reassemble them as a unity linkage group, generating a telomere-to-telomere pseudomolecule. We polished the final assembly with a cycle of Nanopore reads (nanopolish) and nine cycles of Illumina reads (pilon) and 10X-reads mutually to enhance the assembly correctness. Finally, the new assembly consists of eight gapless telomere-to-telomere chromosomes and two organelle genomes of a total length of 198.54 Mbp, L50 of four chromosomes, and N50 of~ 25.77 Mbp (**Table 3.3**).

3.2.4 Comparison between GALA and Hi-C

We used two different chromosome-scale *de novo* assembly techniques independently, the Flye draft assembly beside the traditional Hi-C scaffolding technique as well as GALA. However, as shown above, both methods improved the final assembly and returned chromosome-scale scaffolds, but as we expect, both assemblies do not have the same quality assessment score. For example, Hi-C scaffolding delivers gaped scaffolds with 100 Ns between merged contigs. On the other hand, GALA generates gap-free telomere-to-telomere pseudomolecules. Therefore, consistent with using the Nanopore dataset for GALA assembly, we selected the Hi-C/Nanopore draft, with fewer scaffolds and gaps for a fair and balanced comparison between the two scaffolding methods. Then, we evaluate the contiguity, completeness and correctness in GALA assembly and Hi-C assembly.

The overall outcomes of the contiguity assessment matrices supported the superiority of the GALA assembly over Hi-C. Although the Hi-C and GALA drafts have very close N50 ~ 25 Mbp and L50 of four contigs, GALA assembly comprises ten scaffolds representing the perfect number of pseudomolecules in *C. hirsuta;* eight chromosomes and two organelle genomes. In comparison, the Hi-C assembly comprises eight chromosomes and 27 unanchored contigs. These results pointed that the GALA assembly achieved the perfect contiguity score (Table 3).

To assess the completeness of both drafts, we aligned 1440 conserved Embryophyta genes from the Benchmarking Universal Single-copy Orthologs (BUSCO) to both drafts independently. Busco detected 97.9% of the ortholog genes in GALA assembly, including 96.3 %, 1.6% and 0.2% of single-copy, duplicated and fragmented Busco's, respectively. On the contrary, only 88.4 % of Busco's genes were detected on the Hi-C draft, with a higher

percentage of fragmented and missed genes. Further *K-mers* completeness analyses confirmed the superiority of GALA assembly over Hi-C with a completeness score of 99.36 % and 95.52 % for GALA and Hi-C assemblies, respectively. Finally, GALA assembly achieved the gap-free scale for all chromosomes and organelle genomes. At the same time, the Hi-C draft included 23 gaps; seven gaps originated from Flye assembler and 16 from the HiRise joining process (**Table 3.4**).

	Gala	Hi-C	Ref-chromosomes	Ref-complete
Complete Buscos	1410	1273	1369	1406
Single-copy Buscos	1387	1249	1346	1384
Duplicated Buscos	23	24	23	22
Fragmented Buscos	3	27	7	6
Missing Buscos	27	95	64	28
K-mer Completeness	99.36	95.52	-	98.70
QV	44.14	27.41	-	41.85

Table 3.4: Busco and *Kmer* analysis statistics.

The correctness score was estimated through a *reference-guided* approach. First, we employed minimap2 to align both drafts to the published reference genome of *C. hirsute (ox).* Next, the paftools was hired to call the variants from the new drafts. The Hi-C draft generated 21.75 X variants higher than the GALA assembly, including 10 X of SNPs and 27.39 X of indels (**Supp.Fig.3.5**). Moreover, Merqury was leveraged to calculate the consensus quality score (QV) from *K-mer* analysis, representing a log scaled probability of consensus nucleotide error. Once again, GALA showed a superior score over Hi-C with a QV score of 44.14 for GALA against 27.41 for Hi-C draft. This means that the Hi-C draft has a base accuracy lower than 99.9 %, while the GALA draft accuracy is above 99.99 % (**Table 3.4**)

	***Cardamine hirsuta* (az)**	***Cardamine oligosperma***	***Cardamine resedifolia***
Gnome size (Mb)	201.67	181.17	240.68
Scaffolds	10	10	10
Gaps	0	0	3
N50	26.40	23.11	30.63
L50	4	4	4
Completeness	99.41	98.70	89.85
QV	57.78	51.73	39.95

Table 3.5: GALA assembly statistics.

3.2.5 Comparison between GALA and the reference genome.

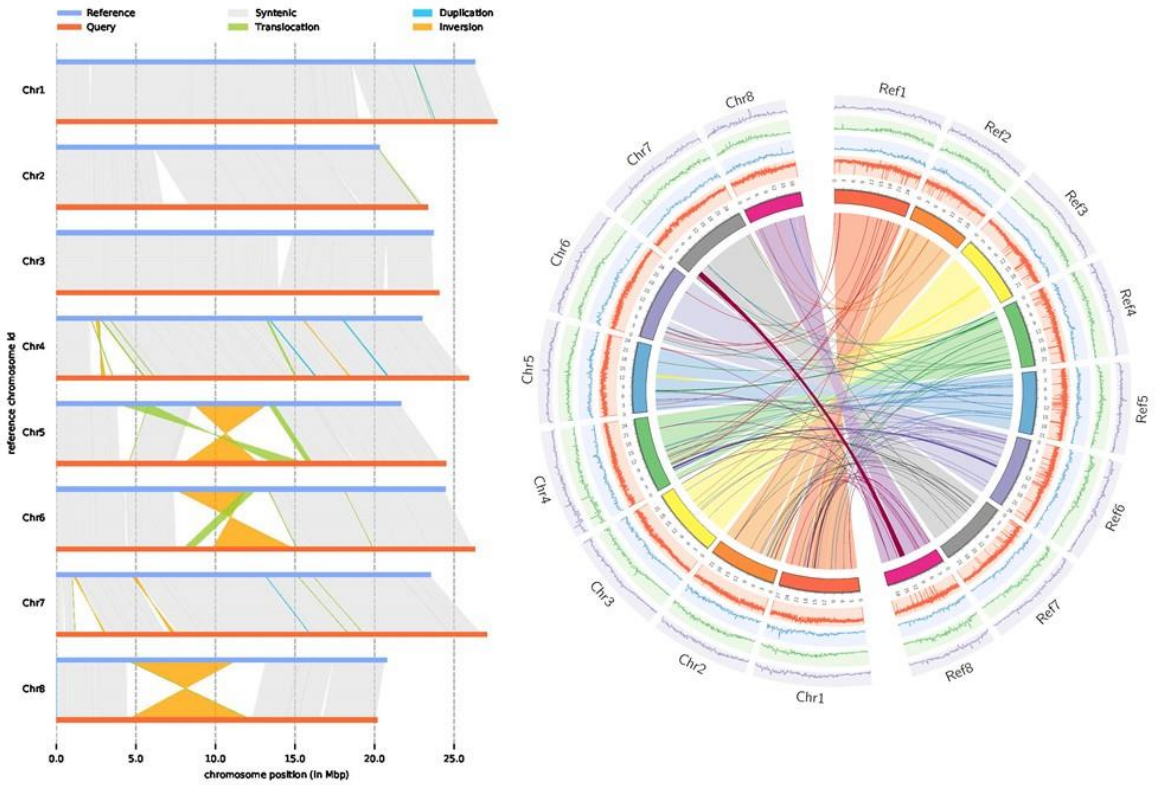
The resequencing and assembly of the *C. hirsuta* genome aim to enhance the published genome assembly quality and release the second version of the genome. Accordingly, we compared the GALA draft and the reference genome in terms of contiguity, completeness and correctness. The reference genome incorporates 614 unanchored contigs/scaffolds over the eight chromosomes and organelle genomes. Additionally, the chromosome sizes of the GALA draft are more extended than the reference chromosomes, except for chromosome 8. The reference genome N50 is ~ 3Mbp lower than GALA's N50 and the L50 of five contigs in the reference genome is bigger than GALA's L50. Therefore obviously, the GALA assembly contiguity is dominating the reference genome. The completeness evaluation showed that the reference genome comprises ~ 7.5 Mbp of unknown Ns, forming thousands of gaps (**Table 3.3**). In addition, the reference genome *K-mers* completeness score is lower than GALA's score, while Busco's results are comparable with three more fragmented and one more missed Busco's in the reference genome. The *K-mers* completeness analysis and reference genome gaps bolstered the GALA assembly completeness over the reference genome (**Table 3.4**).

We used reference-free evaluation approaches to compare the correctness of the GALA assembly and the reference genome in three dimensions; base accuracy, assembly collapsing and mis-joint scaffolds. To estimate the base accuracy score, we leveraged bwa-mem to align 195X of the original Illumina reads dataset - used in reference genome construction - to the reference genome and GALA assembly. The alignment's statistical summary emphasized the GALA assembly's supremacy by mapping 443205 reads more than the mapped reads to the reference genome. As well as, the GALA assembly has only 65.2%, 13.9%, and 25.2% of the total number of the reference genome alignment mismatches, insertions and deletions, respectively. The BCFtools also appropriated to call the variants from both assemblies, showing a significant fall in the called SNPs and indels in the GALA draft compared to the reference genome (**Supp. Table 3.1**). Finally, GALA gained a 2.28 QV score higher than the reference genome in the *K-mer* analysis.

While the variant calling is helpful for the base accuracy assessment, it does not reflect the complete architecture of the assembly repeats. So, we performed a read depth profile analysis for the reference genome and GALA assembly to assess the collapsed regions in both of them based on the sequencing depth excess. The reference genome showed 239 collapsed regions of a total length of 739.5 Kbp, while GALA has only 46 collapsed regions of a total

length of 456.2 Kbp. Thus, the majority of collapsed regions in the reference genome were solved in GALA assembly. At the same time, the blastn results showed that many collapsed regions in chromosomes 1,3,4,6 and 8 are false collapses from chloroplast-like sequences. Notably, the chloroplast genome has coverage over 4000 X. Furthermore, the annotation of the other collapsed regions in the GALA assembly showed that the LTR reach loci and telomeric motifs are the source of collapse, except for the collapsed region in Chr7 (**Supp. Fig. 3.6** and **Supp. Fig. 3.7**).

Finally, we harnessed Syri to reveal the structure variants (SV) between GALA assembly and the reference genome. The analysis reported 27 inversion translocation events and 36 translocation events between the sister chromosomes, comprising 25 events > 10 Kbp and 29 events > 1 Kbp. Furthermore, 72 inversion translocation and 91 translocation events were reported in non-sister chromosomes, holding 14 events > 10 Kbp and 105 events > 1 Kbp (**Supp. Table 3.2**). To confirm these SVs, first, Minimap2 aligned the Nanopore corrected reads to the reference genome; then, we traced the mapped reads in the breakpoints of the 14 translocation events >10 Kbp. We found that the reads in the breakpoints partially mapped to one chromosome and the overhanging part mapped to another chromosome, confirming the translocation event. Also, the Hi-C drafts and preliminary assemblies promoted the GALA assembly architecture in all the translocation events > 10 Kbp. Moreover, 124 duplications and 128 inversion duplication events were reported in non-sister chromosomes with only three duplication events > 10 Kbp (**Fig. 3.2** and **Supp. Fig. 3.8**).



a **b**

Figure 3.2: Comparison between *C.hirsuta (ox)* reference genome and GALA assembly. a) The synteny and intra-chromosomal variants between the reference *C.hirsuta* genome and the GALA assembly. b) circus plot demonstrating the TE-load (purple), Copia-LTRs (green), Gypsy-LTRs (blue) and GC content(orange) histograms, beside the inter-chromosomal variants between the reference *C.hirsuta* genome (Ref) and the GALA assembly (Chr).

3.2.6 *C. hirsuta (az)* genome assembly

We assembled *Cardamine hirsuta* strain Azores by combinatory analysis of two Pacbio datasets. The first dataset derived from SR/Sequel platform, consisted of 3,581,988 reads, containing 25,394,233,997 bases with a read N50 of 11,058 bp and coverage depth of 126 folds. The second dataset was a Hifi dataset, consisting of 2,665,432 reads, containing 27,825,920,767 bases with a read N50 of 10,254 bp and coverage depth of 139 folds (**Supp. Table 3.3 and Supp. Fig. 3.9**).

We assembled each dataset using two assemblers, Canu and Miniasm, for the RS/sequel dataset, as well as Canu and Hifiasm for the Hifi dataset assembly. The overall results showed high contiguity in the Hifi drafts compared to the RS/sequel dataset drafts in terms of N50 and

L50. The Hifiasm assembled Hifi reads into contigs with N50 of 24.61 Mbp and L50 of 5 contigs, while the Canu/Hifi draft has N50 of 11.40 Mbp and L50 of 8 contigs. On the other hand, for the RS/sequel dataset, the Canu draft has N50 of 6.94 Mbp and L50 of 9 contigs. In contrast, the Miniasm draft comprises N50 of 4.12 Mbp and L50 of 15 contigs.

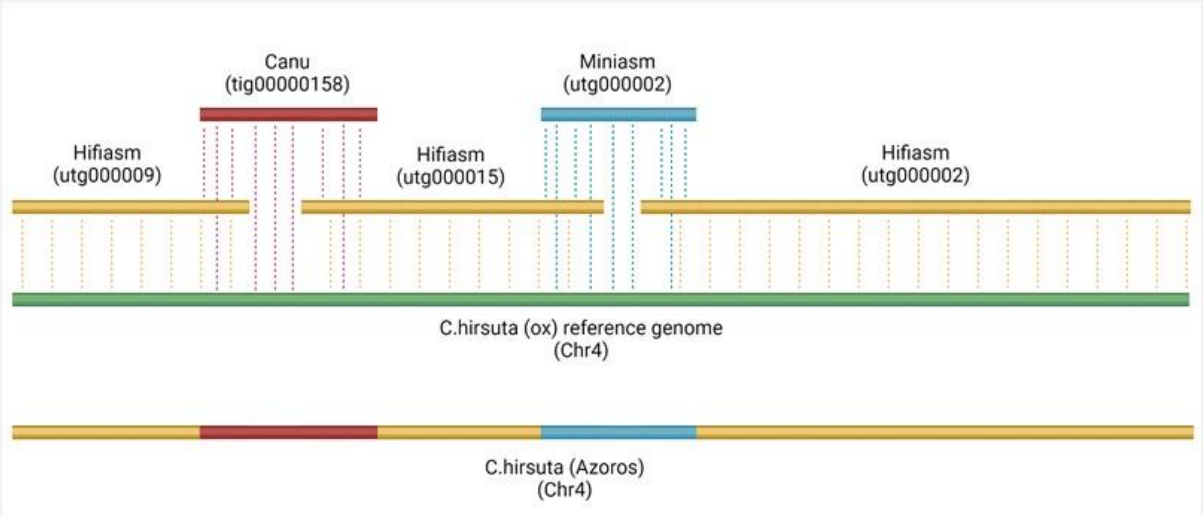


Figure 3.3: demonstration of the complementary assembly of *C. hirsuta* chromosome.

After a filtration step for Hifi assemblies, we implemented GALA on the four drafts. The MDM module detected various misassemblies in the preliminary drafts ranging from 2 in the Hifi drafts to 45 in the Miniasm draft. After that, we applied the CCM module on the four misassembly-free drafts and the *C.hirsuta* (ox) genome. GALA modelled the *C. hirsuta (*Azores) drafts into ten independent linkage groups, seven represent complete chromosomes and the remaining three represent the last chromosome (Chr 4). The Hifiasm draft comprised one contig for each linkage group. To assemble Chr4, we followed an integrative assembly approach between the Hifiasm draft and two contigs from the RS/sequel dataset drafts. The contig tig00000158 of length 1.035 Mbp from Canu draft can fill the gap between (utg000009 and utg000015). Furthermore, the contig utg000068 of length 1.42 Mbp from Miniasm assembly can fill the gap between (utg000015 and utg000002) (**Fig. 3.3**). To assemble the organelle genomes, we used the LGAM module to separate the Hifi reads mapped to *C. hirsuta* (ox) organelle genomes and assemble 50X of the separated reads into two single-pseudomolecules complete genomes. Finally, we accomplished two rounds of polishing using Hifi read to improve the final draft accuracy.

The final assembly of *C.hirsuta* (Azores) comprises 201,677,004 bp, assembled into eight gapless telomere-to-telomere chromosomes and two organelle genomes with N50 of 26.40 Mbp and L50 of four pseudomolecules (**Table 3.5**). The Busco analysis showed a 96.6 % completeness percentage with 94.9%, 1.7% 1.0% and 2.4% of complete single-copy, duplicated, fragmented and missing Busco genes, respectively. In addition, 99.24% of 41X Illumina reads dataset mapped to the genome with a mismatch rate of 2.07e-03 and 4.27e-05 of indels (**Supp. Table. 3.4**). Consequently, BCFtools called only 618 variants from this dataset, including 368 SNPs and 313 indels. The depth profile showed only nine collapsed regions with a total length of 98.3 Kbp (**Supp. Fig. 3.10**). Finally, the *K-mer* analysis reported a 99.419 % completeness score and QV score of 57.786 (**Table 3.5**).

3.2.7 *C. oligosperma* genome assembly

We generated 59 Gb of *C. oligosperma* Hifi long reads and 45 Gb of Pacbio RS reads, comprising 164X and 124X, respectively (**Supp. Table 3.3** and **Supp. Fig. 3.11**). The Canu and Flye tools were used to assemble both datasets individually; also, we used the Hifiasm to assemble the Hifi dataset. While the N50 and L50 are comparable in drafts derived from the same tool, the Hifiasm draft showed the best N50 of 13.15 Mbp. In addition, we noticed that Canu/Hifi and Hifiasm drafts have a significant difference in the assembled genome size. While the genome size ranged between 176.5 Mbp and 181.8 Mbp in the Flye drafts and Canu/RS draft, the Canu/Hifi draft has a total length of 259.4 Mbp and the Hifiasm draft has 296.7 Mbp.

Three Hi-C sequencing libraries holding 178X coverage were constructed. We implemented the HiRise pipeline on Flye/RS draft, which has the lowest number of contigs. As a result, 85.20 % of the reads aligned to the target genome, including 20.52 % mapped on different contigs, as a source of linking/scaffolding information. The HiRise scaffolding module created 97contig-joins and 25 final scaffolds, including eight gapped pseudomolecules at the chromosome-scale and 17 organelles/unanchored contigs. Consequently, the N50 increased to 22.80 Mbp, the L50 decreased to four contigs and the maximum scaffold length reached 24.22 Mbp.

Finally, we executed GALA on the four drafts derived from Pacbio RS and Hifi datasets to achieve gap-free chromosome-scale assembly. The MDM module identified many misassemblies ranging from four in the Canu/Hifi draft to 21 in the Canu/Rs draft. The CCM module modelled the error-free drafts to nine linkage groups. Seven of them represent individual chromosomes and the last two in chromosome arm size. So, we merge them into one

linkage group. Then, we confirmed our scaffolding hypothesis by mapping the Hi-C draft against the Canu/Hifi draft. Lastly, the integrative assembly of each linkage group using both datasets generated a gap-free chromosome-scale assembly for all chromosomes. Then, we followed the previously described pipeline in *C.hirsuta* (Azores) assembly to assemble the organelle genomes and polish the final assembly draft.

The *C. oligosperma* genome assembly holds eight gapless telomere-to-telomere pseudomolecules and two complete organelle genomes, comprising 181,172,011 bp with N50 of 23.11 Mbp and L50 of four chromosomes (**Table 3.5**). The GALA assembly is superior to the HiRise assembly in terms of genome size, N50 and gaps. The Busco completeness analysis identified 96.7 % Busco's, including 95.2% complete single-copy and 2.5% missing orthologs. Moreover, aligning 179X of Illumina reads to the final draft recorded a 96.58% mapping ratio with a mismatch rate of 1.766e-03 and indel rate of 1.69e-04 (**Supp. Table. 3.4**). In addition, the depth profile recorded only 19 collapsed regions > 1000 bp with a total length of 322.3 Kbp (**Supp. Fig. 3.12**). Finally, the *K-mer* analysis reported a 98.703 % completeness score and QV score of 51.733 (**Table 3.5**).

3.2.8 *C. resedifolia* genome assembly

We carried out the *C. resedifolia* genome assembly from 56× coverage of high-quality Hifi sequences with reads N50 of 12.7 Kbp and 110X Hi-C sequences (**Supp. Table 3.3** and **Supp. Fig. 3.13**). First, three preliminary assemblies were constructed from the Hifi dataset using Canu, Hifiasm and Flye assemblers, yielding initial contig sets with N50 of 15.3 Mbp, 13.6 Mbp and 3.50 Mbp for Hifiasm, Canu and Flye drafts, respectively. The total length of this preliminary assembly ranged between 236.3 Mbp in Flye draft and 285.3 Mbp in the Canu draft. While Flye draft has the lowest number of contigs, Hifiasm and Canu drafts have better N50 and L50. So, we used the Hi-C sequences to enhance the contiguity of the Flye draft. Therefore, we organized the contigs of the Flye draft into a scaffold-level genome assembly using the Juicer assembly platform, which employed bwa-mem to align the Hi-C reads to the target genome. Then, Juicer executed two iterations of misjoin detection and assembly module. Consequently, the N50 increased to 26.6 Mbp and the L50 decreased to four.

To achieve gap-free chromosome-scale assembly, we ran GALA on a *reference-guided* data separation mode using the three preliminary assemblies and the Hi-C draft as a reference. First, the CCM modelled the input drafts into nine scaffolding groups, representing seven complete chromosomes and two chromosome arms. Then, we merged the two chromosome

arms groups in one linkage group and executed the LGAM module using Canu and Hifiasm. Finally, we successfully assembled five gap-free complete chromosomes. In contrast, the last three chromosomes were assembled to a gap-free chromosome-arm scale and scaffolded to chromosome-scale pseudomolecules using telomere motif analysis. Eventually, we assembled the organelle genomes and polished the final assembly draft to increase the base accuracy using the described pipeline in *C.hirsuta* (Azores) assembly.

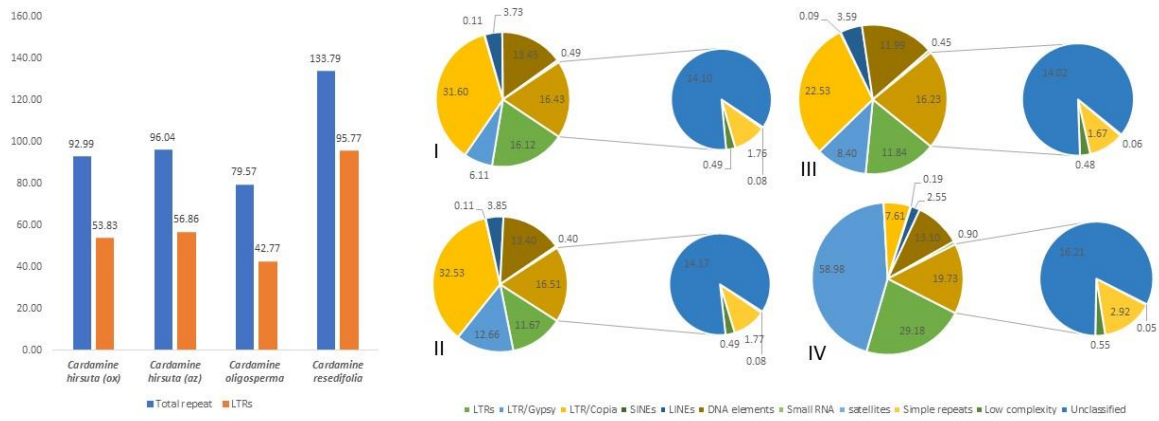
In summary, the GALA assembly for the *C. resedifolia* genome consisted of 240.685 Mbp, representing two complete organelle genomes, five gapless telomere-to-telomere chromosomes and three chromosomes with only a single centromeric gap (Chr1, Chr5 and Chr8). The *K-mer* analysis revealed a 90% completeness ratio and a 39.95 consensus score (**Table 3.5**). Further Busco completeness analysis identified 96.6% of the Busco genes and reported 2.4% missing genes (**Supp. Table. 3.4**). Ultimately, the depth profile showed 29 collapsing events in the assembled genome > 1000 bp with a total length of 413.5 Kbp (**Supp. Fig. 3.14**).

3.2.9 Comparison between the 3 species and the new ox genome

Transposon element (TE) richness is the primary element of plant genome size variation. Hence, we dissected the repeated-elements load and categories in the assembled genomes. As expected, consistent with the genome size, the *C. resedifolia* has the heaviest repeat-load of 133,799,245 bp, forming 55.59% of the total genome size. In contrast, the *C. oligosperma* has the lightest repeat load of 79,5 Mbp, representing 43,92% of the assembled sequences. The *C. hirsuta (az)* has ~ a 3 Mbp repeat-load more than *C. hirsuta (ox)*, which holds ~ 93 Mbp repeated sequences. These repeated elements include ~ 7 % of unclassified elements, < 2% of long interspersed nuclear elements (LINEs) and < 7% of DNA elements in all assembled genomes. The long terminal repeat (LTR) retrotransposon is the most abundant component of the repeat elements, representing 39.79 %, 28.19 %, 27.11 % and 23.61 % of the repeat-load of the *C. resedifolia, C. hirsuta (az), C. hirsuta (ox)* and *C. oligosperma,* respectively. Among them, Copia and Gypsy super-families dominated the LTR load in the four assembled genomes (**Table 3.6; Fig 3.4**).

	Cardamine hirsuta (ox)		*Cardamine hirsuta (az)*		*Cardamine oligosperma*		*Cardamine resedifolia*	
	Length	**Percentage**	**Length**	**Percentage**	**Length**	**Percentage**	**Length**	**Percentage**
Total repeat	92.99	46.84	96.04	47.62	79.57	43.92	133.79	55.59
SINEs	0.11	0.06	0.11	0.06	0.09	0.06	0.19	0.08
LINEs	3.73	1.94	3.85	1.92	3.59	1.98	2.55	1.06
DNA elements	13.45	6.78	13.40	6.64	11.99	6.62	13.10	5.45
LTRs	53.83	27.11	56.86	28.19	42.77	23.61	95.77	39.79
Small RNA	0.49	0.25	0.40	0.20	0.45	0.25	0.90	0.38
satellites	0.08	0.04	0.08	0.04	0.06	0.04	0.05	0.02
Simple repeats	1.76	0.89	1.77	0.88	1.67	0.92	2.92	1.21
Low complexity	0.49	0.25	0.49	0.25	0.48	0.27	0.55	0.23
Unclassified	14.10	7.10	14.17	7.03	14.02	7.74	16.21	6.73

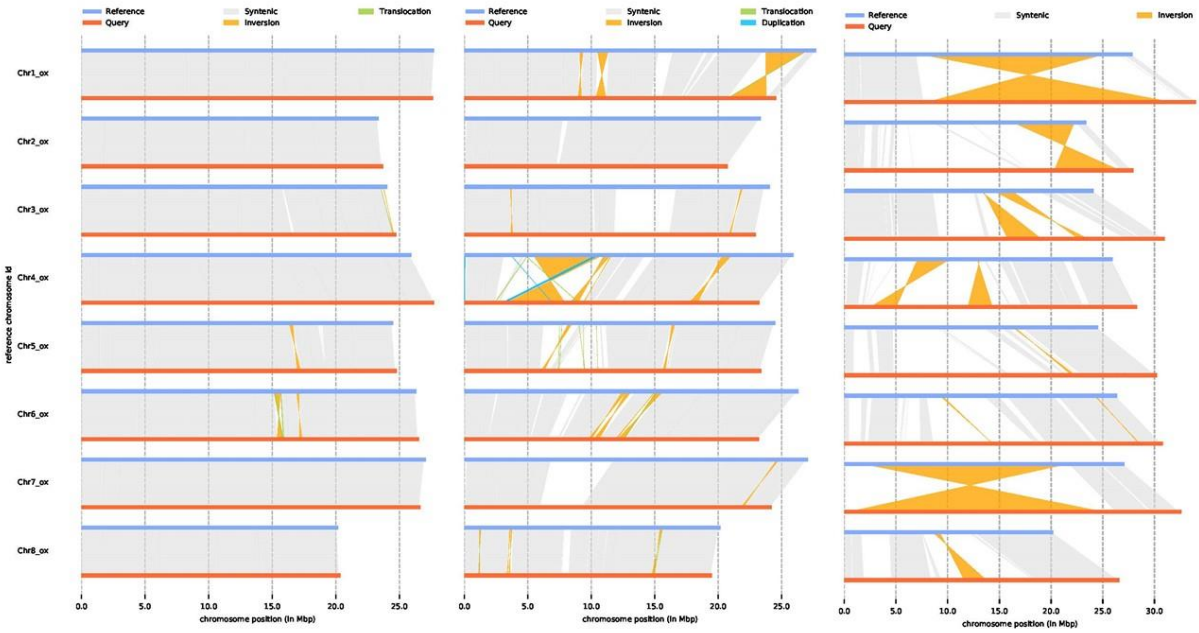
Table 3.6: Repeat load of GALA assemblies.



a **b**

Figure 3.4: Repeat analysis**:** a) comparing the total TE-load and LTR-load among the four assemblies under investigation. b) The TE composition for I) *C. hirsuta* (ox), II) *C. hirsuta (*az*)*, III) *C. oligosperma* and IV) *C. resedifolia.*

Further, we accompanied a comparative genomic approach to infer the karyotype differences and collinearity between the constructed chromosome-scale genomes in this study. Therefore, we used the new *C. hirsuta (ox)* genome assembly as a reference in a whole-genome pairwise alignment against *C. hirsuta (az), C. oligosperma* and *C. resedifolia.* Moreover, *C. oligosperma* was used as a reference against *C. resedifolia.* However, while 81% of the *C. hirsuta (az)* genome was mapped to the reference, only 36% of the *C. oligosperma* was mapped (mapping quality >19). On the other hand, 7.85 % and 7.05% of the *C. resedifolia* genome mapped to *C. hirsuta (ox)* and *C. oligosperma*, respectively. Then, we employed Syri to identify the synteny blocks and rearrangement events. As a result, the *C. hirsuta (az)* showed overall collinearity agreement with *C. hirsuta (ox)* with only four intra-chromosomal translocations ranging from 51.9 Kbp to 17 Kbp in chromosomes 6 and 3. We also detected four inversion events of lengths 425.6 Kbp, 282.6 Kbp, 200.9 Kbp and 34.6 Kbp in chromosomes 6, 5, 6 and 3, respectively. Besides, 98.66 mapping identity percentage, 1.33 Mbp of SNPs and 8.25 Mbp of indels (**Fig. 3.5**).



a **b** **c**

Figure 3.5: The synteny information and intra-chromosomal variants between *C. hirsuta* (ox) genome and a) *C. hirsuta (*az*),* b) *C. oligosperma* and c) *C. resedifolia.*

The *C. oligosperma* genome showed nine intra-chromosomal translocations ranging between 37.6 Kbp and 10.4 Kbp and three inter-chromosomal translocations ranged between 17.5 Kbp and 10.3 Kbp on chromosomes 5/4 and 7/8. However, 20 inversion events shaped the main karyotype differences between *C. oligosperma and C. hirsuta.* The most extended two events were in chromosomes 4 and 1 of length 4.63 Mbp and 3.04 Mbp, respectively, while the remaining events ranged between 825.2 Kbp and 50.4 Kbp, spreading on almost all chromosomes except chromosome 2 (**Fig.3.5**). Even though the low fraction of the *C. resedifolia* mapped sequences to *C. oligosperma and C. hirsuta*, we can detect several karyotype differences in the aligned fraction. For example, a 10 Kbp inter-chromosomal translocation was identified in chromosome 2 with *C. hirsuta* and *C. oligosperma* chromosomes 4 and 5, respectively. Also, we can identify 11 inversion events with *C. hirsuta* and 10 inversion events with *C. oligosperma* with two massive inversions of length 24.5 Mbp and 23 Mbp in chromosomes 7 and 1, respectively (Fig.5; Supp. Fig 15).

3.3 Discussion

The recent progress in sequencing technologies promotes the availability of heterogeneous sequencing data for the same species. However, incorporating these different data sources into a single pipeline to produce a gap-free assembly implies a computational challenge. In this study, we used heterogeneous datasets from Pacbio, Nanopore and Hi-C to generate a high-quality gap-free telomere-to-telomere assembly for *C. hirsuta (ox)*, *C. hirsuta (az)*, and *C. resedifolia*, in addition to the first assembly of *C. oligosperma* genome. The consensus quality (QV) score of *C. hirsuta* and *C. oligosperma* ranged between 44 and 57, while the golden standard plant reference genome TAIR-10 has a QV score of 41.4 (Naish et al. 2021). These genomes will be a valuable source for biological and comparative developmental studies in the family Brassicaceae.

We demonstrated the strength of the GALA assembler to provide the second draft of the *C. hirsuta (ox)* genome assembled from telomere-to-telomere T2T. While the preliminary assemblies are fragmented and admitted inter-chromosomal misassemblies, GALA resolved the misassemblies and the LGAM module revealed the gap-free chromosome-scale assembly. Remarkably, the new draft comprises 198.5 Mbp in eight chromosomes. In comparison, the reference genome has only 183 Mbp in the eight chromosomes and 15 Mbp unanchored fragments. Moreover, the new genome assembly resolved 14 inter-chromosomal translocations of length bigger than 10 Kbp. In the reference genome, 78.9 Mbp of the TE sequences were reported (Gan et al. 2016). In contrast, we can annotate 92.99 Mbp of TE elements in our new assembly. Also, the new assembly showed improved completeness and correctness scores over the previously published genome. Although the new genome spanned the repetitive regions and assembled the eight centromeres and solved several collapses in the reference genome, we still have ~ 450 Kbp of collapsed repeats in the assembly. These collapses are abundant in highly repetitive regions even with long-read technology (Michael et al. 2018; Kolmogorov et al. 2019). However, the assembly collapsing and expansion can be handled through the synergy of experimental and computational efforts in downstream polishing analysis (Miga et al. 2019; Bzikadze and Pevzner 2020). On the other hand, the *C. hirsuta (az)* genome revealed better completeness and correctness scores with an overall collinearity agreement with the oxford genome beside a comparable TE-load and lowered repeat collapsing ratio.

A highly fragmented *C. resedifolia* genome assembly was reported in 2020, holding 42,839 contigs and 192.8 Mbp (Rellstab et al. 2020). In contrast, we reported a high contiguity chromosome-scale assembly of the *C. resedifolia* genome, acquiring 240 Mbp and eight chromosomes with only three gaps in Chromosomes 1, 5 and 8. In addition, our assembly has

a highly significant N50 of 30 Mbp, while the previously reported assembly has an N50 of 48.5 Kbp. Although the overall QV score is lower than the *A. thaliana* TAIR-10 score, further investigation showed that the five gapless telomere-to-telomere chromosomes have a QV score > 40. Finally, we reported the T2T *C. oligosperma* reference genome assembly completely free of gaps beside a QV reaches 51.7 and N50 of 23.1 Mbp. To the best of our knowledge, this is the first assembly of the *C. oligosperma* genome.

The availability of gapless chromosome-scale reference genomes facilitates the accessibility of novel biologically relevant sequence variation (Alkan et al. 2011; Mantere et al. 2019). For example, previous research reported an inability to access and resolve complex genomic structures and rearrangements in gaped and fragmented reference genomes (Michael et al. 2018; Tyson et al. 2018; Audano et al. 2019). Furthermore, the gap-free assemblies with high consensus scores present a potential opportunity to identify unsolved functionally effective polymorphisms in different taxa and populations (Jayakodi et al. 2020; Barchi et al. 2021; Ma et al. 2021). With the affordability of heterogeneous sequencing resources and proper consolidating computational frameworks, e.g., GALA, we can access complex genomic regions, addressing several biological queries beyond laborious experimental strategies.

The genome size expansion or contraction mechanisms correlated with the transposable elements, intergenic and intronic activities (Hu et al. 2011). Our findings inferred that the TE-load, especially the LTR families Copia and Gypsy, are the major components of genome size differences between the assembled species. For example, the *Arabidopsis thaliana* genome holds 17 % TEs (Buisine et al. 2008), while all genomes under this study containing TEs ranged between 43 % and 55 %. In contrast, a 440 Mbp tetraploid *C. enshiensis* genome comprising a remarkably higher TE-load of 61.4 % (Huang et al. 2021). However, a deeper understanding of the distribution and mutational events on TE and LTR retrotransposons is essential for a comprehensive understanding of the processes behind collinearity and genome size changes.

3.4 Methods

3.4.1 Plant datasets:

3.4.1.1 *Cardamine hirsuta* (ox)

This study used 125X Pacbio/RS, 88X Nanopore dataset, 164X Chicago reads, 195X Illumina reads and 10X genomics datasets.

3.4.1.2 *Cardamine hirsuta* (az)

This study used 126X Pacbio/RS, 139X Pacbio/Hifi dataset and 41X Illumina reads datasets.

3.4.1.3 *Cardamine oligosperma*

This study used 93 Pacbio/RS, 147X Pacbio/Hifi dataset, 178X Chicago reads and 185X Illumina reads datasets.

3.4.1.4 *Cardamine resedifolia*

This study used 55X Pacbio/Hifi dataset, 110X Chicago reads and 130X Illumina reads datasets.

3.4.2 Preliminary genome assembly

The Pacbio and Nanopore raw reads were corrected separately using Canu 1.8 (Koren et al. 2017) self-correction and trimming module. Then we hired Canu, Flye (Kolmogorov et al. 2019), Miniasm (Li 2016), Wtdbg2 (Ruan and Li 2020) and Necat/Mecat (Xiao et al. 2017) assemblers with the default parameters to assemble each dataset. Two rounds of racon/minimap (Vaser et al. 2017) consensus correction module were performed to increase the base accuracy of Minimap drafts. Finally, we employed Quiver (Chin et al. 2013) to polish the Pacbio preliminary drafts and Nanopolish (Simpson et al. 2017) to polish the Nanopore assemblies. The Hifi datasets were assembled using HiCanu(Nurk et al. 2020b), Flye (Kolmogorov et al. 2019) and Hifiasm (Cheng et al. 2021) with the default parameters.

3.4.3 *Cardamine hirsuta* (ox) gap-free chromosome-scale assembly

Ten preliminary assemblies generated from the corrected reads of Pacbio and Nanopore datasets were chosen for GALA implementation. We used 'Minimap2 -x asm5' to map the preliminary assemblies against each other and "Minimap2 -ax map-ont' to map the Nanopore reads to the misassembly-free drafts. First, GALA classified them into 13 linkage groups, integrating the Pacbio raw-reads/Flye draft merge them into 11 linkage groups. Next, the telomere motif analysis merges two chromosome-arm scale linkage groups into a single linkage group. Finally, we used Flye, Miniasm and Necat assemblers with the Nanopore dataset to run LGAM.

We mapped the raw Nanopore reads to the final assembly using 'bwa-mem -x ont2d' (Li and Durbin 2009). Then we used Nanopolish to call the base consensus. Furthermore, the Supernova assembler (Weisenfeld et al. 2017) was used to assemble a 10x genomics dataset. Then, we used Minimap2 to map the assembled 10x data to the final genome. Finally, we used ten rounds of mutual polishing using 10x genomics assembly and Pilon an Illumina reads polisher software (Walker et al. 2014).

3.4.4 *Cardamine hirsuta* (Az) chromosome-scale assembly

We filtered out the small contigs (<100Kbp) of the Canu/hifi and Hifiasm drafts. Then we used the filtered drafts beside Canu and Miniasm contigs derived from the Pacbio RS platform to implement GALA on the four drafts. Then, we employed Minimap2 to map the contigs tig00000158 and utg000068 from Canu and Miniasm drafts against Chr4 linkage groups from the Hifiasm draft, merging the overlaps among them closed the gaps in Chr4. Finally, we run two rounds of Hifi reads polishing using ‘Minimap2 -x asm20’ as an aligner and unpublished polisher we developed.

3.4.5 *Cardamine oligosperma* chromosome-scale assembly

After filtering out the small contigs (<100Kbp) of the Canu/hifi and Hifiasm drafts, we implemented GALA on five drafts; Canu and Flye drafts derived from the Pacbio RS platform and the three Hifi dataset assemblies. GALA classified them into nine chromosomal linkage groups. The telomere motif analysis merges two chromosome-arm scale linkage groups into a single linkage group. Finally, we used the Hifiasm assembler to run the LGAM module with the Hifi database. Two rounds of Hifi reads polishing were executed using the previously mentioned pipeline.

3.4.6 *Cardamine resedifolia* chromosome-scale assembly

Firstly, we applied the Juicer (Durand et al. 2016) Hi-C assembly platform on the Flye draft derived from the Hifi dataset. Briefly, the pipeline starts with restriction enzyme (DpnII) sites detection on the subject genome. Then, it employed bwa-mem to map the Hi-C Chicago reads against the draft assembly. Finally, we ran the default parameters of the mis-joint detection and scaffolding modules.

GALA was executed, using the Hi-C draft as a reference with the Hifi drafts as preliminary assemblies. Finally, Canu and Hifiasm were used to implement LGAM. Then, two rounds of Hifi reads polishing were executed using the previously mentioned pipeline.

3.4.7 Organelle genome assembly

We employed the fastq-sample (https://github.com/dcjones/fastq-tools) to extract 50X reads of the chloroplast and mitochondrial reads linkage groups. Then we used the LGAM to assemble both of them individually.

3.4.8 Assembly quality control

The contiguity assessment was assessed using our script (https://github.com/mawad89/assembly_stats). The assembly completeness was assessed using the Benchmarking Universal Single-Copy Orthologs (BUSCO V.3) (Seppey et al. 2019) with the dataset Embryophyta_odb9. Then, for correctness assessment, we used bwa-mem to map Illumine datasets for the assembled species to the final assembly. Then, we collect the mapping statistics from samtools-stats (Li et al. 2009). Finally, we call the variants and collect the variant calling statistics using BCFtools (Li 2011).

For *de novo* completeness and correctness evaluation, the Meryl software (Miller et al. 2008) was hired to build a *K-mer* database. Then, we used Merqury (Rhie et al. 2020), a genome assembly *K-mer* based evaluation software, to evaluate the *K-mer* completeness score and the base consensus score (QV).

For collapsing evaluation, we mapped the primary dataset used for the final assembly of each genome using Minimap2. Then, we collect the depth information using samtools depth and mark all the regions with depth= average depth*2 as collapsed regions. Finally, we used python to plot the average value of window size 5000 bp.

3.4.9 Repetitive element prediction and annotation

We used the RepeatModeler (version open-2.0.1) for *de novo* identification of repetitive sequences in the assembled genomes. Next, the classified repeats from all species were combined to the Brassicaceae repeat library obtained from the Repbase database (Bao et al. 2015), concluding a final repeat library. In a final step, we used this repeat library with RepeatMasker (version open-4.0.9) to annotate transposable elements TE and other repetitive sequences in all assembled genomes.

Furthermore, the LTR retrotransposons were *de novo* predicted for each genome using LTRharvest (Ellinghaus et al. 2008). Then, we annotated the identified LTR using two complementary procedures to reduce false positives. First, LTRdigest (Steinbiss et al. 2009) was used to identify candidate retrotransposons by searching for known protein domains from the Pfam protein family database. Second, the remaining unknown candidates were then classified using RepeatClassifier module.

Pfam family	**LTR superfamily**
gag_pre-integrs	**Copia**
RVT_2	**Copia**
DUF4219	**Copia**
Retrotran_gag_2	**Copia**
Retrotran_gag_3	**Copia**
Chromo	**Gypsy**
Transposase_28	**Gypsy**
zf-CCHC_4	**Gypsy**
Asp_protease_2	**Gypsy**
Asp_protease	**Gypsy**
RVT_3	**Gypsy**
Ty3_capsid	**Gypsy**
zf-RVT	**Gypsy**
RVP_2	**Gypsy**
ATHILA	**Gypsy**
gag-asp_proteas	**Gypsy**
RVT_1	**Gypsy**

3.4.10 karyotype and collinearity analysis

The ‘Minimap2 -ax asm5’ was hired to map the final assembly of *C. hirsuta (ox*) against the published reference genome. Further, we mapped the *C. hirsuta (az), C. oligosperma* and *C. resedifolia* against the new *C. hirsuta (ox)* genome derived from GALA. Moreover, *C.resedifolia was* aligned against *C.oligosperma.* Finally, we used Syri (Goel et al.

2019) with the options '--no-chrmatch --all --nosnp' to identify the synteny blocks, collinearity, and karyotype differences.

References

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363-376.

Arabidopsis Genome I. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796-815.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**: 663-675 e619.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.

Barchi L, Rabanus-Wallace MT, Prohens J, Toppino L, Padmarasu S, Portis E, Rotino GL, Stein N, Lanteri S, Giuliano G. 2021. Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. *Plant J* **107**: 579-596.

Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, Genete M, Berrabah W, Chevre AM, Delourme R et al. 2018. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants* **4**: 879-887.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* **53**: 474-485.

Buisine N, Quesneville H, Colot V. 2008. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics* **91**: 467-475.

Bzikadze AV, Pevzner PA. 2020. Automated assembly of centromeres from ultra-long error-prone reads. *Nat Biotechnol* **38**: 1309-1316.

Chang C, Bowman JL, Meyerowitz EM. 2016. Field Guide to Plant Model Systems. *Cell* **167**: 325-339.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**: 170-175.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563-569.

Domanska D, Kanduri C, Simovski B, Sandve GK. 2018. Mind the gaps: overlooking inaccessible regions confounds statistical testing in genome analysis. *BMC Bioinformatics* **19**: 481.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**: 95-98.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**: 18.

Gan X, Hay A, Kwantes M, Haberer G, Hallab A, Ioio RD, Hofhuis H, Pieper B, Cartolano M, Neumann U et al. 2016. The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. *Nat Plants* **2**: 16167.

Goel M, Sun H, Jiao WB, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**: 277.

Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK, Jr., Maiti R, Chan AP, Yu C, Farzad M, Wu D et al. 2005. Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol* **3**: 7.

Hay AS, Pieper B, Cooke E, Mandakova T, Cartolano M, Tattersall AD, Ioio RD, McGowan SJ, Barkoulas M, Galinha C et al. 2014. Cardamine hirsuta: a versatile genetic system for comparative studies. *Plant J* **78**: 1-15.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* **43**: 476-481.

Huang C, Ying H, Yang X, Gao Y, Li T, Wu B, Ren M, Zhang Z, Ding J, Gao J et al. 2021. The Cardamine enshiensis genome reveals whole genome duplication and insight into selenium hyperaccumulation and tolerance. *Cell Discov* **7**: 62.

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A et al. 2020. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**: 284-289.

Jiao WB, Schneeberger K. 2020. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun* **11**: 989.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**: 540-546.

Koornneef M, Meinke D. 2010. The development of Arabidopsis as a model plant. *Plant J* **61**: 909-921.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-736.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* **40**: D1202-1210.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987-2993.

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103-2110.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Ma Z, Zhang Y, Wu L, Zhang G, Sun Z, Li Z, Jiang Y, Ke H, Chen B, Liu Z et al. 2021. High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. *Nat Genet* doi:10.1038/s41588-021-00910-2.

Mantere T, Kersten S, Hoischen A. 2019. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet* **10**: 426.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat Commun* **9**: 541.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA et al. 2019. Telomere-to-telomere assembly of a complete human X chromosome. doi:10.1101/735928 %J bioRxiv: 735928.

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

Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, Roest Crollius H, Salse J. 2015. Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol* **16**: 262.

Murigneux V, Rai SK, Furtado A, Bruxner TJC, Tian W, Harliwong I, Wei H, Yang B, Ye Q, Anderson E et al. 2020. Comparison of long-read methods for sequencing and assembly of a plant genome. *Gigascience* **9**.

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Lambing C, Kuo P, Yelina N, Hartwick N, Colt K et al. 2021. The genetic and epigenetic landscape of the Arabidopsis centromeres. *bioRxiv*.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30**: 1291-1305.

Perumal S, Koh CS, Jin L, Buchwaldt M, Higgins EE, Zheng C, Sankoff D, Robinson SJ, Kagale S, Navabi ZK et al. 2020. A high-contiguity Brassica nigra genome localizes active centromeres and defines the ancestral Brassica genome. *Nat Plants* **6**: 929-941.

Rellstab C, Zoller S, Sailer C, Tedder A, Gugerli F, Shimizu KK, Holderegger R, Widmer A, Fischer MC. 2020. Genomic signatures of convergent adaptation to Alpine environments in three Brassicaceae species. *Mol Ecol* **29**: 4350-4365.

Rhie A McCarthy SA Fedrigo O Damas J Formenti G Koren S Uliano-Silva M Chow W Fungtammasan A Kim J et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**: 737-746.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**: 245.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**: 155-158.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**: 227-245.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407-410.

Slankster EE, Chase JM, Jones LA, Wendell DL. 2012. DNA-Based Genetic Markers for Rapid Cycling Brassica Rapa (Fast Plants Type) Designed for the Teaching Laboratory. *Front Plant Sci* **3**: 118.

Steinbiss S, Willhoeft U, Gremme G, Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res* **37**: 7002-7013.

Tamokou JDD, Mbaveng AT, Kuete V. 2017. Chapter 8 - Antimicrobial Activities of African Medicinal Spices and Vegetables. In *Medicinal Spices and Vegetables from Africa*, doi:https://doi.org/10.1016/B978-0-12-809286-6.00008-X (ed. V Kuete), pp. 207-237. Academic Press.

Tang H, Lyons E, Town CD. 2015. Optical mapping in plant comparative genomics. *Gigascience* **4**: 3.

Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. 2018. MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res* **28**: 266-274.

Vaser R, Sovic I, Nagarajan N, Sikic M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737-746.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27**: 757-767.

Wotzel S, Andrello M, Albani MC, Koch MA, Coupland G, Gugerli F. 2021. Arabis alpina: a perennial model plant for ecological genomics and life-history evolution. *Mol Ecol Resour* doi:10.1111/1755-0998.13490.

Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, Luo F, Xie Z. 2017. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* **14**: 1072-1074.

Chapter4

MRDA: a computational framework for chromosome-by- chromosome assembly of metagenomes

MRDA: a computational framework for chromosome-by-chromosome assembly of metagenomes

Abstract

The quality of genomes constructed from the metagenome datasets influences the downstream analysis significantly. The 3rd generation sequencing improved the assembly of microbial communities compared to the highly fragmented assembly from NGS. However, constructing circular and complete microbial genomes from complex ecosystems is a challenging process. Here, we provide MRDA, a metagenome a computational framework for *reference-guided* data separation and chromosome-by-chromosome *de novo* assembly, which leverages the chromosome-by-chromosome assembly concept to address the main challenges of the assembly of complex communities. To validate the performance of our framework, we benchmarked MRDA using a synthetic dataset and two human stool datasets. MRDA overcomes the main challenges of metagenome assembly, revealing better results than the existing standard *de novo* assemblers in circularization and contiguity. Finally, we proved that with the availability of proper representative genomes dataset, chromosome-by-chromosome assembly is a robust solution to recover highly contagious genomes from complex communities and ecosystems.

4.1 Introduction

Microbial communities have a vital influence on all live dimensions, including human health, agriculture, fuel production, food science and even climate changes (Human Microbiome Project 2012; Boock et al. 2019; Cavicchioli et al. 2019; Hou et al. 2021). Unfortunately, around 99% of the potential microorganisms are unculturable (Ling et al. 2015). Even cultured microorganisms, examining the pure culture of individual microbial species or strain, cannot reflect their behaviour in complex microbial communities (Gould et al. 2018). Therefore, metagenomics provides revolutionized cultural-independent approaches to a comprehensive understanding of the microbial ecosystem's integrative interactions and circumvents the unculturable microbes (Nakamura et al. 2016). Several pipelines have been developed to address the functional and taxonomical diversity of a microbial community from their environment directly (Langille et al. 2013; Uritskiy et al. 2018). The next-generation sequencing (NGS) technology accelerates and accumulates metagenome data generation, demanding analysis and interpretation. The metagenomic sequencing data can be categorized

into two types: amplicon sequencing and whole metagenome shotgun sequencing (Bertrand et al. 2019; Gupta et al. 2019). The amplicon-based approach aims to dissect a particular microbial ecosystem's composition and taxonomical diversity using operational taxonomic units (OTU), e.g., 16SrRNA, 18SrRNA and ITS. The shotgun-based approach overcomes the constraints of OTU sequencing, and provide an opportunity to reveal the putative functional profile of a distinct microbial community (Lapidus and Korobeynikov 2021).

Retrieving circular and complete metagenome-assembled genomes is an outstanding goal of the metagenome shotgun sequencing approach (Chen et al. 2020). So far, short-read sequences are the most widely used technology in metagenome analysis. Several tools have been developed to assemble metagenome datasets and recover individual fragmented genomes and contigs from culturable and unculturable species (Li et al. 2016; Nurk et al. 2017). The long-read technology has the likelihood to improve the contiguity of the highly fragmented microbial genomes constructed by short read (Bishara et al. 2018). Although the implementation and development of several long-read assemblers on metagenome datasets, increasing the contiguity of long-read metagenome assembly is facing significant difficulties. For example, the deficiency of proper DNA extraction protocols for metagenomic samples adversely affects the amount and length of the recovered sequences (Driscoll et al. 2017). On the other hand, the uneven abundance of species-load in the microbial community, the inter-genomic and intra-genomic heterogeneity and repeats reflected an algorithmic and computational challenge comparing to the assembly of a single-cultured microbial genome (Peng et al. 2012; Kolmogorov et al. 2020).

The final assembly of metagenome samples comprising a mixture of contigs from different microbial species. So, several *reference-guided* and reference-free binning algorithms had been developed to cluster the contigs into representing individual species (Wu et al. 2016; Nissen et al. 2021). The *reference-guided* metagenome assembly is beneficial for recovering a high number of complete and circular microbial genome assemblies (Cepeda et al. 2017; Guyomar et al. 2018). Despite the advantages of the *reference-guided* approach, it is biased against novel species and those without a closely related reference genome. Here, we proposed a *reference-guided* data separation framework to facilitate chromosome-by-chromosome *de novo* assembly as well as taxonomical composition profile module of metagenome long-reads datasets. I describe the new algorithm and benchmark it using a sequencing dataset from a synthetic population and two human stool datasets.

4.2 Results

4.2.1 MRDA graph overview

The Principle of chromosome-by-chromosome assembly is inspired by single-cultured microbial genome assembly. The GALA algorithm proved the advantages of applying chromosome-by-chromosome assembly on multi-chromosome eukaryotic genomes (Awad and Gan 2020). We introduce a *reference-guided* data separation and chromosome-by-chromosome *de novo* assembly module to overcome the nonuniform distribution of reads abundance. In addition, our module exploits the chromosome-by-chromosome assembly assumption, which robustly disentangles the graph of the shared conserved sequences between closely related species.

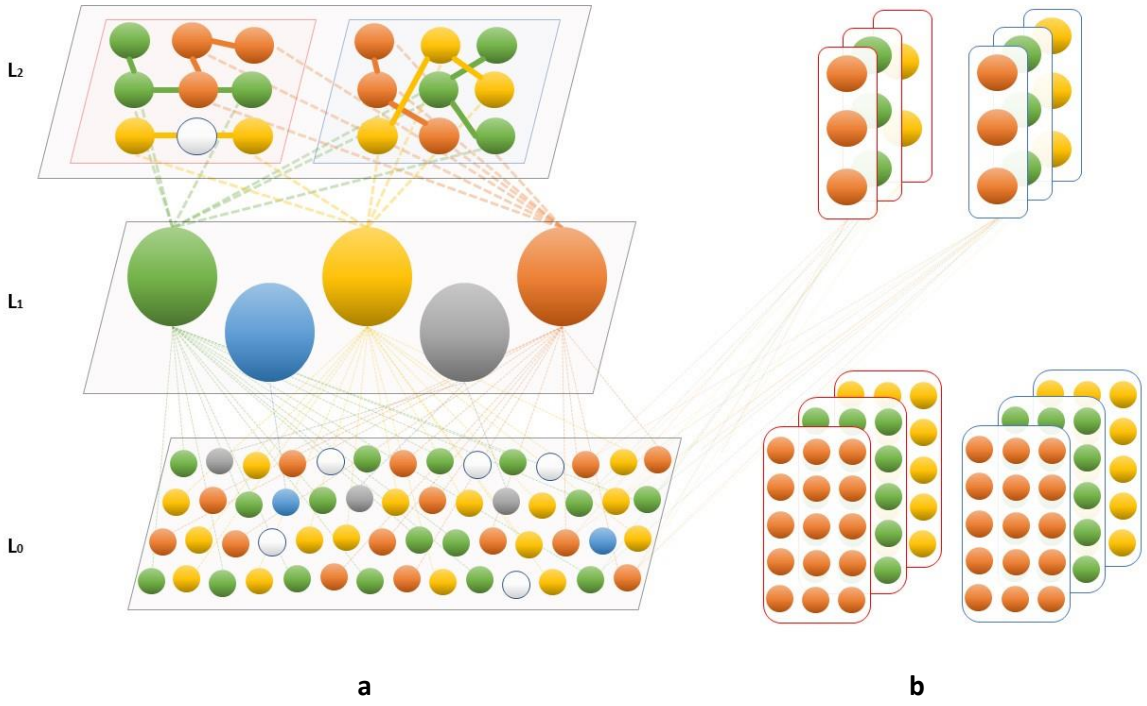


Figure 4.1: Illustration of MARDA triple-layer graph a) The raw reads encoded in L_0, the representative genomes encoded in L_1 and the preliminary assemblies encoded as a disjoint sublayer in L_2. Only L_2 encodes an intra-sublayer and inter-layer edges, while L_0 and L_1 encode only inter-layer edges. B) contig (upper-raw) and reads (lower-raw) species-specific linkage groups.

MRDA exploits information from microbial reference genome dataset and preliminary assembly/assemblies to cluster raw reads into species-level linkage groups as an input for chromosome-by-chromosome *de novo* assembly to construct highly continuous genomes. Firstly, we selected ProGenomes v2.1 as a representative reference genomes dataset. Then as

in GALA, we picked a bunch of metagenomic *de novo* assembly tools to generate preliminary assemblies. We modelled the raw read, preliminary assemblies and representative genomes dataset as a triple-layer graph. The bottom layer (L_0) encoded the raw reads. The middle layer (L_1) outlined the representative genomes dataset with a single node for each reference genome, even a fragmented reference genome is fine. As L_0, this layer has only inter-layer edges with L_0 and upper layer L_2. The upper layer (L_2) is for the preliminary assemblies, which contains *N* distinct horizontal disjoint sub-layers in the same level without inter-sub-layer edges among them. Each sub-layer represents a preliminary assembly. In contrast to L_0 and L_1, each sub-layer in L_2 crypted two types of edges, inter-layer edges and intra-sublayer edges between the nodes (contigs) within the sublayer (**Fig. 4.1**).

We used the constructed multi-layer graph to operate two levels of *reference-guided* data separation. In the first level, we exploit the representative genome inter-layer edges to classify the reads and the contigs into different bins, representing individual species. The raw reads bins were utilized to generate a taxonomical composition profile for the investigated dataset. In the next binning level, the investment of the raw reads layer links to species-specific contig bins produces species-specific raw-reads linkage groups. Finally, we utilize GALA's module LGAM to perform chromosome-by-chromosome *de novo* assembly directly on the species-specific raw-reads linkage groups and merging them with the first level read bins if necessary.

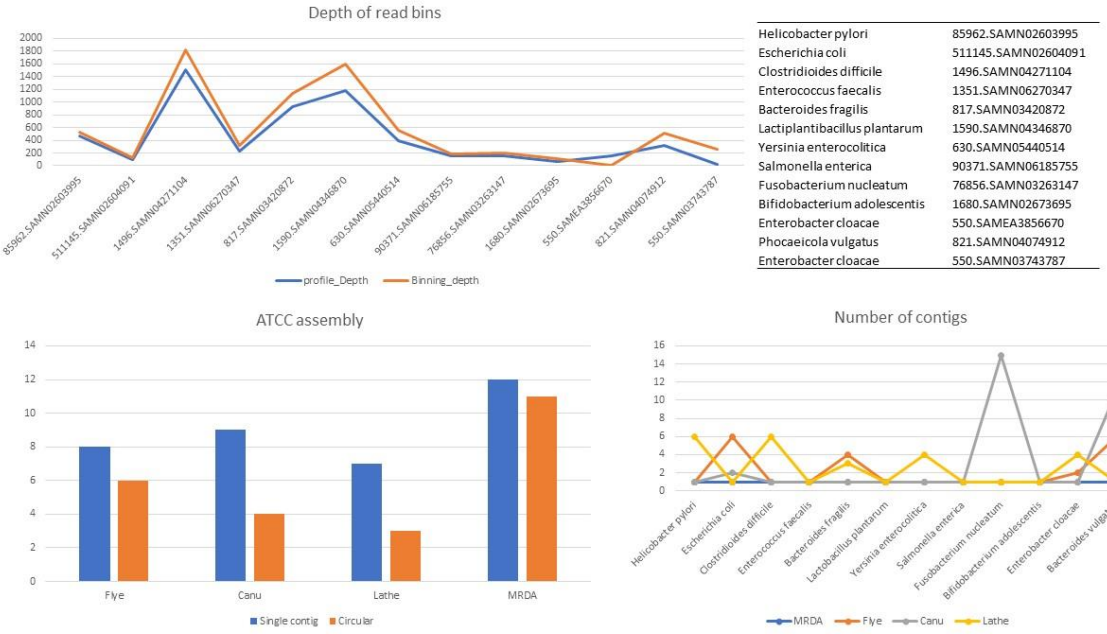
4.2.2 Assembly of ATCC synthetic mixture

To assess the effectiveness of the metagenome assembly module (MRDA), we used a publicly available 30 GB of Nanopore reads from a synthetic bacterial mixture comprising 12 ATCC Gram-positive and Gram-negative standard species. However, Canu and metaFlye were authorized to generate the preliminary assemblies. Canu assembled 105 contigs of total length 48.86 Mbp with N50 3.58 Mbp. In comparison, metaFlye assembler generated 57 contigs of total length 47.58 Mbp and N50 of 3.78 Mbp. Among the assembled contigs, metaFlye yielded eight species as a single contig with six circular molecules. While Canu returned nine single-contig bacterial chromosomes, but only four chromosomes were circularized.

Next, we applied our metagenome assembly module MRDA using the two preliminary assemblies. As a result, 65 and 33 contigs of Canu and metaFlye were clustered into 12 species-specific linkage groups. The raw read taxonomical profile proved the presence of the 12 species in the sequencing dataset without any tangible contamination, with a depth ranging between

1513 X in *Clostridioides difficile* and 71 X in *Bifidobacterium adolescentis*. To avoid the false mapping to closely related species in the representative genomes dataset, we mapped the reads against the 12 reference genomes. Appropriately, the species read depth increased to 1746 X in *Clostridioides difficile* and 95 X in *Bifidobacterium adolescentis* (**Fig. 4.2**). Then, we used the LGAM module to assemble each species-specific raw-reads linkage group individually. LGAM successfully assembled 11 species to circular chromosomes, while *Bacteroides vulgatus* were assembled into a single linear contig of length 5.14 Mbp (**Table 4.1** and **Fig. 4.2**). Thus, the MRDA metagenome assembly module outperformed the preliminary assemblers and the lathe metagenomic assembly pipeline, which assembled seven species as a single contig, including only three circular chromosomes (Moss et al. 2020).

a



Helicobacter pylori	85962.SAMN02603995
Escherichia coli	511145.SAMN02604091
Clostridioides difficile	1496.SAMN04271104
Enterococcus faecalis	1351.SAMN06270347
Bacteroides fragilis	817.SAMN03420872
Lactiplantibacillus plantarum	1590.SAMN04346870
Yersinia enterocolitica	630.SAMN05440514
Salmonella enterica	90371.SAMN06185755
Fusobacterium nucleatum	76856.SAMN03263147
Bifidobacterium adolescentis	1680.SAMN02673695
Enterobacter cloacae	550.SAMEA3856670
Phocaeicola vulgatus	821.SAMN04074912
Enterobacter cloacae	550.SAMN03743787

b **c**

Figure 4.2: a) The depth profile of the ATCC synthetic mixture from the first binning (blue) and second binning (orange) phases. The table showed the scientific name of the representative genome. b) The number of circular and single contig assembled genomes in metaFlye, Canu, Lathe and MRDA. c) The number of assembled contigs for each species using different assemblers.

4.2.3 Assembly of human stool samples

We further investigate MRDA module potency on two publicly available Nanopore human stool datasets previously assessed with lathe pipeline (Moss et al. 2020). The samples P2-A (6.1 GB) and P2-B (7.6 GB) were collected from the same individual, with 15 months between the collection of both samples and read N50 of 3 kbp for both datasets. However, as a preparation step, both datasets were assembled using Canu and metaFlye. The Canu assembly has contigs N50 of 307 Kbp and 241 Kbp, total assembly sizes of 82.1 Mbp and 84.5 Mbp and a total number of assembled contigs of 2946 and 3659 for samples P2-A and P2-B, respectively. In contrast, the Flye assembly has a shorter N50 of 186 Kbp and 107 Kbp, a more extended genome size of 89.6 Mbp and 102.7 Mbp and a lower number of assembled contigs 1366 and 1857 for P2-A and P2-B, respectively. Next, we used the two preliminary assemblies from Canu and metaFlye for MRDA implementation.

The read taxonomical profile of the sample P2-A recorded 23 taxa with a total depth >5 % and reference coverage > 65 %. Among these taxa, *Ruminococcus torques* and *Faecalibacterium praunsitizii* are represented by two different length strains, while *Clostridium sp.* is represented by three different taxa. MRDA clustered the Canu and metaFlye drafts into 21 linkage-groups shared the 20 taxa with depth > 8 X in addition to extra taxa of *Faecalibacterium praunsitizii* and *Clostridium sp.* from Canu and metaFlye, respectively. Finally, the LGAM module assembled eight species to high-quality circular chromosomes, including *Ruminococcus torques*, *Veillonella dispar*, *Prevotella copri*, *phascolarctobacterium sp*. and *Faecalibacterium praunsitizii*, which aligned to fragmented references on the representative genomes dataset. Further analysis proved that *Bacteroides vulgatus* assembled to a single linear chromosome of length 4.9 Mbp despite being the most abundant species on the read taxonomical profile with 313 X coverage (**Table 4.1** and **Fig. 4.3**). On the other hand, the lathe pipeline reported only two circular chromosomes from this dataset. In comparison, metaFlye and Canu preliminary assemblies circularized only *Dialister invisus* and assembled two species to a single linear chromosome.

Additionally, MRDA taxonomical profile determined 26 taxa in the P2-B sample under the default parameters. Like P2-A, two different size strains represented *Faecalibacterium praunsitizii,* while five taxa represented *Clostridium sp.,* including two uncultured taxa. Interestingly, the MRDA contig binning algorithm clustered the Canu and metaFlye assemblies to 17 and 23 linkage groups, respectively. The low coverage of several taxa in the P2-B sample

undermined the Canu performance, while metaFlye generated highly fragmented contigs from low coverage taxa. However, both assemblers successfully assembled all taxa over 10 X coverage. Eventually, we successfully assembled four circular bacterial chromosomes and two single contig linear chromosomes from this dataset. In addition, the LGAM module generated a circularized 4.85 Mbp *Bacteroides vulgatus* genome for the first time in this study (**Table 4.1**; **Fig. 4.3**). Again, our metagenome *reference-guided* data separation and chromosome-by-chromosome *de novo* assembly module MRDA confirmed its superiority over the preliminary assemblers and the lathe metagenomic assembly pipeline, which yielded only a single circular chromosome from each of them.

Overall, MRDA retrieved 12 circularized bacterial chromosomes from the two human stool datasets representing ten distinct species. In contrast, Canu, Flye and lathe pipeline retrieved 2, 2 and 3 circular chromosomes from the same datasets. Notably, even when we fail to assemble the genome into a single contig, MRDA increases the contiguity over the preliminary assemblers. For example, MRDA assembled the *Bacteroides ovatus* genome into three contigs of length 3.6 Mbp, 2.5 Mbp and 0.76 Mbp. While Canu and metaFlye assembled it into 7 and 14 contigs with N50 1.8 Mbp and 0.78 Mbp, respectively (**Fig. 4.3**).

representative ID	Species	Sample	Genome-size (Mbp)	Circularization	Circularization-confirmation
85962.SAMN02603995	Helicobacter pylori	ATCC	1.71	+	Assembler + mapping test
511145.SAMN02604091	Escherichia coli	ATCC	4.68	+	Assembler + mapping test
1496.SAMN04271104	Clostridioides difficile	ATCC	4.20	+	Assembler + mapping test
1351.SAMN06270347	Enterococcus faecalis	ATCC	3.28	+	Assembler + mapping test
817.SAMN03420872	Bacteroides fragilis	ATCC	5.26	+	Assembler + mapping test
1590.SAMN04346870	Lactiplantibacillus plantarum	ATCC	3.34	+	Assembler + mapping test
630.SAMN05440514	Yersinia enterocolitica	ATCC	4.59	+	Assembler + mapping test
90371.SAMN06185755	Salmonella enterica	ATCC	4.64	+	Assembler + mapping test
76856.SAMN03263147	Fusobacterium nucleatum	ATCC	2.24	+	Assembler + mapping test
1680.SAMN02673695	Bifidobacterium adolescentis	ATCC	2.03	+	Assembler + mapping test
821.SAMN04074912	Phocaeicola vulgatus	ATCC	5.12	-	Assembler + mapping test
550.SAMN03743787	Enterobacter cloacae	ATCC	5.35	+	Assembler + mapping test
592028.SAMN00008820	Dialister invisus DSM	P2B	1.92	-	Assembler + mapping test
821.SAMN04074912	Phocaeicola vulgatus	P2B	4.84	+	Assembler + mapping test
1262914.PRJEB757	Phascolarctobacterium sp.	P2B	2.37	+	Assembler + mapping test
292800.SAMN04621620	Flavonifractor plautii	P2B	3.47	+	Assembler + mapping test
28117.SAMN05717213	Alistipes putredinis	P2B	2.72	-	Assembler + mapping test
537011.SAMN00008847	Prevotella copri	P2B	3.81	+	Assembler + mapping test
39778.SAMN04324900	Veillonella dispar	P2A	2.12	+	mapping test
592028.SAMN00008820	Dialister invisus	P2A	1.92	+	Assembler + mapping test
748224.SAMN00189147	Faecalibacterium cf. prausnitzii	P2A	2.60	+	mapping test
821.SAMN04074912	Phocaeicola vulgatus	P2A	4.99	-	Assembler + mapping test
649756.SAMN02725021	Anaerostipes hadrus	P2A	2.86	+	Assembler + mapping test
1262914.PRJEB757	Phascolarctobacterium sp	P2A	2.33	+	Assembler + mapping test
515619.SAMN02603082	[Eubacterium] rectale	P2A	3.34	+	Assembler + mapping test
33039.SAMEA3545353	[Ruminococcus] torques	P2A	3.40	+	Assembler + mapping test
537011.SAMN00008847	Prevotella copri	P2A	3.81	+	Assembler + mapping test

Table 4.1: Single contig assembled genomes

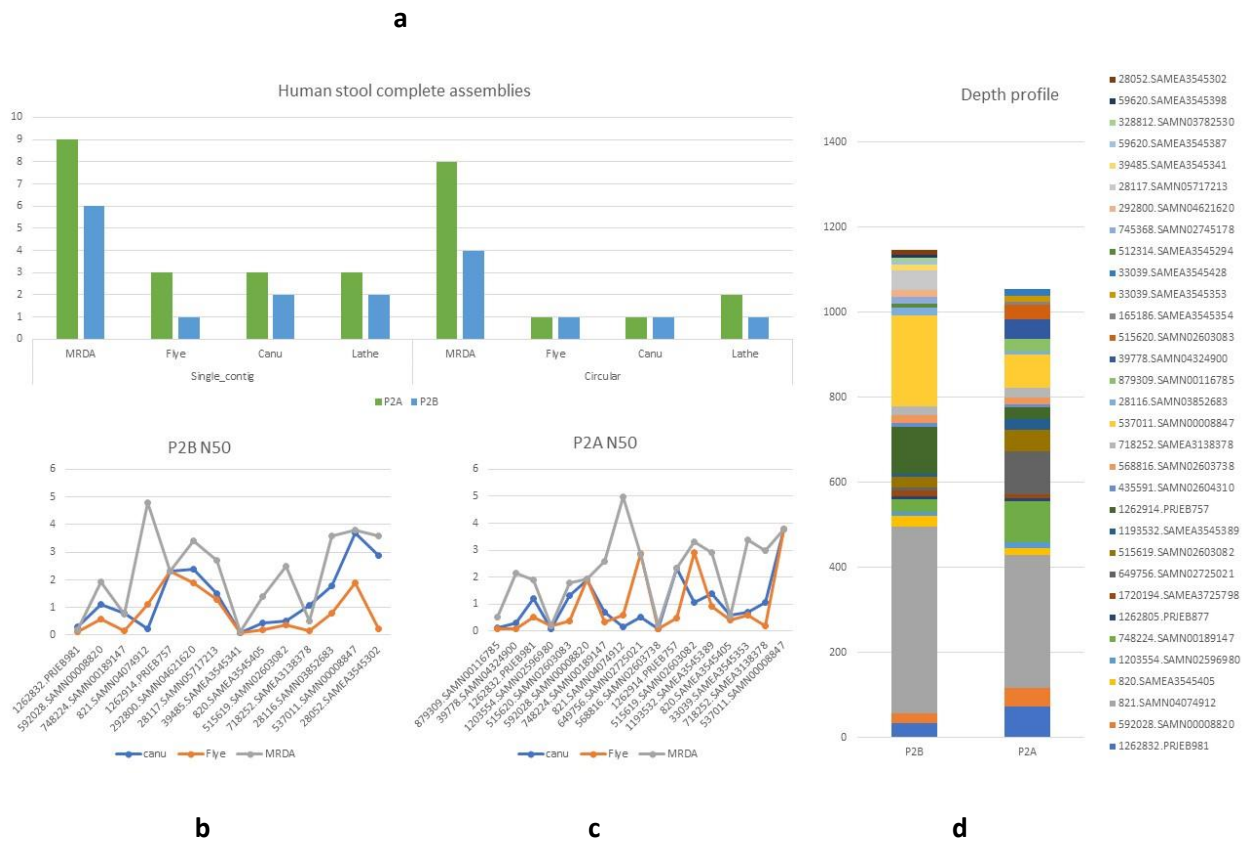


Figure 4.3: a) The assembled circular and single contig genomes from the human stool samples. b) N50 of the assembled genomes from P2B sample. c) N50 of the assembled genomes from P2A sample. d) The depth profile of P2A and P2B samples taxonomical composition.

4.3 Discussion

The Nanopore real-time sequencing technology provides a promising approach to direct sequencing and studying the complex microbial communities in their ecosystems. Several *de novo* assemblers had been developed to handle the assembly of single and multi-chromosome species. Unfortunately, most of these tools are not adapted to handle metagenomic data, giving a significant difference in the performance of metagenome assembly (Latorre-Perez et al. 2020). Although the *reference-guided* metagenome assembly returned better results in short-reads studies (Cepeda et al. 2017), there is a lack of studies to develop long-read metagenome *reference-guided* assemblers and compare the performance of the *de novo* and *reference-guided* assembly state-of-arts for long-read metagenomic data. This study exploits the chromosome-by-chromosome assembly strategy to establish the MRDA module as a long-read metagenome *reference-guided* data separation and *de novo* genome assembler.

MRDA affords substantial improvements over the current classical long-read metagenome assembly workflows. The benchmarked of MRDA, lathe, Canu and metaFlye assemblers on the

ATCC standard species synthetic mixture explained the advantage of using a *reference-guided* data separation approach and chromosome-by-chromosome *de novo* assembly in MRDA by retrieving all the genomes in the mixture as a single contig with 91 % circularized chromosomes. In contrast, using the common *de novo* assembly workflows yielded 66% single contig genomes and 50% circularized chromosomes in the best case. The fragmented chromosomes recovered by the classical *de novo* assembly workflows may affect the structural and functional downstream analysis for distinct communities (Tsai et al. 2016).

Although the MRDA is superior over the classical workflows in recovering high-quality genomes from the human stool datasets, we noticed that the abundance of high divergence strains from the same species with sufficient depth undermines the assembly process (Kolmogorov et al. 2020). For example, the profusion of Clostridium strains on the human stool datasets adversely affects the raw reads binning algorithm, accordingly undermining the chromosome-by-chromosome assembly. Thus, methods that deconvolve the strain composition from the long-read metagenome sequences datasets are obliged to enhance the contiguity of the chromosome-by-chromosome assembly. Additionally, the base accuracy of the preliminary assemblies is a vital factor for the contig binning algorithm. For example, MRDA successfully detected and clustered the *Flavonifractor plautii* contigs in Canu preliminary assembly but failed in the metaFlye draft because of the low mapping quality and identity percentage. Therefore, polishing the preliminary assemblies before MRDA implementation will promote the *reference-guided* data separation results.

Finally, MRDA performance improved the recovery of complete and circular microbial genomes from complex ecosystems, facilitating the study of diverse microbial compositions over global populations. It also promotes examining the potential phenotypes of distinct microorganisms, even uncultured organisms. Furthermore, MRDA authorizes segregation and deep dissection of genotypic structure composition for a definite group, taxa or species correlated with distinct disease or biological function in a particular biosystem. We anticipate MRDA influences in metagenome assembly contiguity will further expedite the functional and structural variant resolution among microbial communities.

4.4 Methods

4.4.1 Representative genomes dataset

Choosing the representative reference genome is a crucial factor in the *reference-guided* assembly approach. Therefore, we selected the recently published ProGenomes v2.1 database (Mende et al. 2020) as a default representative genomes dataset. The dataset consisted of 12221 genomes comprising eight different habitat representative genome subsets.

1. The aquatic representative genomes subset comprises 2420 taxa from 2472 genome projects.
2. The food-associated representative genomes subset comprises 302 taxa from 314 genome projects.
3. The fresh-water representative genomes subset comprises 267 taxa from 272 genome projects.
4. The host-associated representative genomes subset comprises 3229 taxa from 3443 genome projects.
5. The host plant-associated representative genomes subset comprises 381 taxa from 400 genome projects.
6. The sediment-mud representative genomes subset comprises 1181 taxa from 1192 genome projects.
7. The soil representative genomes subset comprises 2502 taxa from 2571 genome projects.
8. The disease-associated representative genomes subset comprises 11788 taxa from 12221 genome projects.

The user can use the default dataset or one of the previous subsets.

4.4.2 The Multilayer graph composition

We built a multi-layer graph consisted of 3 layers G= (V, E, L=3), where V is a set of vertices, L is a set of layers and E $\subseteq$ V x V x L is a set of edges. The first layer L_0= (V_0, E_0) encoded the raw reads as a set of vertices $V_0 \in V$, V_0= {v_{01},v_{02},...v_{0n}} is the set of reads and $E_0 \in$ E, E_0= {V_0 x L_0 xV_1 x L_1} $\cup$ { V_0 x L_0 xV_2 x L_2} is the set of inter-layer edges. The second layer L_1= (V_1, E_1) encoded the representative genomes as a set of vertices $V_1 \in V$, V_1= {v_{11},v_{12},...v_{1n}} is the set of genomes, $v_{1i} \in V_1$, v_{1i}= {c_1,c_2,...c_n} a set of contigs/scaffolds in the representative genome and $E_1 \in$ E, E_1= {V_0 x L_0 xV_1 x L_1} $\cup$ { V_1 x L_1 xV_2 x L_2} is the set of inter-layer edges. The third layer L_2= (V_2, E_2, S), is a composite layer consisting of disjoint sublayers S= {P_1, P_2,...P_n}, each sublayer representing a preliminary assembly. Each sublayer encoded a set of

vertices $V_2P_i \in V_2 \in V$, $V_2P_i = \{v_2p_{i1}, v_2p_{i2}, \ldots v_2p_{in}\}$ and $E_2 \in E$, $E_2 = \{ V_0 \times L_0 \times V_2 \times L_2\} \cup \{ V_1 \times L_1 \times V_2 \times L_2\} \cup \{V_2P_i \times L_2P_i \times V_2P_i \times L_2P_i \}$ is the set of inter-layer and intra-sublayer edges.

4.4.3 The data separation Algorithm

In the first step of the data separation algorithm, we pooled all nodes in L_0 and L_2, which connected to the same node in V_1 into a taxa-specific linkage group. Then we used these raw read linkage groups to generate the taxonomical profile. First, we used reads and contigs linkage groups to create a new representative genomes dataset. Next, we used the new dataset to rebuild and reanalyse the graph to extract the taxa-specific contig linkage groups. Eventually, we merged all the raw-reads linkage groups connected to the same taxa-specific contig linkage groups into a taxa-specific raw reads linkage group.

4.4.4 Final assembly

The reads within a linkage group were assembled using LGAM (Awad and Gan 2020) with third-generation assembly tools in normal mode, e.g., Flye (Kolmogorov et al. 2019), Necat (Xiao et al. 2017), Canu (Nurk et al. 2020a) and Miniasm (Li 2016). Finally, we mapped the newly assembled contigs against one of the preliminary assemblies recorded in the same linkage group and removed the contamination and false duplicated copies.

4.4.5 Metagenome assembly

We employed metaFlye v.2.4 (Kolmogorov et al. 2020) and Canu v.2 (Nurk et al. 2020a) with the default parameters to generate all preliminary metagenomic assemblies. In addition, we supplied the genome size option with 50 Mbp for the synthetic dataset and 100 Mbp for the human stool samples.

4.4.6 MRDA implementation

We used the complete representative genomes dataset and the preliminary assemblies from Canu and metaFlye to apply MRDA. The pipeline started by producing an indexing file for the representative genomes dataset. Next, we hired Minimap2 (Li 2018) to align the raw reads and the preliminary assemblies to the representative genomes. Then, we used the Samtools depth (Li et al. 2009) to generate the depth file for the raw read mapping file. MRDA used the cov_dep module to recover the taxonomical profile composition. Next, MRDA ran the paf_analysis module to bin the preliminary contigs into taxa-specific linkage groups. Then, MRDA exploits the shared taxa between the read taxonomical profile and contig linkage groups to extract distinct representative

genomes dataset. Finally, MRDA leveraged Minimap2 to align the preliminary assemblies to the new representative genomes dataset, producing taxa-specific raw-reads linkage groups.

We borrowed the LGAM module from GALA to assemble the taxa-specific raw-reads linkage groups into a single circular microbial chromosome.

4.4.7 Assembly assessment

The final assemblies mapped to their reference genomes using Minimap2 (-x asm5) to detect the number of assembled contigs mapped to the reference genome and ignore the contaminated and duplicated contigs if they appeared.

References

Awad M, Gan X. 2020. GALA: gap-free chromosome-scale assembly with long reads. *bioRxiv*.

Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvornicic M, Soldo JP, Koh JY, Tong C et al. 2019. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol* **37**: 937-944.

Bishara A, Moss EL, Kolmogorov M, Parada AE, Weng Z, Sidow A, Dekas AE, Batzoglou S, Bhatt AS. 2018. High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat Biotechnol* doi:10.1038/nbt.4266.

Boock JT, Freedman AJE, Tompsett GA, Muse SK, Allen AJ, Jackson LA, Castro-Dominguez B, Timko MT, Prather KLJ, Thompson JR. 2019. Engineered microbial biofuel production and recovery under supercritical carbon dioxide. *Nat Commun* **10**: 587.

Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, Behrenfeld MJ, Boetius A, Boyd PW, Classen AT et al. 2019. Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* **17**: 569-586.

Cepeda V, Liu B, Almeida M, Hill CM, Koren S, Treangen TJ, Pop M. 2017. MetaCompass: Reference-guided Assembly of Metagenomes. *bioRxiv*.

Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF. 2020. Accurate and complete genomes from metagenomes. *Genome Res* **30**: 315-333.

Driscoll CB, Otten TG, Brown NM, Dreher TW. 2017. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Stand Genomic Sci* **12**: 9.

Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Korasidis N, Gavryushkin A, Carlson JM, Beerenwinkel N, Ludington WB. 2018. Microbiome interactions shape host fitness. *Proc Natl Acad Sci U S A* **115**: E11951-E11960.

Gupta S, Mortensen MS, Schjorring S, Trivedi U, Vestergaard G, Stokholm J, Bisgaard H, Krogfelt KA, Sorensen SJ. 2019. Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. *Commun Biol* **2**: 291.

Guyomar C, Delage W, Legeai F, Mougel C, Simon J-C, Lemaitre C. 2018. Reference guided genome assembly in metagenomic samples %+ Institut de Génétique, Environnement et Protection des Plantes (IGEPP) %+ Scalable, Optimized and Parallel Algorithms for Genomics (GenScale). In *RECOMB 2018 - 22nd International Conference on Research in Computational Molecular Biology*, pp. 1 %8 2018-2004-2021, Paris, France.

Hou S, Wolinska KW, Hacquard S. 2021. Microbiota-root-shoot-environment axis and stress tolerance in plants. *Curr Opin Plant Biol* **62**: 102028.

Human Microbiome Project C. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207-214.

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL et al. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**: 1103-1110.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**: 540-546.

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R et al. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814-821.

Lapidus AL, Korobeynikov AI. 2021. Metagenomic Data Assembly - The Way of Decoding Unknown Microorganisms. *Front Microbiol* **12**: 613791.

Latorre-Perez A, Villalba-Bermell P, Pascual J, Vilanova C. 2020. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. *Sci Rep* **10**: 13588.

Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, Yamashita H, Lam TW. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**: 3-11.

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103-2110.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, Mueller A, Schaberle TF, Hughes DE, Epstein S et al. 2015. A new antibiotic kills pathogens without detectable resistance. *Nature* **517**: 455-459.

Mende DR, Letunic I, Maistrenko OM, Schmidt TSB, Milanese A, Paoli L, Hernandez-Plaza A, Orakov AN, Forslund SK, Sunagawa S et al. 2020. proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res* **48**: D621-D625.

Moss EL, Maghini DG, Bhatt AS. 2020. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* **38**: 701-707.

Nakamura K, Iizuka R, Nishi S, Yoshida T, Hatada Y, Takaki Y, Iguchi A, Yoon DH, Sekiguchi T, Shoji S et al. 2016. Culture-independent method for identification of microbial enzyme-encoding genes by activity-based single-cell sequencing using a water-in-oil microdroplet platform. *Sci Rep* **6**: 22259.

Nissen JN, Johansen J, Allesoe RL, Sonderby CK, Armenteros JJA, Gronbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O et al. 2021. Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* **39**: 555-560.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**: 824-834.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *bioRxiv*.

Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420-1428.

Tsai YC, Conlan S, Deming C, Program NCS, Segre JA, Kong HH, Korlach J, Oh J. 2016. Resolving the Complexity of Human Skin Metagenomes Using Single-Molecule Sequencing. *mBio* **7**: e01948-01915.

Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**: 158.

Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**: 605-607.

Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, Luo F, Xie Z. 2017. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* **14**: 1072-1074.

Chapter5

Discussion

Discussion

The development of third-generation sequencing technologies undoubtedly empowers efforts to obtain more contiguous and complete genome assemblies. Short-read assemblies are highly fragmented, while the long-read assemblies contain more extended contigs with higher genome N50. For example, the preliminary assemblies of *C.elegans* and *Cardamine hirsuta* generated in chapters 2 and 3 (**Supp. Table 2.1 and Table 3.2**) have a higher N50 than the previously released short-read assemblies (Desai et al. 2013; Thompson et al. 2015; Gan et al. 2016). Unfortunately, the higher error rate of pacbio and Nanopore reads adversely influences the correctness of produced genomes even after long-read polishing procedures (Koren et al. 2019). Thus, short-read technologies are still playing a notable role in genome assembly pipelines to polish the final drafts and increase the base accuracy (Nurk et al. 2020a). Recently, Pacbio introduced the low error rate Hifi long-reads to bridge the error rate's gap between short-read and long-read technologies (Hon et al. 2020; Lang et al. 2020), thus to some extent reducing the need for short read assembly polishing.

Results from preliminary assemblies confirmed that the different assembly algorithms and tools have various contiguity, completeness, and correctness performances, which also differ from dataset to dataset. Consequently, there is no optimal standard assembly pipeline fit or reflect the best performance for different species or even datasets (Jayakumar and Sakakibara 2019). Therefore, we developed GALA, a gap-free chromosome-scale assembler, to take advantage of combining and organizing constructed contigs by different algorithms. GALA modelled the preliminary assemblies and raw reads in a scalable multilayer graph to solve the assembly errors and produce chromosome-scale assemblies. Our method validated a significant improvement in final assembly contiguity over the current tools.

Most preliminary assemblies suffered from inter-chromosomal misassemblies, affecting the downstream analysis and scaffolding pipelines (Molina-Mora et al. 2020). The current *de novo* misassembly detection state-of-art relies on generating an additional source of information, e.g., 10X genomics (Jackman et al. 2018), Hi-C (Ghurye et al. 2019) or bionano mapping (Yuan et al. 2017), which is costly and sometimes labor and computationally intensive. Thus, we developed a misassembly detection module (MDM), leveraging the misassembly variation in multiple preliminary assemblies organized in a multilayer graph to identify and resolve the chimeric contigs in a reference-free detection manner. The MDM module identified and

effectively resolved the misassembly generated from wide-range assembly algorithms. Furthermore, tracing the fully and partially mapped reads to the misassembled loci in the preliminary assemblies of *C. hirsuta* confirmed the efficiency of MDM.

The GALA's contig clustering module (CCM) proved a highly significant advancement in assigning contigs to correlated chromosomes without any external source of linkage information compared to various scaffolding approaches (**Fig 2.3**). While the Hi-C and various mapping scaffolding techniques produced many unanchored contigs alongside the gapped scaffolds (Zhang et al. 2019), the CCM module can anchor all the contigs in our experiments to their chromosomal linkage groups, except three centromeric contigs on the *Arabidopsis thaliana* genome. Furthermore, the assembly errors, structure variants, chromosomal rearrangements and unplaced contigs of reference genomes undermine the classical *reference-guided* scaffolding strategy (Alonge et al. 2019), whereas CCM can overcome these challenges. The CCM can use the reference genome like a genetic map to support the contig assigning procedure by ignoring the reference errors in the MDM module, which usually comes from the contradictory alignment of gaped regions and unanchored contigs or chromosomal rearrangements between different strains and cell lines (Lischer and Shimizu 2017).

Chromosome flow-sorting is a highly expensive, time-consuming and labor-intensive process affording a lossless reduction of genomic complexity. The sorting techniques require highly complicated protocols to resolve the optical properties, e.g., light scatters and fluorescence of the target chromosomes (Dolezel et al. 2012). Nevertheless, chromosome sorting was used to assemble human, Wheat and many other genomes (International Wheat Genome Sequencing 2014; Tomaszkiewicz et al. 2016; Kuderna et al. 2019). GALA introduced the LGAM module to provide a computational framework for *in-silico* chromosome sorting to apply a chromosome-by-chromosome assembly. The LGAM module successfully reduced the complexity of the assembly graph by reducing the total number of edges in the graph by ~ 30% compared to the whole genome assembly in *C.elegans* (**Fig 2.6**). The concept of chromosome-by-chromosome assembly provides a promising strategy to untangle the shared sequences of the assembly graph, avoiding chimeric contig construction and assembling highly complex genomes.

The current assembly of *C.elegans* strain VC2010 was produced using a *reference-guided* approach and contained two unclosed gaps (Yoshimura et al. 2019). However, GALA successfully closed all the gaps and produced a gapless telomere-to-telomere *de novo* assembly of *C.elegans* strain VC2010 (**Fig 2.3b**). Moreover, our evaluation analysis proved that the GALA

draft has better accuracy and lower collapsing compared to the published draft and the gap-free N2 reference genome (**Table 2.1**). In agreement with the released Pacbio and Nanopore assemblies, the GALA draft has a remarkable 2Mbp genome size extension over the genome size of the gap-free N2 reference genome (Tyson et al. 2018; Yoshimura et al. 2019). GALA also improved the contiguity of the published assembly of *Arabidopsis thaliana* accession KBS-Mac-74. While the most continuous draft of this accession comprised 62 contigs with N50 of 12 Mbp and longest contig of length 14 Mbp (Michael et al. 2018), GALA assembled ten gap-free chromosome arms and only three unanchored centromeric contigs with 2Mbp extension on longest contig size and N50. Interestingly, as evidence of solving collapses in the published assembly, the unpolished draft of the GALA assembly has a ten Mbp genome size extension over the unpolished published draft and three Mbp over the polished published draft.

The genuineness of inferences and the predictive ability of biological studies rely on the availability of genetic materials and expanding the base of model systems. Thus, constructing gapless chromosome-scale assemblies has been an outstanding goal for computational biologists in the last two decades. However, the *Cardamine hirsuta* holds a crucial position as a model plant in comparative developmental studies (Hay et al. 2014), particularly after the first genome assembly draft release (Gan et al. 2016; McKim et al. 2017). In this study, we implemented GALA on Pacbio and Nanopore heterogeneous datasets to construct the second genome assembly draft. GALA generated a very high-quality T2T assembly for the oxford strain and performed better than the classical Hi-C scaffolding technique (**Table 3.3; Table 3.4; Supp. Fig. 3.5**). Importantly, the GALA draft attached all sequences to the gap-free chromosomes, proving higher contiguity, completeness and base accuracy score over the reference genome. Moreover, the new draft resolved several inter-chromosomal, intra-chromosomal variants and collapses in the reference genome (**Fig. 3.2**).

Furthermore, GALA produced a gap-free telomere-to-telomere assembly of the C.*hirsuta* Azores strain genome with a 3 Mbp genome size expansion over the Oxford reference strain genome. To improve the accessibility of high-quality genetic materials for *C.hirsuta's* closely related species and improve the comparative studies inferences, GALA assembled the first draft genome of *C. oligosperma* as a gap-free telomere-to-telomere assembly. Further, we generated a high-quality and complete *C. resedifolia* genome assembly comprising 42 Mbp missing sequences from the published assembly (Rellstab et al. 2020). Undoubtedly, these assemblies would contribute towards a deep dissection of interspecies variations and developmental stages, improving the resolution of interspecies and intraspecies comparative

investigations. For example, the accurate assembly of these species facilitates the prediction of karyotypic differences and the TE-load role on the genome size differences.

The uneven existence of microbial species in the metagenome samples, aw well as the interspecies and intraspecies heterogeneity, makes the metagenome assembly a highly complex computational process (Kolmogorov et al. 2020). Therefore, we developed MRDA, a metagenome *reference-guided* data separation and *de novo* assembler, to take advantage of combining the *reference-guided* binning and chromosome-by chromosome *de novo* assembly approach. MRDA modelled the preliminary assemblies, representative genomes dataset and raw reads in a scalable multilayer graph to provide a taxonomical composition profile and recover high-quality circular bacterial genomes from metagenomic samples. Our method authorized a notable improvement in final assembly contiguity and the number of recovered circular genomes over the current tools.

Metagenome binning is a vital procedure for controlling the downstream analysis. MRDA's data separation algorithm effectively handles the multi-species correlation contigs. At the same time, the chromosome-by-chromosome assembly excludes the incorrectly mapped reads and contigs from the final assembly, which guarantees the authenticity of downstream analysis inferences and calculations. On the other hand, the absence of reference genomes for an existing species or a novel organism in the sample dataset is the primary constraint of using a *reference-guided* metagenome assembly approach (Cepeda et al. 2017). However, MRDA has the capacity to overcome this through the reassembly of reads aligned to unclassified contigs in the preliminary assembly. Consequently, in this study, MRDA recovered the highest number of complete single-molecule microbial genomes from the synthetic mixture and human stool samples compared to the standard *de novo* assembly state-of-art (**Table 4.1**).

5.1 Limitations and future perspective:

The MDM module utilizes contradictory information from various preliminary assemblies to identify the genuine chimeric contigs. However, the efficiency of the misassembly detection process is affected by the number of preliminary assemblies and algorithm diversity in the multilayer graph. Therefore, the fallacious detection of a single position influences the capability of CCM and chromosome-by-chromosome assembly. Usually, the misassembled loci show sudden and extreme change of the read-depth profile (Muggli et al. 2015). Accordingly, incorporating the read-profile of *C.hirsuta's* Nanopore dataset into the MDM module improves the MDM performance and decreases the preliminary assembly's diversity dependability. So, optimizing the

MDM algorithm to combine the read-profile will improve the performance and usability of GALA.

GALA showed a high capacity to assign the contigs to their correlated linkage groups even in low coverage circumstances. However, the LGAM module failed to produce gap-free scaffolds in three cases. First, in case of low coverage, GALA shifts to gaped assembly as a consequence of current *de novo* assembly tools performance. The second case is the absence of raw reads due to the sequencing limitations. Eventually, the assembly of very long centromeres is also limited due to the effectiveness and optimization of current *de novo* assembly tools, requiring specific tools to assemble them, e.g., CentroFlye (Bzikadze and Pevzner 2020). Therefore, we believe that developing a new tool optimized for the single chromosome assembly will advance the performance of LGAM.

MRDA relies on representative genomic information as an efficient strategy for clustering the closely related contigs into a linkage group. However, contigs from different bacterial strains may align to the same representative genome and cluster together. The structure variants, intraspecies heterozygosity and sequence divergence between various strains undermine the chromosome-by-chromosome assembly. Therefore, based on the fact that contigs from the same strain will have an even depth, incorporating the contig depth profile to MRDA's clustering algorithm would improve the clustering algorithm. Moreover, we seek to develop a strain-level clustring algorithm by leveraging the SNP information of MRDA's species-specific read bins.

5.2 Summary and conclusion:

In this dissertation, we developed GALA, a scalable gap-free chromosome-scale assembler. GALA successfully incorporates heterogeneous data from third-generation sequencing technologies and sources of scaffolding information in a multilayer graph to achieve gap-free chromosome-scale assembly. The GALA pipeline started by leveraging the preliminary assembly's contradictory alignments to detect and resolve the misassemblies through the MDM module. Next, the CCM module cluster the contigs into linkage groups, each representing a chromosome/scaffold. Finally, GALA implemented chromosome-by-chromosome assembly to simplify the assembly graph and conduct a gap-free chromosome-scale assembly. Finally, GALA implementation achieved gapless telomere-to-telomere assembly of *C.elegans, C.hirsuta, C. oligosperma,* five chromosomes of *C. resedifolia* and seven human chromosomes. Moreover, we achieved gap-free chromosome-arm scale assembly of the *A.thaliana* genome, three chromosomes

of *C. resedifolia*, four human chromosomes and the long arm of the five acrocentric human chromosomes.

We also developed MRDA, a metagenome *reference-guided* data separation and chromosome-by-chromosome *de novo* assembly module. MRDA uses a triple-layer graph to model a representative genomes dataset with preliminary assemblies and raw reads to construct circular and complete genomes from metagenome datasets. We demonstrated the advantages of using a chromosome-by-chromosome assembly strategy with synthetic and real metagenome samples to recover complete microbial genomes in comparison to the current traditional *de novo* assembly pipelines. Moreover, we proved the contiguity improvements of the recovered genomes from MRDA over the fragmented genomes of the preliminary assemblies.

In the era of telomere-to-telomere assembly, we believe that GALA and the chromosome-by-chromosome assembly strategy provide a promising solution to simplify the assembly graph of complex genomes, expanding the high-quality genomic resources for comparative studies. This is largely because the flexibility of multilayer graph approach enables the effective leveraging of diverse sequencing technologies. On the other hand, the MRDA provides an opportunity to accurately investigate the microbial community's genomic composition. This is mainly because using a representative reference genomes dataset improves the contig binning algorithm, while the unmapped contigs indicate new species existence.

References

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* **20**: 224.

Bzikadze AV, Pevzner PA. 2020. Automated assembly of centromeres from ultra-long error-prone reads. *Nat Biotechnol* **38**: 1309-1316.

Cepeda V, Liu B, Almeida M, Hill CM, Koren S, Treangen TJ, Pop M. 2017. MetaCompass: Reference-guided Assembly of Metagenomes. *bioRxiv*.

Desai A, Marwah VS, Yadav A, Jha V, Dhaygude K, Bangar U, Kulkarni V, Jere A. 2013. Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PLoS One* **8**: e60204.

Dolezel J, Vrana J, Safar J, Bartos J, Kubalakova M, Simkova H. 2012. Chromosomes in the flow to simplify genome analysis. *Funct Integr Genomics* **12**: 397-416.

Gan X, Hay A, Kwantes M, Haberer G, Hallab A, Ioio RD, Hofhuis H, Pieper B, Cartolano M, Neumann U et al. 2016. The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. *Nat Plants* **2**: 16167.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol* **15**: e1007273.

Hay AS, Pieper B, Cooke E, Mandakova T, Cartolano M, Tattersall AD, Ioio RD, McGowan SJ, Barkoulas M, Galinha C et al. 2014. Cardamine hirsuta: a versatile genetic system for comparative studies. *Plant J* **78**: 1-15.

Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC et al. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* **7**: 399.

International Wheat Genome Sequencing C. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* **345**: 1251788.

Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J, Jones SJM et al. 2018. Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* **19**: 393.

Jayakumar V, Sakakibara Y. 2019. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief Bioinform* **20**: 866-876.

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL et al. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**: 1103-1110.

Koren S, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2019. Reply to 'Errors in long-read assemblies can critically affect protein prediction'. *Nat Biotechnol* **37**: 127-128.

Kuderna LFK, Solis-Moruno M, Batlle-Maso L, Julia E, Lizano E, Anglada R, Ramirez E, Bote A, Tormo M, Marques-Bonet T et al. 2019. Flow Sorting Enrichment and Nanopore Sequencing of Chromosome 1 From a Chinese Individual. *Front Genet* **10**: 1315.

Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, Tan Y, Li X, Lai Q, Han L et al. 2020. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *Gigascience* **9**.

Lischer HEL, Shimizu KK. 2017. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics* **18**: 474.

McKim SM, Routier-Kierzkowska AL, Monniaux M, Kierzkowski D, Pieper B, Smith RS, Tsiantis M, Hay A. 2017. Seasonal Regulation of Petal Number. *Plant Physiol* **175**: 886-903.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat Commun* **9**: 541.

Molina-Mora JA, Campos-Sanchez R, Rodriguez C, Shi L, Garcia F. 2020. High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. *Sci Rep* **10**: 1392.

Muggli MD, Puglisi SJ, Ronen R, Boucher C. 2015. Misassembly detection using paired-end sequence reads and optical mapping data. *Bioinformatics* **31**: i80-88.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *bioRxiv*.

Rellstab C, Zoller S, Sailer C, Tedder A, Gugerli F, Shimizu KK, Holderegger R, Widmer A, Fischer MC. 2020. Genomic signatures of convergent adaptation to Alpine environments in three Brassicaceae species. *Mol Ecol* **29**: 4350-4365.

Thompson OA, Snoek LB, Nijveen H, Sterken MG, Volkers RJ, Brenchley R, Van't Hof A, Bevers RP, Cossins AR, Yanai I et al. 2015. Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856. *Genetics* **200**: 975-989.

Tomaszkiewicz M, Rangavittal S, Cechova M, Campos Sanchez R, Fescemyer HW, Harris R, Ye D, O'Brien PC, Chikhi R, Ryder OA et al. 2016. A time- and cost-effective strategy to sequence mammalian Y Chromosomes: an application to the de novo assembly of gorilla Y. *Genome Res* **26**: 530-540.

Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. 2018. MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res* **28**: 266-274.

Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, Smith CL, Edgley ML, Rougvie AE, Fire AZ et al. 2019. Recompleting the Caenorhabditis elegans genome. *Genome Res* **29**: 1009-1022.

Yuan Y, Bayer PE, Scheben A, Chan CK, Edwards D. 2017. BioNanoAnalyst: a visualisation tool to assess genome assembly quality using BioNano data. *BMC Bioinformatics* **18**: 323.

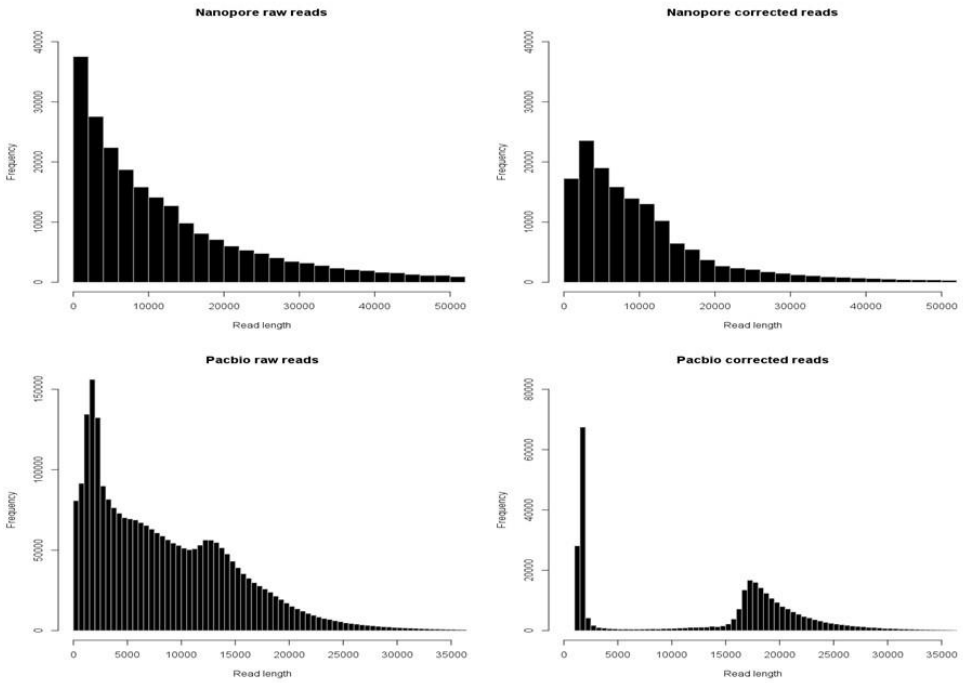
Zhang X, Zhang S, Zhao Q, Ming R, Tang H. 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* **5**: 833-845.

Supplementary

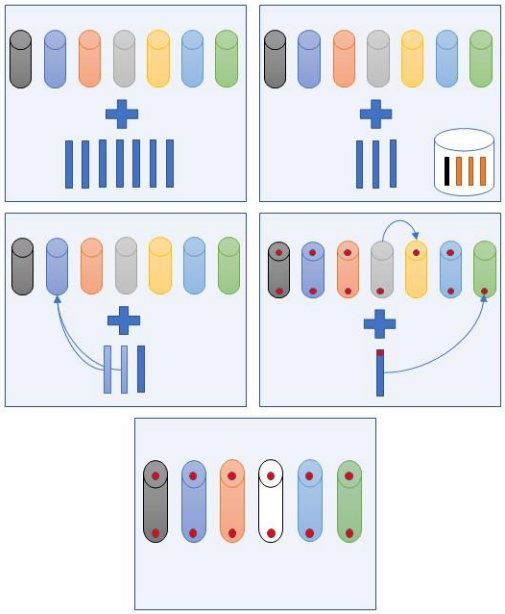
Chapter2

	Number of scaffolds				N50 of the assembly			
Coverage (X)	Flye	GALA without Hi-C	Flye/Hi-C	GALA/Hi-C	Flye	GALA without Hi-C	Flye/Hi-C	GALA/Hi-C
20	652	96	192	14	260,780	1,851,699	1,735,613	15,966,384
30	374	41	149	18	659,879	4,165,642	1,867,656	8,477,738
40	68	26	44	17	2,281,700	6,249,032	8,122,814	14,150,196
50	60	28	45	15	3,047,053	5,364,198	6,220,146	14,282,330
60	51	33	38	22	3,568,950	5,196,967	6,552,095	14,132,688
70	50	23	41	16	4,016,141	6,275,636	5,443,425	14,186,604
80	47	27	41	20	4,159,244	5,389,667	5,896,181	6,825,305
90	46	27	40	15	4,044,388	5,879,863	4,783,642	14,178,668
100	42	28	35	16	4,209,404	6,851,827	6,558,522	14,175,551
150	53	26	43	16	3,620,622	6,588,460	6,557,682	15,337,460

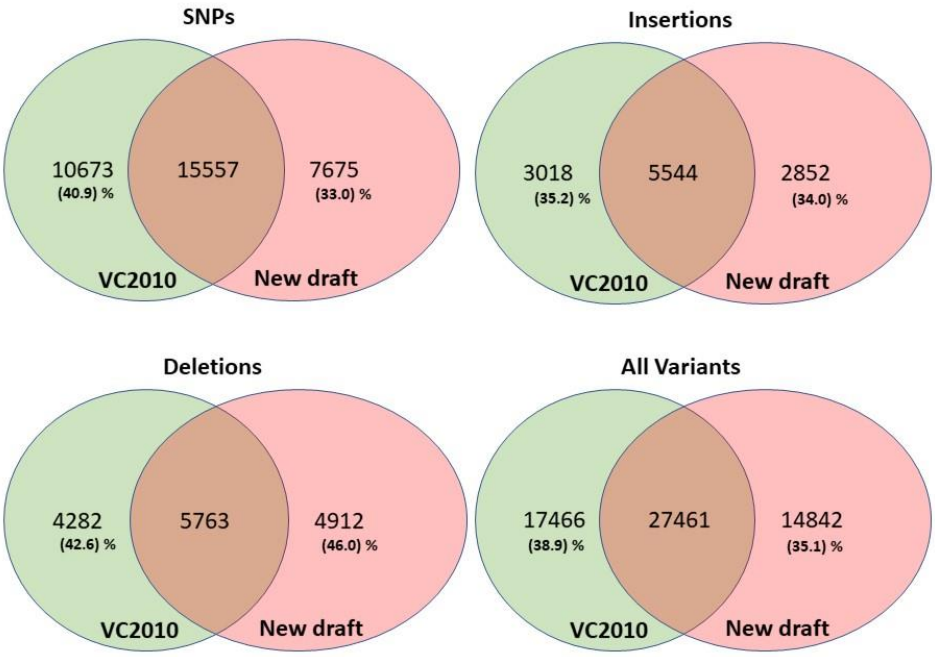
Supplementary Table 1. The statistics of *C. elegans* assemblies using different coverages of Pacbio Sequencing data by Flye and GALA with or without Hi-C data. The statistics for gapped assemblies are shown in blue.



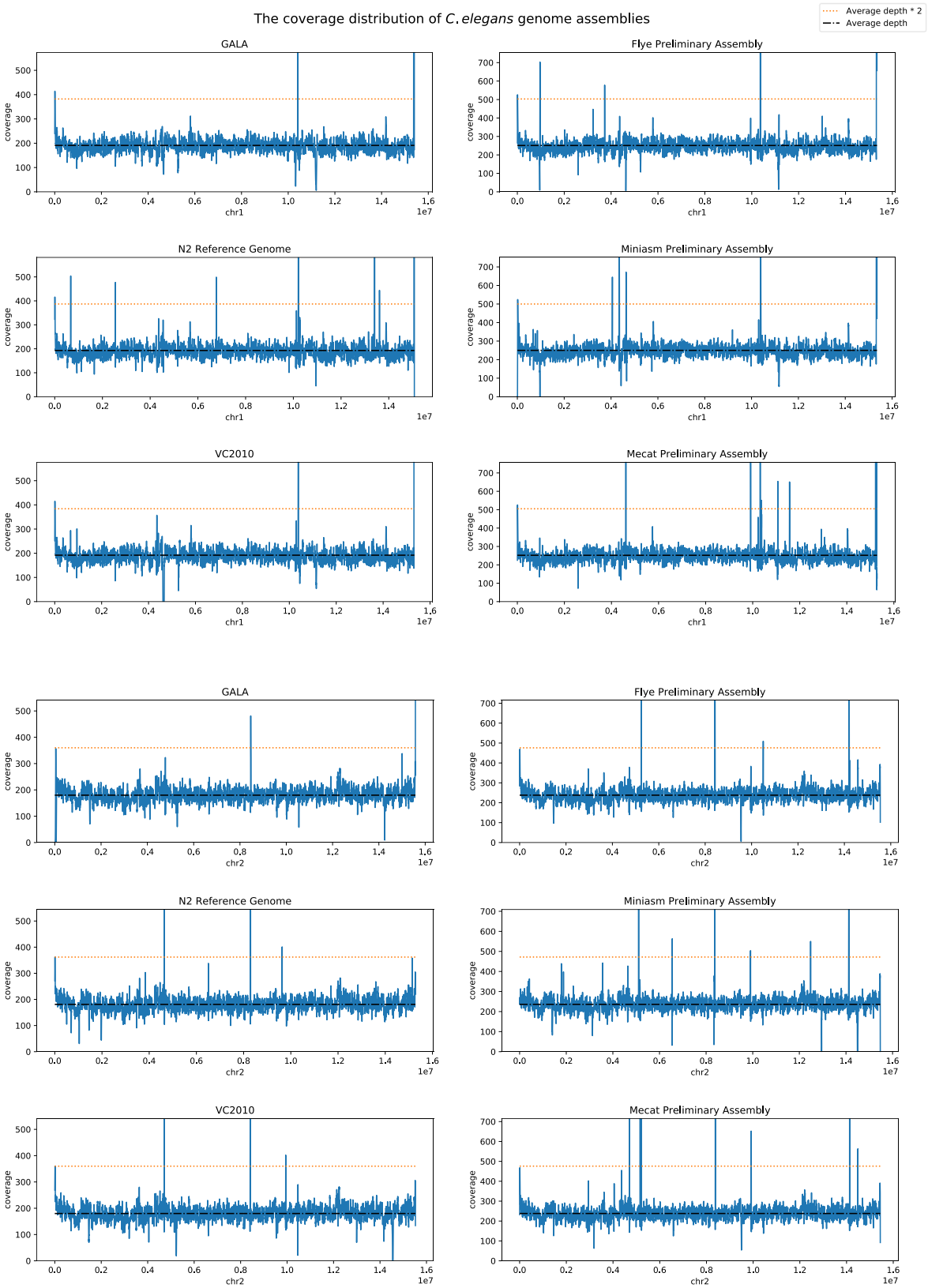
Supplementary Figure 1. The distributions of the length of the reads used for the assembly of *C. elegans* genome. (a) Nanopore raw reads. (b) The self-corrected Nanopore reads by canu (Koren et al. 2017). (c) Pacbio raw reads. (d) The self-corrected reads by canu (Koren et al. 2017).

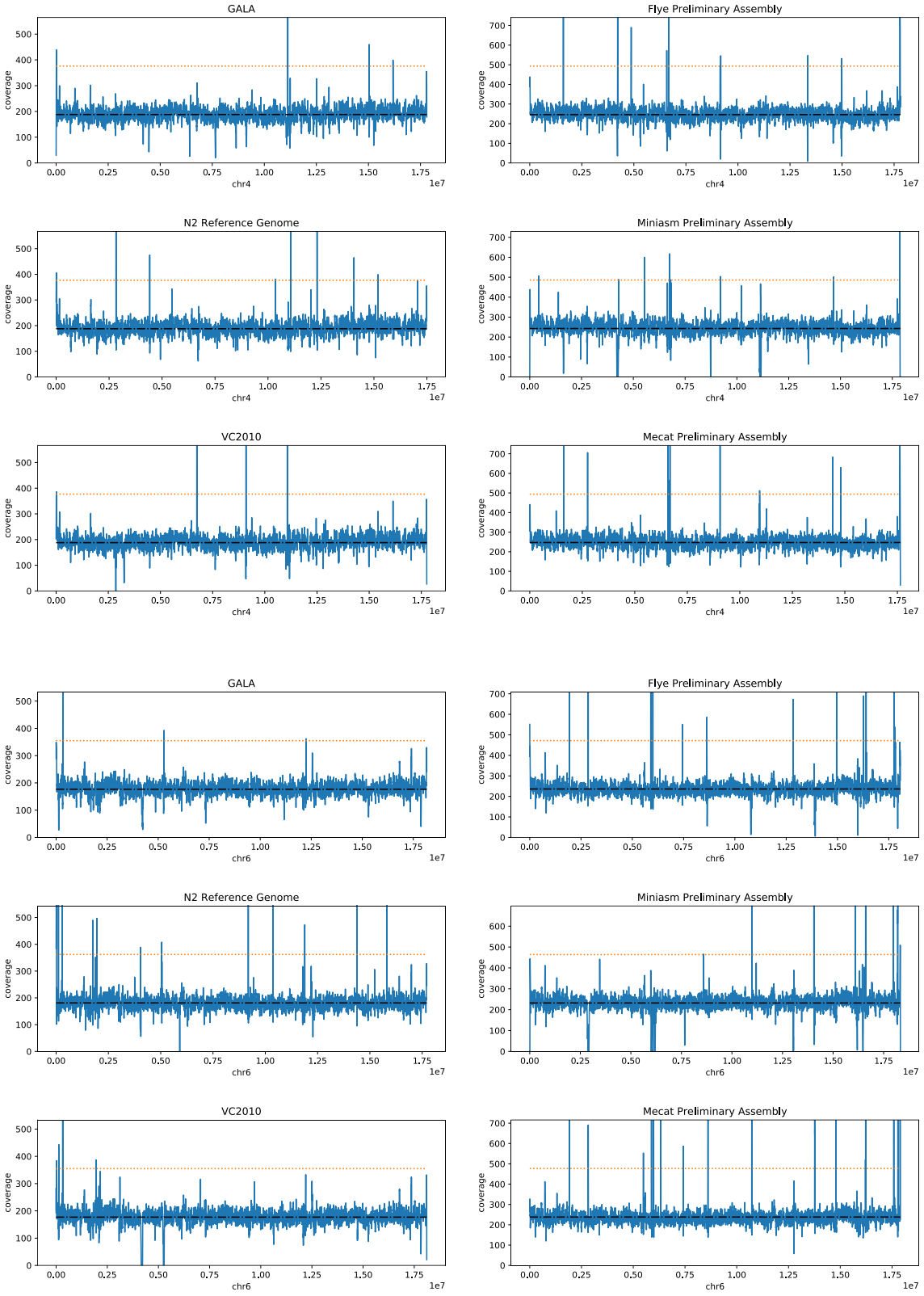


Supplementary Figure 2. The telomeric motif based analyses to merge several linkage groups in *C. elegans*. (a) GALA was applied on preliminary assemblies and raw reads, and produced seven scaffolding groups and seven short continuous contigs. (b) NCBI-blast showed that three contigs are from bacterial contamination (orange), a 13 kbp mitochondrial genome (black). (c) Two of the remaining contigs were anchored to one of the linkage group by Miniasm/Nanopore assembly. (d) Four linkage groups had telomere motif at both terminals indicating the complete chromosome; two (grey and yellow) had only one telomere and were from a single chromosome indicated by their sizes; the last group (green) had only one telomere and the missing telomere appeared on the remaining short contig. (e) The chromosome-by-chromosome assembly successfully assembled each linkage group.

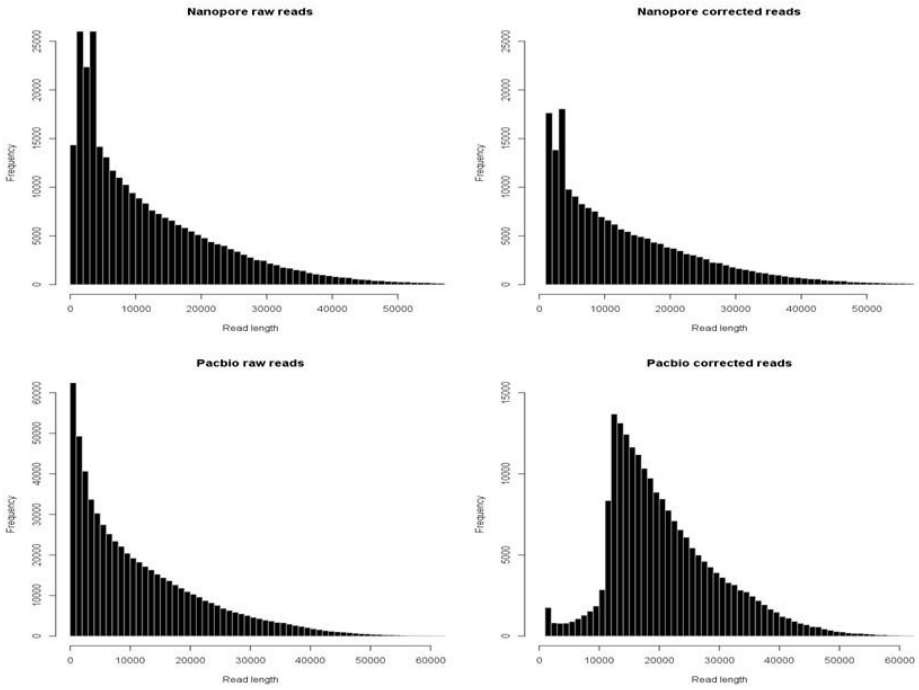


Supplementary Figure 3. Variant calling of *C.elegans* VC2010 and our new assembly against N2 reference. BWA (Li and Durbin 2009) used for mapping and Denom (Gan et al. 2011) used for variant calling.

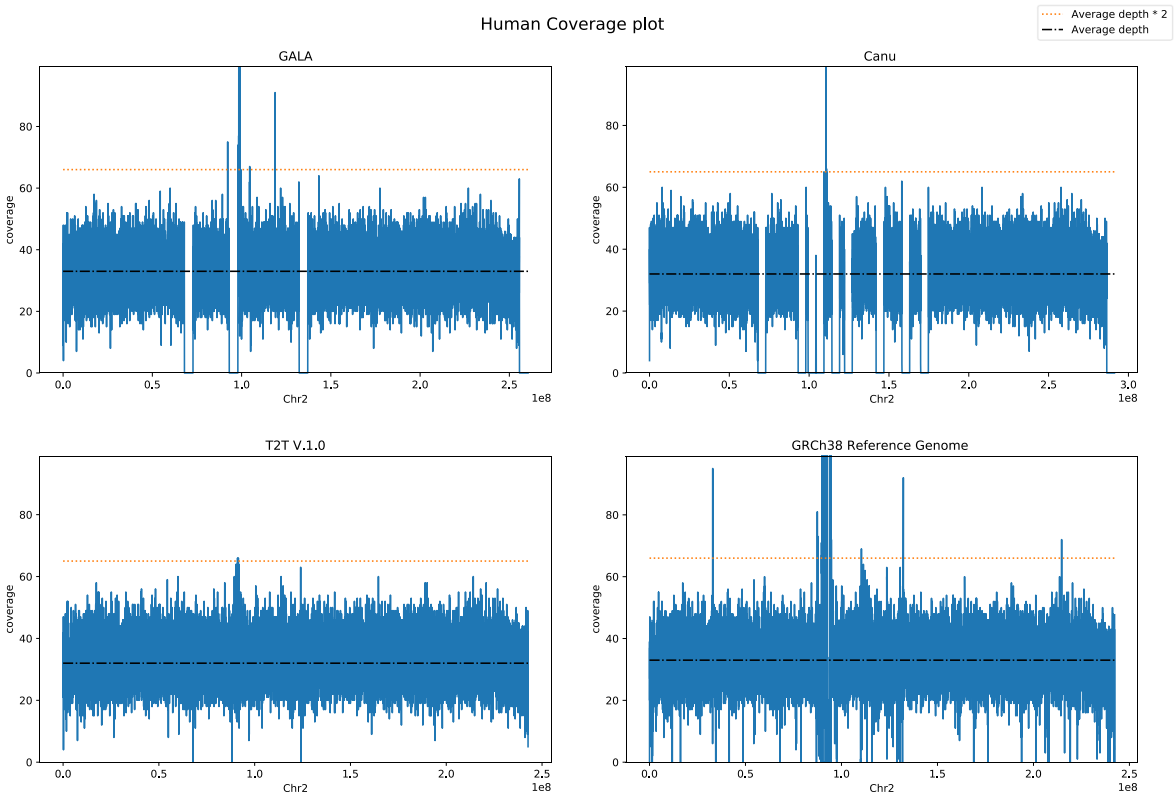

The coverage distribution of C. elegans genome assemblies
Average depth * 2
Average depth
GALA
Flye Preliminary Assembly
N2 Reference Genome
Miniasm Preliminary Assembly
VC2010
Mecat Preliminary Assembly
GALA
Flye Preliminary Assembly
N2 Reference Genome
Miniasm Preliminary Assembly
VC2010
Mecat Preliminary Assembly
coverage
chr1
chr2
1e7

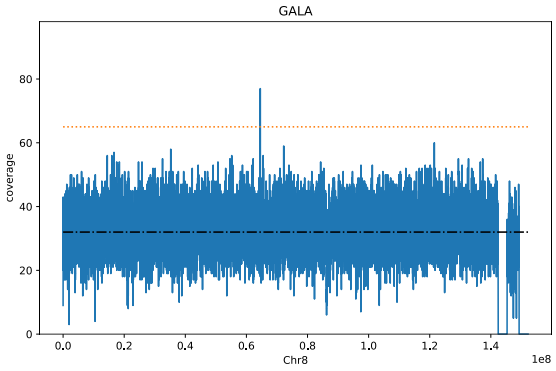


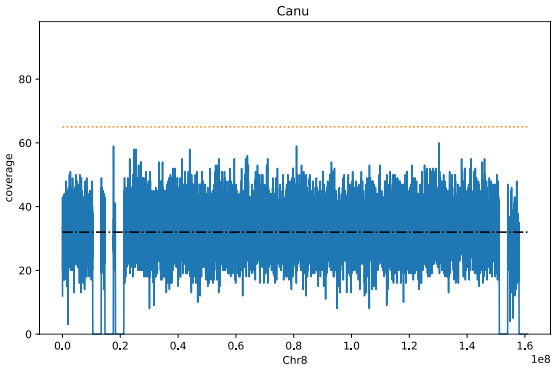
Supplementary Figure 4. Distributions of the depth-of-coverage when aligning the raw Pacbio reads from *C. elegans* genome to the Gala assembly, N2 reference genome, VC2010 assembly, Flye preliminary assembly, Miniasm preliminary assembly and Mecat preliminary assembly. For simplicity, only chr1, chr2, chr4 and chr6 are shown here. The GALA assembly shows better performance than the preliminary assemblies and N2 reference genome and are comparable to the VC2010 assembly.

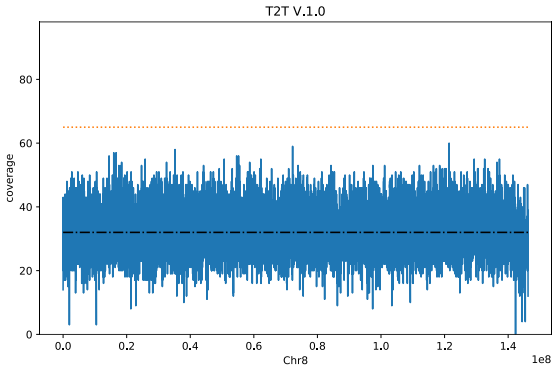


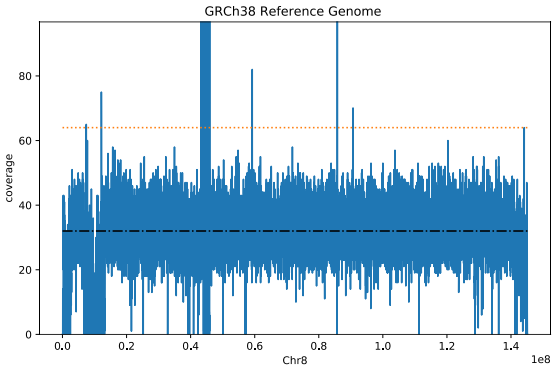
Supplementary Figure 5. The distributions of the length of the reads used for the assembly of *A. thaliana* genome. (a) Nanopore raw reads. (b) The self-corrected Nanopore reads by canu (Koren et al. 2017). (c) Pacbio raw reads. (d) the self-corrected Pacbio reads by canu (Koren et al. 2017).

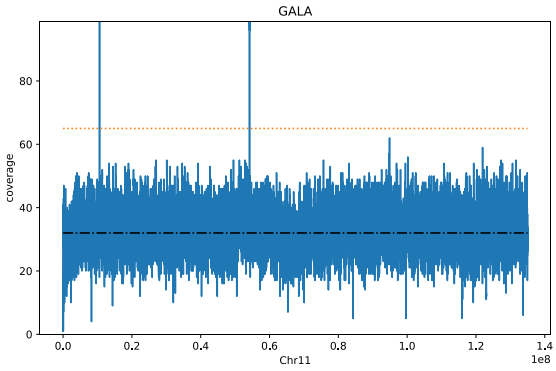


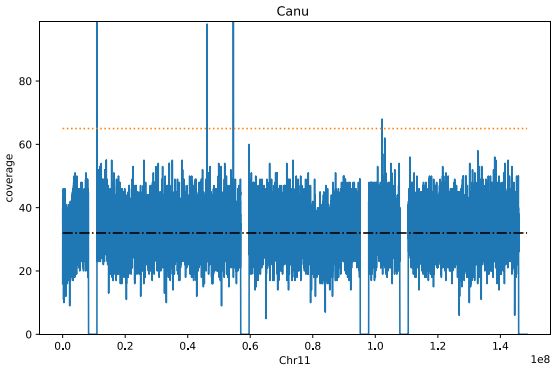

GALA
coverage
Chr8
1e8

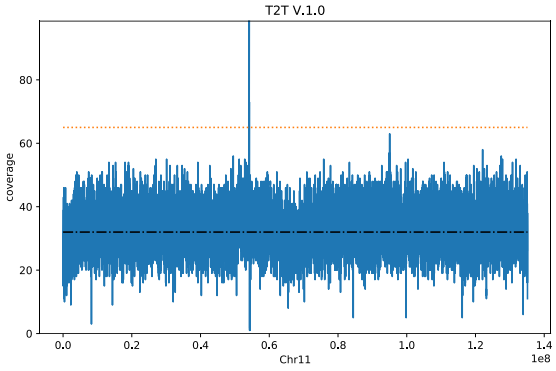

Canu
coverage
Chr8
1e8

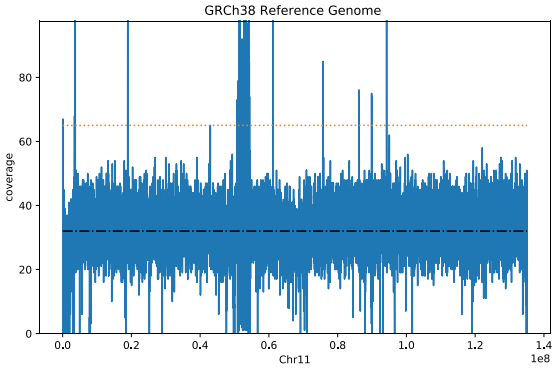

T2T V.1.0
coverage
Chr8
1e8

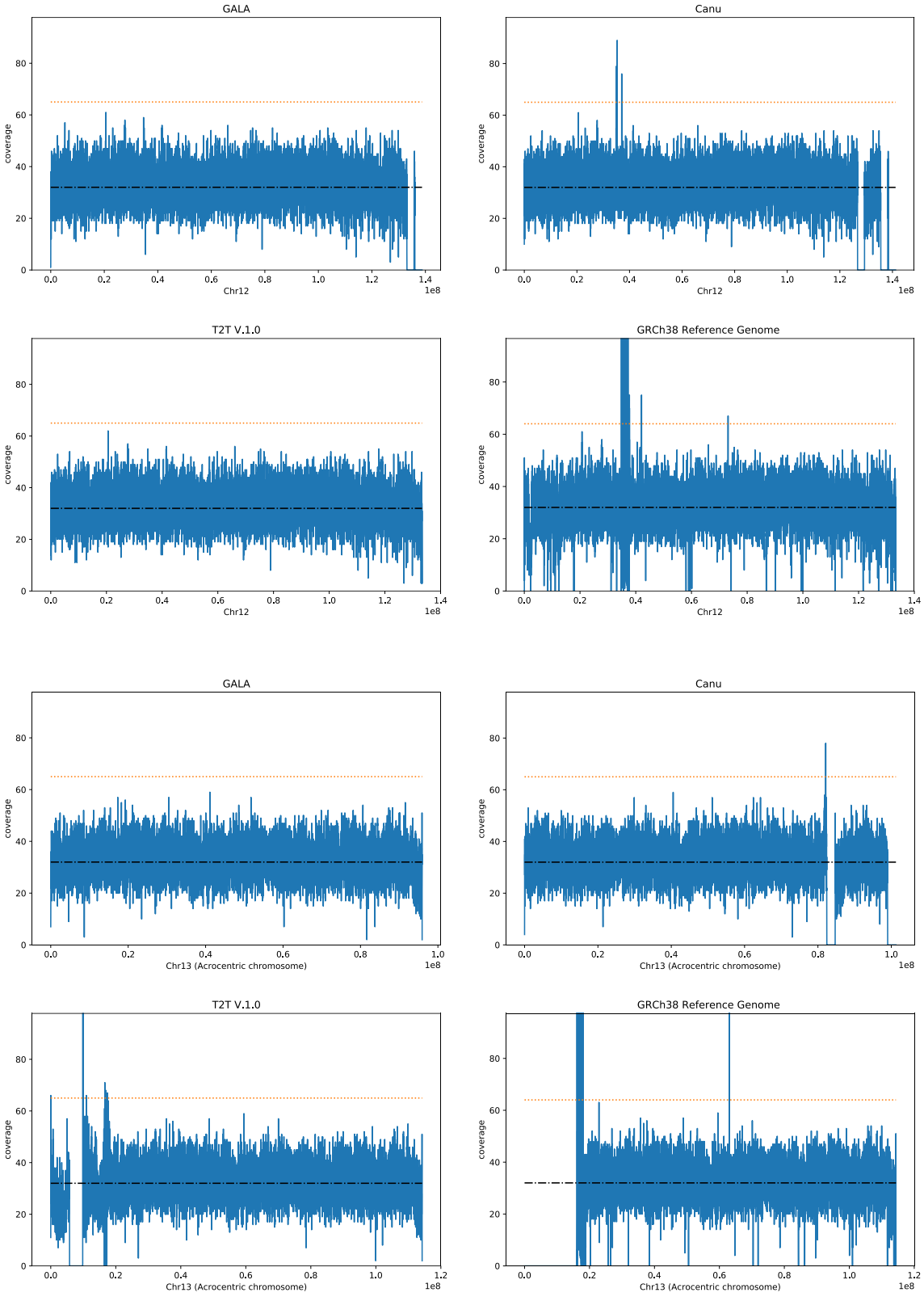

GRCh38 Reference Genome
coverage
Chr8
1e8

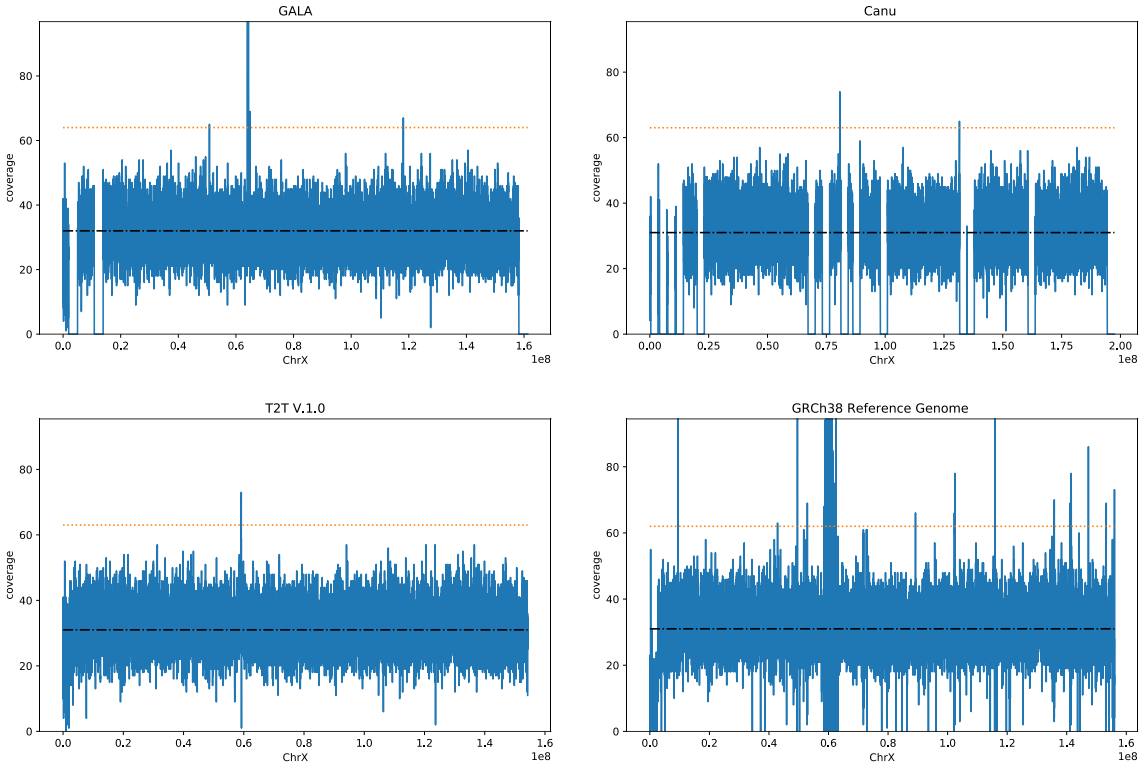

GALA
coverage
Chr11
1e8


Canu
coverage
Chr11
1e8

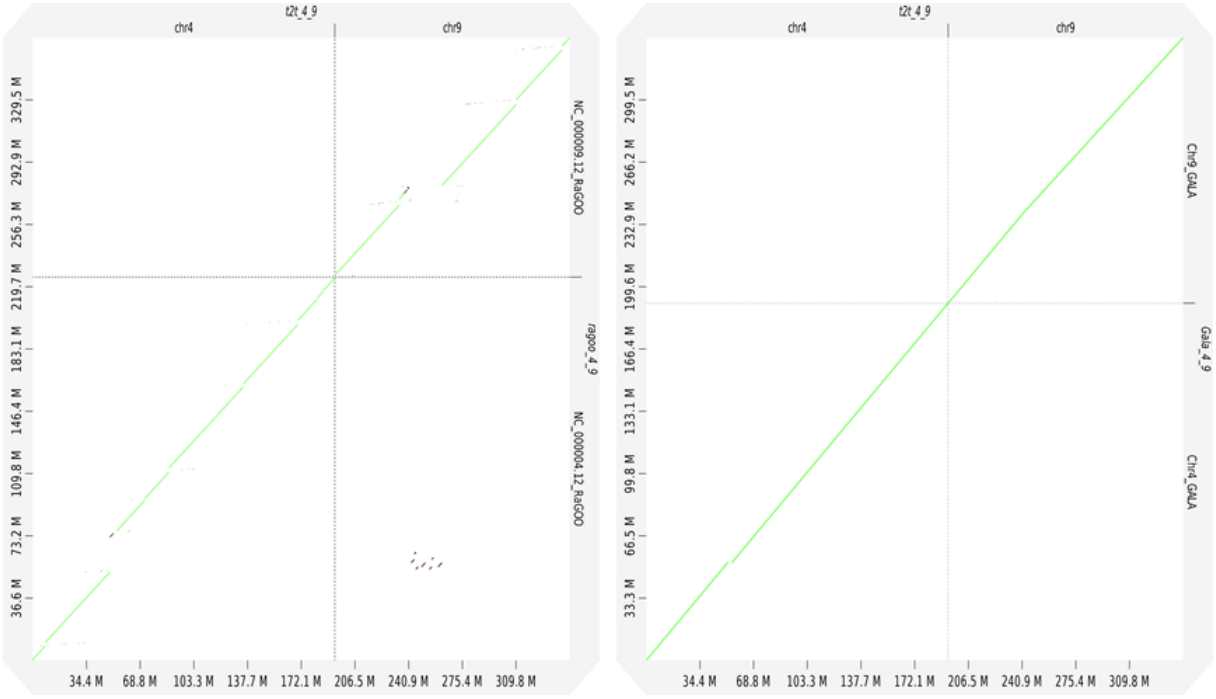

T2T V.1.0
coverage
Chr11
1e8


GRCh38 Reference Genome
coverage
Chr11
1e8


GALA
Canu
T2T V.1.0
GRCh38 Reference Genome
coverage
Chr12
Chr13 (Acrocentric chromosome)
1e8

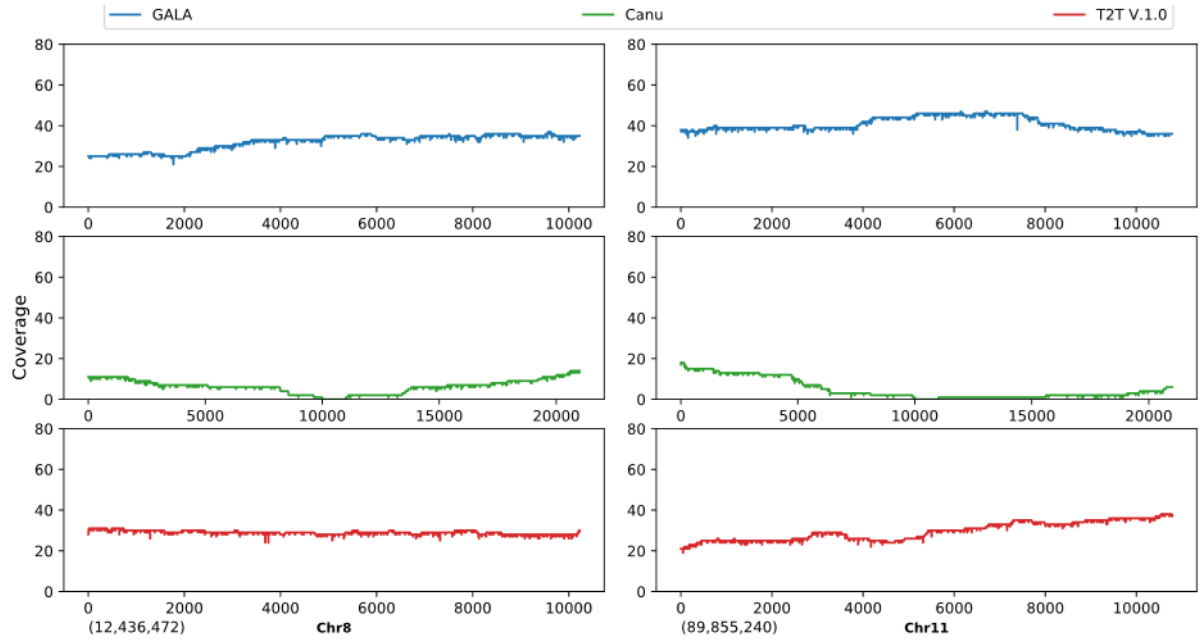


Supplementary Figure 6. Distributions of the depth-of-coverage when aligning the raw HiFi reads to the Gala assembly, final HiCanu assembly, T2T v1.0 assembly, GRCh38 human reference genome. For simplicity, only chr2, chr8, chr11, chr12, and chrX are shown here. The GALA assembly has very few gaps and shows comparable performance to the Canu assembly and the T2T v1.0 assembly. Note the final HiCanu assembly here is the one suggested by Nurk *et al* in (Nurk et al. 2020b) (by filtering out the contigs <50Kbp in "HiCanu 20kb HiFi" at https://obj.umiacs.umd.edu/marbl_publications/hicanu/chm13_20k_hicanu_hifi.fasta.gz).

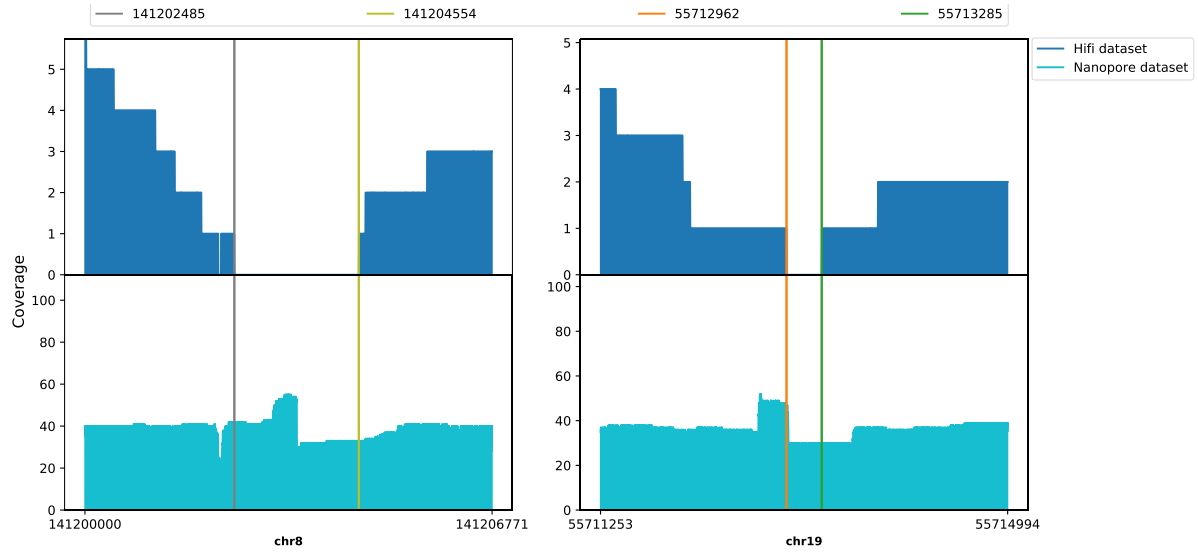


(a) **(b)**

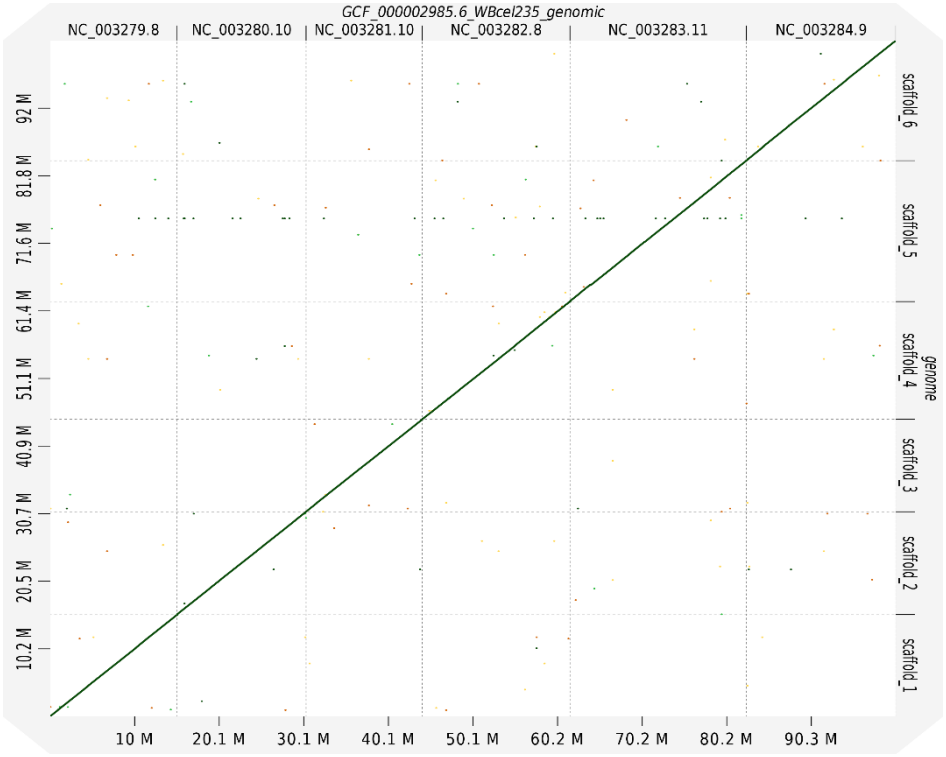
Supplementary Figure 7. Dotter plot of the alignments of chr4 and chr9 in a) the *reference-guided* scaffolding and gap-filling of preliminary HiCanu assembly and b) GALA assembly.



Supplementary Figure 8: Two regions from the GALA assembly and the HiCanu assembly. Reads alignment of the GALA assembly, the HiCanu assembly and the T2T assembly indicated that GALA successfully assembled the two regions in chr8 and chr11 as good as the T2T assembly. While the preliminary assembly of HiCanu contain duplicated sequences around the contigs break point.



Supplementary Figure 9. Two regions where GALA failed to produce gap-free assembly. The upper subplots show the depth of raw HiFi reads, which are used by GALA for assembly, to the reference genome. Gaps indicate the missing of raw sequencing reads. The lower subplots show the depth of Nanopore reads from the same cell line (the sequencing data is from https://github.com/nanopore-wgs-consortium/CHM13 and is not used for our assembly), indicating the gaps are not caused by the divergence of the genome.



Supplementary Figure 10. Dotter plot of the alignments of our assembly to the *C. elegans* N2 reference genome using D-Genies (Cabanettes and Klopp 2018).

References:

Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**: e4958.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT et al. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**: 419-423.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-736.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30**: 1291-1305.

Supplementary

Chapter3

Test	item	Ref	Gala
Mapping states	Mapped reads	385454375	385897580
	Unmapped reads	4567333	4124128
	Mismatches	246397127	160738433
	Deletions	12902511	2579827
	Insertions	5946006	826499
	Indels	6956505	1753328
Variant calling stats	Total variants	114629	26559
	SNPs	101355	24253
	Indels	13274	2306
	Multiallelic sites	82	17
	Multiallelic SNPs	36	5

Supplementary Table 1: *C. hirsuta* reference genome and GALA assembly mapping and variant calling stats for 195X Illumina read dataset.

Ref-Chr	Start	End	Gala-Chr	Start	End	Length	Variant
Chr8	11354616	13185689	Chr7	1178144	3047547	1869404	translocation
Chr3	13875384	14904741	Chr5	12215122	13218590	1003469	invTranslocation
Chr6	12592804	13254050	Chr4	4356480	5033974	677495	translocation
Chr4	2185913	2446608	Chr5	7280286	7543062	262777	invTranslocation
Chr1	2092526	2211107	Chr6	4182992	4301523	118532	invTranslocation
Chr8	17382780	17442974	Chr7	415097	470619	55523	invTranslocation
Chr4	19304561	19334578	Chr1	9711746	9740173	28428	invTranslocation
Chr3	10950279	10961695	Chr5	2301653	2313404	11752	translocation
Chr3	20681296	20692467	Chr7	12057974	12069113	11140	translocation
Chr3	17524856	17535746	Chr2	7522448	7533416	10969	invTranslocation
Chr2	10912373	10922959	Chr6	3596777	3607504	10728	invTranslocation
Chr4	9226850	9237155	Chr3	13925600	13936071	10472	translocation
Chr1	13984849	13995258	Chr3	18532372	18542672	10301	translocation
Chr6	7031994	7041963	Chr3	18532129	18542377	10249	translocation

Supplementary Table 2: The GALA assembly translocation and inversion translocation events > 10Kbp.

		Pacbio RS/Sequel		Hifi
		Raw	corrected	
Cardamine hirsuta (az)	**Total bases (Gb)**	25.39	7.29	27.82
	Total reads	3581988	584299	2665432
	Read N50 (Kb)	11	12.97	10.2
	Read mean (kb)	7	12.47	10.4
	Read L50	833122	222909	1143468
	Coverage	126	36	139
Cardamine oligosperma	**Total bases (Gb)**	22.43	7.50	29.57
	Total reads	1862865	355987	1993732
	Read N50 (Kb)	17.4	21	14.8
	Read mean (kb)	12	21	14.8
	Read L50	489943	146279	936714
	Coverage	124	42	164
Cardamine resedifolia	**Total bases (Gb)**	-	-	13.40
	Total reads	-	-	1055683
	Read N50 (Kb)	-	-	12.7
	Read mean (kb)	-	-	12.6
	Read L50	-	-	498735
	Coverage	-	-	56

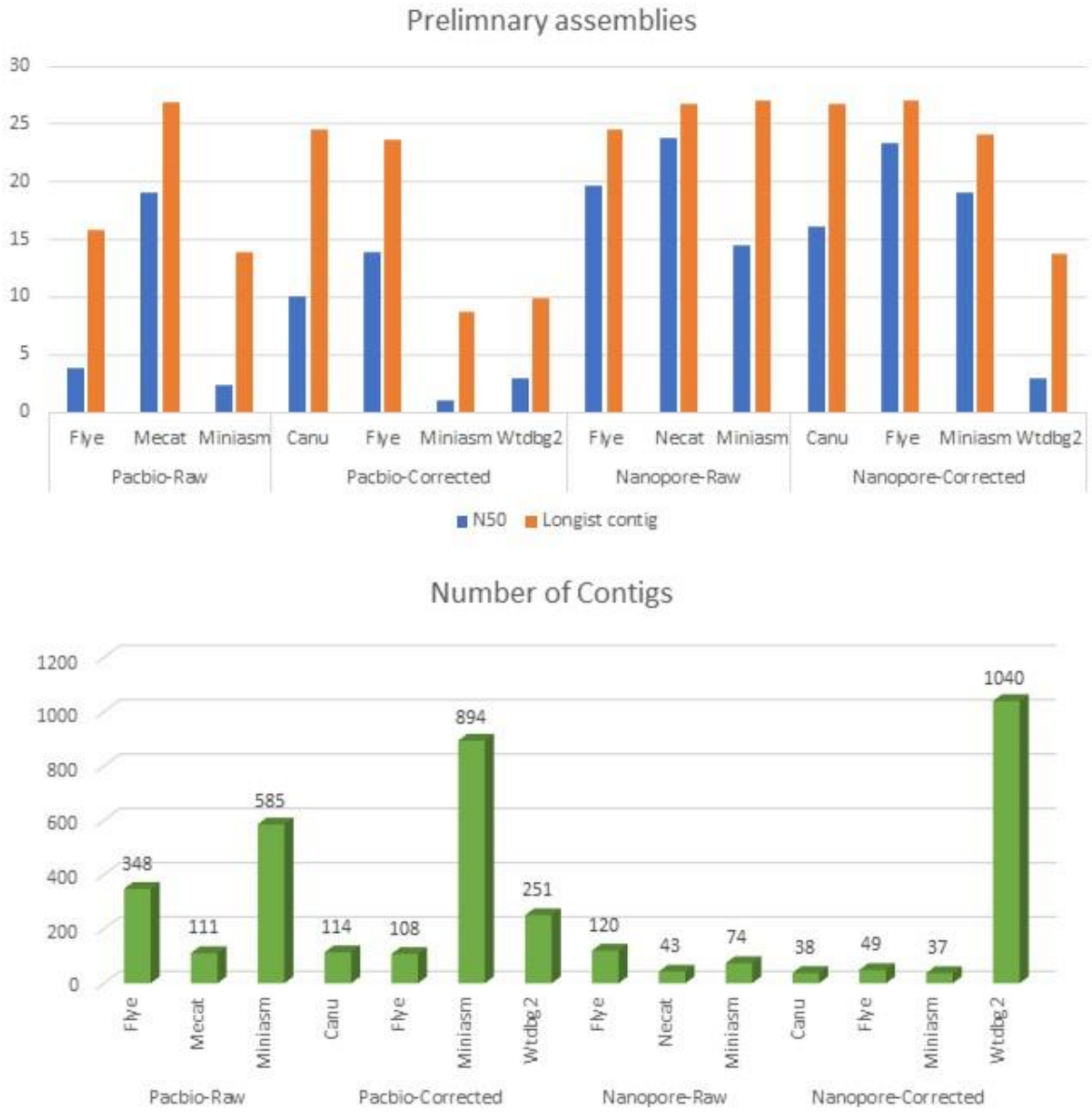
Supplementary Table 3: Pacbio reads statistics.

		Cardamine hirsuta (az)	*Cardamine oligosperma*	*Cardamine resedifolia*
Mapping	Mapped reads	201587316	313324182	313425263
	Unmapped reads	1013520	11095136	20929205
	Mismatches	30316888	55173098	145228083
	Deletions	392323	1339744	10402199
	Insertions	149212	761063	6753394
	Indels	541535	2100807	17155593
busco	Complete	1391 (96.6)	1392 (96.7)	1391 (96.6)
	Single-copy	1376 (94.9)	1371 (95.2)	1358 (94.3)
	Duplicated	24 (1.7)	21 (1.5)	33 (2.3)
	Fragmented	14 (1.0)	11 (0.8)	14 (1.0)
	Missing	35 (2.4)	37 (2.5)	35 (2.4)

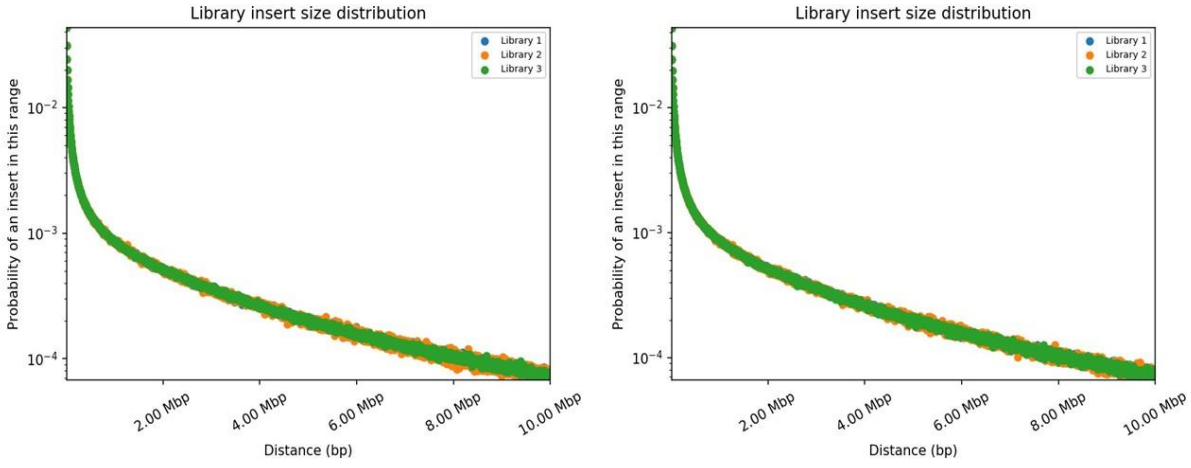
Supplementary Table 4: Mapping and Busco stats.

	Cardamine hirsuta (ox)		*Cardamine hirsuta (az)*		*Cardamine oligosperma*		*Cardamine resedifolia*	
	Length	**Percentage**	**Length**	**Percentage**	**Length**	**Percentage**	**Length**	**Percentage**
Copia	31.60	15.92	32.53	16.12	22.53	12.43	7.61	3.16
Gypsy	6.11	3.07	12.66	6.27	8.40	4.63	58.98	24.50

Supplementary Table 5: Copia and Gypsy LTR's superfamilies load.

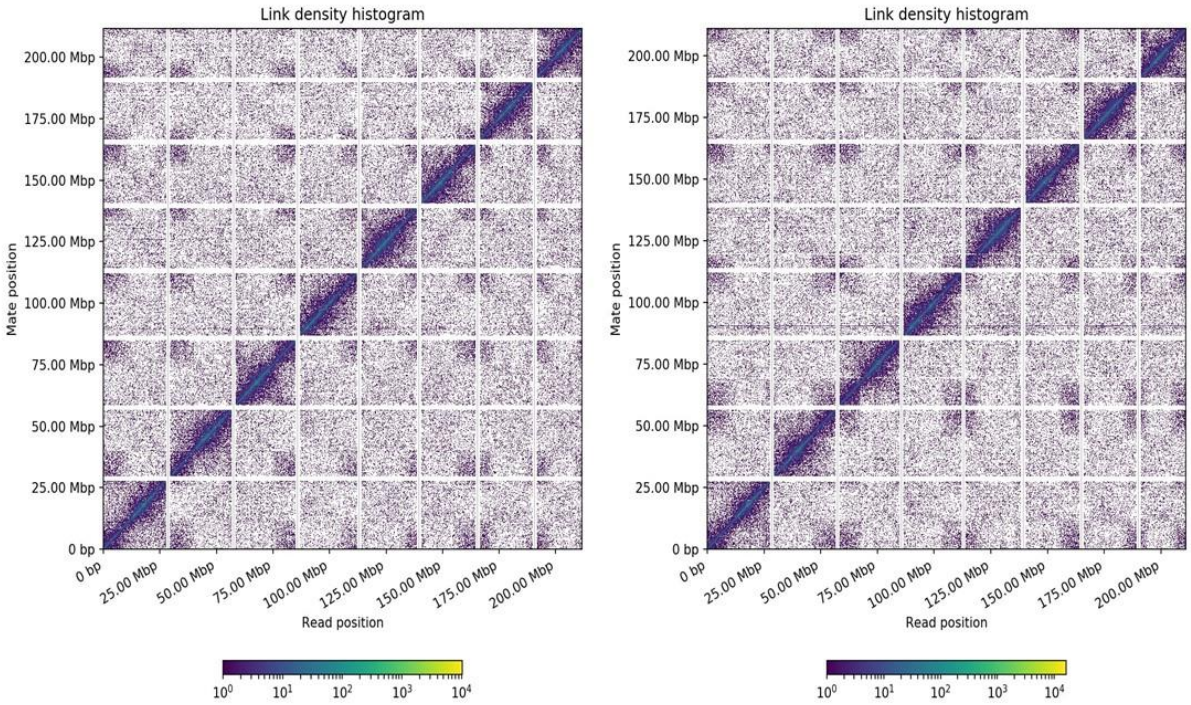


Supplementary Figure 1:Statistics of *C.hirsuta* (ox) preliminary assemblies.



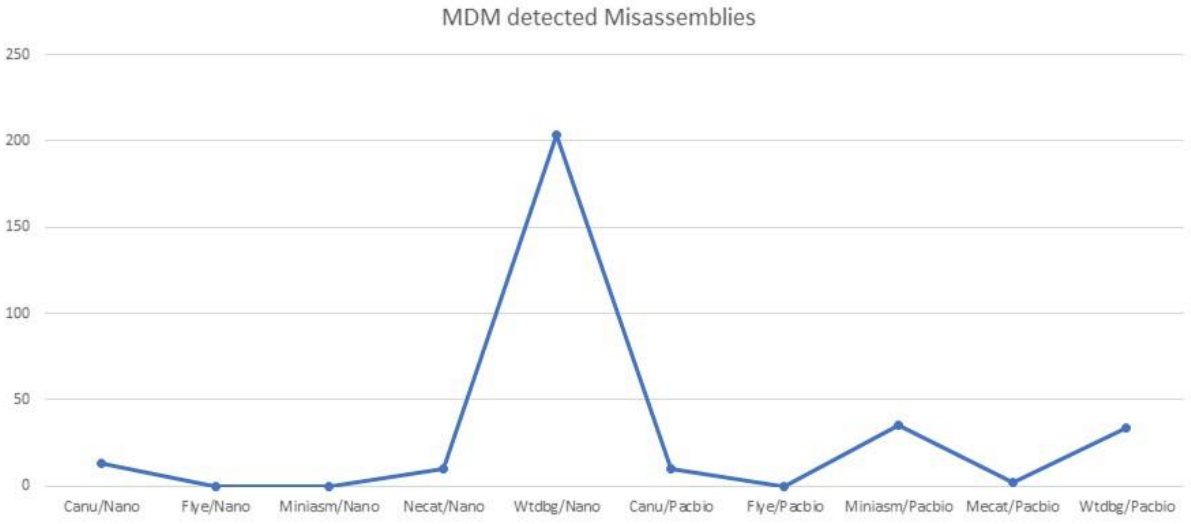
(A) **(B)**

Supplementary Figure 2: A) Insert size distribution on Flye/Nanopore draft. B) Insert size distribution on Flye/Pacbio draft

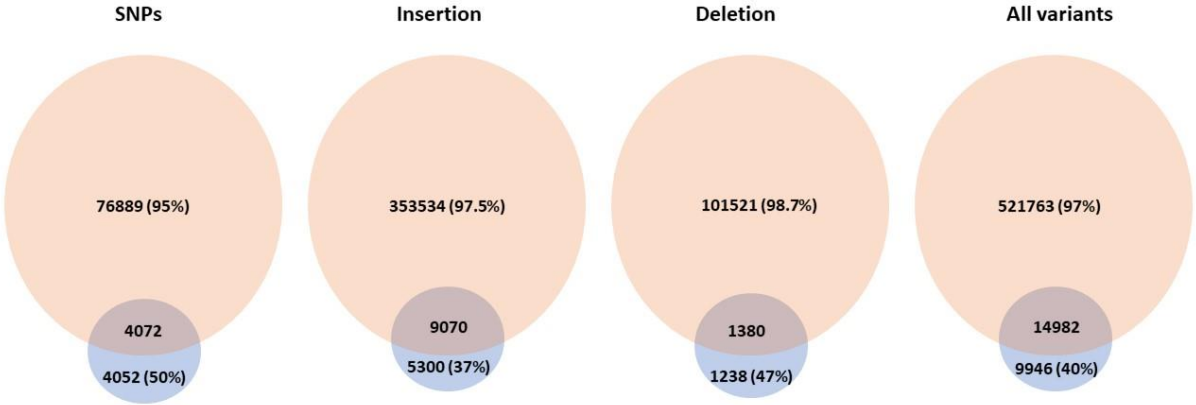


(A) **(B)**

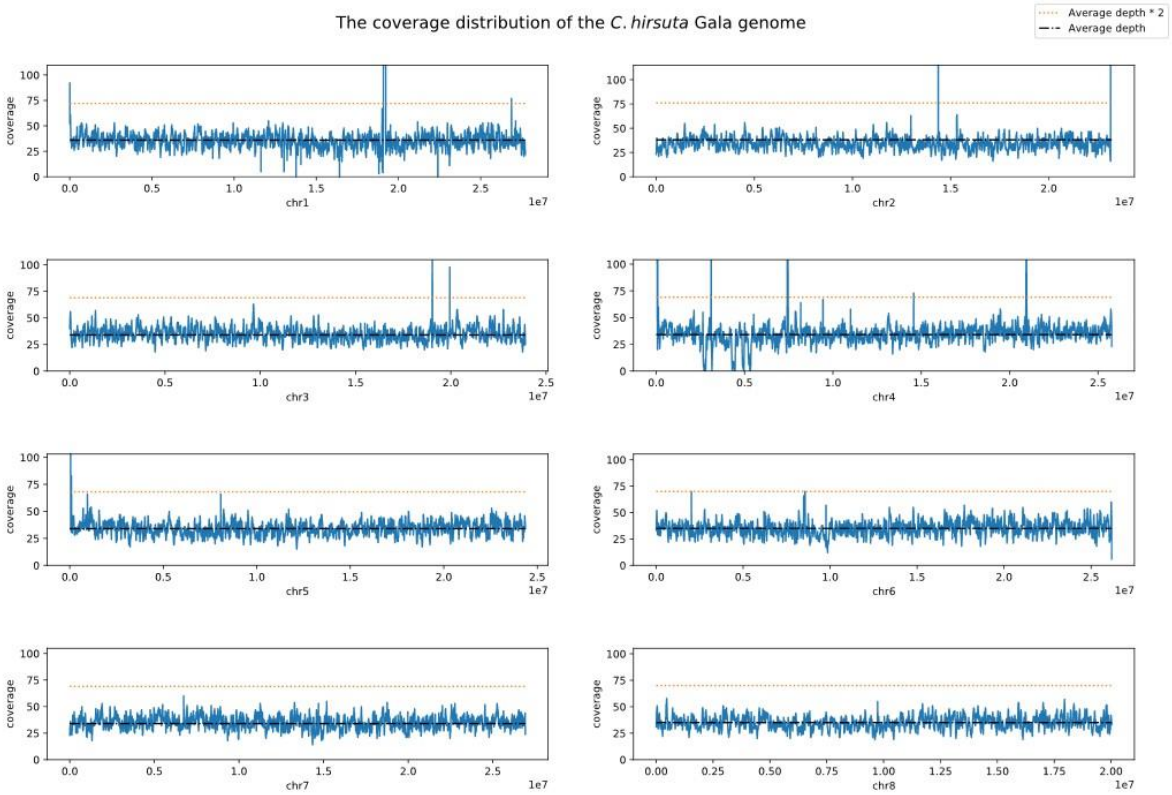
Supplementary Figure 3: A) Link density histogram on Flye/Nanopore draft. B) Link density histogram on Flye/Pacbio draft



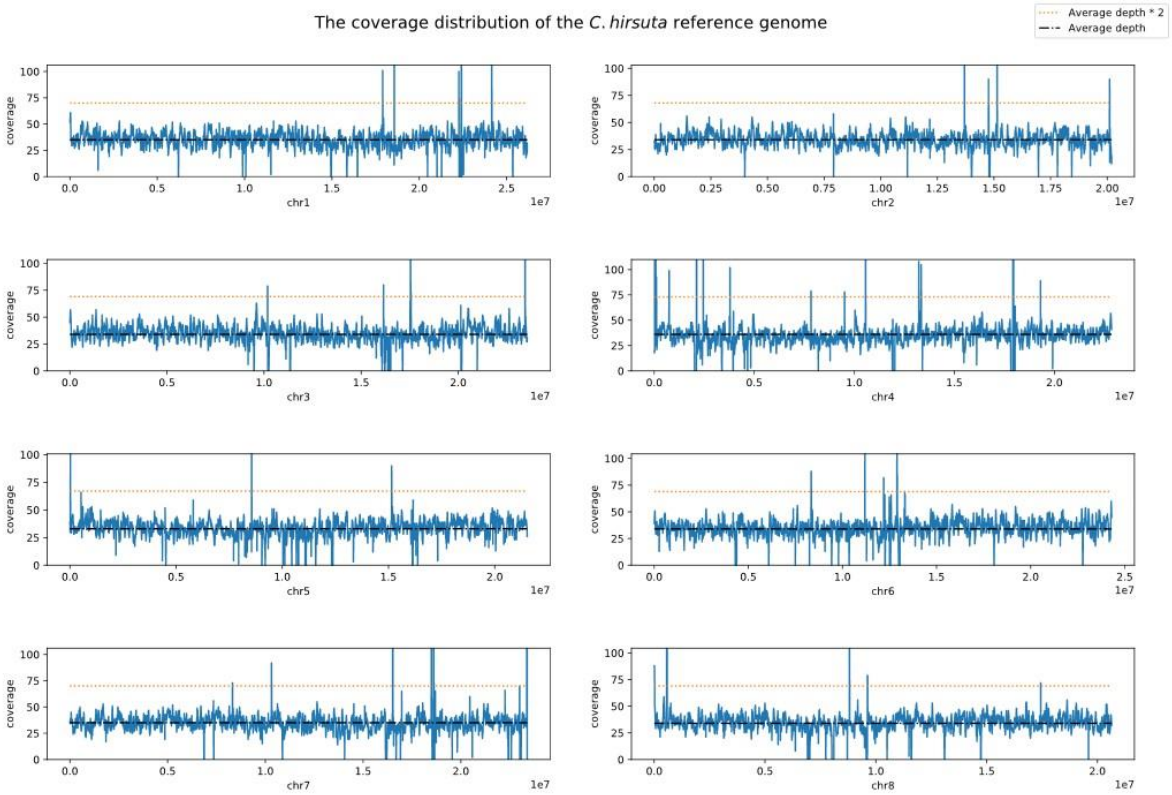
Supplementary Figure 4: Number of misassemblies detected on preliminary assemblies using MDM module.



Supplementary Figure 5: Number of variants in HiC/Nanporedraft (orange) and GALA draft (blue).

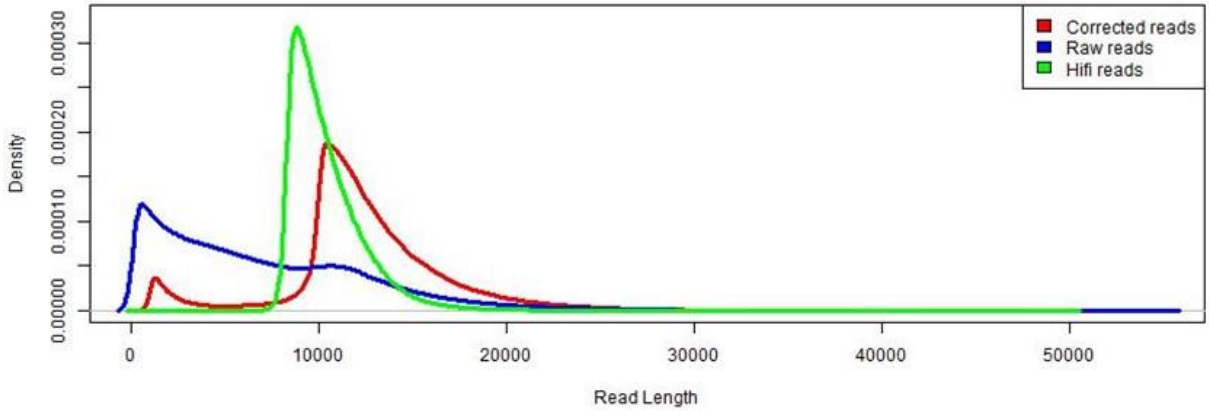


Supplementary Figure 6:

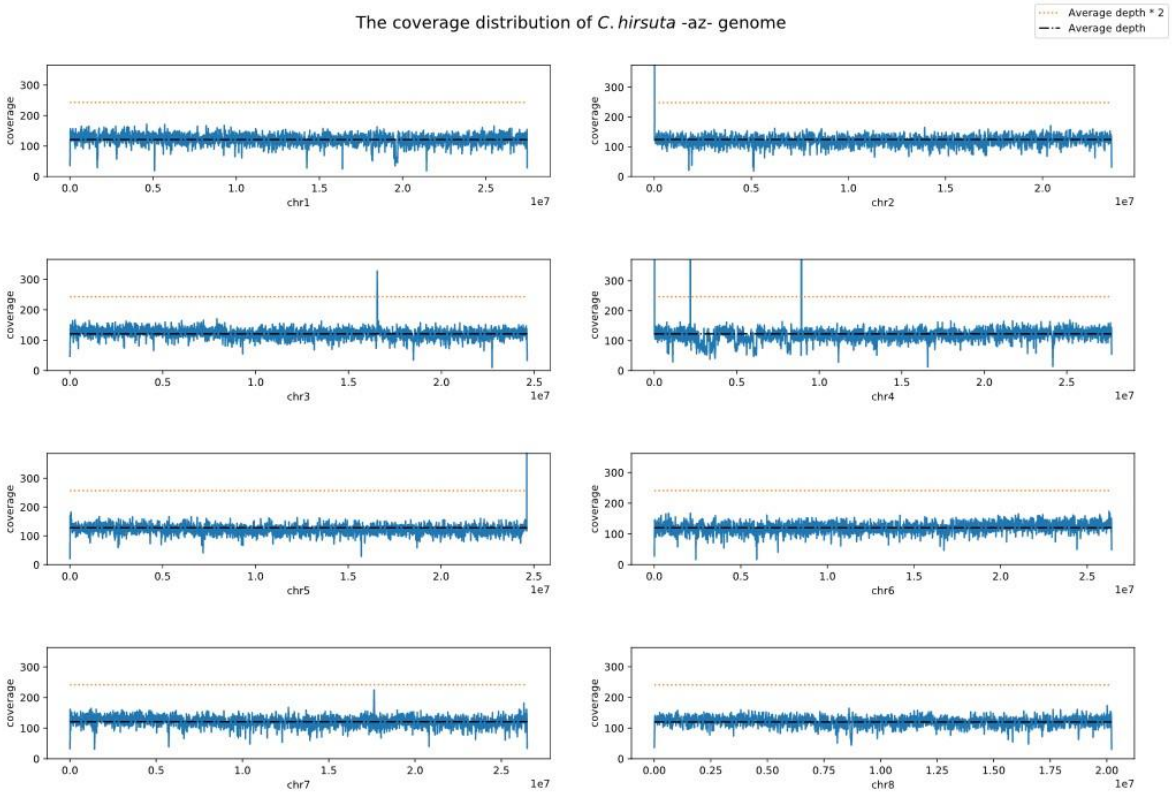


Supplementary Figure 7:

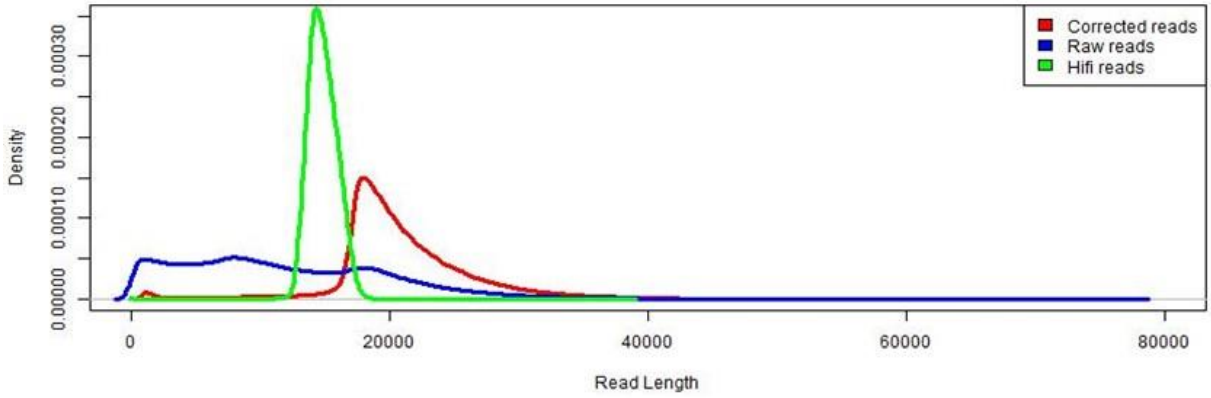
***Supplementary Figure 8: C. hirsuta* (ox) chromosomes ideogram. The red numbers represent the GALA assembly chromosomes size. While the black represents the reference genome chromosomes size.**



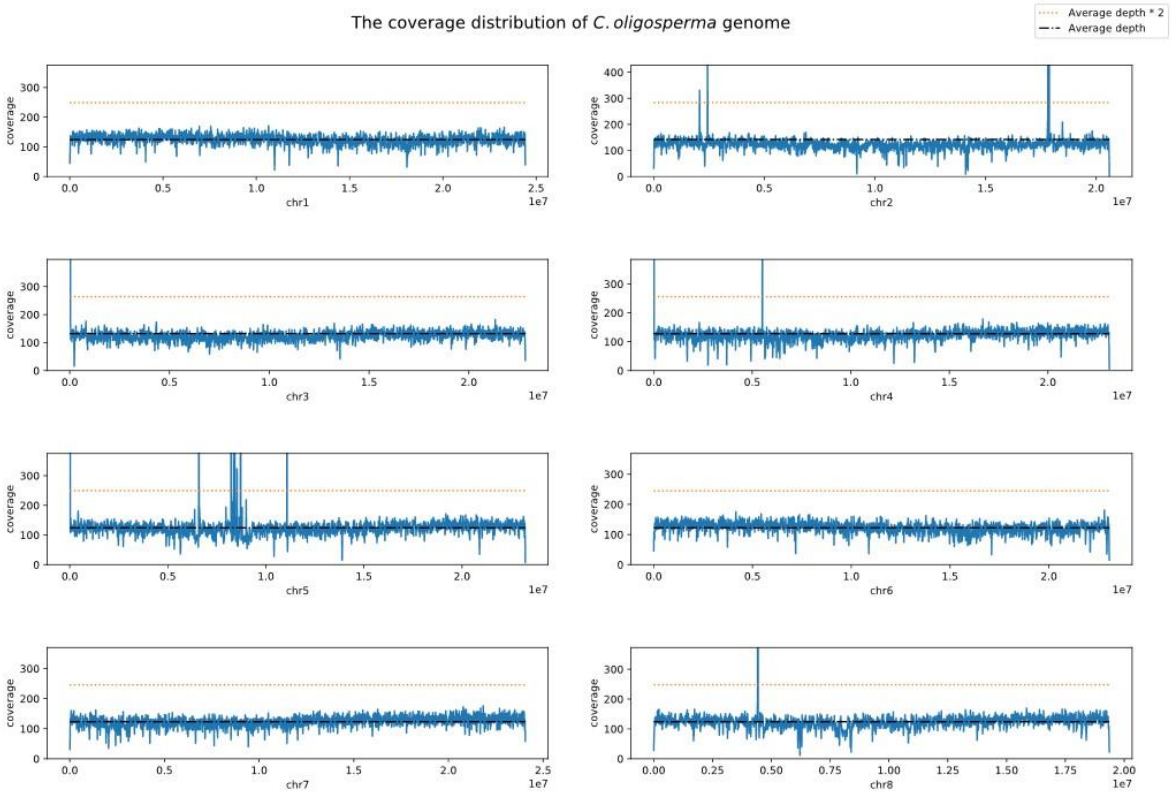
Supplementary Figure 9: *Cardamine hirsuta* (Az) long-reads datasets length distribution.



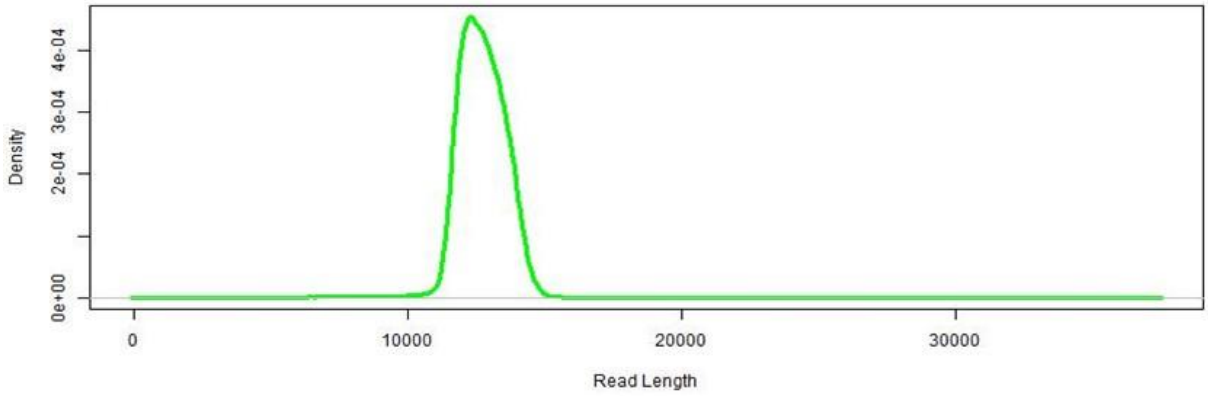
Supplementary Figure 10:



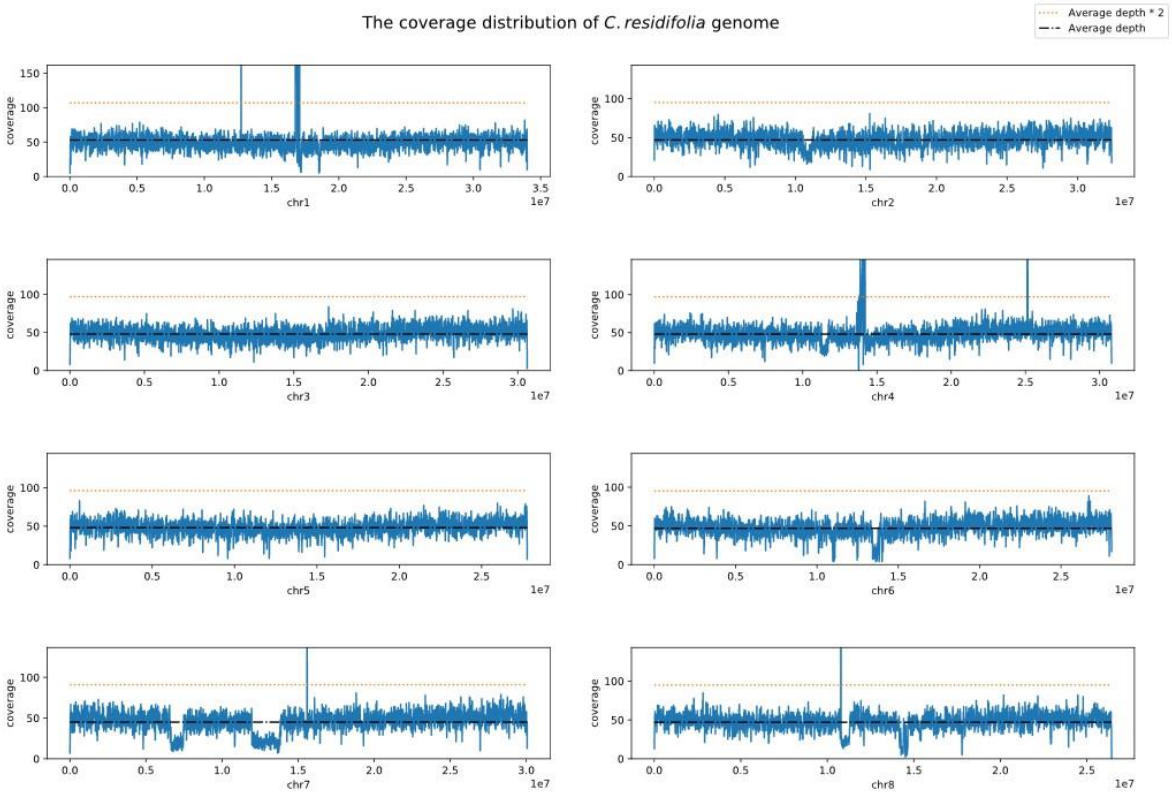
Supplementary Figure 11: *Cardamine oligosperma* long-reads datasets length distribution



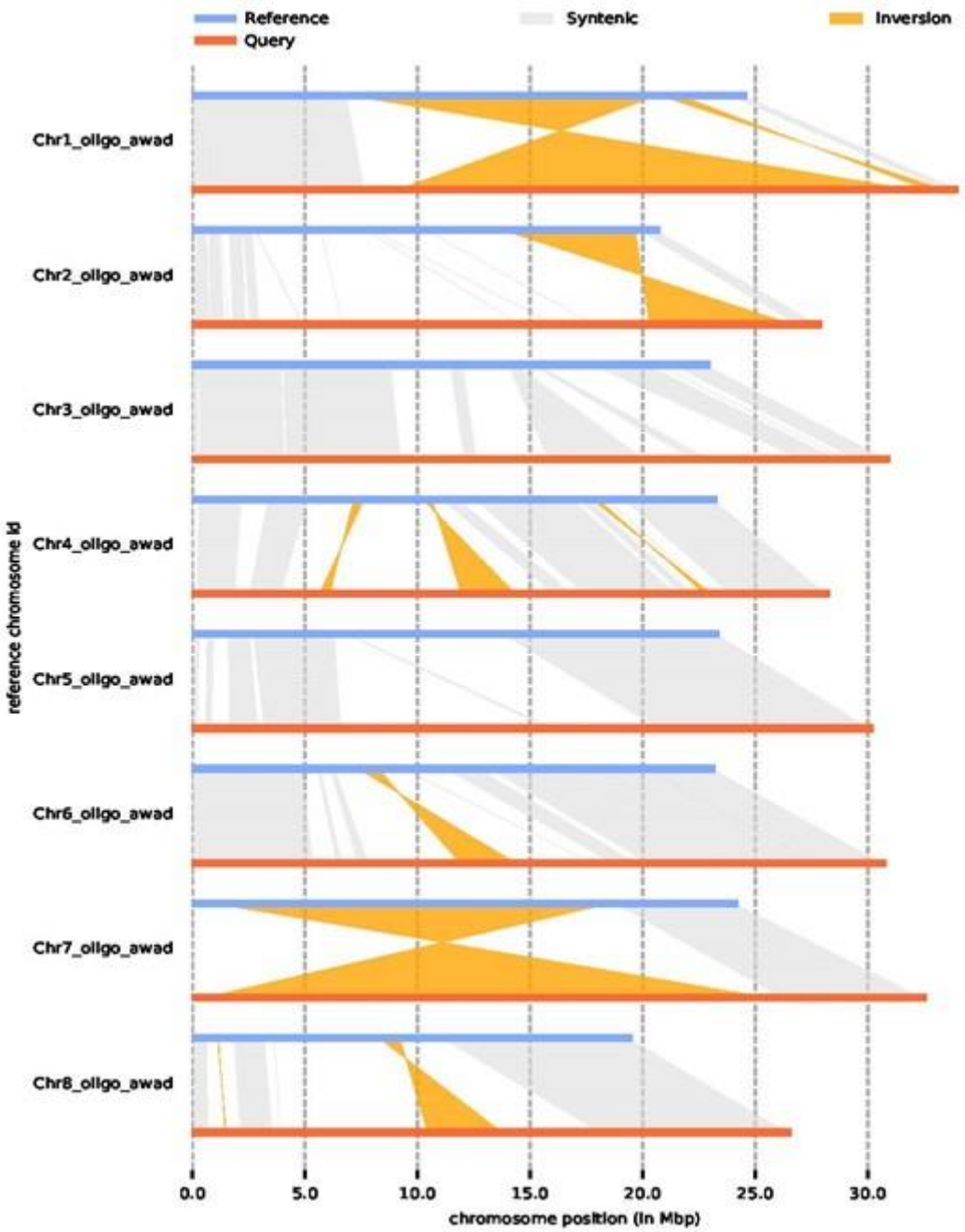
Supplementary Figure 12:



Supplementary Figure 13:*C. resedifolia* Hifi dataset length distribution



Supplementary Figure 14:



Supplementary Figure 15: The synteny information and intra-chromosomal variants between *C. oligosperma* and *C. resedifolia.*

Erklärung zur Dissertation
gemäß der Promotionsordnung vom 12. März 2020

Diese Erklärung muss in der Dissertation enthalten sein.
(*This version must be included in the doctoral thesis*)

„Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken demWortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universitätzur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen undeingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung desPromotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universitätzu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten undder schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dortgenannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ichversichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht."

Teilpublikationen:

6.12.2021
Mohamed Awad

Datum, Name und Unterschrift