




## RESOURCE ARTICLE

# An exploration of assembly strategies and quality metrics on the accuracy of the rewarewa (*Knightsia excelsa*) genome

Ann M. McCartney<sup>1,2</sup>  | Elena Hilario<sup>2,3</sup> | Seung-Sub Choi<sup>1,2,4</sup> | Joseph Guhlin<sup>2,5</sup> |  
Jessica M. Prebble<sup>2,6</sup> | Gary Houlston<sup>2,6</sup> | Thomas R. Buckley<sup>1,2,4</sup>  | David Chagné<sup>2,7</sup> 

<sup>1</sup>Manaaki Whenua - Landcare Research, Auckland, New Zealand

<sup>2</sup>Genomics Aotearoa, Dunedin, New Zealand

<sup>3</sup>The New Zealand Institute for Plant and Food Research (Plant & Food Research), Sandringham, New Zealand

<sup>4</sup>School of Biological Sciences, The University of Auckland, Auckland, New Zealand

<sup>5</sup>University of Otago, Dunedin, New Zealand

<sup>6</sup>Manaaki Whenua Landcare Research, Lincoln, New Zealand

<sup>7</sup>Plant & Food Research, Fitzherbert, Palmerston North, New Zealand

## Correspondence

Ann M. McCartney, Manaaki Whenua - Landcare Research, Saint Johns, Auckland, New Zealand.

Email: ann.mccartney2@mail.dcu.ie

## Funding information

This work was supported by New Zealand's Ministry of Business, Innovation and Employment (MBIE) Strategic Science Investment Fund (SSIF) "Genomics Aotearoa" programme ([www.genomics-aotearoa.org.nz](http://www.genomics-aotearoa.org.nz)).

## Abstract

We used long read sequencing data generated from *Knightsia excelsa*, a nectar-producing Proteaceae tree endemic to Aotearoa (New Zealand), to explore how sequencing data type, volume and workflows can impact final assembly accuracy and chromosome reconstruction. Establishing a high-quality genome for this species has specific cultural importance to Māori and commercial importance to honey producers in Aotearoa. Assemblies were produced by five long read assemblers using data subsampled based on read lengths, two polishing strategies and two Hi-C mapping methods. Our results from subsampling the data by read length showed that each assembler tested performed differently depending on the coverage and the read length of the data. Subsampling highlighted that input data with longer read lengths but perhaps lower coverage constructed more contiguous, kmers and gene-complete assemblies than short read length input data with higher coverage. The final genome assembly was constructed into 14 pseudochromosomes using an initial FLYE long read assembly, a RACON/MEDAKA/PILON combined polishing strategy, SALSA2 and ALLHiC scaffolding, JUICEBOX curation, and *Macadamia* linkage map validation. We highlighted the importance of developing assembly workflows based on the volume and read length of sequencing data and established a robust set of quality metrics for generating high-quality assemblies. Scaffolding analyses highlighted that problems found in the initial assemblies could not be resolved accurately by Hi-C data and that assembly scaffolding was more successful when the underlying contig assembly was of higher accuracy. These findings provide insight into how quality assessment tools can be implemented throughout genome assembly pipelines to inform the *de novo* reconstruction of a high-quality genome assembly for nonmodel organisms.

## KEYWORDS

*de novo* assembly, endemic, Hi-C, methods, New Zealand, next generation sequencing, Oxford Nanopore, proteaceae, quality metrics, rewarewa

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

## 1 | BACKGROUND

It is of critical importance that an optimal genome assembly strategy is used to maximize the impact, effectiveness and accuracy of resulting pseudochromosome-scale *de novo* reference genomes. As long read sequencing data become more affordable, the integration of a multitude of next generation sequencing (NGS) platforms is becoming standard for generating near-complete *de novo* genome assemblies. The construction of an accurate *de novo* assembly is crucial to facilitating investigations of species evolution (Lewin et al., 2018; Rhie, McCarthy, et al., 2020; Rhie et al., 2020) and organism diversity (Gurdasani et al., 2016), and to informing health and disease treatments in fields such as cancer treatment programmes (Berger & Mardis, 2018) and vaccine development (Prachi et al., 2013). To cater for the synergistic nature of different types of sequencing data, the research field of genome assembly is moving quickly, and new methods are becoming more flexible, accurate and efficient. Genome assembly software incorporates sophisticated algorithms built to deal with a multitude of sequencing data types; for instance, accounting for the different base calling accuracies of Oxford Nanopore Technology (ONT) (<5% error rate), PacBio Single Molecule Real-Time (SMRT) (<1% error rate) and Illumina short paired-end (PE) reads (<0.1% error rate). They also allow a multitude of parameter specifications to cater for various genome architectures. For example, CENTROFLYE (Bzikadze & Pevzner, 2019) was designed for centromere assembly, and CHLOROEXTRACTOR (Ankenbrand et al., 2018) was developed to assemble chloroplastic genomes from whole genome sequencing (WGS) data. Through different error correction and consensus approaches, these programs use noisy ONT data to construct contig assemblies which can be scaffolded to generate high-quality assemblies, but it is not generally clear what type of data are required or the volume necessary to generate the "optimal" assembly, or indeed what combination of software one should use given the available types and volumes of data.

Despite a thorough investigation of the computational resource performance of long read assemblers by Wick and Holt (2019), published data on the optimization of read length and depth in the context of the most commonly used long read assemblers (NECAT, WTDBG2/RedBean [WTD], CANU, FYLE, FALCON and SHASTA) is limited. Although the underlying long read data used by these toolkits are shared, their methods for error correction, assembly and consensus generation differ greatly. For instance, FYLE (Kolmogorov et al., 2019) identifies "disjointigs" and uses these to first resolve the repeat graph in order to construct the final assembly. CANU (Koren et al., 2017) carries out extensive error correction and trimming prior to generating the final assembly using overlap-consensus methods based on string graph theory (Myers, 2005). NECAT (Chen et al., 2020) acts similarly to CANU albeit using a more progressive correction and assembly strategy. In contrast, WTD (Ruan & Li, 2020) uses only a single round of consensus by a fuzzy DeBruijn algorithm (Zerbino & Birney, 2008) that is based on initial short read assembly algorithms that have been adjusted to accommodate the base calling inaccuracies of noisy long reads. The SHASTA (Shafin et al., 2020a) algorithm

maximizes computational efficiency through the identification of reduced marker kmers to initially find overlaps and then build the consensus sequence.

Gaining an understanding of each assembler's advantages and shortcomings is an important consideration prior to assembly to form a more educated assembly strategy and ultimately resulting in a genome assembly sufficient for individual project needs. Quantitative metrics to track the accuracy and completeness of the assembly must be performed as often as possible throughout the workflow. In the past, appropriate nonmanual methods of genome accuracy assessment have been limited, particularly with regard to scaffolding steps using proximity-guided methods like Hi-C (Lieberman-Aiden et al., 2009). Recently, more advanced quantitative toolkits have become available, such as kmer completeness (MERQUY; Rhie, Walenz, et al., 2020), Long terminal repeat retrotransposons Assembly Index (LAI; Ou et al., 2018), mapping rate and highly conserved gene completeness (BUSCO; Simão et al., 2015). However, an isolated selection of assembler without factoring the input data and downstream post-processing steps is insufficient, as the tools used for these steps are also important considerations to generate an optimal genome assembly.

The identification and correction of misassemblies, or "polishing," is determined by the initial assembler and the algorithm used, but comprehensive analyses of the impact of different polishing strategies on genome accuracy are scarce. Assembler algorithms act differently during contig construction; thus, the initial assembly accuracy they produce before polishing is not always a fair indication of the metrics that will be obtained afterwards. Iterative polishing steps increase assembly accuracy after each step so that reads previously unable to map due to error or misassembly in the initial assembly become mappable, leading to a more accurate consensus assembly. Polishers are placed in two categories: "Sequencer bound" or "General." Both NANOPOLISH (Loman et al., 2015) and MEDAKA (Technologies, 2018) are examples of sequencer-bound polishers that utilize raw signal information, while RACON (Vaser et al., 2017) and PILON (Walker et al., 2014) are examples of general polishers that are applicable to any sequencing platform. To obtain a better understanding of polishing and post-assembly processing performance, initial contig assemblies generated from a selection of ONT assemblers must be tested using a combination of polishing strategies.

Three main methods are commonly used for scaffold ordering and orientation to generate chromosome-level assemblies. Traditionally, linkage maps made of thousands of genetic markers obtained from large segregating progenies were used to anchor assembly contigs to linkage groups (Linsmith et al., 2019). However, this method can be expensive and can give false orientations due to inaccuracies in marker orientation and ordering due to genotyping errors. Synteny-based approaches can be used when a closely related high-quality genome is available. However, all results obtained via these strategies are heavily biased toward the provided reference assembly, and any unique translocations or re-orderings will be lost. Further, errors in the provided reference assembly can cascade into further projects. Recently, proximity ligation methods

have become a more cost-effective and less biased (Peichel et al., 2017) approach for generating chromosome-level assemblies. The Hi-C method is commonly used for scaffolding genomes (Lightfoot et al., 2017; Thrimawithana et al., 2019). Hi-C data are generated by cleaving chromatin using restriction endonucleases and ligating only fragments that are close in 3D chromosomal space. The underlying premise is that the closer two fragments the more linkage markers they will share. Hi-C scaffolding algorithms take advantage of interactions at contig ends to orient and order scaffolds. However, many chromosome-level assemblies generated using Hi-C are littered with inaccurate contig placements due to shorter contigs that contain interactions spanning their entire length, inhibiting the ability of Hi-C software to effectively orient and order these contigs accurately (Burton et al., 2013). Traditionally, Hi-C software are built for homozygous diploid genome assemblies and are heavily reliant on the accuracy of the reference assembly provided and many require a priori knowledge of chromosome number such as LACHEISIS (Burton et al., 2013) and ALLHIC (Zhang et al., 2019). Two tools are commonly employed for Hi-C scaffolding: SALSA2 (Ghurye et al., 2019) and ALLHIC (the latest version of LACHEISIS). The effects of input assembly on Hi-C mapping rate and the performance of such software must also be evaluated.

*Knightia excelsa* (rewarewa) is a nectar-producing tree of the family Proteaceae, endemic to Aotearoa. Despite its size (>1,660 species; Christenhusz & Byng, 2016), the Proteaceae has received minimal attention from genome researchers, probably due to most diversity being restricted to the southern hemisphere as well as the nut-producing macadamia tree being the only species within this family of significant worldwide economic interest. A genome assembly of *Macadamia integrifolia* has been developed (Nock et al., 2020), the information from which is used for genome-informed breeding (O'Connor et al., 2020). Very little genetic information is available for *K. excelsa*; however, karyotype analysis indicated it is a diploid species with  $n = 14$  chromosomes (Hair & Beuzenberg, 1958). Rewarewa is the basis of a burgeoning honey industry in Aotearoa. Most of its honeys are produced from traditional land owned by Aotearoa's Indigenous Peoples, Māori. Rewarewa is considered "taonga" by Māori, meaning this tree is treasured and under their "kaitiaki" or guardianship. To this end, an ethical framework is necessary for managing samples and data during the project, as has been performed for other taonga species (Marshall et al., 2015; Morgan et al., 2019).

The objective of this research was to investigate the impact of sequence volume and depth on genome assembly accuracy using *K. excelsa* as a model (Figure 1) whilst also generating a high-quality reference genome for *K. excelsa* using all of the sequencing data available. Subject to Māori consent, Illumina PE (61×), ONT (52×) and Hi-C data were obtained, with genome coverage estimates based on flow cytometry size estimates. Software for contig assembly, polishing and Hi-C scaffolding were evaluated, and quality metrics were measured at each step. Initial contig assemblies were generated from five long read assemblers across four subsampled sets of ONT data (reads >5 kb, >10 kb, >22 kb, >30 kb). Furthermore, assembly methods were optimized using all available ONT data to be used for the

high-quality genome assembly of *K. excelsa*. All assemblies produced were corrected using a combination of long and short read polishing tools (Figure 1). After this, the effectiveness of each ONT assembly method on chromosomal construction was assessed through Hi-C scaffolding using two software packages, SALSA2 and ALLHIC. These tools were systematically implemented across all four read subsamples, and the accuracy of each assembly was quantitatively assessed and compared. For assemblies produced using all ONT data, conservation of macrosynteny was tested against macadamia linkage maps (Langdon et al., 2020) in order to identify the optimal *K. excelsa* genome assembly that could be generated from our data.

## 2 | MATERIALS AND METHODS

### 2.1 | Sample collection

*Knightia excelsa* (rewarewa) is an endemic tree of Aotearoa, mostly found on the North Island, and common in coastal, lowland and lower montane habitats. This evergreen tree species can grow up to 30 m tall, and bears dark green serrated leathery leaves and dense racemes of red flowers. *K. excelsa*'s genome size was estimated to be 1.15 pg per 1C using flow cytometry.

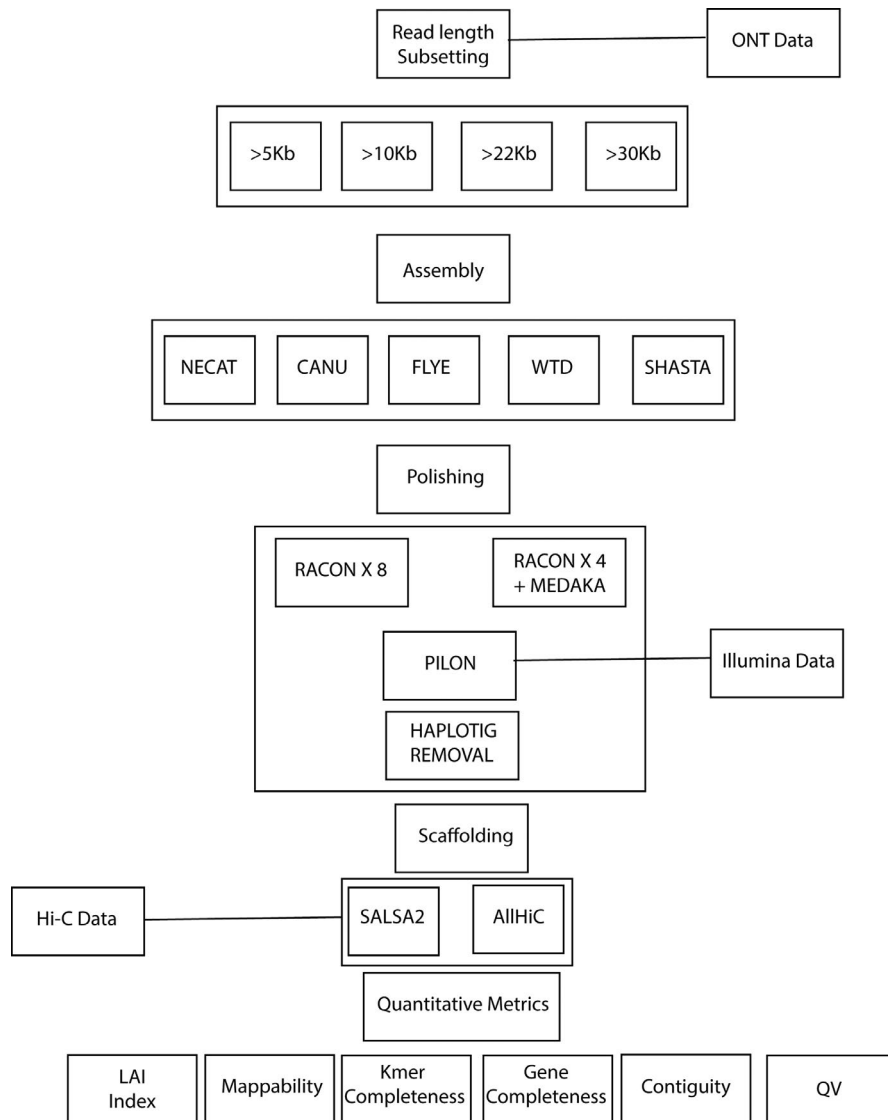
The single *K. excelsa* tree selected for this project grows in the Warawara Forest, Northland, Aotearoa (Lat.  $\pm 4$  m 35°22'1.9"S, Long.  $\pm 4$  m 173°16'5.4"E, altitude 440 m). The leaves were collected in November 2018, authorized by the Te Rarawa Anga Mua and the Komiti Kaitiaki for Warawara Ngahere. At the time of collection, the tree was about 3 m tall, growing in full sun, isolated from other trees, and colonizing a bulldozed site along with *Lycopodiella cernuua*, *Leucopogon fasciculatus* and *Blechnum novae-zelandiae* beneath it. The tree has deep magenta flowers and was fruiting at the time of sample collection. The tree had two trunks from the same base. One trunk was 4 cm in diameter at 1.35 m above the ground and the other was 2 cm in diameter at 1.35 m above the ground. The combined cross-sectional area at breast height was 15.7 cm<sup>2</sup>. The leaves sampled were undamaged leaves without visible fungal infections that ranged in size from 8 to 12 cm long by 2 to 3 cm wide. Two leaf samples were collected (~20 and ~30 g).

The leaves were collected aseptically and packed in a sealable plastic bag, placed inside a Styrofoam box with crushed ice, and protected from ice burn by a stack of paper towels. The sample was delivered within 2 days after collection and stored at -80°C upon arrival at the laboratory.

### 2.2 | Nuclear genomic DNA extraction

#### 2.2.1 | Nuclei isolation

Nuclear genomic DNA was extracted from isolated nuclei as described previously (Hilaro, 2018; Naim et al., 2012) with the following modifications regarding the homogenization method, the



**FIGURE 1** Overview of the genome assembly workflow used for assessing the effect of read length and data volume and the workflow used for optimal *Knightia excelsa* genome assembly construction

type of lysis buffer and its ratio to the number of nuclei obtained. The leaf sample (20 or 30 g) was ground with liquid nitrogen in a precooled large mortar. The freeze/grinding cycle was repeated three times until a fine powder was obtained. The complete nuclei isolation buffer (plus sodium metabisulphite,  $\beta$ -mercaptoethanol and 0.5% Triton X-100) was poured into a 1-L beaker with stirrer. The powdered sample was added gradually and stirred until completely dissolved. The homogenate was filtered through two layers of Miracloth (Merck) over a funnel. The nuclei were collected by low-speed centrifugation and washed twice with the nuclei isolation buffer (with sodium metabisulphite only). The final nuclei pellet was stored without any liquid at  $-80^{\circ}\text{C}$  until used for DNA extraction.

### 2.2.2 | DNA extraction

The nuclear genomic DNA was extracted with a cetyl trimethylammonium bromide (CTAB)-based buffer as described previously (Hilario, 2018; Naim et al., 2012) with the following modifications: The isolated nuclei were lysed with 15 ml of CTAB buffer and 100  $\mu\text{l}$

proteinase K (20 mg  $\text{ml}^{-1}$ ). After the lysis incubation, the sample was extracted with an equal volume of chloroform/iso-amyl alcohol (24:1), precipitated with ethanol and the DNA collected by centrifugation. The DNA pellet was washed with 10 ml 70% ethanol, centrifuged again and dissolved in 200  $\mu\text{l}$  TE buffer. The quality of the DNA was assessed by spectrophotometry (Nanodrop) and electrophoresis separation (standard and pulse field gel electrophoresis). The amount of DNA was estimated by fluorometry (Qubit high-sensitivity dsDNA kit). The average yield of nuclear genomic DNA per gram of leaf sample was 1  $\mu\text{g}$ . The quality parameters were  $A_{260/280} = 2.0$ ,  $A_{260/230} = 1.88$ , Qubit/Nanodrop  $\sim 0.5$ , a concentration of 164 ng  $\mu\text{l}^{-1}$  and an average fragment size of 50 kbp.

### 2.3 | NGS library preparation

#### 2.3.1 | Short insert Illumina sequencing library

The generation of paired-end Illumina data was required to remove errors by polishing the initial noisy ONT-based genome assemblies.

Eight reactions of 500 ng of nuclear genomic DNA each were set up for preparing the short insert Illumina library with the NEBNext Ultra FS II DNA library kit as described by the vendor with the following parameters: the fragmentation, end repair and deoxyadenylation incubation was 3.75 min (fragments ranging from 200 to 1,000 bp). After USER digest, all the reactions were combined and split into five tubes. The library was left size selected with AMPure XP beads at 0.4× ratio followed by another left side selection at 0.2× ratio. The DNA was eluted from both bead fractions (0.4× and 0.4×/0.2×) in 30 µl TE buffer and the concentration estimated by fluorometry. A cycle test was performed with 5 ng of each size-selected library (0.4× and 0.4×/0.2×) amplified 4, 6, 8, 10 or 12 times with NEBNext Ultra II Q5 Master mix, and the Illumina universal and index primers. Ten cycles produced the optimal amplicon size after a dual size selection (0.77×/0.61×) from the 0.4× size-selected library fraction (average amplicon size: 473 bp). Four reactions from this library fraction were set up under these conditions, pooled, dual size-selected, quality checked and sent to our service provider (Custom Science, New Zealand) to be sequenced.

### 2.3.2 | Long-range sequencing library (Hi-C)

The long-range Hi-C sequencing library was prepared with isolated nuclei as starting material. The nuclei enrichment method is similar to the protocol described above but with extra steps to remove contaminants and large particle debris with polyvinylpyrrolidone (PVPP) and Percoll gradients, respectively. The Hi-C library was prepared with a combination of kits and in-house methods. The nuclei crosslinking, quenching, washing, lysis and chromatin normalization steps were performed according to the Dovetail Genomics Hi-C kit. The chromatin lysate was bound to AMPure XP beads and washed with five sets of 1 ml Wash buffer (Dovetail Genomics Hi-C kit). Chromatin fragmentation and biotinylation were performed with the Fragmentation buffer and Fragmentation Enzyme mix from the Phase Genomics Hi-C kit for plants version 1.0. Once the digestion was completed, the captured chromatin was washed twice with Wash buffer (Dovetail Genomics). Intramolecular ligation was performed in 500 µl of 1× T4 DNA ligase buffer (Invitrogen) and 10 units of T4 DNA ligase (Invitrogen). The ligation was performed at 16°C in a thermomixer (Eppendorf) at 1,250 rpm overnight. The ligation mixture was discarded, and the crosslink reversal was performed by adding 50 µl 1× CutSmart buffer (New England Biolabs) and 20 µg proteinase K (Qiagen) and incubated at 55°C for 15 min followed by 45 min at 68°C at 1,250 rpm. The released DNA was transferred to a new tube and purified with AMPure XP beads at 2× ratio. The DNA was eluted in 150 µl 10 mM Tris-HCl pH 8 and the biotinylated molecules captured with Dynabeads M280 (Invitrogen) according to the manufacturer's protocol but using 150 µl Bead Binding buffer (Phase Genomics) for coupling the biotinylated molecules to the beads and continuation with the Phase Genomics Hi-C kit for plants protocol. The amplified library was size selected by agarose gel electrophoresis followed by an AMPure XP double size selection (0.77×/0.64×). The average fragment size of the selected amplicons was 500 bp.

The size-selected amplicons were assessed by capillary electrophoresis (Fragment Analyzer) and showed an average fragment size of 441 bp, at 1.5 ng µl<sup>-1</sup> and 4.7 nm. The amplicons were sequenced (150-bp PE reads) and delivered 221,731,503 raw PE reads, and 66.96 Gb.

### 2.3.3 | PromethION oxford nanopore sequencing

The PromethION libraries were prepared by the contracted service provider (Custom Sciences) with ~50 µg of nuclear genomic DNA preparation described above. Here, four libraries were made generating a total of 176,417,984,645 read bases.

## 2.4 | Genome assembly and assessment

### 2.4.1 | Initial quality assessment and subset generation of oxford nanopore reads

All data sets were base called using GUPPY flip flop software package (Appendix S1) and quality assessed using the FASTQC raw reads for quality assessment. To understand the impact of data volume and coverage on ONT assembly, read subsetting was carried out using the PORECHOP software package. The data were subsampled by read length into four read subsamples: >5-kb reads only (52×), >10-kb reads only (50×), >22-kb reads only (33×) and >30 kb (23×) reads only. These values were selected in order to retain sufficient sequencing depth within each subset.

### 2.4.2 | Oxford nanopore assembly

Five long-read assemblers were used: CANU, FYLE, WTD, SHASTA and NECAT (for parameters and versions used see Appendix S1). In order to further understand the effects of polishing strategies on assembly accuracy, combinations of polishings methods were examined and haplotigs were purged. These include general polishing strategies: RACON with four rounds (RX4) of polishing only, RACON with eight rounds of polishing both before (RX8) and after PILON (RX8\_SR) polishing and haplotig purging (RX8\_SR\_PH). A sequencer-specific strategy alone was also included: MEDAKA only (M) polishing, as well as combined polishing approaches: MEDAKA with four iterations of RACON polishing both with (M\_RX4) and without PILON polishing (M\_RX4\_SR) and haplotig purging (M\_RX4\_SR\_PH). Each assembly was initially quality checked using QUAST, BUSCO and LAI.

### 2.4.3 | Hi-C mapping

The Hi-C data set was filtered using the Phase Genomics filtration guidelines (<https://phasegenomics.github.io/2019/09/19/hic-align-ment-and-qc.html>). The data successfully passed all quality assessment analysis requiring no additional filtration. The data were

mapped to each generated ONT contig set using BWA MEM and scaffolding was carried out by SALSA2 and ALLHIC (see Appendix S1 for parameters and versions used).

#### 2.4.4 | Hi-C assembly quantitative quality assessment

Each Hi-C assembly kmer spectrum profile was assessed through MERYL and consensus accuracy and completeness were analysed using the MERQURY toolkit. Map back rates were also used to assess the quality of each assembly using SAMTOOLS (Cock et al., 2015) flagstat statistics (see Appendix S1 for parameters used). All assemblies were additionally compared using the LAI index which assesses the LTR repeat completeness of plant genomes specifically. On top of this, assemblies were compared through QUASt (Gurevich et al., 2013) and gene completeness examined through a BUSCO (*embryophyta\_obd9*) identification of complete single copy, duplicated, fragmented and missing genes. Additionally, we used contact map manual inspection by PRETEXTMAP and PRETEXTVIEW (<https://github.com/wtsi-hpag/PretextView>).

#### 2.4.5 | Utilization of *Macadamia* linkage maps for QC

Nine linkage maps accompanying 64-bp DartSeq reads were downloaded from the Southern Cross University data repository (<http://dx.doi.org/10.25918/5dc2589924ca2>). These reads were aligned using BLASTN to three Hi-C assemblies (the best assemblies selected based on quantitative assembly accuracy metrics) and only unique hits of >90% identity were used. These markers were mapped to each assembly using ALLMAPS (Tang et al., 2015a) (see Appendix S1 for versions and parameters).

### 2.5 | Computational resources

The majority of analyses were carried out on the New Zealand eScience Infrastructure high-performance computer on the Mahuika partition. The Mahuika partition consists of a Cray CS400 Cluster High Performance Computer with 8,424 × 2.1-GHz Intel Broadwell cores and 30 Tb of memory along with IBM ESS Disk and SSD storage. For computational efficiency each assembly was run using minimal requirements (See Appendix S1). CANU assemblies were performed at the University of Otago's Biochemistry Servers, which have 1 Tb of memory, and 8× Intel(R) Xeon(R) CPU E7-8860 v4 with 18 cores each, and two threads per core.

## 3 | RESULTS

### 3.1 | Sequencing data

In total, 2.3 million ONT sequencing reads were obtained totalling 52.5 Gbp of data and with a read N50 of 28 kbp. Table 1 indicates the initial summary read statistics for the ONT data. The significance

**TABLE 1** Basic statistics of Oxford Nanopore Technologies sequencing data for *Knightsia excelsa*

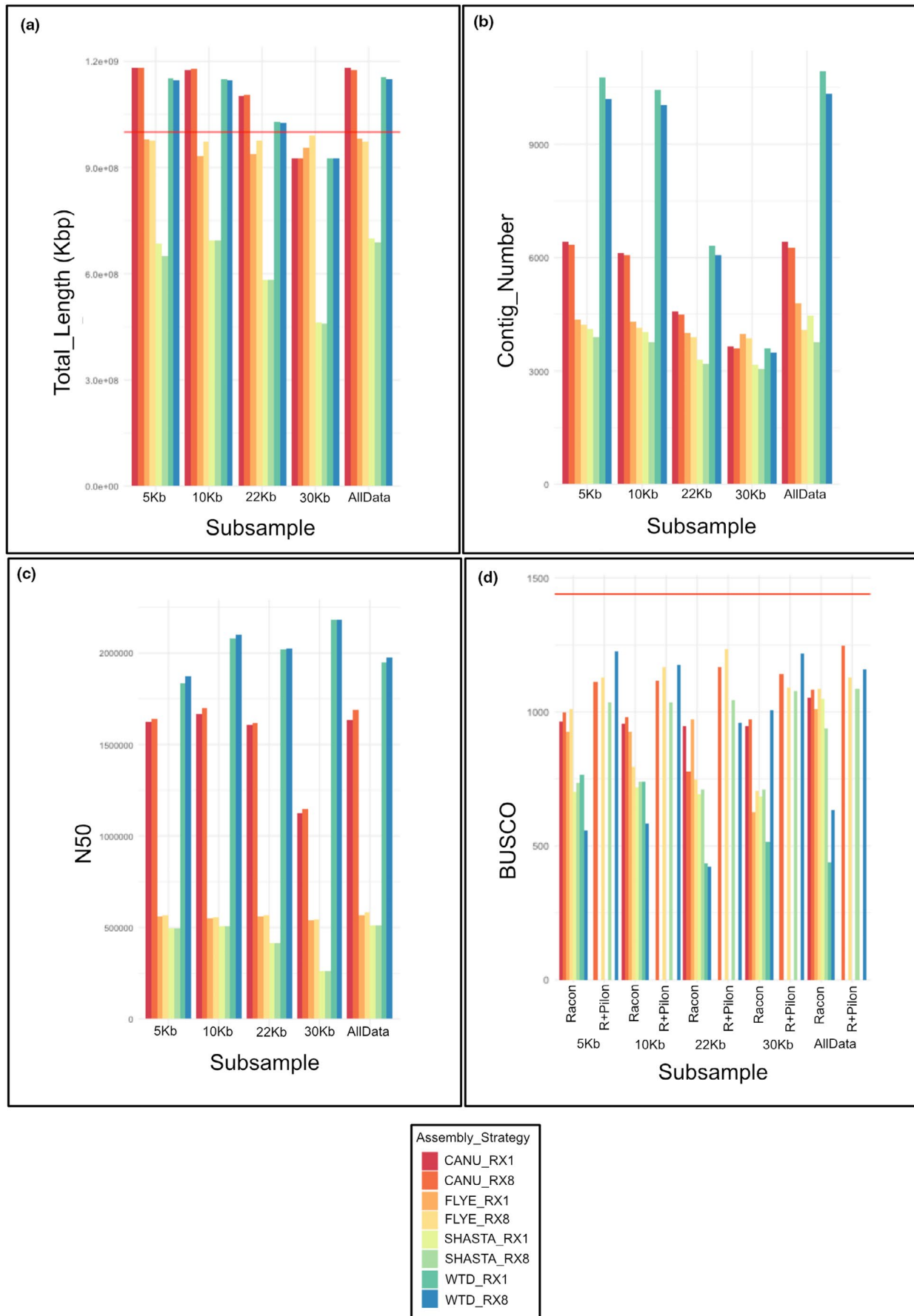
Statistics	<i>Knightsia excelsa</i>
Total number of sequences	2,314,274
Total length	52,588 Mb
Longest sequence	229 kb
Shortest sequence	55 b
Mean length	22 kb
Median length	19 kb
N10	53 kb
N50	28 kb
N90	13 kb

of base-calling was assessed both before and after base-calling using MINION QC (Lanfear et al., 2019) (Appendix S2). A significant increase in overall Q score was achieved and specifically for longer read lengths. Hi-C data produced from the Phase Genomics kit and Illumina PE sequencing yielded 443 million reads in total (67 Gb of data). Short read WGS data were obtained and consisted of 407 million PE Illumina reads. Through kmer counting ( $k = 21$ ) by the GENOMESCOPE software (Ranallo-Benavidez et al., 2020) a genome size of 0.95 Gb and heterozygosity of 0.1%–1.0% was estimated.

### 3.2 | In-depth critique of ONT assembler performance and exploration of the impact of iterative polishing

An assessment of the performance of five long read assemblers, NECAT, CANU, SHASTA, FYLE and WTD, was carried out and initial contig sets generated. ONT data were then split into four subsamples based on read length: >5 kb, >10 kb, >22 kb, >30 kb. Assembly performance was compared across these subsamples to assess how read length might affect the performance of individual long-read assemblers. Furthermore, ONT assemblies were generated using all available ONT data (All\_Data) to facilitate the optimal high-quality Rewarewa genome assembly. For the purposes of initial comparisons all assemblies generated both by read length subsamples and by All\_Data will be compared.

First, the output from iterative long-read polishing using RACON was examined to explore potential effects on assembly accuracy. Overall, when N50, contiguity and total length were considered, the first round of polishing always showed significant improvement, but additional rounds of polishing had a marginal increase on genome accuracy (Figure 2). Interestingly, the NECAT assembly generated for All\_Data appears collapsed after two rounds of polishing, with a drastic reduction in contig number and total assembly length below flow cytometry estimations (1 Gbp). NECAT failed to complete for all other read length subsamples despite adjusting parameters to accommodate low coverage thresholds and therefore was not included in further performance comparison analysis. SHASTA-generated assembly metrics remain consistent in the >10-, >22- and >30-kb read



**FIGURE 2** A comparison of the performance of five ONT assemblers and iterative polishing by Racon across assemblies generated by all read length subsamples and those generated from All\_Data. A comparison is given of contig number, N50 and total length of a single Racon polishing in comparison to eight rounds of polishing

subsamples with iterative polishing having little effect, but total genome length slightly reduced in >5-kb and All\_Data subsamples. WTD assembly metrics remained robust against polishing for all read subsamples, apart from the >10-kb subsample which encountered a total length expansion.

Gene completeness was assessed; generally, iterative long read polishing increased the number of complete genes identified across assemblers (Figure 2d), except for CANU and FYLE's >22-kb subsample assembly that experienced a 169 and 224 complete gene reduction, respectively, and FYLE's >10-kb assembly that showed a 135 reduction. SHASTA experienced a complete gene reduction of 158, 10 and 208 in >10-, >22- and >5-kb read subsamples, respectively, and WTD only experienced a reduction of 111 genes in the assembly constructed by the All\_Data. Across all assemblies the accuracy of Illumina data increased the gene completeness through PILON polishing. Interestingly, the ONT assemblies that experienced a reduction in gene completeness score after iterative long read polishing incurred the greatest increase in score after short read polishing, with CANU and FYLE >22-kb subsample experiencing an increase of 388 and 286 genes, respectively, and SHASTA >5-, >10- and >22-kb read subsamples gaining 669, 593 and 536 genes, respectively.

In terms of total length and contiguity (Figure 2a,b), FYLE's performance appeared the most robust, with total length and N50 values remaining consistent (Figure 2c), but contiguity was increased in the >30-kb subsample (smaller number of contigs). WTD, SHASTA and CANU appeared to perform much better with longer read lengths, based on the lower number of contigs and increased N50 for the >30-kb subsample. However, the >30-kb subsample genome had a total length that was below the length expected from flow cytometry estimates. This may have been the result of low depth of coverage in this read length subsample and may be improved with an increase in data volume within this read length subsample.

### 3.3 | Assessment of general and sequencing-specific polishing strategies on ONT assemblies

The previous section highlighted the importance of considering all quantitative metrics when implementing polishing strategies, with BUSCO scores improving significantly while the effect on N50, total length and contiguity was marginal. Here, the effect of both general and sequence-specific polishing strategies was examined (Figure 3).

The >5- and >10-kb subsamples performed consistently across CANU, FYLE, WTD and SHASTA assemblers with regard to total length in strategies combining RACON and MEDAKA (M\_RX4, M\_RX4\_SR\_PH and M\_RX4\_SR) (Figure 3a,b). Both FYLE and WTD obtained a contig set that was the most representative of the expected genome size, with no bias toward general or sequencer-specific polishing identified. However, when gene completeness is considered, all assemblers showed a benefit from using polishing strategies that incorporate MEDAKA (M, M\_RX4, M\_RX4\_SR\_PH and M\_RX4\_SR).

Analyses using the >22-kb subsample indicated that read length and depth of coverage enabled the WTD assembler to more accurately represent the total genome size while retaining gene completeness

in comparison to that with the >5- and >10-kb subsamples—again no polisher bias was apparent (Figure 3c). WTD assemblies were less contiguous to those constructed by FYLE and CANU, whose bias toward a combined polishing (M\_RX4, M\_RX4\_SR\_PH and M\_RX4\_SR) remained consistent to that constructed with >5- and >10-kb subsamples.

The CANU >30-kb subsample outperformed all other assemblers with regard to gene completeness and there was a clear bias towards polishing strategies that include MEDAKA. The >30-kb subsample SHASTA and WTD assembly performance was compromised from the lack of read depth, as highlighted by the continual reduction of total assembly size as read length cut offs increase and both assemblers remain consistent across polishers. FYLE's performance at this read length remained consistently biased toward MEDAKA-incorporated strategies (Figure 3d).

All\_Data SHASTA and WTD assemblies failed to represent an accurate total length despite becoming more contiguous. Both assemblers experienced the same shortcomings when polished using RACON only, MEDAKA only and combined strategies. CANU's total genome length suffered despite good performance with regard to gene completion and retained its bias towards MEDAKA polishing. FYLE's performance was relatively constant across all subsets. Interestingly, although an integrated RACON/MEDAKA polishing strategy still performed better here, RACON only polished assemblies performed better than across other read length subsamples (Figure 3e).

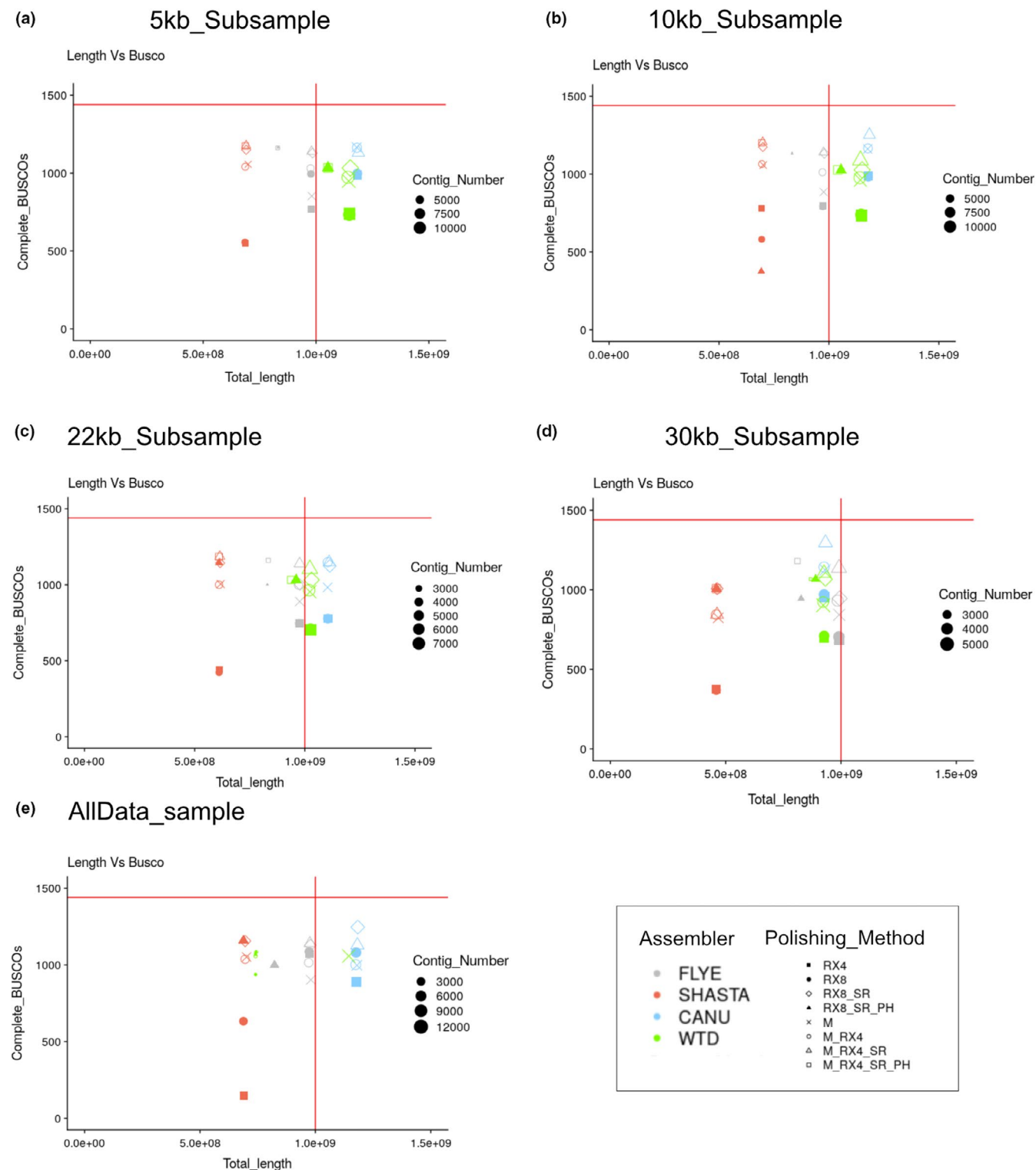
Assemblers performed differently to haplotig purging, with diploid-aware assemblers such as CANU experiencing a reduction in gene completeness score when this technique was implemented across all read subsamples. Overall, FYLE retained gene completeness and increased its contiguity in response to haplotig purging, but across each read subset the total genome size experienced a reduction to below what was expected for the genome. WTD responded similarly to FYLE across all read subsets whilst SHASTA experienced varying shifts in performance over all read subsets.

Across both general and sequencer-specific polishing, WTD assemblies appeared the most fragmented, with expanded genome total lengths, and no bias toward polishing strategy was identified. In contrast, all SHASTA assemblies appeared highly contiguous, although they had unexpectedly small total lengths and were unaffected by polisher. FYLE and CANU appeared to perform best with FYLE performing equally well across all polishing methods with regard to genome size representation. These results highlight a clear bias toward polishing strategies that incorporate MEDAKA as opposed to those utilizing RACON alone, and this suggests that polishing methods specific to the sequence platform utilize have a superior performance than general polishers that are not platform-specific.

### 3.4 | Analysing the effect of ONT data volume and read length on the accuracy of Hi-C scaffolding

To assess the impact of the underlying read length, coverage and genome assembly quality on Hi-C scaffolding, each ONT assembly constructed using subsampled data (>5-, >10-, >22- and >30-kb read





**FIGURE 3** A comparison of polishing strategy performance on four read subsets and All\_Data across contig sets generated by four long assemblers: (a) >5-kb read length subsample, (b) >10-kb read length subsample, (c) 22-kb read length subsample, (d) >30-kb read length subsample, and (e) All\_Data. The vertical red line illustrates the estimated genome size, and horizontal red line highlights maximum number of BUSCO genes within the *embryophyta\_odb9* data set. Increasing data point size indicates an increase in contig number within the assembly

samples; RACON+pilon polished assemblies only) was taken and further scaffolded by ALLHiC and SALSA2. Each scaffolded assembly was quality assessed as summarized in Table 2.

These results (Table 2) demonstrated the inability of Hi-C scaffolding to effectively reduce the high contig number found across all WTD initial ONT assemblies. ALLHiC failed to complete scaffolding

**TABLE 2** Quantitative quality assessment of the impact of two Hi-C mapping strategies on four read length subsampled genome assemblies of *Knighthia excelsa* produced across four long read assemblers

Readlength subsample	WTD		SHASTA		FYLE		FYLE		CANU		CANU	
	ALLHIC	SALSA	ALLHIC	SALSA	ALLHIC	SALSA	ALLHIC	SALSA	ALLHIC	SALSA	ALLHIC	SALSA
5 kb												
Contig number	Failed	8,982	407	2,870	263	4,217	956	4,606	956	4,606	956	4,606
Total length	Failed	1,152,532,074	690,899,518	691,149,212	985,207,495	984,812,095	1,191,929,616	1,192,353,516	1,191,929,616	1,192,353,516	1,191,929,616	1,192,353,516
% total busco groups	Failed	72	77	72	77	72	77	72	77	72	77	72
Largest contig	Failed	17,068,456	582,650,745	3,947,786	531,814,962	3,831,468	1,182,335,195	10,264,999	1,182,335,195	10,264,999	1,182,335,195	10,264,999
Largest contig/total length (%)	Failed	1.48	84.33	0.57	53.97	0.38	99.19	0.86	99.19	0.86	99.19	0.86
Post HiC map contig number (%)	Failed	11.99	89.51	68.04	90.83	0	84.89	27.23	84.89	27.23	84.89	27.23
10 kb												
Contig number	Failed	8,843	456	2,872	240	2,712	818	4,328	818	4,328	818	4,328
Total length	Failed	1,151,665,616	699,678,152	699,871,915	979,671,329	980,085,429	1,188,083,588	1,188,447,988	1,188,083,588	1,188,447,988	1,188,083,588	1,188,447,988
% total busco groups	Failed	83	79	83	79	83	79	83	79	83	79	83
Largest contig	Failed	17,422,921	570,551,806	5,383,323	538,803,058	4,235,289	1,179,420,319	15,560,476	1,179,420,319	15,560,476	1,179,420,319	15,560,476
Largest contig/total length (%)	Failed	1.51	81.54	0.76	54.99	0.43	99.27	1.3	99.27	1.3	99.27	1.3
Post HiC map contig number (%)	Failed	11.94	87.9	23.81	94.21	34.63	86.5	28.6	86.5	28.6	86.5	28.6
22 kb												
Contig number	1,202	5,039	109	2,349	182	2,493	300	2,833	300	2,833	300	2,833
Total length	1,030,932,843	1,031,061,243	615,777,790	615,939,025	978,593,846	979,009,846	1,113,578,981	1,114,095,581	1,113,578,981	1,114,095,581	1,113,578,981	1,114,095,581
% total busco groups	67	83	67	83	67	83	67	83	67	83	67	83
Largest contig	1,004,627,000	17,271,930	298,131,008	4,769,417	511,135,815	4,839,984	1,080,398,592	15,560,476	1,080,398,592	15,560,476	1,080,398,592	15,560,476
Largest contig/total length (%)	97.44	1.67	48.41	0.77	52.23	0.49	97.02	1.39	97.02	1.39	97.02	1.39
Post HiC map contig number (%)	16.88	80.17	26.15	96.57	36.19	93.32	36.19	36.96	93.32	36.19	93.32	36.96
30 kb												
Contig number	508	2,870	14	2,329	202	3,867	340	2,174	340	2,174	340	2,174
Total length	932,695,601	932,810,701	846,518,332	463,135,811	994,307,362	993,940,862	934,936,230	935,385,930	934,936,230	935,385,930	934,936,230	935,385,930
% total busco groups	68	85	68	85	68	85	68	85	68	85	68	85
Largest contig	919,896,502	14,211,288	463,071,347	2,763,082	504,576,793	4,219,938	869,314,329	7,345,167	869,314,329	7,345,167	869,314,329	7,345,167
Largest contig/total length (%)	98.62	1.52	54.7	0.59	50.74	0.42	92.98	0.78	92.98	0.78	92.98	0.78
Post HiC map contig number (%)	17.92	85.479	23.76	99.54	94.77	0	7.31	39.34	94.77	0	7.31	39.34

on the >5- and >10-kb read length subsamples and only a 16% and 17% reduction in contig number was achieved for >22- and >30-kb subsamples, respectively. ALLHIC-WTD >30- and >22-kb subsample assemblies consisted of a single “mega” scaffold that contained >97% of the total length (Figure 4d), which is not consistent with the expected karyotype for *Knightsia excelsa*. Despite producing less contiguous assemblies, SALSA2 scaffolding using the initial WTD assemblies appeared more accurate with optimal performance resulting from using the >30-kb subsample, achieving a kmer completeness value of 83%, gene completeness score of 85% and an 85% reduction in contig number observed, and gave a contig length distribution (Figure 5c) more similar to the known karyotype.

Similarly, the problem of SHASTA ONT reduced assembly length relative to the genome size estimated by flow cytometry was not resolved by Hi-C scaffolding using either SALSA2 or ALLHIC. ALLHIC scaffolding, although greatly reducing contig number, produced a suspicious “mega” scaffold similar to the distribution found in WTD scaffolded assemblies (Figure 4d). Again, SALSA2 scaffolds constructed using the >30-kb subsample appeared to be the most accurate, with a gene completeness of 85% and a 99% reduction in contig number and no “mega” scaffold (Figure 4c). Unfortunately, due to the poor total length of the initial ONT assembly provided by SHASTA, the kmer completeness score of the final assembly remained low at 55%, therefore drastically under-utilizing the data provided.

ALLHIC performed optimally when utilizing more robust initial ONT assemblies generated by FYLE across each subsample (Table 2), and although a “mega” scaffold still persists throughout each subset, its size reduced to ~50% of the total genome length (Figure 4c,d). Interestingly, the >5- and >10-kb subsampled assemblies represented more of the data, with higher gene completeness values and a kmer completeness score of 91% for both, in comparison to >22- and >30-kb subsamples that had lower gene completeness scores and a kmer completeness value of only ~80%. Comparatively, SALSA2 scaffolding failed to reduce contig numbers in the >5- and >30-kb subsamples and only achieved a 34% and 36% reduction in the >10- and >22-kb read subsamples, respectively. Although SALSA2 assemblies have lower contiguity in comparison to that of ALLHIC, they consistently outperformed it in terms of gene completeness.

Finally, ALLHIC scaffolding for all CANU ONT assemblies yielded a suspicious contig length distribution, with >92% of the total genome size being placed on a single mega-scaffold. Gene completeness values were also poor with all samples below 80%, which was surprising considering the scaffolds generated were the most kmer-complete, peaking at 93% across samples. Again, the optimal initial CANU assembly from the >30-kb read subsample generated the optimal assembly after SALSA2 scaffolding, with gene completeness of 85% and 93% kmer completeness. In this case, SALSA2 also reduced contig numbers by 39%.

Due to ALLHIC requiring prior knowledge of karyotype to inform the pseudochromosome construction it performed well with higher contiguity and longer scaffolds. However, on comparison of the scaffold length distribution, SALSA2 generated more uniform scaffold lengths across subsamples whereas ALLHIC tended to generate

assemblies with one single mega-scaffold and a multitude of much shorter scaffolds, which is not in agreement with the known karyotype of *K. excelsa*.

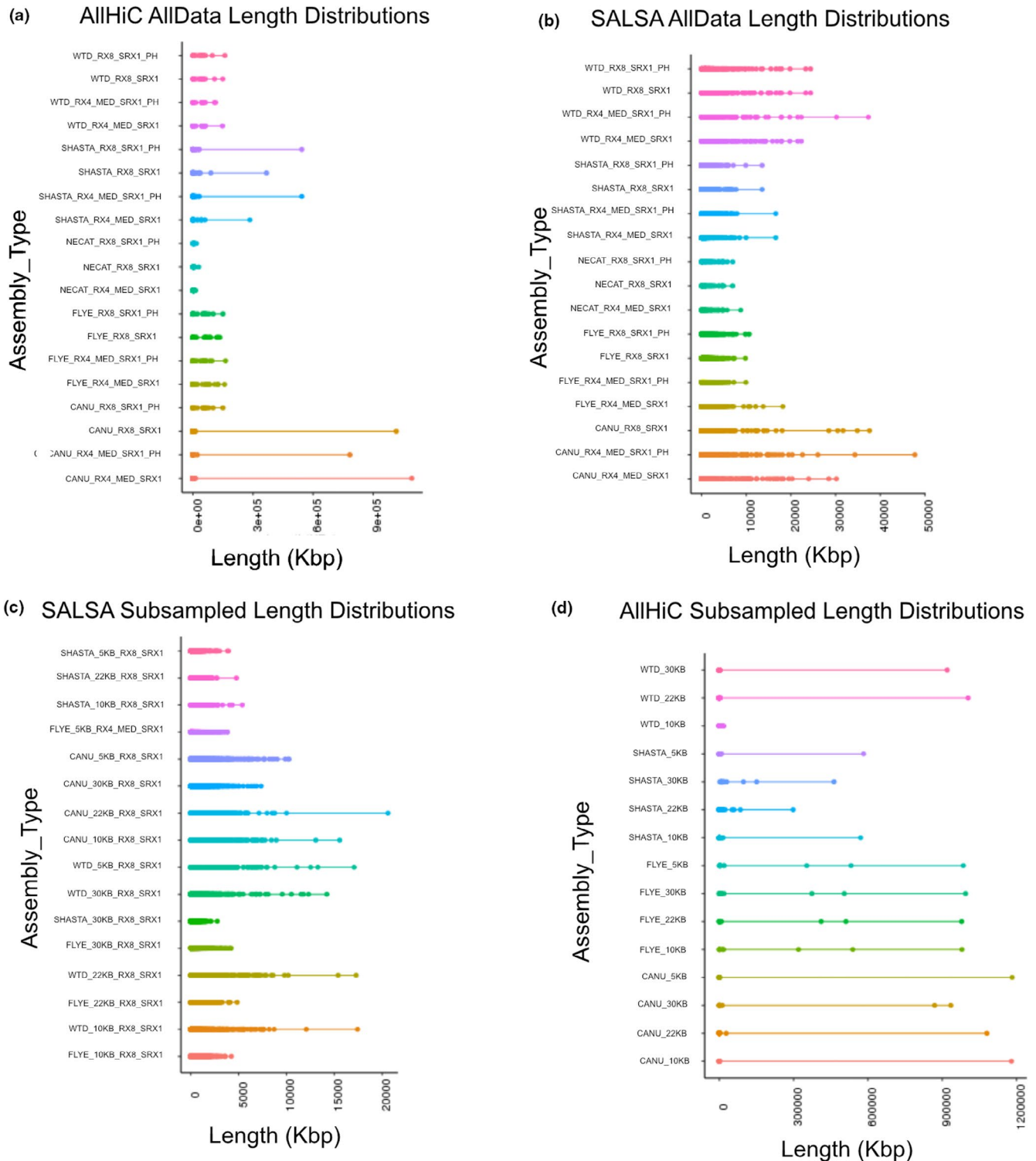
### 3.5 | Optimizing Hi-C scaffolding strategies in preparation for pseudochromosome construction using all ONT data

In the previous section we determined the effect of data volume and sequence length on Hi-C strategies. Here, we optimize the optimal Hi-C scaffolding procedure for the 18 assemblies generated utilizing all available ONT data. Hi-C data were mapped, and duplicates again filtered in accordance with Phase Genomics quality assessment guidelines (Appendix S3). Each assembly assessment indicated that the library preparation would sufficiently inform the underlying assembly. However, for *de novo* assembly scaffolding, high-quality read pairs between contigs are crucial and those found within contigs are uninformative. During QC it was found that WTD and NECAT assemblies had a reduced intercontig percentage of reads pairs when compared to that of CANU, FYLE and SHASTA each peaking at 19%, 21% and 22%, respectively. Each of the eighteen polished ONT assembly constructs were scaffolded using two software packages, SALSA2 and ALLHIC. After initial scaffolding a series of quantitative quality assessments including gene completeness, total length assessment, map back rate, kmer spectra analysis, kmer completeness profiling, consensus accuracy and LAI were calculated (Figure 5).

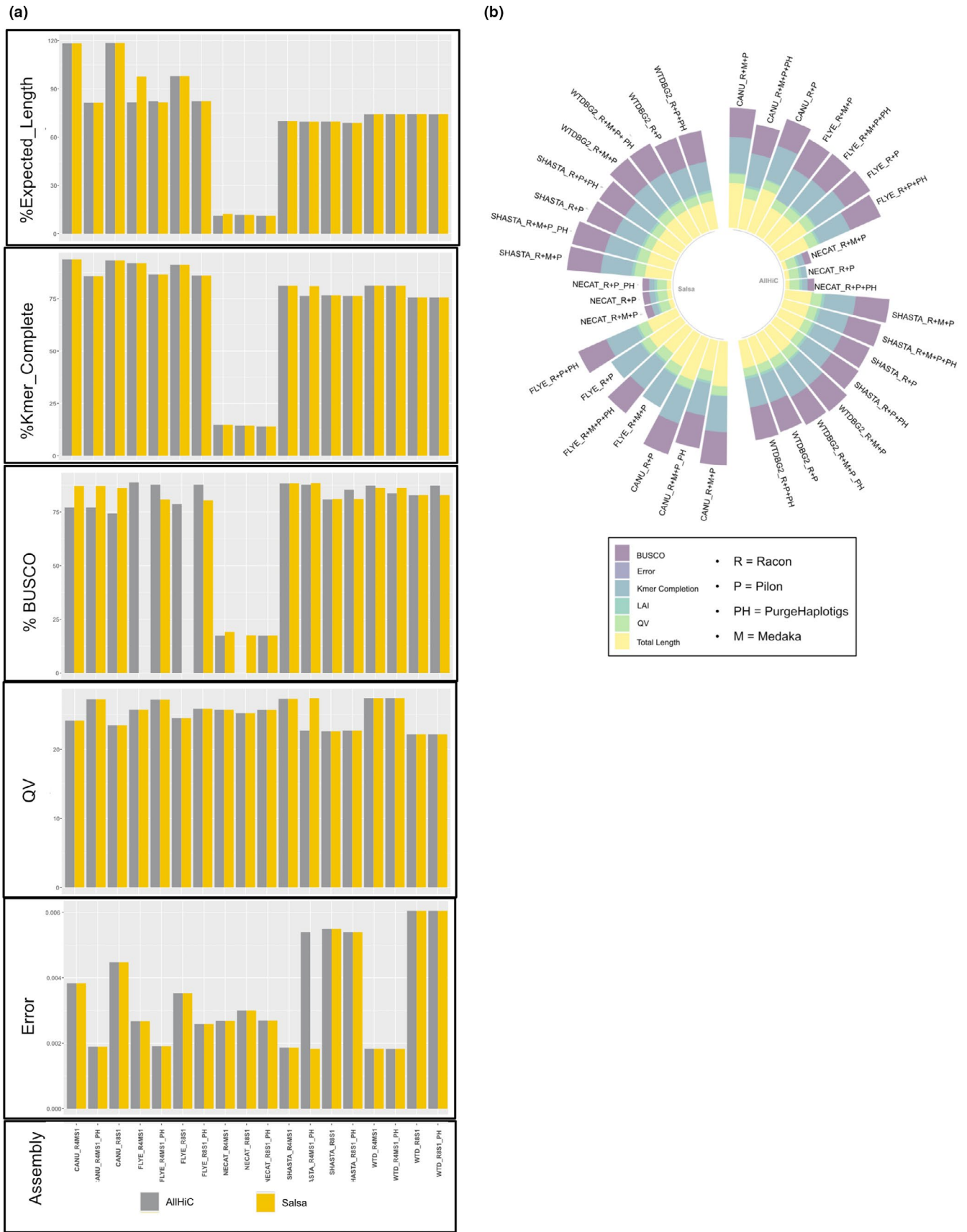
Figure 3 highlights the impact of highly fragmented assemblies on scaffolding pipelines. For example, in WTD contigs sets the total genome size and overall kmer completeness scores fall short after Hi-C data mapping. Overall, CANU assemblies performed well and scaffolding on assemblies based on a RACON/MEDAKA/PILON polishing strategy outperformed those produced by RACON/PILON-only polishing. CANU-based scaffold sets by both ALLHIC and SALSA2 were 93% kmer-complete. However, read map back rates suggested SALSA2 utilized 7% more input data when compared to scaffold sets produced by ALLHIC. The CANU-SALSA2 strategy also outperformed with regard to genome completeness with scaffold sets containing 77% complete genes in comparison to CANU-ALLHIC scaffolds having only 70% complete genes. All FYLE scaffold sets perform well with regard to kmer completeness. FYLE-SALSA2 and FYLE-ALLHIC mappings with RACON/PILON-only polishing produced more accurate total genome lengths but had lower gene completeness scores. Through quantitative metric assessment across all 18 polished assemblies, two assemblies were selected for further analyses before pseudochromosome construction: FYLE/MEDAKA/RACON/PILON/SALSA2 and FYLE/MEDAKA/RACON/PILON/ALLHIC.

### 3.6 | Verification of Hi-C scaffolding using synteny with nine macadamia genetic maps

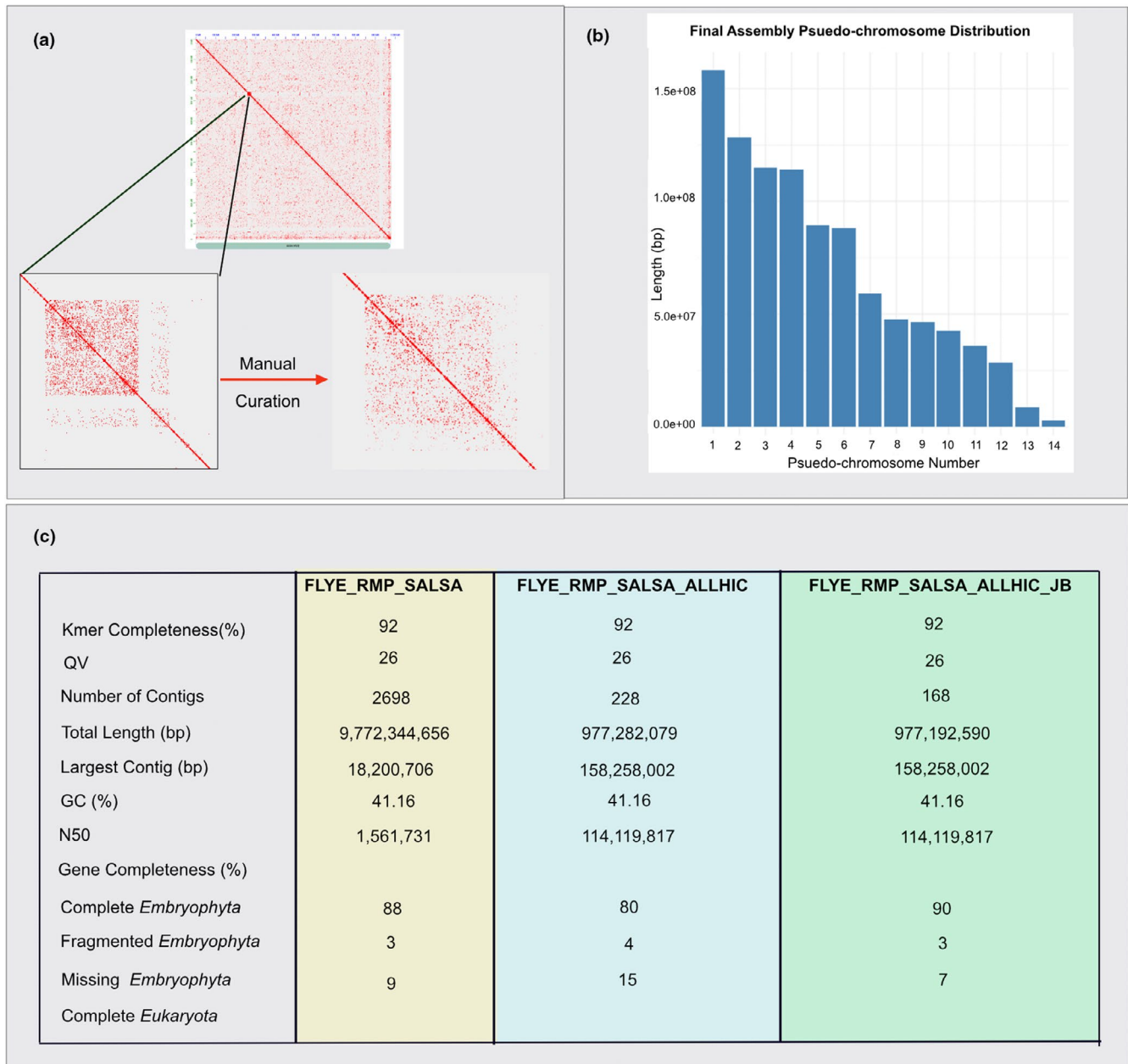
As mentioned in the previous section based on quantitative metric assessment (Figure 5), FYLE/MEDAKA/RACON/PILON/SALSA2 and



**FIGURE 4** An extensive quantitative quality assessment of scaffold sets produced by both SALSA2 and ALLHiC on ONT assemblies. (a) Scaffold lengths for assemblies generated using read length subsampled data by four long read assemblers across four read lengths and Hi-C data mapped using SALSA2. (b) Scaffold lengths for genomes produced using data subsampled by read length by four long read assemblers and Hi-C data mapped using ALLHiC. (c) Scaffold lengths for assemblies produced using all read length data by five long read assemblers utilizing two alternative polishing strategies and Hi-C data mapped using SALSA2. (d) Scaffold lengths for assemblies produced using all read lengths by five long read assemblers utilizing two alternative polishing strategies and Hi-C data mapped using ALLHiC



**FIGURE 5** A quantitative metric comparison of *Knightia excelsa* genome assemblies generated using all available ONT data after both SALSA2 and ALLHiC Hi-C mapping. (a) Quantitative metrics of Hi-C assemblies generated using ALLHiC and SALSA2. (b) A summary of LAI, kmer completeness, base error rate, consensus accuracy (QV), total length, and gene completeness (BUSCO)



**FIGURE 6** Pseudochromosome assembly curation and validation using both quantitative metrics, karyotype evaluation and manual curation. (a) The contact map generated for FLYE/RACON/MEDAKA/PILON/SALSA2/ALLHIC assembly and zooms in on the misassembly both before and after manual correction. (b) Scaffold lengths of the 13 pseudochromosomes and the longest two additional scaffolds. (c) A panel of quantitative statistics generated to compare each scaffolding iteration

FLYE/MEDAKA/RACON/PILON/ALLHIC assemblies were selected for further validation using 14 linkage groups generated for the macadamia genome (Nock et al., 2020). Macadamia was selected as it belongs to the Proteaceae and shares a karyotype of 14 chromosomes with *K. excelsa*. Using BLASTN, unique markers were identified (mean =227 unique markers identified per map) across nine maps. These unique markers were then mapped using ALLMAPS (Tang et al., 2015b) to the two *K. excelsa* assemblies and the order and orientation of scaffolds were visually examined for synteny (Appendix S4). Whole genome alignments were constructed to compare each macadamia-informed assembly to its original Hi-C assembly. From this it was clear that

scaffolds generated using SALSA2 shared a greater proportion of synteny with macadamia when compared to ALLHIC scaffolds, although they were less contiguous. To further assess accuracy, the location of the telomere motif “TTAAGGG” was identified in each assembly using EMBOSS (Rice et al., 2000) and visual constructions created using CHROMOMAP (Anand, 2019) (Appendix S4). These analyses indicate that both Hi-C scaffolders were unable to accurately represent telomere sequences, although SALSA2 scaffolds generated a more accurate assignment than scaffolds constructed using ALLHIC.

Overall, the analyses confirmed that the FLYE/MEDAKA/RACON/PILON/SALSA2 assembly outperformed the scaffolds produced by ALLHIC

with respect to the orientation and ordering of scaffolds and accuracy of regions of complexity.

### 3.7 | Pseudochromosome-level assembly construction

Based on a combination of the quantitative metrics and the linkage group validation outlined in previous sections, the FYLE/MEDAKA/RACON/PILON/SALSA2 assembly was selected for further scaffolding using the ALLHIC package. By specifying the expected karyotype, ALLHIC binned contigs into pseudochromosomes, and metrics of 92% kmer completion, a QV score of 26 and a total length of 0.97 Gb were obtained. However, although a higher level of contiguity was achieved, gene completeness scores dropped to 80% from 87%. On inspection of the ALLHIC contact map produced using the PRETEXT software package a misassembly was identified and rectified through manual intervention using JUICEBOX (Figure 6a). This manual curation resulted in a 10% increase of gene completeness whilst retaining the contiguity provided by ALLHIC. The final *K. excelsa* assembly (Table 3) had a 90% and 97% gene complete using the *embryophyta* and *eukaryota* databases, respectively, and an N50 of 114 Mb (Figure 6c), a karyotype similar to that expected for the species (Figure 6b) and which is available through <https://doi.org/10.7931/paqq-kk20>.

## 4 | DISCUSSION

Here, we describe the impact of data volume and coverage on assembly, polishing and scaffolding workflows through analyses of four ONT read length subsamples. We also outline how to optimize a workflow for "optimal" pseudochromosomal assembly construction

using *Knightsia excelsa* as an exemplar. It is crucial that an appropriate assembler and sequencing strategy is selected prior to data generation, in order to maximize the use of both the data volume and read length during assembly to meet the goal of constructing an assembly optimal for individual project needs. The performance of five ONT assemblers across four read length subsets (reads >5 kb only, >10 kb only, >22 kb only, >30 kb only) were investigated. By subsampling the input ONT data we examined the impact of both data coverage and volume on quantitative metrics and how these metrics can be used to inform the best assembly approach, facilitating the reconstruction of the optimal initial ONT contigs for further Hi-C scaffolding. Furthermore, this highlights the use of using a range of quantitative metrics when dealing with different data volumes and qualities to generate high-quality nonmodel organism genome assemblies. We separately assessed the optimal assembly workflow utilizing all available ONT data for the generation of the first high-quality genome assembly for Rewarewa.

### 4.1 | Long read assembler performance with iterative polishing for assemblies generated using subsampled ONT data

In order to compare assembly performance, quantitative metrics such as total length, contiguity and N50 for each assembly were generated (Figure 2). SHASTA was built for quick assembly construction and was originally developed for the human genome, with 11 human genomes assembled in 9 days on a single computer node. This was made possible by strategic read length encoding, reduced marker representation and heuristics. However, although being fast, SHASTA requires a large amount of RAM, 1–2 Tb for the human genome (Shafin et al., 202b), which is not always readily available and compromises assembly accuracy over contiguity and total length. Figure 2 highlights SHASTA'S

TABLE 3 Pseudochromosome length of the final *Knightsia excelsa* genome assembly

Pseudochromosome number	Length (bp)	
1	158,258,002	
2	128,337,228	
3	115,057,523	
4	114,119,817	
5	89,465,163	
6	87,992,687	
7	58,927,204	
8	47,592,350	
9	46,680,506	
10	42,698,337	
11	36,056,839	
12	28,555,625	
13	8,702,882	
14	2,822,772	
Sum of pseudochromosome length	965,266,935	965 Mb

shortcoming with assemblies across all read lengths unable to represent the expected total length of *K. excelsa* (1 Gb) whilst consistently achieving high gene completeness scores.

Interestingly, Figure 2 illustrates this assembler's dependency on a high depth of coverage for optimal performance, with subsamples that include additional read depth, for example >5-kb assemblies being of better quality than those with a lower depth of coverage such as >30 kb. *SHASTA* appears to positively respond to iterative *RACON* polishing with continuous improvements in contiguity found across subsamples when implemented (Figure 2). When alternative polishing strategies were tested across subsampled data sets *SHASTA* appeared unbiased with both general and sequencer-specific polishers performing equally well in the >30- and 22-kb subsamples whilst *MEDAKA*-based approaches seem to outperform general approaches in the >5- and >10-kb subsample (Figure 3).

*WTD* produces highly fragmented assemblies in comparison to all other assemblers when shorter read lengths are included, with enhanced N50 and reduced number of contigs occurring when only longer read lengths are provided (Figure 2). This result is due to its underlying algorithm having only a single consensus step and is reiterated by a significant improvement of contiguity, gene completeness and N50 after *RACON* and *PILON* polishing, which has been identified as an issue in other plant species assemblies such as *Acer yangbiense* (Yang et al., 2019).

Consistent with results found for prokaryotes (Wick & Holt, 2019), we demonstrated *WTD*'s decreased performance at lower read depths with assemblies produced by >30-kb read subsample failing to span the expected total length, which was not rescued by iterative polishing. *WTD* was the only assembler identified without a bias toward a *MEDAKA*-based polishing strategy, with *RACON*-based strategies also performing well.

*CANU*'s optimal performance is reached when only longer read lengths are provided, and performance is compromised when additional shorter read length data are added. This clearly shows a preference by this assembler for read length over depth of coverage, supporting claims made by the developers that only >20 $\times$  coverage is required for accurate assembly. Algorithmically, *CANU* contains extensive rounds of error correction and consensus, and the developers do not suggest additional long read polishing. Thus, as expected, the post-assembly iterative polishing shown in Figure 2 has the least effect on these assemblies when compared to all other assemblers, as the initial assemblies generated have substantially fewer errors to correct. This finding does not appear to be specific to *RACON* and *MEDAKA* long read polishers only, as these minimal effects have also been identified by other long read polishing tools. For example, it has been shown that by polishing bacterial assemblies generated by *CANU* using *NANOPOLISH* an increase in errors found in the assembly occurred when compared to short-read polishing alone (Goldstein et al., 2019). Similar results are represented in Figure 2, as implementing the short-read polishing recommended by developers achieved a substantial gene completeness score across all subsamples.

*FYLE* achieved the most robust performance, with the assemblies generated not significantly impeded by the addition or exclusion of

certain read lengths or read depths. This result has been demonstrated for bacterial genome assemblies whereby the assembler performs well at <10 $\times$  coverage and in *Eucalyptus pauciflora* genome assembly (Wang et al., 2020) where *FYLE* performs consistently well when >1-kb read lengths are subsampled when compared to >35-kb read length subsamples. Iterative *RACON* polishing has a marginal beneficial effect (Figure 2) and across all read lengths a combined *MEDAKA* and *RACON* polishing strategy yields the most enhanced genome assembly (Figure 3).

Overall, this analysis highlights the advantages and shortcomings of various assemblers and provides use cases for each. *SHASTA*, when given higher coverage data, is an incredibly powerful assembler that runs quickly (Wick & Holt, 2019) and generates extremely accurate contigs. However, *SHASTA* is not robust with regard to data volume, as without sufficient read depth this assembler performs suboptimally and fails to generate complete assemblies. It could still be useful at a lower depth of coverage for the purposes of complete and fast gene identification, particularly for large genomes. In comparison, *CANU* is the slowest running assembler, due to its extensive pre-assembly error correction and trimming steps. However, length can be prioritized over depth when using this assembler, and the incorporation of shorter read lengths may even result in suboptimal results. The advantage of this assembler for more advanced users is the ability to modify parameters, although this may not be appropriate for novices. *FYLE* is the most robust of the assemblers tested, with results across subsamples appearing consistent. This assembler may be an attractive tool for most data volumes and particularly for novice usage, as minimal parameter adjustments are required with the single caveat of a user-defined genome size.

## 4.2 | An assessment of the impact of read length and data volume on Hi-C mapping performance

To assess the impact of initial assembly quality on Hi-C mapping performance, Hi-C data, generated using the Phase Genomics kit, were mapped to the initial assemblies generated across the >5-, >10-, >22- and >30-kb subsamples. The contigs were scaffolded using two commonly used software packages, *ALLHiC* and *SALSA2*. *ALLHiC* uses "pruning" and "optimization" steps to produce allele-aware scaffolds, although it requires a priori knowledge of the chromosome number. *SALSA2* uses the ONT assembly graph in order to assess assembly accuracy prior to Hi-C scaffolding and does not require karyotype information. Assessing assembly constructs from these two scaffolding software programs allowed not only a comparison of informed (*ALLHiC*) and noninformed (*SALSA2*) strategies but also the performance of a software that corrects misassemblies prior to scaffolding to a software that scaffolds based on the input assembly alone. When comparing scaffold performance, it was important to integrate quantitative metrics such as N50, total length along with intrinsic karyotype information.

Amongst all four read length subsamples, the resulting assemblies from the *SALSA2*-generated scaffolds were of greater accuracy



than those produced by ALLHIC (Table 2). SALS2 generated assemblies that were both more gene complete and of a total length closer to the expected genome size. These findings are further supported by Figure 4(c,d) that highlight suspicious scaffold length distributions constructed by ALLHIC, a result of over-assembly. This over-assembly could be a consequence of the homozygous *K. excelsa* sample resulting in a reduced long-range interaction signal (Zhang et al., 2019) and other more heterozygous genomes may indeed perform better.

ALLHIC fails to construct scaffolds from highly fragmented ONT assemblies, and this is highlighted in its inability to construct scaffolds for >5- and 10-kb read subsample assemblies produced by WTD, which contain 10,317 and 10,043 contigs respectively. SALS2 circumvents this issue by removing all contigs <1,000 bp prior to Hi-C assembly.

Through short contig removal, SALS2 achieves scaffolds utilizing all WTD subsamples and overall performs better when longer reads are supplied, although the >30-kb read subsample SALS2 scaffolding slightly underestimated the total length. As previously mentioned, all SHASTA-generated assemblies, although gene-complete, underestimate the total length of the genome and despite Hi-C scaffolding the assembly failed to gain coverage of the entire genome but retained gene completeness score. For instance, the SHASTA assembly constructed using the >30-kb subsample and scaffolded with SALS2's gene completeness was 85%, which was the highest score achieved when compared to all other ONT assemblers, but the total length was less than half the expected length. Similarly, SALS2 performs optimally for both the CANU and FYLE assembly constructed using the >30-kb subsample as it does with SHASTA.

From the analyses highlighted in Table 2 it is clear that the input assembly does substantially affect the accuracy of Hi-C scaffolding, as the issue of fragmentation found in ONT assemblies produced by WTD profoundly affected the scaffolding process, generating assemblies of low contiguity even after Hi-C scaffolding. Furthermore, the lack of genome length coverage of the initial SHASTA assemblies was not resolved with the addition of Hi-C data. Again, FYLE appears more robust than other assemblers to scaffolding software, although ALLHIC still produces a suspicious read length distribution, suggesting over-assembly in individual subsamples. CANU assemblies perform suboptimally using the ALLHIC scaffold but the high-quality initial >30-kb read subsample ONT assembly appears to remain the superior assembly after SALS2 scaffolding.

### 4.3 | A pseudochromosome-length near-complete genome assembly for Proteaceae using all ONT data

In order to generate a high-quality *K. excelsa* genome the initial 18 ONT contig sets generated using all of the available ONT data were used and quality metrics were obtained (Figure 2). These contigs sets were scaffolded by both ALLHIC and SALS2 and compared (Figure 5). Overall, this analysis highlighted the importance of the accuracy of the underlying ONT assembly as errors found in these assemblies (WTD, SHASTA and NECAT) were unresolved by further scaffolding with

Hi-C data. This result is consistent with the accuracy of the initial ONT contigs produced by both FYLE and CANU maintaining superior quantitative metrics after scaffolding (Figure 4a,b). Both CANU and FYLE Hi-C assemblies retained a high consensus quality (QV), kmer completion scores peaking at 91% and 93% respectively and map back rates peaking at 84% for both (Figure 5). Furthermore, SHASTA, NECAT and WTD all failed to produce reliable scaffolded assemblies, suffering from collapsed genome lengths, low kmer completeness, and poor mapping back rates and therefore were not considered for further analyses.

Interestingly, FYLE-based Hi-C assemblies appear to have a higher degree of gene completeness in comparison to CANU with only a slightly smaller total assembly length than expected from flow cytometry. FYLE initial assemblies also appear robust to different scaffolding strategies with similar results across both ALLHIC and SALS2 being achieved. Focusing on N50, contiguity, gene completeness scores and total length alone lead to misleading conclusions about genome accuracy being drawn as ALLHIC appeared more contiguous and had similar total lengths and gene completeness scores when compared to SALS2 assemblies. Through the integration of scaffold length distributions as a quality metric, the accuracy of these assemblies could be more thoroughly evaluated as the karyotype shows all 14 chromosomes are of similar length (Hair & Beuzenberg, 1958) and this should be represented in the scaffolds produced after Hi-C data integration. Here, SALS2 had a more realistic length distribution whilst ALLHIC generated assemblies with length distributions inconsistent with the karyotype.

FYLE-ALLHIC assemblies produced high-quality metrics with a gene completeness of 88%, N50 of 66.6 Mbp, kmer completeness of 91% and a Hi-C read map-back rate of 80% (Figure 5). However, total length is lower than expected at 816 Mbp, and evidence of over-assembly was identified with 50% of the genome being placed on a single chromosome (Figure 4a). Comparatively, FYLE-SALS2 also performed well, with a gene completeness score of 87%, a total length of 977 Mbp, kmer completeness of 84%, a Hi-C read map-back rate of 84% and a chromosome length distribution in line with what is expected for this species (Figure 5). However, in this case contiguity suffered with an N50 of only 1.56 Mbp being obtained. Due to the superiority of the overall quantitative metrics obtained for FYLE, ONT assemblies scaffolded with ALLHIC and SALS2 were selected for further inspection.

Both assemblies were further validated for structure and orientation accuracy through comparison to macadamia linkage maps. This analysis highlighted the accuracy and orientation of both scaffold sets. Despite the FYLE-SALS2 being less contiguous, the metrics suggested more accurate scaffolds (Figure 4b) and therefore this assembly underwent pseudochromosome reconstruction through an additional round of ALLHIC scaffolding. After pseudochromosome reconstruction, contiguity was increased with an N50 of 114 Mbp, whilst retaining a high-quality total length and kmer completion, although gene completeness scores dropped by 8% (Figure 6c). In order to assess this, the data were manually curated using JUICER and JUICEBOX. Here, a single misassembly was detected and manually

curated (Figure 6c) resulting in a genome that is 91% kmer-complete, 97.5 Mbp in length, 90% gene-complete (99% complete if considering Eukaryota data set) and has an N50 of 114 Mbp (Figure 6c).

This assembly is the first near-complete genome sequence for the Proteaceae clade and will provide invaluable information to the honey production industry in Aotearoa New Zealand, but also provides a reference for other Proteaceae in this clade.

## 5 | CONCLUSIONS

Our long read and Hi-C-based assemblies of *Knightsia excelsa* could potentially be useful as a benchmarking resource to be utilized regularly on release of new ONT assembly and Hi-C scaffolding tools. This will allow the continuous assessment of performance of new genomic packages across both read length and read depth. Furthermore, this could enhance the genomics community's ability to make a more educated *de novo* genome assembly pipeline prior to assembly whilst also giving information on the data volume required. In future it will be important that more assemblers, polishing mechanisms and Hi-C scaffolders are investigated and benchmarked. Finally, the *K. excelsa* assembly produced here will be used to assess the genomic diversity of rewarewa across its natural range in Aotearoa New Zealand, in collaboration with Māori agribusinesses involved in the honey industry.

## ACKNOWLEDGEMENTS

The authors thank the Te Rarawa iwi (tribe) for supporting this research and approving publication of this paper. The authors also thank members of Genomics Aotearoa's High Quality Genomes project, support from New Zealand eScience Infrastructure (NeSI) and Amali Thrimawithana (Plant & Food Research) and Abdul Batten (AgResearch) for useful comments on the manuscript. Leaf samples were collected by Peter Bellingham (Manaaki Whenua – Landcare Research) with field support from Mike White (Northland Regional Council). The Department of Conservation helped with collecting logistics and permits (CA-31615-OTH).

## AUTHOR CONTRIBUTIONS

A.M., T.R.B. and D.C. conceived the study. J.M.P. collected samples and G.H. coordinated the engagement with Te Rarawa Anga Mua and the Komiti Kaitiaki for Warawara Ngahere. E.H. performed the laboratory work. A.M. analysed and interpreted the data. S.S.C., C.W. and J.G. contributed to Hi-C and scaffolding analysis. A.M. and D.C. wrote the manuscript with input from all authors. D.C. and T.R.B. coordinated the project.

## AVAILABILITY OF SOURCE CODE AND REQUIREMENTS

See Appendix S1.

## DATA AVAILABILITY STATEMENT

Permission from representatives of the Indigenous Peoples (Māori) was obtained for using the plant material used for this study. Further

studies using this material, raw sequencing data and final genome assembly will require consent from the Māori iwi (tribe) who exercises guardianship for this material according to Aotearoa New Zealand's Treaty of Waitangi and the international Nagoya protocol on the rights of indigenous peoples. Raw and analysed data are available through the Manaaki Whenua Landcare Research data repository (<https://doi.org/10.7931/paqq-kk20>) with managed access. Access to these data will require permission from representatives of the Te Rarawa iwi (tribe).

## ORCID

Ann M. McCartney  <https://orcid.org/0000-0003-3191-3200>

Thomas R. Buckley  <https://orcid.org/0000-0002-3076-4234>

David Chagné  <https://orcid.org/0000-0003-4018-0694>

## REFERENCES

- Anand, L. (2019). *chromoMap: An R package for Interactive Visualization and Annotation of Chromosomes*.
- Ankenbrand, M., Pfaff, S., Terhoeven, N., Qureschi, M., Gündel, M., Weiß, C., Hackl, T., & Förster, F. (2018). chloroExtractor: extraction and assembly of the chloroplast genome from whole genome shotgun data. *Journal of Open Source Software*, 3(21), 464.
- Berger, M. F., & Mardis, E. R. (2018). The emerging clinical relevance of genomics in cancer medicine. *Nature Reviews Clinical Oncology*, 15(6), 353–365. <https://doi.org/10.1038/s41571-018-0002-6>
- Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R. L., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31(12), 1119–+. <https://doi.org/10.1038/nbt.2727>
- Bzikadze, A. V., & Pevzner, P. A. (2019). centroFlye: Assembling centromeres with long error-prone reads. *bioRxiv*, 38, 772103. <https://doi.org/10.1101/772103>
- Chen, Y., Nie, F., Xie, S.-Q., Zheng, Y.-F., Bray, T., Dai, Q., Wang, Y. X., Huang, Z., Wang, D. P., He, L. J., Luo, F., Wang, J. X., Liu, Y. Z., Xiao, C. L., & Xiao, C.-L. (2020). Fast and accurate assembly of Nanopore reads via progressive error correction and adaptive read selection. *bioRxiv*. <https://doi.org/10.1101/2020.02.01.930107>
- Christenhusz, M. J. M., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, 261(3), 201–217. <https://doi.org/10.11646/phytotaxa.261.3.1>
- Cock, P. J. A., Bonfield, J., Chevreur, B., & Li, H. (2015). SAM/BAM format v1.5 extensions for *de novo* assemblies. *bioRxiv*, 20024. <https://doi.org/10.1101/020024>
- Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Computational Biology*, 15(8), e1007273. <https://doi.org/10.1371/journal.pcbi.1007273>
- Goldstein, S., Bekka, L., Graf, J., & Klassen, J. L. (2019). Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics*, 20(1), 23. <https://doi.org/10.1186/s12864-018-5381-7>
- Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., Doumatey, A. P., ... & Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, 517(7534), 327–332. <https://doi.org/10.1038/nature13997>
- Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUASt: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>

- Hair, J. B., & Beuzenberg, E. J. (1958). Contributions to a chromosome atlas of New Zealand flora - 1. *New Zealand Journal of Science*, 1, 617–628.
- Hilario, E. (2018). *Plant nuclear genomic DNA preps. protocols.io*. Retrieved from <https://doi.org/10.17504/protocols.io.rncd5aw>
- Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5), 540–546. <https://doi.org/10.1038/s41587-019-0072-8>
- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5), 722–736. <https://doi.org/10.1101/gr.215087.116>
- Lanfear, R., Schalamun, M., Kainer, D., Wang, W., & Schwesinger, B. (2019). MinIONQC: Fast and simple quality control for MinION sequencing data. *Bioinformatics*, 35(3), 523–525. <https://doi.org/10.1093/bioinformatics/bty654>
- Langdon, K. S., King, G. J., Baten, A., Mauleon, R., Bundock, P. C., Topp, B. L., & Nock, C. J. (2020). Maximising recombination across macadamia populations to generate linkage maps for genome anchoring. *Scientific Reports*, 10(1), 5048. <https://doi.org/10.1038/s41598-020-61708-6>
- Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., ... Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17), 4325–4333. <https://doi.org/10.1073/pnas.1720115115>
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), 289–293. <https://doi.org/10.1126/science.1181369>
- Lightfoot, D. J., Jarvis, D. E., Ramaraj, T., Lee, R., Jellen, E. N., & Maughan, P. J. (2017). Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biology*, 15(1), 74. <https://doi.org/10.1186/s12915-017-0412-4>
- Linsmith, G., Rombauts, S., Montanari, S., Deng, C. H., Celton, J.-M., Guérif, P., Liu, C., Lohaus, R., Zurn, J. D., Cestaro, A., Bassil, N. V., Bakker, L. V., Schijlen, E., Gardiner, S. E., Lespinasse, Y., Durel, C.-E., Velasco, R., Neale, D. B., Chagné, D., ... Bianco, L. (2019). Pseudo-chromosome-length genome assembly of a double haploid "Bartlett" pear (*Pyrus communis* L.). *Gigascience*, 8(12), giz138–giz138. <https://doi.org/10.1093/gigascience/giz138>
- Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8), 733–735. <https://doi.org/10.1038/nmeth.3444>
- Marshall, C. W., Chagné, D., Deusch, O., Gruenheit, N., McCallum, J., Bergin, D., Lockhart, P. J., & Wilcox, P. L. (2015). A DNA-based diagnostic for differentiating among New Zealand endemic *Podocarpus*. *Tree Genetics & Genomes*, 11(4), <https://doi.org/10.1007/s11295-015-0888-4>
- Morgan, E. R., Perry, N. B., & Chagne, D. (2019). Science at the intersection of cultures - Maori, Pakeha and manuka. *New Zealand Journal of Crop and Horticultural Science*, 47(4), 225–232. <https://doi.org/10.1080/01140671.2019.1691610>
- Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, 21, 79–85. <https://doi.org/10.1093/bioinformatics/bti1114>
- Naim, F., Nakasugi, K., Crowhurst, R. N., Hilario, E., Zwart, A. B., Hellens, R. P., Taylor, J. M., Waterhouse, P. M., & Wood, C. C. (2012). Advanced engineering of lipid metabolism in *Nicotiana benthamiana* using a draft genome and the V2 viral silencing-suppressor protein. *PLoS One*, 7(12), e52717. <https://doi.org/10.1371/journal.pone.0052717>
- Nock, C. J., Baten, A., Mauleon, R., Langdon, K. S., Topp, B., Hardner, C., Furtado, A., Henry, R. J., & King, G. J. (2020). Chromosome-scale assembly and annotation of the Macadamia genome (*Macadamia integrifolia* HAES 741). *G3: Genes|genomes|genetics*, 10(10), 3497–3504. <https://doi.org/10.1534/g3.120.401326>
- O'Connor, K., Hayes, B., Hardner, C., Nock, C., Baten, A., Alam, M., & Topp, B. (2020). Genome-wide association studies for yield component traits in a macadamia breeding population. *BMC Genomics*, 21(1), 199. <https://doi.org/10.1186/s12864-020-6575-3>
- Ou, S., Chen, J., & Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research*, 46(21), e126. <https://doi.org/10.1093/nar/gky730>
- Peichel, C. L., Sullivan, S. T., Liachko, I., & White, M. A. (2017). Improvement of the threespine stickleback genome using a Hi-C-based proximity-guided assembly. *Journal of Heredity*, 108(6), 693–700. <https://doi.org/10.1093/jhered/esx058>
- Prachi, P., Donati, C., Masciopinto, F., Rappuoli, R., & Bagnoli, F. (2013). Deep sequencing in pre- and clinical vaccine research. *Public Health Genomics*, 16(1–2), 62–68. <https://doi.org/10.1159/000345611>
- Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, 11(1), 1432. <https://doi.org/10.1038/s41467-020-14998-3>
- Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A. M., Gedman, G., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L. S., Lee, C., June, K. B., Kim, J., Bista, I., Smith, M., Haase, B., ... Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv*, 592, 737–746. <https://doi.org/10.1038/s41586-021-03451-0>
- Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality and phasing assessment for genome assemblies. *bioRxiv*. <https://doi.org/10.1101/2020.03.15.992941>
- Rice, P., Longden, I., & Bleasby, A. (2000). EMBOS: The European molecular biology open software suite. *Trends in Genetics*, 16(6), 276–277. [https://doi.org/10.1016/s0168-9525\(00\)02024-2](https://doi.org/10.1016/s0168-9525(00)02024-2)
- Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 17(2), 155–+. <https://doi.org/10.1038/s41592-019-0669-3>
- Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., ... Paten, B. (2020a). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, 38(9), 1044–1053. <https://doi.org/10.1038/s41587-020-0503-6>
- Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., ... Paten, B. (2020b). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, 33. <https://doi.org/10.1038/s41587-020-0503-6>
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., Schnable, P. S., Lyons, E., & Lu, J. (2015a). ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology*, 16(1), 3. <https://doi.org/10.1186/s13059-014-0573-1>
- Tang, H. B., Zhang, X. T., Miao, C. Y., Zhang, J. S., Ming, R., Schnable, J. C., Schnable, P. S., Lyons, E., & Lu, J. (2015b). ALLMAPS: robust

- scaffold ordering based on multiple maps. *Genome Biology*, 16(3). <https://doi.org/10.1186/s13059-014-0573-1>
- Technologies, O. N. (2018). *Medaka*.
- Thrimawithana, A. H., Jones, D., Hilario, E., Grierson, E., Ngo, H. M., Liachko, I., Sullivan, S., Bilton, T. P., Jacobs, J. M., Bicknell, R., David, C., Deng, C., Nieuwenhuizen, N., Lopez-Girona, E., Tobias, P. A., Morgan, E., Perry, N. B., Lewis, D. H., Crowhurst, R., & Schwinn, K. E. (2019). A whole genome assembly of *Leptospermum scoparium* (Myrtaceae) for mānuka research. *New Zealand Journal of Crop and Horticultural Science*, 47(4), 233–260. <https://doi.org/10.1080/01140671.2019.1657911>
- Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5), 737–746. <https://doi.org/10.1101/gr.214270.116>
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11), e112963. <https://doi.org/10.1371/journal.pone.0112963>
- Wang, W., Das, A., Kainer, D., Schalamun, M., Morales-Suarez, A., Schwessinger, B., & Lanfear, R. (2020). The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing de novo assemblies. *Gigascience*, 9(1), <https://doi.org/10.1093/gigascience/giz160>
- Wick, R. R., & Holt, K. E. (2019). Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research*, 8, 2138. <https://doi.org/10.12688/f1000research.21782.2>
- Yang, J., Wariss, H. M., Tao, L., Zhang, R., Yun, Q., Hollingsworth, P., Dao, Z., Luo, G., Guo, H., Ma, Y., & Sun, W. (2019). De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan Province, China. *Gigascience*, 8(7), <https://doi.org/10.1093/gigascience/giz085>
- Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821–829. <https://doi.org/10.1101/gr.074492.107>
- Zhang, X., Zhang, S., Zhao, Q., Ming, R., & Tang, H. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants*, 5, <https://doi.org/10.1038/s41477-019-0487-8>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** McCartney AM, Hilario E, Choi S-S, et al. An exploration of assembly strategies and quality metrics on the accuracy of the rewarewa (*Knightia excelsa*) genome. *Mol Ecol Resour.* 2021;21:2125–2144. <https://doi.org/10.1111/1755-0998.13406>