
The draft nuclear genome assembly of *Eucalyptus pauciflora*: new approaches to comparing *de novo* assemblies

Weiwen Wang^{1*^}, Ashutosh Das^{1,2^}, David Kainer¹, Miriam Schalamun^{1,3},
Alejandro Morales-Suarez⁴, Benjamin Schwessinger¹, Robert Lanfear^{1*}

1. Research School of Biology, the Australian National University, Canberra, Australia

2. Department of Genetics and Animal Breeding, Faculty of Veterinary Medicine, Chittagong Veterinary and Animal Sciences University, Chittagong, Bangladesh

3. Institute of Applied Genetics and Cell Biology, University of Natural Resources and Life Sciences, Vienna, Austria

4. Department of Biological Sciences, Macquarie University, Sydney, Australia

[^] Equal contribution

* Corresponding authors: wei.wang@anu.edu.au and rob.lanfear@anu.edu.au

Email:

Weiwen Wang: wei.wang@anu.edu.au

Ashutosh Das: ashutosh.das@cvasu.ac.bd

David Kainer: dkainer@outlook.com

Miriam Schalamun: miriam.schalamun@gmail.com

23 Alejandro Morales-Suarez: eder-alejandro.morales-suar@hdr.mq.edu.au

24 Benjamin Schwessinger: benjamin.schwessinger@anu.edu.au

25 Robert Lanfear: rob.lanfear@anu.edu.au

26

27

28

29

30

31 **Abstract**

32 **Background**

33 Selecting the best genome assembly from a collection of draft assemblies for the same
 34 species remains a difficult task. Here, we combine new and existing approaches to
 35 help to address this, using the non-model plant *Eucalyptus pauciflora* (snow gum) as a
 36 test case. *Eucalyptus pauciflora* is a long-lived tree with high economic and
 37 ecological importance. Currently, little genomic information for *Eucalyptus*
 38 *pauciflora* is available.

39 **Findings**

40 We generated high coverage of long- (Nanopore, 174x) and short- (Illumina, 228x)
 41 read data from a single *Eucalyptus pauciflora* individual and compared assemblies
 42 from four assemblers with a variety of settings: Canu, Flye, Marvel, and MaSuRCA.
 43 A key component of our approach is to keep a randomly selected collection of ~10%
 44 of both long- and short-reads separate from the assemblies to use as a validation set

with which to assess the assemblies. Using this validation set along with a range of existing tools, we compared the assemblies in eight ways: contig N50, BUSCO scores, LAI scores, assembly ploidy, base-level error rate, computing genome assembly likelihoods, structural variation and genome sequence similarity. Our result showed that MaSuRCA generated the best assembly, which is 594.87 Mb in size, with a contig N50 of 3.23 Mb, and an estimated error rate of ~0.006 errors per base.

Conclusions

We report a draft genome of *Eucalyptus pauciflora*, which will be a valuable resource for further genomic studies of eucalypts. These approaches for assessing and comparing genomes should help in assessing and choosing among many potential genome assemblies for a single species.

Keywords: Long-read assembly; nanopore sequencing; hybrid assembly; genome assessment; assembly comparison; *Eucalyptus pauciflora*; haplotig separation; genome polishing

Data Description

Introduction

Eucalypts are widely distributed in Australia, including three genera *Eucalyptus*, *Corymbia* and *Angophora*, and have around 900 species [1]. *Eucalyptus pauciflora* (*E. pauciflora*) (Fig. 1), also known as snow gum, is a highly variable eucalyptus species that inhabits diverse landscapes in south-eastern Australia [1]. *E. pauciflora* can survive from close to sea level to up to the tree line of the Australian Alps, displaying the broadest altitudinal range in the *Eucalyptus* genera [2-4]. Due to its wide distribution and drought and cold tolerance, *E. pauciflora* is used for carbon offset plantings, ecological restoration, honeybee food source, and also has medicinal uses [1, 5-11]. However, genomic resources for *E. pauciflora* are currently very limited: there exists a single chloroplast genome [12], two sets of microsatellite markers [13, 14], and two nuclear loci used for phylogenetics [15]. The assembly of *E. pauciflora* genome will assist in elucidating the genetic basis of cold tolerance in *Eucalyptus*.

Across the ~900 extant eucalypt species, there are only two genomes published: those for *E. grandis* and *E. camaldulensis* [16, 17]. Both of these genomes were sequenced with a combination of Sanger sequencing and short-read sequencing, and as a result both assemblies are somewhat fragmented. There are 81,246 scaffolds in *E. camaldulensis* assembly [17]. While the *E. grandis* genome is highly contiguous, assembled to chromosome level, it still has 4,941 unplaced scaffolds [16]. New technologies, such as third-generation long-read sequencing, have the potential to

89 produce less fragmented assemblies at a fraction of the cost of previous methods.

90 Nevertheless, many challenges still remain, not least of which is that different genome

91 assembly software, and small changes to the parameters of a single piece of software,

92 can produce substantially different assemblies. In light of this, methods for choosing

93 the most accurate assembly from a set of possible assemblies have become

94 increasingly important.

95

96 Two metrics are commonly used to assess and compare genome assemblies: contig

97 N50 and Benchmarking Universal Single-Copy Orthologs (BUSCO [18],

98 RRID:SCR_015008) scores. The contig N50 is the size of the contig where at least 50%

99 of the assembled nucleotides can be found in contigs of that size or larger. The N50 is

100 a measure of genome contiguity, where a higher N50 suggests a genome that has been

101 assembled into fewer and larger contigs. All else being equal, we should prefer

102 genome assemblies with a larger N50, up to the point where the N50 is equal to the

103 N50 of the chromosomes themselves. Perhaps because of this, the N50 is one of the

104 most widely reported metrics in genome assembly. However, it is important to

105 remember that the N50 measures contiguity, not accuracy. For example, N50 scores

106 may be artificially inflated by incorrectly linking contigs [19, 20]. The BUSCO score

107 estimates the proportion of highly conserved orthologous genes that are present in

108 assemblies. The underlying assumption is that there exists a certain set of highly

109 conserved single-copy genes, the vast majority of which we should expect to observe

110 in single copies in any given haploid genome assembly. BUSCO scores provide a very

111 useful measure of genome assembly completeness (a component of accuracy), and in
112 principle we should prefer genome assemblies with BUSCO scores closer to 100%.
113 One limitation of BUSCO scores is that they assess only a very small proportion of
114 the genome, typically around 1000 highly conserved genes which represent less than
115 1% of the total genome. Furthermore, by their nature these protein-coding regions of
116 the genome tend to be among the easiest to assemble because they are usually
117 single-copy regions of high complexity. Hence, assemblies can have very similar
118 BUSCO scores even if they differ considerably in their assembly of the non-BUSCO
119 genomic regions, which means that it is sometimes difficult to use BUSCO scores to
120 distinguish among competing assemblies [21]. In this study, we complement these
121 commonly-used measures with a range of other metrics to assess and compare
122 genome assemblies, and we use these measures to choose the best draft assembly of *E.*
123 *pauciflora*.

124

125 One measure we propose is the assembly ploidy: the proportion of the genome that is
126 represented by haploid contigs. One important problem in genome assembly is that
127 we commonly represent the genome of diploid (or polyploid) organisms as a haploid
128 sequence. Traditionally, genome projects would alleviate this problem by sequencing
129 highly inbred individuals [22, 23], thus reducing the discrepancy between the diploid
130 individual and the haploid representation. However, as genome assembly has become
131 more commonplace, we often want to assemble the genomes of highly heterozygous
132 individuals. For example, heterozygosity in *Eucalyptus* is around 1% [24], and varies

133 substantially along the genome [16]. The consequence of this is that regions of low
134 heterozygosity tend to be assembled into a single collapsed haploid sequence, whereas
135 regions of high heterozygosity tend to be assembled into two haplotypes of the same
136 region, which are usually labelled the ‘primary contig’ (referring to the longer of the
137 two contigs) and the ‘haplotig’ (referring to the shorter of the two contigs) [25].
138 Although there has been some progresses in estimating truly diploid assemblies [25,
139 26], most assemblers still produce primary contigs and haplotigs without labelling
140 them as such [27, 28]. Crucially, unidentified haplotigs may cause issues in the
141 downstream analyses, because many analyses assume that we have a haploid
142 representation of the genome. Because of this, we propose a novel and simple (but
143 imperfect) metric to measure the assembly ploidy, which is simply the ratio of the
144 assembly size to the estimated haploid genome size. If the aim is to produce a haploid
145 representation of a genome, then an assembly ploidy of 1 is preferable (i.e. the
146 assembly size should equal the estimated haploid genome size). If the aim is to
147 produce a diploid representation of a genome, then an assembly ploidy of 2 is
148 preferable (i.e. the assembly size should be double the estimated haploid genome size).
149 One limitation of this metric is that it is sensitive to errors in the estimation of haploid
150 genome size, and it is also sensitive to errors in genome assembly (e.g. highly
151 incomplete assemblies) that might affect the numerator. Nevertheless, in combination
152 with other measures, we show below that the assembly ploidy provides a useful
153 metric with which to compare genome assemblies.

154

155 We also apply a suite of measures designed to provide a genome-wide assessment of
156 contiguity and accuracy that can complement the widely-used contig N50 and
157 BUSCO scores. The advantages of these measures lie in the fact that they assess more
158 of the genome than BUSCO scores, though each also has its limitations. The first
159 measure is the long-terminal repeat (LTR) assembly index, or LAI [21]. The LAI
160 score is the proportion LTR sequences in the genome that are intact, and is
161 independent of genome size and repeat content. In general, a higher LAI score
162 suggests a more contiguous and complete assembly [21]. The second measure we use
163 is the base-level error rate evaluated by remapping independent sets of long and short
164 validation reads (around 10% of all reads, randomly selected) to the assembly.
165 Previous studies have evaluated the base-level error rate by remapping all reads to the
166 assembly [29, 30]. Here, we use validation reads which are not involved in the
167 assembly, in order to avoid any possible biases introduced by validating an assembly
168 with the same data that was used to produce it. For a perfect assembly in which the
169 ploidy of the entire assembly matches the ploidy of the individual, a lower base-level
170 error rate is preferable, with a theoretical minimum of the error rate of the sequencing
171 technology (e.g. ~0.3% for raw Illumina reads [31], and ~10-15% for raw Nanopore
172 reads [32, 33]). For a haploid representation of a diploid assembly, the minimum
173 possible base-level error rate will be higher, because by necessity a haploid
174 representation of a heterozygous site will not match approximately half of the reads.
175 In this case, the theoretical minimum base-level error rate is the sum of the error rate
176 of the sequencing technology and half of the heterozygosity. The third measure is the

177 computing genome assembly likelihoods (CGAL) score [20]. The CGAL score is the
178 likelihood of an assembly calculated from a model that accounts for errors in reads,
179 read coverage across the assembly, and the proportion of reads that do not contribute
180 to the assembly. A higher likelihood suggests that a genome assembly is a better
181 representation of the truth. The fourth measure we use is the number of structural
182 variants detected when re-mapping our long validation reads to assemblies. As with
183 the base-level error rate, if the ploidy of the assembly matches the ploidy of the
184 individual, then the theoretical minimum of this metric is the structural error rate
185 introduced into sequencing reads by the sequencing technology. For a haploid
186 representation of a diploid genome, the theoretical minimum is the sum of the error
187 rate of the technology plus half of the structural heterozygosity. These two quantities
188 are rarely known, but nevertheless, a very high structural error rate of validation reads
189 mapped to a haploid assembly may indicate cases in which the assembly has a large
190 proportion of incorrectly linked contigs. The final measure is the genome sequence
191 similarity of each assembly when compared to all other assemblies. This measure
192 does not provide any information relative to an underlying truth, but it may help to
193 identify significant differences between otherwise plausible genome assemblies that
194 can aid in choosing the best assembly. The selection of the best assembly should
195 consider all measures together.

196

197 Here, we used long- and short-reads to create a draft haploid assembly of the *E.*
198 *pauciflora* genome. We use the metrics we describe above to compare a range of

199 assemblies from a range of different assemblers. We performed different assemblies
200 with long-read-only assemblers (Canu (Canu, RRID:SCR_015880) [34], Flye [35]
201 and Marvel [36]) and hybrid assembler MaSuRCA (MaSuRCA, RRID:SCR_010691)
202 [37], using long-read datasets with different minimum read lengths in each case (1 kb
203 and 35 kb).

204

205

206 **Sample collection, DNA sequencing and quality control**

207 We collected leaves from the single *E. pauciflora* tree near Thredbo, Kosciuszko
208 National Park, New South Wales, Australia (36° 29' 39.58" N, 148° 16' 58.73" E) in
209 March 2016 (for Illumina sequencing) and June 2017 (for MinION sequencing). We
210 stored leaves at 4°C when transported them to the laboratory.

211

212 For long-read sequencing, we extracted high molecular weight genomic DNA from
213 leaves following a protocol optimized for *Eucalyptus* nanopore sequencing [38]. We
214 prepared ONT 1D ligation libraries according to the manufacturer's protocol
215 (SQK-LSK108) and sequenced the reads using MinKNOW v1.7.3 with R9.5
216 flowcells on a MinION sequencer. We performed basecalling with Albacore v.2.0.2
217 (Albacore, RRID:SCR_015897). This resulted in 12,584,100 raw long-reads (106.96
218 Gb) with average read length of 8.5 kb. We removed adapters from long-reads with
219 Porechop v0.2.1 (Porechop, RRID: SCR_016967) [39]. Next, we trimmed bases with
220 quality <10 on both ends of the reads using NanoFilt (NanoFilt, RRID:SCR_016966)

[40] and discarded reads shorter than 1 kb after trimming. This recovered 96.66 Gb of long-read data comprising 7,711,141 filtered reads with an average read length of 12.53 kb (minimum 1 kb and maximum ~150 kb). Given an estimated genome size of 500 Mb (see below), this represents a coverage of 193x.

For short-read sequencing, we extracted genomic DNA from freeze-dried leaves using a CTAB protocol [41] followed by purification with a Zymo kit (Zymo Research Corp). We constructed TruSeq Nano libraries with an insert size of 400 bp using protocol provided by Illumina, then sequenced the reads (paired-end 150 bp) using an Illumina HiSeq2500 platform (Illumina Inc., San Diego, CA). This Illumina sequencing generated 506,840,789 paired raw reads (152.05 Gb). We used BBduk v37.31 (BBmap, RRID:SCR_016965) [42] to remove adapters and to trim both sides of raw short-reads which quality was lower than 30. We discarded filtered reads with a length under 50 bp. Around 122.69 Gb short-read data containing 414,697,585 paired reads were left, representing 246x coverage with an estimated genome size of 500 Mb (see below).

Genome size and heterozygosity estimation

We used GenomeScope (GenomeScope, RRID:SCR_017014) [43] and SGA-preqc (SGA, RRID:SCR_001982) [44] to estimate the *E. pauciflora* genome size. We first generated a 32-mer distribution using Jellyfish v1.1.12 (Jellyfish, RRID:SCR_005491) [45] from all of our short-reads, then ran GenomeScope using this 32-mer distribution

with a maximum k-mer coverage of 1000x. This gave a genome size estimate of 408.16 Mb (Additional file 1: Fig. S1), which is lower than expected for other *Eucalyptus* species [16, 17]. However, it is known that genomic repeats can lead to underestimation of genome sizes from uncorrected kmer distributions [46], and the *Eucalyptus* genome is repeat-rich, for example around 50% of genome was annotated as repeats in *E. grandis* [16], suggesting that 408.16 Mb may be a significant underestimate of the genome size. SGA-preqc estimates genome size from k-mer distributions that are corrected to attempt to better account for repeat content, in line with this, SGA-preqc gave a genome size estimate of 529.40 Mb. Because of this, we expect that the SGA-preqc genome size is likely to be more accurate, and in what follows we assume that the *E. pauciflora* genome size is roughly 500 Mb. This suggests that the *E. pauciflora* genome may be around ~30% smaller than that of the other two sequenced *Eucalyptus* species, *E. grandis* (691.43 Mb) [16] and *E. camaldulensis* (654.92 Mb) [17]. However, the genome sizes of *E. grandis* and *E. camaldulensis* may be overestimated due to the assembly and scaffolding of both haplotypes at heterozygous regions.

259

260 **Creation of assembly and validation datasets**

We separated our long-read and short-read data into assembly dataset (~90% of reads) and validation dataset (~10% of reads) by randomly assigning the trimmed and filtered reads into the two datasets. The assembly dataset comprised 86.94 Gb of long-read data (174x coverage) and 114.10 Gb of short-read data (228x coverage).

265 The validation dataset comprised and 9.67 Gb of long-read data (19x coverage) and
266 8.59 Gb of short-read data (17x coverage).

267

268 **Genome assembly**

269 Here, we compared five long-read-only assemblies and two hybrid assemblies. For
270 each combination of data and genome assembler, we followed the same genome
271 assembly pipeline. We first used the assembler to produce an initial assembly.
272 Following this, we identified and removed contigs from contaminant sequences, and
273 then polished the resulting assembly. We then identified and removed haplotigs from
274 the assembly, and finally re-polished each assembly after haplotig removal. To select
275 the best assembly, we calculated the contig N50 with Quast [19], BUSCO scores with
276 BUSCO, and LAI scores using the LTR_retriever pipeline [47]. After mapping the
277 long- and short- validation reads to the final assemblies (using Ngmlr [48] for the
278 former and Bowtie2 (Bowtie2, RRID:SCR_016368) [49] for the latter), we calculated
279 the base-level error rate using Qualimap [50] the structural variant error rate using
280 Sniffles [48], and CGAL scores using CGAL. Finally, we performed whole genome
281 alignment between different assemblies with NUCmer module of MUMmer [51].

282

283 Oxford Nanopore reads tend to have error rates of ~10-15%, which can make
284 assembly of uncorrected reads very challenging. To alleviate this, we first corrected
285 the long-reads assembly dataset with Canu v1.6 with default parameters except for
286 setting corMinCoverage to 8, meaning that read correction would only be applied

287 where at least 8 reads overlapped. We deemed this reasonable given the very high
288 coverage of our data (174x). We then put the corrected long-read datasets into two
289 sets for assembly. The first dataset contained all corrected long-reads, such that the
290 minimum read length was 1 kb (174x of coverage). The second dataset contained all
291 corrected reads longer than 35 kb (~40x of coverage). We refer to these datasets as the
292 1 kb and the 35 kb datasets, respectively.

293

294 We attempted six long-read-only assemblies and two hybrid assemblies. Assemblies
295 solely with long-read data were performed on corrected reads of two read lengths (1
296 kb and 35 kb) using three long-read assemblers: Canu v1.6 and v1.7, Flye v2.3.5 and
297 Marvel v1.0. The Marvel assembly with 1kb dataset was not feasible because it
298 required more disk space than we had available, resulting in five successful long-read
299 only assemblies. We used MaSuRCA v3.2.6 to perform hybrid assemblies with both
300 read length datasets (1 kb and 35 kb) each combined with the short-read dataset. In
301 what follows, we refer to these assemblies as Canu_1kb, Canu_35kb, Flye_1kb,
302 Flye_35kb, Marvel_35kb, MaSuRCA_1kb and MaSuRCA_35kb. We used default
303 settings in all assemblers, and an estimated genome size of 500 Mb where this setting
304 was required. For Canu assemblies, the 1 kb dataset was assembled using Canu v1.6,
305 whereas the 35 kb dataset was assembled using Canu v1.7. We did not repeat the
306 Canu_1kb assembly after Canu v1.7 was released, because we no longer had
307 sufficient computational resources. The chloroplast genome and mitochondrial
308 genome were removed from each assembly. For each assembly, we recorded the

runtime in CPU hours, the raw assembly length, and the N50 (Table 1).

Contamination detection

Following initial assembly, we used Blobtools [52] to assess contamination in each genome assembly. To do this, we first generated a hit file for each assembly by searching all contigs against the National Center for Biotechnology Information (NCBI) non-redundant nucleotide database using BLASTN v2.7.1+ (BLASTN, RRID:SCR_001598) [53] (E-value $\leq 1e-20$). We then analysed the hit file for each assembly using Blobtools, which provides taxonomic annotations and other diagnostic plots to detect contamination in raw genome assemblies. The top-hit was streptophyta phylum, comprising 99.72% to 100% of the hits in different assemblies (Additional file 2: Fig. S2), indicating that there was no potential contamination from a non-plant origin in each raw assembly.

Genome polishing

We polished each initial genome assembly in order to improve its accuracy. For the Canu, Flye, and Marvel assemblies (i.e. those built from long-reads only), we polished first with Racon [54] using Ngmlr v0.2.6 using the long-read assembly dataset, and then with Pilon v1.22 (Pilon, RRID:SCR_014731) [55] using Bowtie v2.3.4.1 with the short-read assembly dataset. For the MaSuRCA assemblies, we polished only with Pilon because MaSuRCA is a hybrid assembler, and using error-prone long-reads to polish hybrid assemblies tends to induce more errors rather

331 than remove them (Additional file 3: Table S1).

332

333 We ran each polishing algorithm for multiple iterations until the accuracy of the
334 resulting assembly stopped improving or improving slightly. We assessed the
335 improvements using BUSCO scores and the base-level error rate by re-mapping
336 validation long- and short-reads to each assembly (mapped as above). We evaluated
337 the BUSCO scores using BUSCO v3.0.2 with the embryophyta_odb9 lineage (1440
338 genes in total). Polishing with Racon took between 4 and 12 iterations, and with Pilon
339 between 6 and 10 iterations (Additional file 3: Table S1).

340

341 Polishing with both Racon and Pilon significantly improved all of the raw genome
342 assemblies, measured with base-level errors in long- and short- reads, and with
343 BUSCO scores (Additional file 3: Table S1). Polishing with Racon improved
344 long-read base level accuracy by up to 0.83% (in the Marvel_35kb assembly),
345 short-read base level accuracy by up to 1.51% (also in the Marvel_35kb assembly),
346 and the BUSO completeness scores by up to 30.76% (in the Flye_35 assembly).
347 Polishing with Pilon further improved the long-read base level accuracy by up to 0.40%
348 (in the Marvel_35kb assembly), the short-read base level accuracy by up to 1.41% (in
349 the Flye_35kb assembly), and the BUSO completeness scores by up to 24.44% (in the
350 Flye_1kb assembly).

351

352 **Assembly ploidy and haplotig removal**

Comparison of the polished genome assemblies revealed large variation in assembly size (Table 2). We calculated the assembly ploidy of each assembly as above, assuming a genome size of 500 Mb. The assembly ploidy ranges from 1.12 (Flye_35kb assembly) to 1.79 (Canu_1kb assembly) (Table 2), suggesting that the Canu_1kb assembly is close to a diploid assembly (i.e. ~80% of the genome is represented by two contigs) and that the Flye_35kb assembly is close to a haploid assembly (i.e. only ~12% of the genome is represented by two contigs). To attempt to produce haploid representations of the genome from all assemblies, we used Purge Haplotigs [28] and a custom pipeline, which we call gene conservation informed contig alignment (GCICA) (script available on github from [56]) to find and remove haplotigs from all the assemblies (Fig. 2A).

Purge Haplotigs assigns contigs to primary contigs and haplotigs depending on both coverage information generated by long-read mapping and pairwise alignments of all contigs. To run Purge Haplotigs, we first mapped the long-read assembly dataset to each polished assembly using Ngmlr v0.2.6, and then separated the contigs into primary contigs and haplotigs with default settings. 8% to 29% of each genome assembly was annotated as haplotigs, and removing these haplotigs reduced the assembly ploidy from 1.12 – 1.79 to 1.03 – 1.29 (Table 2).

The high assembly ploidy for some assemblies after running Purge Haplotigs suggested that these assemblies retained haplotigs that covered up to 29% of the

375 genome. We therefore further filtered possible haplotigs using a custom approach,
376 GCICA. If a pair of contigs comprise a primary contig and a haplotig, we would
377 expect most of regions of the haplotig to be very similar to that of the primary contig.
378 To find putative pairs of primary contigs and haplotigs, we therefore looked for pairs
379 of contigs with similar gene content, and then examined these pairs in more detail. To
380 do this, we first mapped the nucleotide sequences of all *E. grandis* genes to all contigs
381 in an assembly using BLASTN v2.7.1+. If >70% of mapped markers in a contig could
382 also be mapped to another contig, and at least 80% of sequence of the smaller contig
383 could be aligned to the other contig (detecting with NUCmer module of MUMmer
384 v4.0.0beta2), we considered these two contigs as a putative primary contig and
385 haplotig pair. We then examined the alignments of all such pairs by eye and removed
386 any pairs in which the smaller contig appeared to be completely contained within the
387 larger, i.e. in which the smaller contig was an unambiguous haplotig. This process
388 identified a further ~2% of each assembly as haplotigs (Table 2).

389
390 Following removal of haplotigs, we re-evaluated each assembly using BUSCO scores
391 (Fig. 2). We noted that, depending on the genome assembly, the number of complete
392 BUSCO genes sometimes dropped and sometimes increased slightly after removing
393 haplotigs (Fig. 2B). We hypothesised that BUSCO scores could drop either because
394 haplotig removal mistakenly removed a contig that was not a haplotig, or because
395 haplotig removal correctly removed a haplotig which contained a more conserved
396 representation of a BUSCO gene. BUSCO scores could increase because they are

based on E-value scores of alignments, which may be affected by the total length of the assembly. To attempt to alleviate some of these potential issues, we re-polished all of the genome assemblies with multiple rounds of Pilon using the short-read assembly dataset, as above. BUSCO scores recovered across all assemblies with additional Pilon polishing (Fig. 2B). As expected, the number of duplicated BUSCO genes decreased substantially (~50%-70%) after haplotigs were removed from the assemblies and this did not change substantially after additional polishing (Fig. 2C and Additional file 4: Table S2). Together, these results suggest that our haplotig removal pipelines largely succeeded in removing haplotigs, although some haplotigs likely remain if the true genome size is around 500 Mb (Fig. 2A).

407

408 **Assessment of assembly quality with eight measures**

After haplotig removal and polishing, we considered the primary contigs of each assembly as the final assembly, and evaluated each of the final assembly in using the eight statistics we describe above: contig N50, BUSCO scores, LAI scores, assembly ploidy, base-level error rate, CGAL scores, structural variation and genome sequence similarity (Table 3 and Fig. 4).

414

Comparison of the eight metrics we used suggested that the MaSuRCA_35kb assembly was likely to be the most accurate assembly overall and that the Marvel_35kb assembly was the least accurate. However, we note that the MaSuRCA assembly did not receive the best scores for all metrics, suggesting that the choice of

419 which assembly to use will sometimes be question-specific. Also, in most of cases,
 420 performances of the two MaSuRCA assemblies are very similar.
 421
 422 N50 scores varied from 295 kb (Flye_1kb) to 3.2 Mb (MaSuRCA_35kb), with Flye
 423 achieving notably lower N50 values than the other assemblers (Table 3). BUSCO
 424 scores ranged from 1180 complete genes (81.94%, Marvel_35kb) to 1362 complete
 425 genes (94.58%, MaSuRCA assemblies), although all assemblies except the
 426 Marvel_35kb assembly had scores >92%. The MaSuRCA_35kb assembly also
 427 achieved the highest LAI score (9.31), which was substantially higher than the best
 428 assembly from any other assembler (Canu_1kb, LAI score: 7.04). The lowest LAI
 429 score (3.77) was observed in Marvel_35kb assembly. The assembly ploidy was the
 430 closest to one for the Flye assemblies (e.g. 1.03 for the Flye_35kb assembly vs. 1.19
 431 for the MaSuRCA_35kb assembly). Although these scores have to be interpreted with
 432 caution, because the true genome size remains unknown, they are to some extent
 433 corroborated by the lower number of duplicated BUSCO genes in the assemblies with
 434 the lower assembly ploidy (e.g. 90 duplicated BUSCO genes in the Flye_35kb
 435 assembly, vs. 200 in the MaSuRCA_35 assembly). Nevertheless, given that gene
 436 duplication is common in *Eucalyptus* species, all such measures need to be interpreted
 437 with some caution, since the BUSCO genes themselves could be duplicated in the *E.*
 438 *pauciflora* genome. Taken together, these four metrics suggest that the
 439 MaSuRCA_35kb assembly is the most complete, most contiguous, and among the
 440 most accurate of the assemblies we produced.

441

442 The other three metrics assess the entirety of every assembly, and also suggest that the
443 best assemblies for our data are produced by MaSuRCA (Table 3). The MaSuRCA
444 assemblies (1kb and 35kb) had the lowest error rates (0.006 errors per base for
445 short-read mapping and 0.166 for long-read mapping in both assemblies), and the
446 smallest total number of structural variants estimated from the long validation reads
447 (4017 structural variants for the MaSuRCA_35KB assembly). Flye tended to perform
448 the worst on these metrics, although we note that these results will be affected by the
449 fact that the MaSuRCA assemblies contain more duplicated genome regions (see
450 above), which will tend to reduce the estimated error rates and number of structural
451 variants, because duplicated regions can accurately represent heterozygous variants
452 that will be present in the reads. CGAL ranked MaSuRCA assemblies as the best (1kb:
453 lnL -1774303 and 35kb: lnL -1790386) as the best, and the Marvel_35kb assembly as
454 the worst (lnL -4450742).

455

456 Finally, to further investigate the different assemblies, we compared the genome
457 sequence similarity between different assemblies using NUCmer module of MUMmer
458 v4.0.0beta2 (Fig. 4), with the minimum identity set to 75. Notably, around 10% of the
459 sequence of Canu/Flye/MaSuRCA assemblies failed to align to Marvel_35kb
460 assembly (Fig. 4), which, along with the low genome completeness (BUSCO scores)
461 of the Marvel_35kb assembly (Table 3), suggest that the Marvel_35kb assembly may
462 contain many more small duplicated regions than other assemblies. In turn, these

463 duplicated regions may explain the fact that Marvel_35kb assembly has the lowest
464 genome completeness but not the smallest genome size compared to other assemblies
465 (Table 3). Other assemblies have rough 98% - 99% of similarity to each other.

466

467 Based on the eight metrics we used above (Table 3), we suggest that the
468 MaSuRCA_35kb assembly represents the most accurate representation of the *E.*
469 *pauciflora* genome. We note, though, that the Flye assembler only took 1-3% of
470 runtime of the other assemblers used in this paper (Table 1), and produced genome
471 assemblies that were of similar quality to the MaSuRCA_35kb assembly in many
472 respects. The Marvel_35kb assembly received the worst scores on many metrics, and
473 also appears to be missing roughly ~10% of the genome according to BUSCO scores
474 and genome sequence similarity analyses (Table 3).

475

476 **Comparative genome analysis between *E. pauciflora* and *E. grandis***

477 Using the MaSuRCA_35KB assembly, we estimate that the *E. pauciflora* genome is
478 594,871,467 bp in length, with 416 contigs and a contig N50 of 3,235 kb. The genome
479 has up to 0.006 errors per base. Around 94% of complete BUSCO genes were
480 identified in this *E. pauciflora* genome assembly.

481

482 *E. grandis* is the only published *Eucalyptus* genome that is assembled to chromosome
483 level. We therefore compared *E. grandis* with our *E. pauciflora* genome. The *E.*
484 *grandis* contains 691.43 Mb of sequence, roughly 16% larger than the *E. pauciflora*

genome. We compared these two genome assemblies using the NUCmer module of MUMmer v4.0.0beta2 to perform whole genome alignment as described above. This alignment shows that the *E. pauciflora* genome assembly covers just 61.56% of the *E. grandis* genome sequence, leaving approximately 265 Mb of the *E. grandis* genome sequence not covered by the *E. pauciflora* assembly, and 113 Mb of the *E. pauciflora* assembly not covered by the *E. grandis* assembly. Despite this, the coverage of the *E. pauciflora* assembly when mapped to the 11 chromosome-scale scaffolds of the *E. grandis* genome is fairly constant (Fig. 5A), suggesting either that many of these differences result from small errors in both assemblies, and/or from relatively small-scale differences in the underlying genomes.

To examine whether the differences between *E. pauciflora* and *E. grandis* could be explained by their repeat content, we annotated repetitive elements of *E. pauciflora* and *E. grandis* with RepeatMasker v4.0.7 [57]. Although the repeats of *E. grandis* have been annotated before [16], we reannotated them here to enable us to make a direct comparison of the repeat content using an identical pipeline for both genomes. First, we created the custom consensus repeat library using RepeatModeler v1.0.11 [58] with parameter “-engine ncbi”. The classifier was built upon Repbase v20170127 [59]. Then we merged the repeat libraries from RepeatModeler and LTR retrotransposon candidates from LTR retriever to create a comprehensive repeat library as the input for RepeatMasker. We ran the RepeatMasker with “-engine ncbi” model. We used the ‘calcDivergenceFromAlign.pl’ script in RepeatMasker pipeline to

507 calculate the Kimura divergence values, and plotted the repeat landscape with repeats
508 presented in both *E. pauciflora* and *E. grandis* genomes.

509

510 The repeat content of the two genomes is similar. The *E. pauciflora* genome contains
511 44.77% of repetitive elements, compared to 41.22% in *E. grandis*. Retrotransposons
512 account for 29.53% of *E. pauciflora* genome, and 26.94% in *E. grandis*, and DNA
513 transposons account for 6.04% and 4.80% of the genome in *E. pauciflora* and *E.*
514 *grandis*, respectively. The repeat landscapes of the two genomes are also similar,
515 showing roughly two waves of repeat expansion, which is most likely explained by a
516 shared inheritance of most of the repeats in the two genomes (Fig. 5B).

517

518

519 **Conclusions**

520 Here, we report a high-quality draft haploid genome of *E. pauciflora*. It is the first
521 *Eucalyptus* genome assembled with third-generation sequencing reads (Nanopore
522 sequencing), and is the third nuclear genome of *Eucalyptus* species. Due to the
523 economic and ecological importance of *Eucalyptus*, this high-quality genome will
524 support further analysis on *Eucalyptus* and its related species. Additionally, this study
525 will provide useful information for *de novo* plant genome assembly with Nanopore
526 sequencing reads. Finally, the approaches using in this study to assess and compare
527 different assemblies should help in assessing and choosing among many potential
528 genome assemblies

529

530

531

2 Table 1. The statistics information of raw assemblies.

	Long-read^	Short-read	Assembler	Assembly time (CPU hours)*	Length (bp)	contigs	Largest contig (bp)	N50 (bp)	L50	GC	Percent Ns
Canu_1kb	≥1 kb (~174x)	X	Canu	~300,000	871,577,052	2,867	7,123,373	629,835	259	39.18%	0.00%
Canu_35kb	≥35 kb (~40x)	X	Canu	~50,000	825,916,527	2,550	10,153,603	962,598	158	39.18%	0.00%
Flye_1kb	≥1 kb (~174x)	X	Flye	~700	596,007,484	5,930	2,755,662	255,434	652	39.12%	0.00%
Flye_35kb	≥35 kb (~40x)	X	Flye	~500	561,349,738	4,145	2,407,003	352,050	448	39.17%	0.00%
Marvel_35kb	≥35 kb (~40x)	X	Marvel	~28,000	649,061,435	1,181	6,453,759	795,971	182	39.07%	0.00%
MaSuRCA_1kb	≥1 kb (~174x)	~228x	MaSuRCA	~23,000	778,288,575	1,311	12,224,271	1,885,174	95	39.35%	0.04%
MaSuRCA_35kb	≥35 kb (~40x)	~228x	MaSuRCA	~21,000	773,035,614	1,703	8,684,546	1,304,720	146	39.39%	0.09%

3 ^all long-reads were corrected by Canu before assembly. The Canu correction step took around 200,000 CPU hours, which has not been calculated into the assembly runtime.

4 *with around 1 Tb of RAM.

5

6 Table 2. Genome size and assembly ploidy

	Genome size (bp)	Assembly ploidy	Genome size after Purge Haplotigs (bp)	Assembly ploidy	Genome size after Purge Haplotigs and GCICA (bp)*	Assembly ploidy
Canu_1kb	893,781,515	1.79	645,703,255	1.29	622,473,836	1.24
Canu_35kb	847,395,928	1.69	605,520,689	1.21	586,032,599	1.17
Flye_1kb	593,219,654	1.19	529,107,244	1.06	528,619,533	1.06
Flye_35kb	561,597,192	1.12	517,329,093	1.03	517,061,277	1.03
Marvel_35kb	666,317,308	1.33	547,630,224	1.10	537,813,575	1.08
MaSuRCA_1kb	778,307,850	1.56	608,764,671	1.22	594,680,200	1.19
MaSuRCA_35kb	773,071,231	1.55	608,629,204	1.22	595,020,257	1.19

7 *Result before final genome polishing.

8

Table 3. The comparison of final assemblies.

				BUSCO score (1440 genes in total)						LAI scores	Assembly ploidy	Short-read mapping		Long-read mapping		CGAL scores	Structural variants
	Length (bp)	Contig number	Contig N50 (bp)	Complete genes		Duplicated genes		Fragmented genes				Mapping rate	Error rate	Mapping rate	Error rate		
Canu_1kb	622,218,742	895	1,502,325	1,346	93.47%	183	12.71%	23	1.60%	7.04	1.24	96.02%	0.0061	91.73%	0.1661	-1.959E+06	4,243
Canu_35kb	585,785,283	655	2,258,674	1,345	93.40%	138	9.58%	29	2.01%	5.34	1.17	95.52%	0.0066	92.64%	0.1677	-2.226E+06	5,043
Flye_1kb	528,563,896	2,947	295,613	1,344	93.33%	100	6.94%	31	2.15%	5.7	1.06	94.86%	0.0077	93.04%	0.1694	-2.536E+06	7,137
Flye_35kb	516,992,152	2,548	385,290	1,336	92.78%	90	6.25%	31	2.15%	6.5	1.03	94.24%	0.0080	92.34%	0.1699	-2.726E+06	7,458
Marvel_35kb	537,615,613	730	1,202,845	1,180	81.94%	153	10.63%	32	2.22%	3.77	1.08	87.37%	0.0075	85.18%	0.1689	-4.451E+06	5,162
MaSuRCA_1kb	594,528,099	415	3,234,447	1,362	94.58%	201	13.96%	21	1.46%	9.27	1.19	94.91%	0.0060	91.57%	0.1656	-1.774E+06	4,020
MaSuRCA_35kb	594,871,467	416	3,234,549	1,362	94.58%	200	13.89%	21	1.46%	9.31	1.19	94.92%	0.0060	91.49%	0.1655	-1.790E+06	4,017

544 **Availability of supporting data**

545 The *E. pauciflora* genome project was deposited at NCBI under BioProject number
546 PRJNA450887. The whole genome sequencing data are available in the Sequence
547 Read Archive with accession number SRR7153044-SRR7153116. The scripts we used
548 in this paper, including the genome assembly, genome polishing, repeat annotation
549 and genome assessments are available in the Github
550 (<https://github.com/asdcid/Eucalyptus-pauciflora-genome-assembly>).

551

552 **Additional files**

553 **Additional file 1:** A png format with Fig. S1 (GenomeScope result of *E. pauciflora*.)

554 **Additional file 2:** A png format with Fig. S2 (Genome contamination detection.

555 Almost all sequences were matched the sequences in streptophyta phylum group. No
556 contamination was found.)

557 **Additional file3:** A xlsx format with Table S1 (The comparison of polishing results of
558 raw assemblies.)

559 **Additional file4:** A xlsx format with Table S2 (The comparison of polishing result of
560 each genome after haplotig removal.)

561

562 **Abbreviations**

563 BUSCO: Benchmarking Universal Single-Copy Orthologs; CGAL: computing genome
564 assembly likelihoods; *Eucalyptus grandis*: *E. grandis*; *Eucalyptus pauciflora*: *E.*
565 *pauciflora*; the National Center for Biotechnology Information: NCBI; long-terminal

566 repeat: LTR; long-terminal repeat assembly index: LAI.

567

568

569 **Conflict of Interest**

570 The authors declare that they have no competing financial interests.

571

572 **Ethics Statement**

573 *E. pauciflora* leaves were collected a single *E. pauciflora* individual in Thredbo,
574 Kosciuszko National Park, New South Wales, Australia (Latitude -36.49433 ,
575 Longitude 148.282983). The written permission was from the Scientific Licensing
576 office of the Office of Environment and Heritage for New South Wales:
577 www.licence.nsw.gov.au, in accordance with national guidelines in Australia. Tissues
578 were not deposited as voucher specimens.

579

580 **Funding**

581 This research is supported by the Australian Research Council Future Fellowship,
582 FT140100843 to Rob Lanfear and FT180100024 to Benjamin Schwessinger.

583

584 **Author Contributions**

585 AD, DK, RL and WW conceived this project. AMS and RL performed sample
586 collection for Illumina sequencing. AMS extracted genomic DNA, and constructed
587 library for Illumina sequencing. RL and MS carried out sample collection for

588 Nanopore sequencing. MS and BS performed DNA extraction, library preparation,
 589 and Nanopore sequencing. DK performed long-read polishing and Canu 1kb assembly,
 590 whereas AD performed Canu_35kb, Flye_1kb Flye_35kb and Marvel_35kb
 591 assemblies and contamination detection. AD and WW conducted the whole genome
 592 alignment analysis. WW conducted all the remaining analyses. AD, BS, DK, RL and
 593 WW were involved in data interpretation. AD, RL and WW drafted the original
 594 manuscript. RL and WW finalized the manuscript. All authors read and approved the
 595 final manuscript.

596

597

598 **References**

599

- 600 1. Department of Agriculture and Water Resources. Australian forest profiles Eucalypt.
 601 2016.
- 602 2. Williams JE. Biogeographic Patterns of Three Sub-Alpine Eucalypts in South-East
 603 Australia with Special Reference to *Eucalyptus pauciflora* Sieb. Ex Spreng. Journal of
 604 Biogeography. 1991;18 2:223-30.
- 605 3. Boland DJ, Brooker MIH, Chippendale GM, Hall N, Hyland BPM, R.D. J, et al. Forest
 606 trees of Australia. CSIRO, Canberra. 2002.
- 607 4. Gauli A, Vaillancourt RE, Bailey TG, Steane DA and Potts BM. Evidence for local
 608 climate adaptation in early-life traits of Tasmanian populations of *Eucalyptus*
 609 *pauciflora*. Tree Genetics & Genomes. 2015;11:104-15.

-
- 610 5. Cochrane PM and Slatyer RO. Water relations of *Eucalyptus pauciflora* near the
611 alpine tree line in winter. *Tree Physiol.* 1988;4 1:45-52.
- 612 6. Evans JR and Vogelmann TC. Photosynthesis within isobilateral *Eucalyptus*
613 *pauciflora* leaves. *New Phytol.* 2006;171 4:771-82.
614 doi:10.1111/j.1469-8137.2006.01789.x.
- 615 7. Warren CR. Uptake of inorganic and amino acid nitrogen from soil by *Eucalyptus*
616 *regnans* and *Eucalyptus pauciflora* seedlings. *Tree Physiol.* 2009;29 3:401-9.
617 doi:10.1093/treephys/tpn037.
- 618 8. Buckley TN, Turnbull TL, Pfautsch S and Adams MA. Nocturnal water loss in mature
619 subalpine *Eucalyptus delegatensis* tall open forests and adjacent *E. pauciflora*
620 woodlands. *Ecol Evol.* 2011;1 3:435-50. doi:10.1002/ece3.44.
- 621 9. Martorell S, Diaz-Espejo A, Medrano H, Ball MC and Choat B. Rapid hydraulic
622 recovery in *Eucalyptus pauciflora* after drought: linkages between stem hydraulics and
623 leaf gas exchange. *Plant Cell Environ.* 2014;37 3:617-26. doi:10.1111/pce.12182.
- 624 10. Way DA, Holly C, Bruhn D, Ball MC and Atkin OK. Diurnal and seasonal variation in
625 light and dark respiration in field-grown *Eucalyptus pauciflora*. *Tree Physiol.* 2015;35
626 8:840-9. doi:10.1093/treephys/tpv065.
- 627 11. Prior LD, Paul KI, Davidson NJ, Hovenden MJ, Nichols SC and Bowman DJMS.
628 Evaluating carbon storage in restoration plantings in the Tasmanian Midlands, a highly
629 modified agricultural landscape. *The Rangeland Journal.* 2015;37 5:477-88.
630 doi:<https://doi.org/10.1071/RJ15070>.
- 631 12. Wang W, Schalamun M, Morales-Suarez A, Kainer D, Schwessinger B and Lanfear R.

632 Assembly of chloroplast genomes with long- and short-read data: a comparison of
633 approaches using *Eucalyptus pauciflora* as a test case. BMC Genomics. 2018;19
634 1:977. doi:10.1186/s12864-018-5348-8.

635 13. Gauli A, Vaillancourt RE, Steane DA, Bailey TG and Potts BM. Effect of forest
636 fragmentation and altitude on the mating system of *Eucalyptus pauciflora* (Myrtaceae).
637 Australian Journal of Botany. 2014;61 8:622-32. doi:<https://doi.org/10.1071/BT13259>.

638 14. Gauli A, Steane DA, Vaillancourt RE and Potts BM. Molecular genetic diversity and
639 population structure in *Eucalyptus pauciflora* subsp. *pauciflora* (Myrtaceae) on the
640 island of Tasmania. Australian Journal of Botany. 2014;62 3:175-88.
641 doi:<https://doi.org/10.1071/BT14036>.

642 15. Thornhill AH, Crisp MD, Külheim C, Lam KE, Nelson LA, Yeates DK, et al. A dated
643 molecular perspective of eucalypt taxonomy, evolution and diversification. Australian
644 Systematic Botany. 2019;32 1:29-48. doi:<https://doi.org/10.1071/SB18015>.

645 16. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al.
646 The genome of *Eucalyptus grandis*. Nature. 2014;510 7505:356-62.
647 doi:10.1038/nature13308.

648 17. Hirakawa H, Nakamura Y, Kaneko T, Isobe S, Sakai H, Kato T, et al. Survey of the
649 genetic information carried in the genome of *Eucalyptus camaldulensis*. Plant
650 Biotechnology. 2011;28 5:471-80. doi:10.5511/plantbiotechnology.11.1027b.

651 18. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO:
652 assessing genome assembly and annotation completeness with single-copy orthologs.
653 Bioinformatics. 2015;31 19:3210-2. doi:10.1093/bioinformatics/btv351.

-
- 654 19. Gurevich A, Saveliev V, Vyahhi N and Tesler G. QUAST: quality assessment tool for
655 genome assemblies. Bioinformatics. 2013;29 8:1072-5.
656 doi:10.1093/bioinformatics/btt086.
- 657 20. Rahman A and Pachter L. CGAL: computing genome assembly likelihoods. Genome
658 Biol. 2013;14 1:R8. doi:10.1186/gb-2013-14-1-r8.
- 659 21. Ou S, Chen J and Jiang N. Assessing genome assembly quality using the LTR
660 Assembly Index (LAI). Nucleic Acids Research. 2018:gky730-gky.
661 doi:10.1093/nar/gky730.
- 662 22. Slovin JP, Schmitt K and Folta KM. An inbred line of the diploid strawberry *Fragaria*
663 *vesca* f. *semperflorens* for genomic and molecular genetic studies in the Rosaceae.
664 Plant Methods. 2009;5:15. doi:10.1186/1746-4811-5-15.
- 665 23. Yasui Y, Hirakawa H, Oikawa T, Toyoshima M, Matsuzaki C, Ueno M, et al. Draft
666 genome sequence of an inbred line of *Chenopodium quinoa*, an allotetraploid crop
667 with great environmental adaptability and outstanding nutritional properties. DNA Res.
668 2016;23 6:535-46. doi:10.1093/dnares/dsw037.
- 669 24. Arumugasundaram S, Ghosh M, Veerasamy S and Ramasamy Y. Species
670 Discrimination, Population Structure and Linkage Disequilibrium in *Eucalyptus*
671 *camaldulensis* and *Eucalyptus tereticornis* Using SSR Markers. PLOS ONE. 2011;6
672 12:e28252. doi:10.1371/journal.pone.0028252.
- 673 25. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al.
674 Phased diploid genome assembly with single-molecule real-time sequencing. Nat
675 Methods. 2016;13 12:1050-4. doi:10.1038/nmeth.4035.

-
- 676 26. Garg S, Rautiainen M, Novak AM, Garrison E, Durbin R and Marschall T. A
677 graph-based approach to diploid genome assembly. *Bioinformatics*. 2018;34
678 13:i105-i14. doi:10.1093/bioinformatics/bty279.
- 679 27. Pryszcz LP, Németh T, Gácsér A and Gabaldón T. Genome Comparison of *Candida*
680 *orthopsilosis* Clinical Strains Reveals the Existence of Hybrids between Two Distinct
681 Subspecies. *Genome Biology and Evolution*. 2014;6 5:1069-78.
682 doi:10.1093/gbe/evu082.
- 683 28. Roach MJ, Schmidt SA and Borneman AR. Purge Haplotigs: Synteny Reduction for
684 Third-gen Diploid Genome Assemblies. *bioRxiv*. 2018; doi:10.1101/286252.
- 685 29. Schmidt MH, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, et al. De Novo
686 Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing. *Plant*
687 *Cell*. 2017;29 10:2336-48. doi:10.1105/tpc.17.00521.
- 688 30. Costa MD, Artur MA, Maia J, Jonkheer E, Derks MF, Nijveen H, et al. A footprint of
689 desiccation tolerance in the genome of *Xerophyta viscosa*. *Nat Plants*. 2017;3:17038.
690 doi:10.1038/nplants.2017.38.
- 691 31. Schirmer M, D'Amore R, Ijaz UZ, Hall N and Quince C. Illumina error profiles:
692 resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*.
693 2016;17:125. doi:10.1186/s12859-016-0976-y.
- 694 32. Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, et al. de novo assembly
695 and population genomic survey of natural yeast isolates with the Oxford Nanopore
696 MinION sequencer. *Gigascience*. 2017;6 2:1-13. doi:10.1093/gigascience/giw018.
- 697 33. Giordano F, Aigrain L, Quail MA, Coupland P, Bonfield JK, Davies RM, et al. De novo

698 yeast genome assemblies from MinION, PacBio and MiSeq platforms. Sci Rep.
699 2017;7 1:3935. doi:10.1038/s41598-017-03996-z.

700 34. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH and Phillippy AM. Canu:
701 scalable and accurate long-read assembly via adaptive k-mer weighting and repeat
702 separation. Genome Res. 2017;27 5:722-36. doi:10.1101/gr.215087.116.

703 35. Kolmogorov M, Yuan J, Lin Y and Pevzner PA. Assembly of long, error-prone reads
704 using repeat graphs. Nature Biotechnology. 2019; doi:10.1038/s41587-019-0072-8.

705 36. Nowoshilow S, Schloissnig S, Fei JF, Dahl A, Pang AWC, Pippel M, et al. The axolotl
706 genome and the evolution of key tissue formation regulators. Nature. 2018;554
707 7690:50-5. doi:10.1038/nature25458.

708 37. Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL and Yorke JA. The MaSuRCA
709 genome assembler. Bioinformatics. 2013;29 21:2669-77.
710 doi:10.1093/bioinformatics/btt476.

711 38. Schalamun M and Schwessinger B. High molecular weight gDNA extraction after
712 Mayjonade et al. optimised for eucalyptus for nanopore sequencing. Protocolsio 2017.
713 doi:dx.doi.org/10.17504/protocols.io.ka2csge.

714 39. Wick RR: Porechop. <https://github.com/rwick/Porechop>. Accessed 13 Jul 2017.

715 40. De Coster W, D'Hert S, Schultz DT, Cruts M and Van Broeckhoven C. NanoPack:
716 visualizing and processing long-read sequencing data. Bioinformatics. 2018;34
717 15:2666-9. doi:10.1093/bioinformatics/bty149.

718 41. Suarez AM and Rutherford S. gDNA Extraction of Eucalypts pauciflora for full genome
719 sequencing. Protocolsio. 2018. doi:dx.doi.org/10.17504/protocols.io.j7ecrje.

-
- 720 42. BBMap. <http://sourceforge.net/projects/bbmap/>. Accessed 16 Jun 2017.
- 721 43. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.
- 722 GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*.
- 723 2017;33 14:2202-4. doi:10.1093/bioinformatics/btx153.
- 724 44. Simpson JT and Durbin R. Efficient de novo assembly of large genomes using
- 725 compressed data structures. *Genome Res.* 2012;22 3:549-56.
- 726 doi:10.1101/gr.126953.111.
- 727 45. Marcais G and Kingsford C. A fast, lock-free approach for efficient parallel counting of
- 728 occurrences of k-mers. *Bioinformatics*. 2011;27 6:764-70.
- 729 doi:10.1093/bioinformatics/btr011.
- 730 46. Edwards RJ, Tuipulotu DE, Amos TG, O'Meally D, Richardson MF, Russell TL, et al.
- 731 Draft genome assembly of the invasive cane toad, *Rhinella marina*. *Gigascience*.
- 732 2018; doi:10.1093/gigascience/giy095.
- 733 47. Ou S and Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for
- 734 Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* 2018;176
- 735 2:1410-22. doi:10.1104/pp.17.01310.
- 736 48. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al.
- 737 Accurate detection of complex structural variations using single-molecule sequencing.
- 738 *Nat Methods*. 2018;15 6:461-8. doi:10.1038/s41592-018-0001-7.
- 739 49. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat*
- 740 *Methods*. 2012;9 4:357-9. doi:10.1038/nmeth.1923.
- 741 50. Okonechnikov K, Conesa A and Garcia-Alcalde F. Qualimap 2: advanced

-
- 742 multi-sample quality control for high-throughput sequencing data. *Bioinformatics*.
743 2016;32 2:292-4. doi:10.1093/bioinformatics/btv566.
- 744 51. Marcais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL and Zimin A. MUMmer4:
745 A fast and versatile genome alignment system. *PLoS Comput Biol*. 2018;14
746 1:e1005944. doi:10.1371/journal.pcbi.1005944.
- 747 52. Laetsch D and Blaxter M. BlobTools: Interrogation of genome assemblies [version 1;
748 referees: 2 approved with reservations]. *F1000Research*. 2017;6 1287
749 doi:10.12688/f1000research.12232.1.
- 750 53. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
751 BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
752 doi:10.1186/1471-2105-10-421.
- 753 54. Vaser R, Sovic I, Nagarajan N and Sikic M. Fast and accurate de novo genome
754 assembly from long uncorrected reads. *Genome Res*. 2017;27 5:737-46.
755 doi:10.1101/gr.214270.116.
- 756 55. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an
757 integrated tool for comprehensive microbial variant detection and genome assembly
758 improvement. *PLoS One*. 2014;9 11 doi:10.1371/journal.pone.0112963.
- 759 56. W.Wang: Gene conservation informed contig alignment.
760 <https://github.com/asdcid/Gene-conservation-informed-contig-alignment> (2018).
761 Accessed 30 Oct 2018.
- 762 57. Smit A, Hubley R and Green P. RepeatMasker Open-4.0. <http://www.repeatmasker.org>.
763 2015.

-
- 764 58. Smit A and Hubley R. RepeatModeler Open-1.0. <http://www.repeatmasker.org>. 2015.
- 765 59. Bao W, Kojima KK and Kohany O. Repbase Update, a database of repetitive elements
- 766 in eukaryotic genomes. Mob DNA. 2015;6:11. doi:10.1186/s13100-015-0041-9.

767

768

769 **Figure legends**

770 **Figure 1:** The *E. pauciflora* sequenced in this study. This *E. pauciflora* is located in

771 Thredbo, Kosciuszko National Park, New South Wales, Australia (36° 29' 39.58" N,

772 148° 16' 58.73" E).

773 **Figure 2: A.** The length of primary contigs and haplotigs between different

774 assemblies. **B.** The comparison of complete BUSCO genes (1440 in total) between

775 different primary contigs. **C.** The comparison of duplicated BUSCO genes between

776 different primary contigs.

777 **Figure 3:** Structural variation analysis of different assembly primary contigs. Each

778 variant was supported by at least 10 long-reads. **A.** The total event of each structural

779 variances of each assembly. **B.** The insertion event of each assembly. **C.** The

780 translocation event of each assembly. **D.** The Deletion event of each assembly.

781 **Figure 4:** The sequence coverage of whole genome alignment among different

782 assemblies. The sequence coverage was calculated by the length of aligned reference

783 sequence / the total length of reference genome.

784 **Figure 5: A.** The histogram of location and coverage of *E. pauciflora* genome aligned

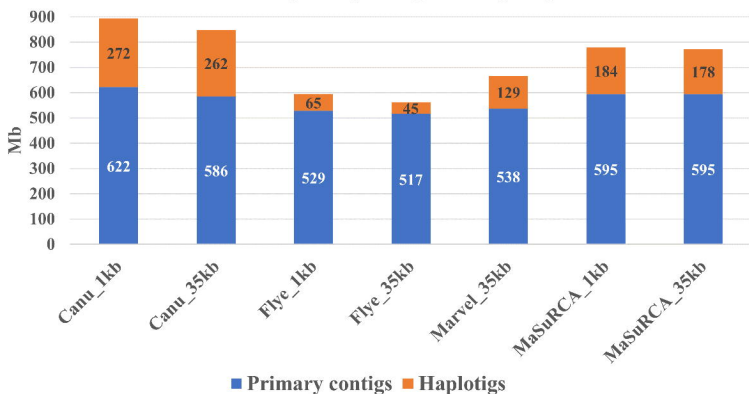
785 to the 11 chromosomes of *E. grandis*. The scale of y-axis is 0x-2x of coverage. Every

786 bar is 1 Mb. The coverage was calculated by the total aligned length of *E. grandis* in
 787 each bar / the length of bar. If a site in *E. grandis* is aligned by *E. pauciflora* twice or
 788 more, this site will be counted twice or more. **B.** Repeat landscape comparison
 789 between *E. pauciflora* and *E. grandis*. Only repeats that are found in both genomes
 790 are shown. Older repeat insertions could accumulate more mutations compared to new
 791 repeat insertions. This leads to older repeat insertions to have accumulated a higher
 792 level of divergence (shown on the right size of the graph).

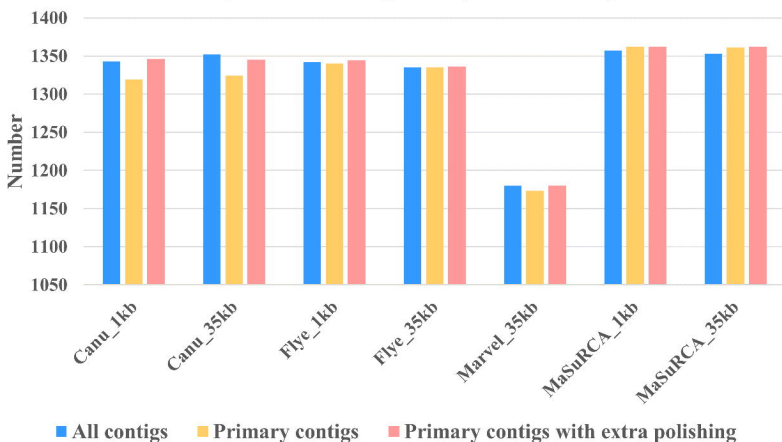


A.

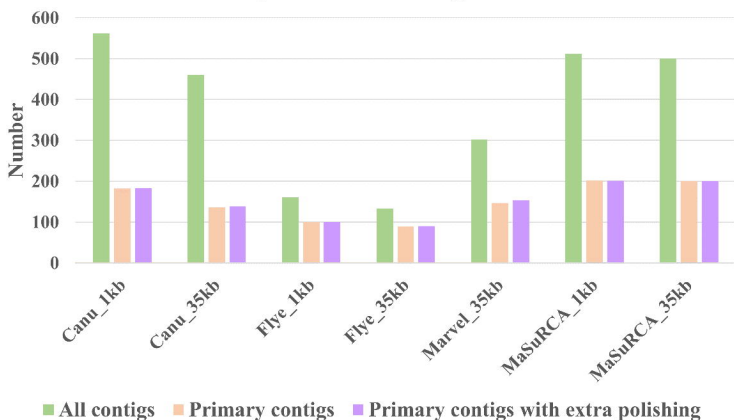
Length distribution between primary contigs and haplotigs

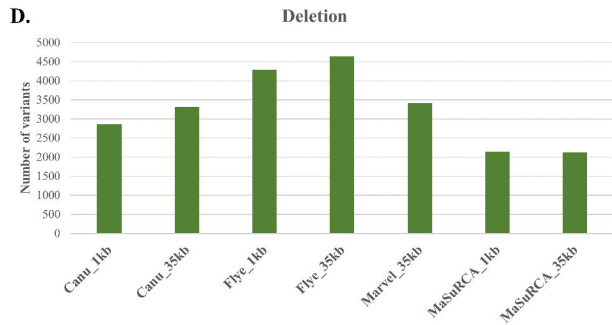
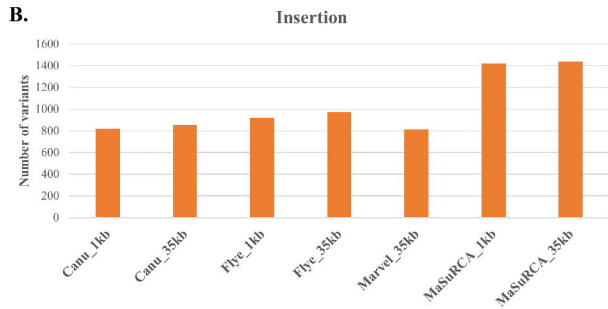
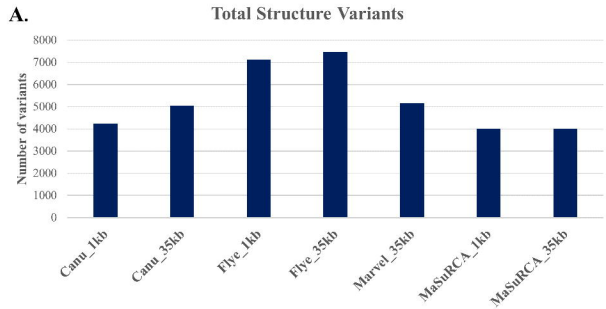
**B.**

Complete BUSCO genes (1440 in total)

**C.**

Duplicated BUSCO genes

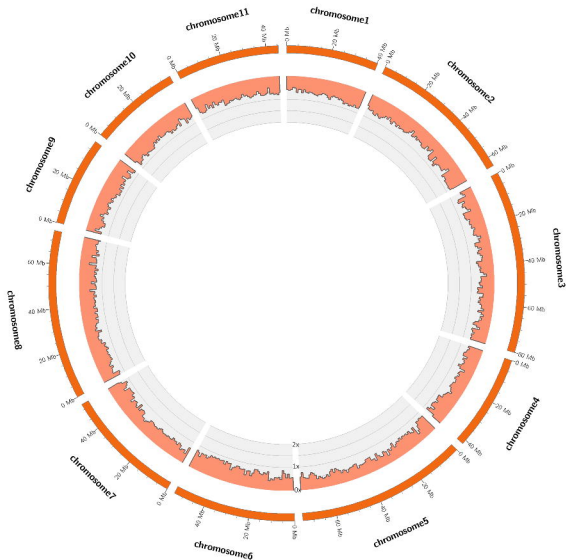




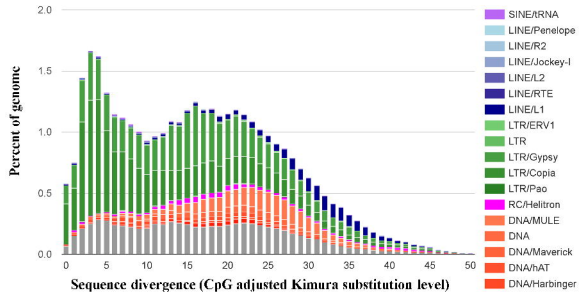
Query Reference	Canu_1kb	Canu_35kb	Flye_1kb	Flye_35kb	Marvel_35kb	MaSuRCA_1kb	MaSuRCA_35kb
Canu_1kb		99.37%	98.82%	98.83%	98.93%	99.12%	99.12%
Canu_35kb	99.24%		98.72%	98.74%	98.85%	99.03%	99.03%
Flye_1kb	99.14%	99.14%		99.24%	99.01%	99.18%	99.18%
flye_35kb	98.98%	99.02%	99.07%		98.84%	99.04%	99.04%
Marvel_35kb	91.77%	91.83%	91.12%	91.14%		91.94%	91.94%
MaSuRCA_1kb	98.72%	98.74%	98.42%	98.47%	98.59%		99.99%
MaSuRCA_35kb	98.73%	98.75%	98.43%	98.48%	98.60%	99.99%	



A.



B.

E. pauciflora*E. grandis*