BMC Genomics

METHODOLOGY ARTICLE Open Access

Assembly of chloroplast genomes with long- and short-read data: a comparison of approaches using *Eucalyptus pauciflora* as a test case

Weiwen Wang[1*], Miriam Schalamun[1,2], Alejandro Morales-Suarez[3], David Kainer[1], Benjamin Schwessinger[1] and Robert Lanfear[1]

Abstract

Background: Chloroplasts are organelles that conduct photosynthesis in plant and algal cells. The information chloroplast genome contained is widely used in agriculture and studies of evolution and ecology. Correctly assembling chloroplast genomes can be challenging because the chloroplast genome contains a pair of long inverted repeats (10–30 kb). Typically, it is simply assumed that the gross structure of the chloroplast genome matches the most commonly observed structure of two single-copy regions separated by a pair of inverted repeats. The advent of long-read sequencing technologies should remove the need to make this assumption by providing sufficient information to completely span the inverted repeat regions. Yet, long-reads tend to have higher error rates than short-reads, and relatively little is known about the best way to combine long- and short-reads to obtain the most accurate chloroplast genome assemblies. Using *Eucalyptus pauciflora*, the snow gum, as a test case, we evaluated the effect of multiple parameters, such as different coverage of long-(Oxford nanopore) and short-(Illumina) reads, different long-read lengths, different assembly pipelines, with a view to determining the most accurate and efficient approach to chloroplast genome assembly.

Results: Hybrid assemblies combining at least 20x coverage of both long-reads and short-reads generated a single contig spanning the entire chloroplast genome with few or no detectable errors. Short-read-only assemblies generated three contigs (the long single copy, short single copy and inverted repeat regions) of the chloroplast genome. These contigs contained few single-base errors but tended to exclude several bases at the beginning or end of each contig. Long-read-only assemblies tended to create multiple contigs with a much higher single-base error rate. The chloroplast genome of *Eucalyptus pauciflora* is 159,942 bp, contains 131 genes of known function.

Conclusions: Our results suggest that very accurate assemblies of chloroplast genomes can be achieved using a combination of at least 20x coverage of long- and short-reads respectively, provided that the long-reads contain at least ~5x coverage of reads longer than the inverted repeat region. We show that further increases in coverage give little or no improvement in accuracy, and that hybrid assemblies are more accurate than long-read-only or short-read-only assemblies.

Keywords: Chloroplast genome, Genome assembly, Polishing, Illumina, Long-reads, Nanopore

* Correspondence: wei.wang@anu.edu.au
[1]Research School of Biology, Australian National University, Canberra, Australia
Full list of author information is available at the end of the article

Background

Chloroplasts are important organelles in algal and plant cells, which generate carbohydrates by photosynthesis [1]. The chloroplast genome provides important information for phylogenetics, population-genetics and species identification [1–8], and is also the focus of genetic engineering because it contains many genes involved in photosynthesis [1].

The chloroplast genome is a double-stranded DNA molecule of around 120 kb – 160 kb in size in most plants, encoding around 100 protein coding genes [9, 10]. The chloroplast genome is usually circular, although some research suggests that it could be linear in certain developmental stages [11, 12]. The structure of chloroplast genome is highly conserved among plants, and usually consists of a long single copy and a short single copy region (the long single copy region is longer and contains more genes than the short single copy region), separated by two identical inverted repeat regions. In some species, one copy of the inverted repeats has been lost during evolution [13]. The length of the inverted repeats usually ranges from 10 to 30 kb [9], although in extreme cases can be as short as 114 bp [14] or as long as 76 kb [15]. There are now more than 1500 chloroplast genomes available in the NCBI organelle genome database.

Accurate genome assembly is a key first step in the study of chloroplast genomes. Initial assemblies of the chloroplast genome relied on Sanger sequencing, which produces highly accurate reads of around 1 kb in length [16–18]. However, Sanger sequencing is expensive and time-consuming. Over the last decade, Sanger sequencing has been largely replaced by short-read sequencing as the primary approach to producing chloroplast genome assemblies [19–21]. Short-read sequencing produces large amounts of data for a relatively low cost, and has a high per-base accuracy. However, it produces DNA fragments that are typically around 50–400 bp in length [22], which can make genome assembly challenging.

A lack of long-range information can limit the accuracy of genome assemblies derived from short-read sequencing data. For example, the two inverted repeats can make it difficult to assemble the chloroplast genome into a single contig, because short-read data rarely contain sufficient long-range information to span an entire inverted repeat (~ 10–30 kb). Most studies which assembled chloroplast genomes from short-read data rely on assemblers designed for whole genome assembly, such as AbySS [23] and SOAPdenovo [24]. These assemblers produce multiple contigs, which were then assembled manually into a single contig according to the structure of existing chloroplast genomes, or by performing additional Sanger sequencing to confirm the conjunctions between the two single copy regions and the inverted repeats [19–21]. Assembling chloroplast genomes by performing synteny alignment with published chloroplast genomes may lead to inaccurate results if the chloroplast genome structure is not conserved (for example, the Chickpea chloroplast genome contains only one inverted repeat region [13]), or if the published chloroplast genome structure contains errors. And when Sanger sequencing is used to confirm an assembly, this removes some of the benefits of using short-reads to assemble chloroplast genomes.

New long-read sequencing technologies have the potential to allow for reference-free assembly of chloroplast genomes by combining some of the best features of Sanger and short-read sequencing. Like short-read technologies, long-read technologies produce large volumes of data for low cost. Technologies such as the MinION sequencing technology from Oxford Nanopore Technologies (ONT) and single-molecule real time sequencing technology from Pacific Biosciences (PacBio) routinely produce single reads longer than 10 kb [22, 25], and even up to 200 kb [25]. It is possible, therefore, for a single read to cover the entire chloroplast genome, or at least a very large section of it, suggesting that it should be feasible to use long-reads to perform reference-free assembly of the chloroplast genome. The main drawback of long-reads is that they have a relatively high per-base error rate of around 10–15% [22, 25], and ONT reads tend to systematically underestimate the length of homopolymer runs [26]. Typically, such errors may be mitigated by polishing the genome after assembly. Genome polishing involves remapping the reads back to the assembly to identify and fix single nucleotide and small structural errors [27, 28]. Chloroplast genome assembly using PacBio data has been reported [29–33], but no study has focused on employing ONT data to assemble chloroplast genomes. It is important to ascertain the best approach to assembling chloroplast genomes with ONT data, because it is increasingly widely used due to its high yield, low cost, and very rapid turnaround times from tissue to data [22, 34].

Hybrid assembly using a combination of long- and short-reads may be the best approach to assembling chloroplast genomes, because it can potentially combine the benefits of the length of long-reads and the accuracy of short-reads. Long-reads could allow the chloroplast structure to be inferred de-novo, rather than inferring it with reference to the structure of a related species. This will be particularly important for chloroplast genomes with atypical structures, such as the Chickpea, which has lost one of the inverted repeat regions [13], or *Selaginella tamariscina* in which the repeats are in the same orientation instead of inverted [35]. Furthermore, long-reads may help avoid errors associated with including reads from nuclear or mitochondrial copies of chloroplast genes into chloroplast assemblies. Parts of the chloroplast genome are frequently transferred into the mitochondrial or nuclear genomes [36–39], such that it can be difficult to tell from whether a short-read which matches a typical chloroplast sequences is of nuclear, mitochondrial, or chloroplast origin. Such contamination may be reduced but perhaps not totally eliminated even when sequencing only purified chloroplast DNA [40]. Introducing long-reads can overcome this problem, because long-reads that are not of chloroplast origin will typically contain nuclear or mitochondrial flanking regions. Despite the potential advantages of long-reads for chloroplast genome assembly, we lack a comparison of the performance of long-read-only, short-read-only, and hybrid assembly approaches for the chloroplast genome. Furthermore, neither the minimum required length of long-reads, nor the optimum combination and coverage of long- and short-reads to obtain accurate chloroplast genome assemblies is known.

In this study, we set out to determine the best approach to chloroplast genome assembly using the chloroplast genome of *Eucalyptus pauciflora* (*E. pauciflora*) as a test case. Eucalypts are widely distributed in Australia, accounting for roughly 75% of the forest areas [41]. *E. pauciflora*, the snow gum, is an evergreen tree found in eastern Australia from close to sea level up to the tree line of the Australian Alps, displaying the broadest altitudinal range in the Eucalypts [42–44]. *E. pauciflora* is of particular interest due to its wide distribution and drought and cold tolerance [45–50]. However, no chloroplast genome for *E. pauciflora* is available. To evaluate the most efficient and accurate way to de novo assemble the chloroplast genome of *E. pauciflora*, we compared long-read-only assemblies, short-read-only assemblies and hybrid assemblies, with large ranges of different coverage and long-read lengths. We used Unicycler [51] to do the short-read-only and hybrid assemblies, because it is designed specifically for the assembly of small circular genomes. For long-read-only assemblies, we compared two assemblers: Hinge [52], an assembler designed for solving the long repeats problem in long-read assemblies of circular genomes; and Canu [53], one of the most popular long-read assemblers currently available. We also assessed the performance of a variety of read-correction methods, assembly software and polishing algorithms, and evaluated how the length of long-reads affects chloroplast genome assembly.

Results

Sequencing

We generated 6.0 Gb of raw long-read data, comprising 705,554 reads with a mean length of 8504 bp. We generated 6.4 Gb of short-read data, comprising 21,114,786,150 bp paired-end reads. After trimming adapters and low-quality bases, and mapping the reads to the reference set (Additional file 1: Table S1), we recovered 0.57 Gb of long-read data comprising 28,777 reads with a mean length of 19,807 bp (minimum 5002 bp and maximum 150,181 bp), and 0.38 Gb of short-read data comprising 2,840,345 paired-end reads. Assuming that the *E. pauciflora* chloroplast genome size is ~ 160 kb, this represents a total coverage of ~ 3,600x for long-reads and ~ 2,400x for short-reads.

Genome assembly comparison

Of the 648 genome assemblies we performed, the best genome assemblies were the hybrid assemblies with at least 20x coverage of both long- and short-reads (Fig. 1). Hybrid assemblies with lower coverage of either data type, and assemblies that relied on only one data type, were of lower quality. In the following, we discussed each of the assembly categories in more detail.

Long-read-only assembly

For long-read-only assemblies, the best assembly was produced by Hinge with 500x long-read coverage and polishing using Racon+Nanopolish (Additional file 2: Figure S1). Neither Hinge nor Canu were able to assemble the entire chloroplast genome into a single contig (Additional file 2: Figure S1a), so all long-read-only assemblies had to be manually curated (see Methods). In general, Canu assembled one or two contigs which failed to cover some regions of the genome or covered some regions twice, whereas Hinge assembled at least two contigs containing many duplicated regions. Canu produced assemblies from all 12 different long-read coverages (5-500x; the second row of Fig. 1a), but only the assemblies generated from 80x and 200x coverage had a total length that was close to the expected 160 kb after manual curation (Fig. 1b and Additional file 2: Figure S1b). Surprisingly, the total length of the Canu assemblies varied widely with relatively minor changes in coverage (E.g. ~ 145 kb at 60x, ~ 160 kb at 80x, and ~ 140 kb at 100x coverage; Fig. 1b and Additional file 2: Figure S1b). Hinge failed to produce any contigs when the coverage was lower than 20x (top row of Fig. 1a), but produced contigs that could be manually curated into near complete chloroplast genomes (with lengths very close to 160 kb) when the coverage was ≥20x (Fig. 1b and Additional file 2: Figure S1b).

Validation of the long-read-only assemblies after manual curation showed that the assemblies produced by Hinge were more accurate than those produced by Canu (Fig. 1c). This is because Hinge was able to assemble contigs covering the entire chloroplast genome. The error rate of mapped validation reads decreased predictably as the long-read coverage increased for both Canu and Hinge assemblies (Fig. 1d and Additional file 2: Figure S1d). For the Hinge assemblies, the mapping rate of the validation reads increased predictably with the

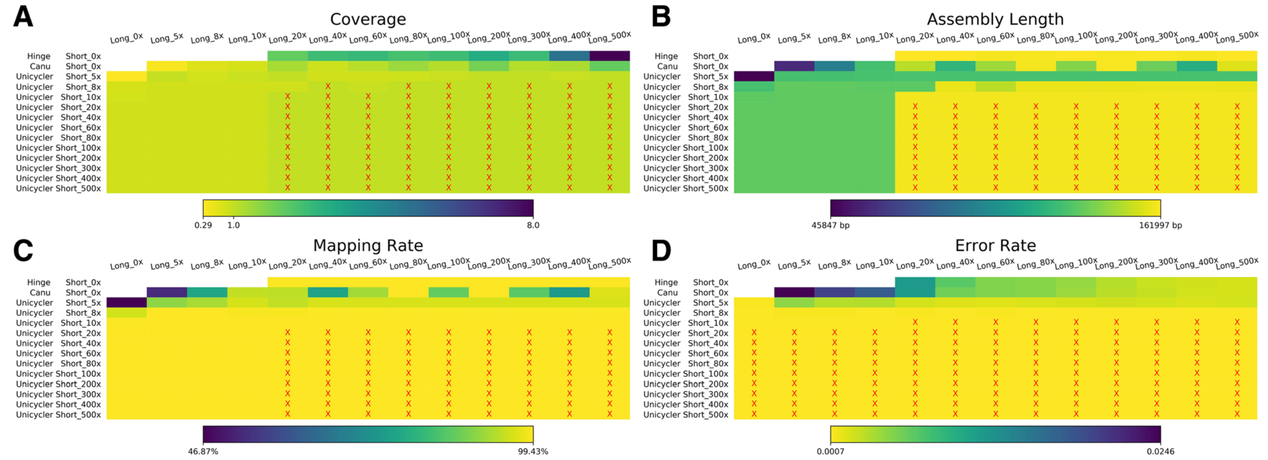

Fig. 1 Comparison of chloroplast genome assemblies. The coverage of the long- and short-reads is shown along the top and left-hand-side of each panel, respectively. The left-hand-side also shows which assembler was used for each row of assemblies in that panel. Hinge and Canu are long-read-only assemblers, whereas Unicycler is short-read-only and hybrid assembler. The Hinge and Canu results in **b**, **c**, **d** were polished by Racon+Nanopolish, and the Unicycler results used Karect-corrected short-reads. **a.** The total coverage of the chloroplast genome across all contigs output by the assembler. Panels marked with a red 'x' contained a single contig covering the whole chloroplast genome. The heatmap indicates the chloroplast genome coverage. **b.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). Panels marked with an 'x' denote assemblies with the expected length, in the range 155,938 bp–155,945 bp. **c.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with a red 'x'. **d.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with a red 'x'

increased coverage of input reads, up to a maximum of 99.41% with 300x or 500x input read coverage (Fig. 1c and Additional file 2: Figure S1c). For the Canu assemblies, the mapping rate of the validation reads was the highest (~ 99.40%) at 80x and 200x input read coverage (Fig. 1c and Additional file 2: Figure S1c), but varied substantially at lower coverage, reflecting the fact that many of the Canu assemblies were missing large portions of the chloroplast genome. These data suggest that there is a complex relationship between assembly accuracy and input read coverage when assembling chloroplast genomes with ONT data using Canu.

Polishing the long-read-only genome assemblies with both Racon and Nanopolish resulted in more accurate assemblies than polishing them with either Racon or Nanopolish alone (Additional file 2: Figure S1c and d). For Canu assemblies, when polishing with Racon alone, assembly accuracy improved from an error rate (mismatches and indels) of ~ 0.0300 per base of the validation reads to ~ 0.0060 per base as coverage increased up to ~100x, but did not improve further at higher coverage. When polishing with Nanopolish or Racon +Nanopolish, assembly accuracy continued to improve up the maximum coverage we examined (500x; Additional file 2: Figure S1c and d), at which the error rate was 0.0022 per base of the validation reads in the best assemblies. The situation was the same in Hinge assemblies. Since the expected error rate of the validation reads likely to be at most 0.0010 (see above), this suggests that at least 0.0012 mismatches or indels per base in the validation reads mapped to the long-read--only assemblies comes from errors in the assemblies themselves.

Short-read correction

For short-read-only and hybrid assemblies, we compared the effect of four different approaches to short-read correction: no correction, Karect, SPAdes, and Karect +SPAdes. When the short-read coverage was ≥20x, we found no meaningful differences between these four conditions (Additional files 3, 4, 5, and 6: Figures S2-S5). When the short-read coverage was lower than 20x, we found that correction with Karect gave slightly higher genome assembly accuracy than the other three conditions (i.e. slightly higher mapping rates and lower per-base error rates with the validation reads; Additional files 3, 4, 5, and 6: Figures S2-S5). Therefore, we focus on comparing short-read-only and hybrid assemblies in which short-reads were corrected using Karect.

Short-read-only assemblies

Short-read-only assemblies with coverage ≥20x produced three complete contigs corresponding to the three major structural regions of the chloroplast genome: the long single copy, the short single copy and the inverted repeat (Additional file 3: Figure S2a). After manual curation (see Methods), in which we assembled the resulting three contigs by hand and polishing, (not duplicate the inverted repeat), the short-read-only assemblies had high mapping rates of the validation reads (99.36%) and low rates of mismatches and indels between the validation reads and the assemblies (0.0007). This error rate is lower than the expected error rate of the validation reads, suggesting that the short-read-only assemblies may contain few or no errors. One small source of error in the short-read-only assemblies, which is not reflected in these statistics, is that in these assemblies ~ 10 bp from the beginning or end of long single copy or short single copy region are sometimes incorrectly assigned to the inverted repeat region, vice versa. This may cause minor errors in assemblies when whole chloroplast genomes are manually assembled from the three contigs (duplicate the inverted repeat) that result from short-read-only assemblies.

Hybrid assemblies

Hybrid assembly with Unicycler performed better than all other assemblies as long as both long- and short-read coverage was at least 20x (Fig. 1 and Additional file 3: Figure S2). In these cases, the entire chloroplast genome was assembled into a single contig without manual curation, and the validation reads showed a mapping rate of 99.43% and an error rate of 0.0007 after polishing. As above, this suggests that these hybrid assemblies likely contained few or no errors. When the long-read coverage was less than 20x, the long-reads made no meaningful difference to the hybrid assemblies when compared to short-read-only assemblies. When coverage of long-reads was ≥20x, all assemblies with the same short-read coverage were exactly the same. These results suggest that the long-read information is useful for confirming the correct conjunctions between the two single copy regions and the inverted repeats in Unicycler. Assemblies using only 5x short-read coverage were missing ~ 40 kb of the chloroplast genome, regardless of the long-read coverage (Additional file 3: Figure S2b). The length of hybrid assemblies with ≥20x coverage of both long- and short-reads were all very similar, differing by at most four bases from a length of 159,942 bp.

Hybrid assembly with different lengths of long-reads

Hybrid assembly performance was highly dependent on the length of the long-reads. Hybrid assembly produced a single contig spanning the entire complete chloroplast genome with as little as 5x long-read coverage and 8x short-read coverage, provided that all the long-reads were longer than 30 kb (Table 1 and Additional files 7, 8, 9, 10, 11 and 12: Figures S6-S11), which is long enough to cover the entire inverted repeat region (~ 26 kb). When long-reads were shorter than 20 kb, the hybrid assembly failed to assemble the entire chloroplast genome into a single contig (Table 1, Additional files 7 and 8: Figures S6 and S7), likely because these reads are shorter than the inverted repeat region. When long-reads were 20–30 kb, hybrid assembly produced a complete assembly with a single contig with at least 20x long-read coverage and at least 8x short-read coverage.

Table 1 The minimum coverage of long-/short-reads required to assemble one contig spanning the entire chloroplast genome

Long-read length range	5-10 kb	10-20 kb	20-30 kb	30-40 kb	40-50 kb	
Long-read	N/A	N/A	20x	5x	5x	5x
Short-read	N/A	N/A	8x	8x	8x	8x

N/A: No one contig spanning the entire chloroplast genome can be assembled under this long-read length range.

Final genome assessment

To choose one assembly as the *E. pauciflora* chloroplast reference genome, we compared all hybrid assemblies with ≥20x long- and short-read coverage. The 81 assemblies fall into six groups of identical assemblies (Fig. 2), all of which have identical mapping and error rates (Fig. 1c and d). All assemblies with the same short-read coverage were identical, regardless of their long-read coverage. Assemblies with 40x, 200x or 500x short-read coverage were identical, as were assemblies with 300x and 400x short-read coverage, while assemblies with 20x, 60x, 80x and 100x short-read coverage were all unique. The differences between these six groups of assemblies fall into just three regions of the chloroplast genome (Fig. 2): (i) an adenine homopolymer site, 9048–9061, containing six variants; (ii) an adenine homopolymer site, 32,043–32,059, in which the 80x short-read coverage assemblies have a small deletion; (iii) a thymine homopolymer site, 51,964–51,975, in which the 20x short-read coverage assemblies have a short deletion.

To distinguish which variant in each of these regions was most likely to be correct, we mapped all of the long- and short-reads to all six assemblies. This revealed that the identical assemblies from the 40x, 200x, and 500x short-read coverage datasets are most likely to be correct. In the first adenine homopolymer site, both short- and long-reads rejected the thymine substitution (which is present in the 20x, 60x, and 80x short-read coverage assemblies, Fig. 2), but neither the short- nor the long-reads provided clear preference for the length of the homopolymer. The long-reads provided little useful information due to their high rate of systematic error around homopolymers. Roughly the same number of reads map successfully to this region regardless of whether the assembly has 14, 15, or 16 adenines in this homopolymer, and in all cases no roughly 10% of short-reads disagree with the length of the homopolymer in the assembly. This may be the result of sequencing error, or mapping error, or heteroplasmies affecting this

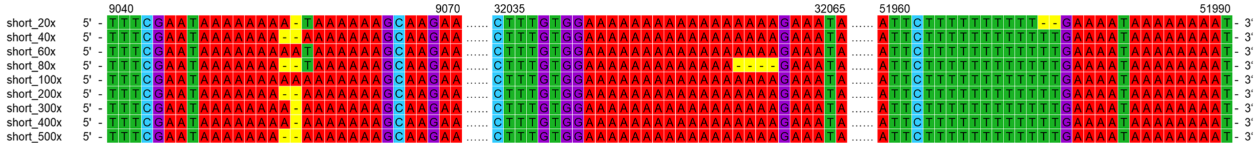
9040
9070
32035
32065
51960
51990
short_20x 5' - TTTCGAATAAAAAAAA-TAAAAAAGCAAGAA ...... CTTTGTGGAAAAAAAAAAAAAAAAAGAAATA ...... ATTCTTTTTTTTTT--GAAAATAAAAAAAAT - 3'
short_40x 5' - TTTCGAATAAAAAAA--AAAAAAAGCAAGAA ...... CTTTGTGGAAAAAAAAAAAAAAAAAGAAATA ...... ATTCTTTTTTTTTTTTGAAAATAAAAAAAAT - 3'
short_60x 5' - TTTCGAATAAAAAAAAATAAAAAAGCAAGAA ...... CTTTGTGGAAAAAAAAAAAAAAAAAGAAATA ...... ATTCTTTTTTTTTTTTGAAAATAAAAAAAAT - 3'
short_80x 5' - TTTCGAATAAAAAAA--TAAAAAAGCAAGAA ...... CTTTGTGGAAAAAAAAAAAAA----GAAATA ...... ATTCTTTTTTTTTTTTGAAAATAAAAAAAAT - 3'
short_100x 5' - TTTCGAATAAAAAAAAAAAAAAAAGCAAGAA ...... CTTTGTGGAAAAAAAAAAAAAAAAAGAAATA ...... ATTCTTTTTTTTTTTTGAAAATAAAAAAAAT - 3'
short_200x 5' - TTTCGAATAAAAAAA--AAAAAAAGCAAGAA ...... CTTTGTGGAAAAAAAAAAAAAAAAAGAAATA ...... ATTCTTTTTTTTTTTTGAAAATAAAAAAAAT - 3'
short_300x 5' - TTTCGAATAAAAAAAA-AAAAAAAGCAAGAA ...... CTTTGTGGAAAAAAAAAAAAAAAAAGAAATA ...... ATTCTTTTTTTTTTTTGAAAATAAAAAAAAT - 3'
short_400x 5' - TTTCGAATAAAAAAAA-AAAAAAAGCAAGAA ...... CTTTGTGGAAAAAAAAAAAAAAAAAGAAATA ...... ATTCTTTTTTTTTTTTGAAAATAAAAAAAAT - 3'
short_500x 5' - TTTCGAATAAAAAAA--AAAAAAAGCAAGAA ...... CTTTGTGGAAAAAAAAAAAAAAAAAGAAATA ...... ATTCTTTTTTTTTTTTGAAAATAAAAAAAAT - 3'

Fig. 2 Comparison of hybrid assembly sequences with ≥20x long- and short-read coverage. Short_20x indicates 20x coverage of short-reads were used in these assemblies (with ≥20x long-read coverage). For hybrid assemblies with ≥20x short-reads, if the long-read coverage was ≥20x, all assemblies with the same short-read coverage was identical. The number at the top is the position

site. For the other two homopolymer sites, the short-reads clearly showed that the deletions in the 20x and 80x coverage assemblies in these regions where likely to be assembly errors (> 50% of short-reads show insertions in both regions). Based on these observations, we selected the genome assembly derived from 500x coverage of both short- and long-read data as the final genome assembly. Despite the uncertainty in the homopolymer at position 9048–9061, we preferred this assembly because it is derived from the most data.

The mapping results suggest that the per-base accuracy of the final genome assembly is very high, and is unlikely to contain any errors. We identified all sites that show a non-reference base or small indel at a frequency of ≥10%, revealing 1016 sites for the mapped long-reads and one for the mapped short-reads. Of these, there are no sites that contained a non-reference allele at ≥50% frequency in either the long- or the short-read mapping. This suggests that the final genome assembly is unlikely to contain any errors. The high number of sites with a non-reference allele at ≥10% for the mapped long-reads is expected given their much higher error rate, and biases associated with the estimation of the length of homopolymer runs. The one site identified with a non-reference base ≥10% in the short-reads was site 51,961 in which 47.71% of bases support a T insertion. This insertion is located in a homopolymer T region (Additional file 13: Figure S12), and so the ONT long-reads can provide little information on whether the insertion is likely to be an assembly error, a true variant (e.g. due to heteroplasmy), sequencing error in the short-reads, or a mapping error (e.g. due to a proportion of short-reads being derived from DNA transferred from the chloroplast to the mitochondrial or nuclear genomes [40]).

E. pauciflora chloroplast genome annotation, and phylogenetic analysis

Our best assembly of the *E. pauciflora* chloroplast genome is a hybrid assembly, which has the best overall quality statistics, and does not suffer from the minor errors at the junctions of the single copy and inverted repeat regions associated with short-read only assemblies (see above). This *E. pauciflora* chloroplast genome is 159,942 bp in size, comprised of two inverted repeat regions of 26,367 bp, a long single copy region of 88,787 bp, and a short single copy region of 18,421 bp. We identified and annotated 131 genes of known function including 37 transfer RNA genes and 8 ribosomal RNA genes (Fig. 3a). All ribosomal RNA genes are located in the inverted repeat regions. Five genes, *psbL*, *infA*, *ycf1* and two copies of *ycf15*, were annotated as pseudogenes.

The phylogenetic analysis places *E. pauciflora* as a sister to a clade comprised of *E. regnans, E. elata,* and *E. sieberi* with high bootstrap support (Fig. 3b). As expected, interrelationships among the other 31 *Eucalyptus* chloroplast genomes are the same as reported previously [5]. The placement of *E. pauciflora* is consistent with previous *Eucalyptus* analyses using Diversity Array Technology markers [54].

Discussion

By comparing a large range of different approaches to chloroplast genome assembly, we show that hybrid assembly with at least 20x coverage of long-reads (containing at least 5x coverage of reads large than the inverted repeat region) and 20x coverage of short-reads is sufficient to assemble the entire chloroplast genome into a single contig with few or no errors. We also show that very similar accuracy can be obtained from short-read-only assemblies, although in this case the genome is assembled into three contigs representing the three main regions of the chloroplast genome. Long-read-only assemblies using ONT data produced multiple contigs and contained many errors, even after polishing with long-read polishing algorithms. We demonstrate how multiple genome assemblies from the same data can be rigorously compared by retaining a subset of reads as a validation set. Finally, by re-mapping both our long- and short-reads to our final assembly, we show that it is likely that our final assembly has few or no errors, and identify one sites in which the lengths of homopolymers were difficult to determine, even with very high coverage of long- and short-reads.

Short-read-only assemblies of the chloroplast genome were highly accurate, but were divided into the three regions of the chloroplast genome – the long single copy, short single copy, and inverted repeat. Such assemblies have two limitations. First, they must

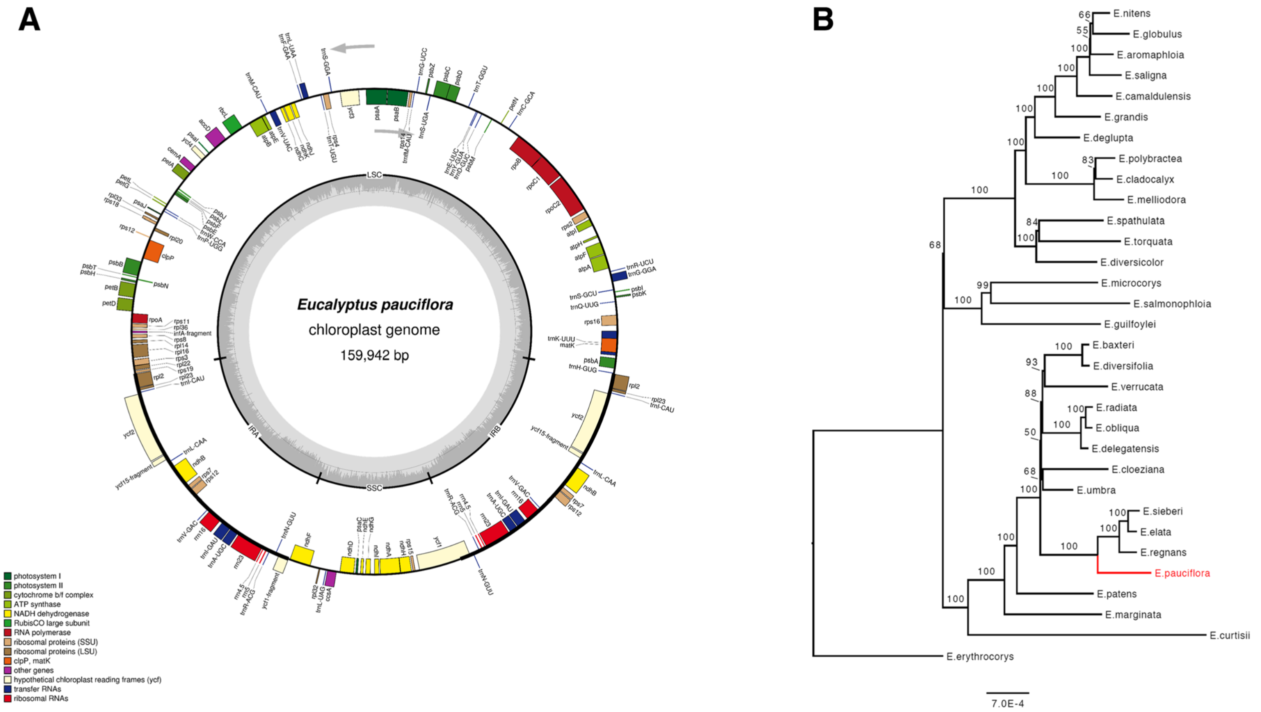

Fig. 3 a. Annotated *E. pauciflora* chloroplast genome. Genes shown on the inside of the circle are transcribed clockwise, whereas genes shown on the outside of the circle transcribed counterclockwise. The grey region in the inside circle shows the GC content across the chloroplast genome. This figure was produced by OGDraw v1.2 [93]. **b.** A phylogenetic tree of 32 *Eucalyptus* taxa based on analysis of full chloroplast genomes

be manually combined into a full-length chloroplast genome, and this relies on comparison to a reference genome. Second, if using Unicycler as the assembler, although it can assemble the three regions of chloroplast genome (can be merged into one contig if we duplicate the inverted region and link them together according to the graphical fragment assembly graph produced by Unicycler), it occasionally placed ~ 10 bp from the end of one region onto the start of another, such that the accuracy of manually-curated short-read-only assemblies may be rather low at the junctions between single copy and inverted repeat regions. The lack of structural resolution will not matter if genome structure is not the focus of a study, or if the genome structure is known in advance. Similarly, the very slightly higher error rate of short-read-only assemblies compared to hybrid assemblies may not matter for some approaches (e.g. phylogenetics, in which such errors could be safely ignored), but may be important for other applications (e.g. population or functional genetics in which variants in these regions may be the focus of a study).

Our study suggests that there are two key advantages to hybrid assemblies compared to short-read-only assemblies. First, the addition of as little as 5x coverage of long-reads of sufficient length that span the inverted repeat (roughly 10–30 kb in most species) results in assembling the entire chloroplast genome into a single contig. This will be beneficial for those who wish to obtain a reference-free estimate of the overall structure of the chloroplast genome (e.g. for researchers interested in structural variation [55–58]), and for those who wish to accurately infer the sequence of the junctions between the major regions of the genome without additional Sanger sequencing. Second, long-reads provide useful information for assessing assembly errors or heteroplasmies in chloroplast genome assembly. The limited length of short-reads (150 bp in this study) will increase the rate at which these reads map incorrectly, particularly given the existing copies of sections of the chloroplast genome in the mitochondrial or nuclear genome [36–39]. For example, 28 chloroplast DNA fragments were found inserted into the rice chromosome 10, covering almost the entire rice chloroplast genome [59]. If the nuclear insertions have accumulated mutations, this could cause some problems with short-read-only chloroplast genome assemblies, and particularly analyses of heteroplasmy. These problems should be much reduced for long-reads, because the extent the and frequency of chloroplast DNA transfer can be inferred from the long reads themselves. Furthermore, analysis of reads that contain chloroplast DNA with nuclear flanking regions should enable nuclear copies of the chloroplast genome to be excluded from analyses, which may help clarify some ongoing debates about the prevalence of heteroplasmy in chloroplast genomes [40, 60, 61].

Our results differ from those reported in a recent study of long-read-only and short-read-only chloroplast genome assemblies, which suggested that there were substantial benefits of long-read-only assembly for chloroplast genomes assembly [29]. That study used PacBio data (we used ONT data here), and compared the resulting assemblies to those generated from Illumina data (as we used here). That study suggested that long-reads were beneficial based on the observation that short-read-only assemblies recovered just ~ 90% of the chloroplast genome assembled into seven contigs, with a relatively high proportion (0.12%) of uncertain sites. We were able to successfully assemble 100% of the chloroplast genome into three contigs (corresponding directly to the three major structural regions of the chloroplast genome) with very high accuracy, given just 20x coverage of short-read data. The difference between these two studies is likely due to recent improvements in genome-assembly algorithms, and in particular the development of Unicycler, which is designed specifically for the assembly of circular genomes from short-read and combined short- and long-read data. These improved genome assembly algorithms mitigate many of the previous limitations of short-read-only assemblies of the chloroplast genome, making the differences between approaches less severe than they were previously.

Assembling highly accurate chloroplast genomes from purely long-reads generated on the ONT platform remains challenging. The long-read-only assemblers we assessed were unable to assemble the genome into a single contig, although they were able to assemble multiple contigs which could subsequently be manually curated to produce full-length chloroplast genomes (Fig. 1 and Additional file 2: Figure S1). However, the single-base accuracy of these genome assemblies (after polishing) remained much higher than those of short-read-only or hybrid assemblies (after polishing). Long-read-only assemblies are likely to be more accurate at the single base level when using PacBio sequencing, because the error profile of PacBio sequencing is less systematically biased than that of ONT sequencing. However, in agreement with our findings, a recent study showed using PacBio sequencing data it was not yet possible to assemble the entire chloroplast genome into a single contig without post-assembly processing [29–31, 62].

Conclusions

Our results show that it is possible to produce complete and highly-accurate chloroplast genome assemblies by combining at least 20x coverage of long- and short-read data, respectively. Given the low cost and simplicity of generating long-read data from the ONT sequencer (MinION) [22, 34], the extremely low cost of producing high-coverage short-read data using Illumina technology, and the potential to multiplex multiple samples on both devices, this provides a clear path towards producing multiple highly-accurate and complete chloroplast genome assemblies for very low cost. The ability to recover the genome in a single contig will also be of great benefit to those interested in assembling chloroplast genomes with atypical structures.

Perhaps the biggest remaining challenge with the hybrid assembly approach is the production of long-reads that span the entire inverted repeat region. We showed that it is necessary to include at least 5x coverage of such reads to gain the benefits of the hybrid assembly approach. In most species, the inverted repeat region is ~ 10–30 kb, and successfully extracting and sequencing DNA containing sufficient intact molecules longer than 30 kb remains difficult. Nevertheless, many groups are now working on establishing high-molecular-weight DNA extraction protocols (e.g. [63, 64]), suggesting that the remaining issues are unlikely to be insurmountable for most species.

Broadly speaking, our analyses may also provide hints on obtaining high-quality assembly of other circular genomes, such as those of other organelles and bacteria.

Methods

The raw data have been submitted to NCBI under accession numbers SRR7153095 (long-read data), and SRR7153063 and SRR7153071 (short-read data). The final genome assembly of the *E. pauciflora* chloroplast genome, and the scripts necessary to reproduce the analyses in this study are available on github at https://github.com/asdcid/Chroloplast-genome-assembly. Subfolders in that repository are referenced throughout the methods where appropriate.

Sample collection and DNA sequencing

Sample collection

Leaves were collected in March 2016 (for Illumina sequencing) and June 2017 (for MinION sequencing) from the same branch of a single *E. pauciflora* individual in Thredbo, Kosciuszko National Park, New South Wales, Australia (Latitude − 36.49433, Longitude 148.282983). Leaves were stored at 4 °C until they were returned to the laboratory.

DNA extraction and quality control for Illumina sequencing

Leaves were freeze-dried, and total DNA was extracted using a CTAB protocol [65], then purified with a Zymo kit (Zymo Research Corp). TruSeq Nano libraries were constructed according to the manufacturer's instructions and whole genome shotgun sequencing was carried out at the ACRF Biomolecular Resource Facility (the Australian National University, Canberra, Australia) on the

Illumina HiSeq-2500 platform using 150 bp paired-end sequencing with a roughly 400 bp insert size. We used BBDuk v37.31 [66] for quality and adapter trimming of Illumina data, removing bases with a quality score < 30 on the left or right side of a read. After trimming, reads shorter than 50 bp were removed, and only paired reads were kept. We used FastQC [67] to perform quality checks of libraries before and after trimming. Scripts for this analysis are available in the 1_pre_assembly/1_quality_control/short_read folder of the github repository.

DNA extraction and quality control for MinION sequencing

High molecular weight DNA was extracted from leaves using a protocol [68] based on Mayjonade's method [69]. Libraries were prepared according to the ONT 1D ligation library protocol (SQK-LSK108). Long-read sequencing was carried out on a MinION sequencer using MinKNOW v1.7.3, and then basecalled using Albacore v2.0.2. We removed adapters from long-reads using Porechop v0.2.1 [70], and trimmed bases with quality < 9 on both sides of reads using Nanofilt v1.2.0 [71]. We discarded reads shorter than 5 kb. Scripts for this analysis are available in the 1_pre_assembly/1_quality_control/long_read folder of the github repository.

Chloroplast read extraction

In both the long- and short-read data, the majority of reads come from non-chloroplast sources, such as the nuclear genome, the mitochondrial genome and other contaminants, because we performed whole genome sequencing on DNA extracted from whole leaf tissues. To facilitate chloroplast genome assembly, we first extracted the chloroplast reads by attempting to align all reads to a dataset of 31 known *Eucalyptus* chloroplast genomes, which we refer to as the reference set (Additional file 1: Table S1). The chloroplast genome is circular, but alignment algorithms rely on linear genomes. Simple linearization of the reference set would risk failing to capture reads that span the point at which the genomes were circularized. To avoid this, we duplicated and concatenated the sequence of each genome in the reference set. In this way, single reads that span the point at which the genome was circularized, or long-reads that span the entire chloroplast genome, would successfully map to the reference set. Short-reads were aligned to the reference set using Bowtie2 v2.2.6 [72] and long-reads were aligned to the reference set using Blasr v5.1 [73]. Scripts for this analysis are available in the 1_pre_assembly/2_cpDNAExtraction folder of the github repository.

Genome assembly

Validation dataset preparation

As this is the first time that the *E. pauciflora* chloroplast genome has been sequenced, and because we want to assess and compare a range of different approaches to genome assembly, we used a subset of our short-read data to validate our genome assemblies. To do this, we randomly selected 100x coverage of paired-end Illumina reads (59,656 pairs of reads in total) from all of the chloroplast paired-end Illumina reads identified above. We excluded these reads from all genome assemblies, and instead used them exclusively to assess and compare the genome assemblies we produced. In general, we should expect that reads in the validation set will map more successfully to chloroplast genomes that are closer to the true genome from which the reads were sequenced. Because of this, we can use the mapping rate and the error rate of mapped validation reads to different genome assemblies to compare different genome assemblies. The mapping rate may not reach 100% even with a perfect genome assembly, because we cannot guarantee that all of reads in the validation set come from the chloroplast genome (e.g. some may come from nuclear copies of chloroplast genes, or from contaminants) and some reads will also contain sequencing errors. Nevertheless, as long as the vast majority of the validation reads come from the chloroplast genome of *E. pauciflora,* it should be the case that the mapping rate of the validation reads will increase monotonically as the accuracy of the chloroplast genome assembly increases. We also expect the error rate of mapped validation reads to decrease monotonically as the accuracy of the genome assembly increases. Short Illumina reads typically have a low error rate of ~ 0.0010 errors per base [74], therefore for a perfect genome assembly we might expect the mapped validation reads to have an error rate of at most ~ 0.0010 errors per base (the true number is likely to be lower than this, because we filtered out reads and bases with low quality scores before constructing the validation set). Assembly errors will increase the error rate in the mapped validation reads. Thus, if the error rate of the mapped reads is greater than ~ 0.0010 errors per base, this suggests that there are errors that cannot be attributed to sequencing error, and are likely to represent assembly errors. If the error rate is less than 0.0010 errors per base, this suggests there are few if any errors in the assembly, although it does not guarantee that the assembly is error free. Scripts for random read selection are available in the 2_assembly/randomSelection folder of the github repository.

Long-read-only assembly

We performed 24 long-read-only assemblies, using all combinations of two assemblers, Hinge and Canu, and 12 different coverages (calculated assuming a 160 kb genome size): 5x, 8x, 10x, 20x, 40x, 60x, 80x, 100x, 200x, 300x, 400x, 500x. Reads for each coverage were selected randomly from the full set of filtered and trimmed long-reads. The assemblers were run with the default settings, except: (i) in Hinge, the read filter cut-off was set to 10, as it failed to assemble reads with default setting; and (ii) in Canu, the correctErrorRate was increased to 0.154 when the coverage was less than 40x according to the recommendation in the Canu manual. Hinge failed to assemble the chloroplast genome with 5x, 8x and 10x long-read coverage. Scripts for this analysis are available in the 2_assembly/longReadOnly folder of the github repository.

Short-read-only assembly

We performed 48 short-read-only assemblies using default settings in Unicycler v0.3.1. These 48 assemblies comprised 12 different coverages (5x, 8x, 10x, 20x, 40x, 60x, 80x, 100x, 200x, 300x, 400x, 500x), each combined with four types of short-read error correction: (i) no correction; (ii) SPAdes correction [75]; (iii) Karect correction [76]; and (iv) Karect +SPAdes correction. A recent study [77] suggested that Karect performs better than other error correction methods, while SPAdes is the correction method built into the Unicycler pipeline. Corrected and uncorrected paired-end reads for the different coverages were randomly selected from the total set of filtered and trimmed short-reads prior to assembly. Scripts for this analysis are available in the 2_assembly/shortReadOnly folder of the github repository.

To test the performance of organelle genome specific assembler, we employed NOVOPlasty [78] to assemble the *E. pauciflora* chloroplast genome with different short-read coverages (5x, 8x, 10x, 20x, 40x, 60x, 80x, 100x, 200x, 300x, 400x, 500x). Following the suggested best practice in the NOVOPlasty documentation, we only removed the adaptors from the short-reads. We used two settings for the“Genome Range”parameter: 120 kb – 200 kb to reflect a case in which the range includes the likely correct genome size (~ 160 KB), and 120 kb – 140 kb to reflect a case where the correct genome size is slightly underestimated. We used the closely related *E. regnans* chloroplast genome as a reference sequence. We found that the final assembly was highly sensitive to the maximum specified genome size, and that NOVOPlasty assemblies showed unpredictable behavior as coverage increased (see Additional file 14: Supplementary results and Additional file 15: Table S2). Because of this, we do not discuss the NOVOPlasty assemblies further, but instead use the results from Unicycler to represent the short-read-only assemblies.

Hybrid assembly

We performed 576 hybrid assemblies combining long- and short-reads. All assemblies were performed using default settings in Unicycler v0.3.1, except for those assemblies which used only Karect to correct the reads, for which we turned off the SPAdes correction in Unicycler). The 576 assemblies comprise all 144 possible combinations of 12 different coverages of both long- and short-read used above (5x, 8x, 10x, 20x, 40x, 60x, 80x, 100x, 200x, 300x, 400x, 500x), repeated for four types of short-read error correction: (i) no correction; (ii) SPAdes correction; (iii) Karect correction; and (iv) Karect +SPAdes correction. We did not perform long-read error correction prior to assembly, because Unicycler is designed to work with raw long-read data, and we polish all genome assemblies using the best available short-read polishing tools. Scripts for this analysis are available in the 2_assembly/hybrid folder of the github repository.

Hybrid assembly with different lengths of long-reads

To investigate how the length of long-reads affects the accuracy of chloroplast genome assemblies, we performed 576 genome assemblies that combined short-reads with long-reads of different lengths. All assemblies were performed using Unicycler v0.3.1. The 576 assemblies comprise all 96 possible combinations of short- (5x, 8x, 10x, 20x, 40x, 60x, 80x, 100x, 200x, 300x, 400x, 500x) and long-read coverage (5x, 8x, 10x, 20x, 40x, 60x, 80x, 100x), repeated for six different length categories for the long-reads: 5–10 kb, 10–20 kb, 20–30 kb, 30–40 kb, 40–50 kb, ≥50 kb. As above, sets of reads were picked randomly from each length specific sub-dataset. Short-reads were corrected by Karect, and long-reads were uncorrected, as above. We did not investigate higher coverage of long-reads because 100x was the maximum coverage category available for the longest subset of long-reads (≥50 kb). Assembly scripts for this analysis are the same as for the hybrid assemblies, except for the input files.

Genome polishing

To facilitate certain comparisons among genome assemblies, we manually assembled the multiple contigs into a single contig from some genome assemblies. For example, the long-read-only assemblies, short-read-only assemblies, and hybrid assemblies with <20x long-read coverage all tended to produce multiple contigs, occasionally with some regions of the chloroplast genome represented more than once. This is, of course, useful information for comparing assembly performance. However, in order to calculate comparable estimates of the per-base error rate from our validation set, we manually removed duplicated regions from these assemblies and then manually combined all remaining contigs together to create a single contig spanning as much of the chloroplast genome as possible. To do this, for assemblies with multiple contigs, we used *E. regnans* (NCBI accession

NC_022386.1) as a reference, which is a close phylogenetic relative of *E. pauciflora*, and created pairwise genome alignments between this reference and the contigs from each assembly using Mummer v3.23 [79], after removing contigs shorter than 5 kb. We then used a python script to remove duplicate regions according to the pairwise genome alignments. Scripts for this analysis are available in the 3_post_assembly/1_same_structure folder of the github repository.

All genome assemblies were polished at the individual base level using the best available tools. Hybrid and short-read-only assemblies were polished using Pilon v1.22 [80], and long-read-only assemblies were polished using Racon [28], Nanopolish v0.8.1 [81], and Racon +Nanopolish v0.8.1. Pilon uses information in short-reads to polish out errors from assemblies, whereas Racon and Nanopolish use information in long-reads. Nanopolish is designed for ONT data, but Racon is designed for ONT and PacBio data. We ran Nanopolish and Pilon for multiple iterations until the polished genome remained unchanged. We ran Racon for ten iterations, because the polished genome continued changing by tiny amounts after 10 iterations during testing. Scripts for this analysis are available in the 3_post_assembly/2_polish folder of the github repository.

Evaluation of different assembly performance

A perfect genome assembly would cover the entire genome in a single contig, with the maximum possible percentage of validation reads mapping successfully with the lowest possible error rate, given the limitations of the accuracy of the sequencing platform and the mapping software. Therefore, to assess and compare our genome assemblies, we considered five key statistics: (i) the number of contigs output by the raw assembly; (ii) the average per-base coverage of the reference genome by the contigs (calculated by the sum of any part of any contig that aligns successfully to the *E. regnans* chloroplast genome divided by the length of the *E. regnans* chloroplast genome), which should equal ~ 1.0 for a perfect assembly but may be much higher than one for assemblies that contain duplicated regions; (iii) the sum of the length of all contigs output by the assembler; (iv) the percentage of the validation reads successfully mapped to the polished genome; (v) the error rate of the validation reads that successfully mapped to the polished genome. As mentioned above, the error rate here is the sum of mismatches and indels (insertions or deletions) between the validation reads and the assemblies, because it is not possible to know in advance whether a mismatch or indel between a read and an assembly derives from an error in the read or the assembly (although this may be determined post-hoc by analysis of the aligned reads to the assembly, see below). We mapped the validation reads to each genome with default settings in Bowtie v2.2.6, and extracted the mapping rate and general error rate from the result using Qualimap v2.2.1 [82]. Scripts for this analysis are available in the 3_post_assembly/3_assembly_quality_control folder of the github repository.

Final genome assessment

To further assess the assembly with lowest error rate, highest mapping rate, and coverage closest to 1.0, we attempted to estimate the number of possible errors and heteroplasmic sites in that assembly using all of the long- and short-read data available to us. To do this, we mapped all of the short-reads with Bowtie v2.2.6, and long-reads using Ngmlr [83] (because Blasr failed to produce BAM file for ONT data) to the assembly after duplicating and concatenating it as above, and compared the mapped reads at each site by visualising the alignments in IGV [84], and by using samtools v1.5 [85] and varScan v2.4.0 [86] to look for mismatches and indels between the short-reads and the assembly, and Nanopolish v0.8.1 to look for mismatches and indels between the long-reads and the assembly. We considered that a true assembly error would be likely to produce consistent mismatches or indels between the assembly and both the long- and short-reads. Both technologies have their limitations which may produce well-supported mismatches or indels between the assembly and one read type. For example, short-reads that derive from regions of the nuclear or mitochondrial genome that are similar to the chloroplast genome may map with high probability to the chloroplast genome, but this mapping is far less likely to occur with long-reads, which in our dataset are a minimum of 5 kb. Similarly, ONT reads are known to contain biases with respect to the length of homopolymer regions, but this bias should not occur in short-reads. Thus, if the frequency of the non-reference base of a mismatch or indel is similar between the short- and long-read mappings, then this is most likely to indicate either an assembly error or a heteroplasmic site. Scripts for this analysis are available in the 3_post_assembly/4_SNP_call folder of the github repository.

Gene annotation

We used GeSeq [87] to annotate the best assembly, followed by ARAGORN v1.2.38 [88] to annotate the tRNAs. Since GeSeq failed to detect the very short exons, genes containing such exons, such as *petB* and *petD*, were manually annotated based on the annotations in the reference set of chloroplast genomes.

Phylogenetic analysis

We added the *E. pauciflora* chloroplast genome estimated here to the reference set of chloroplast genomes used above, giving a dataset of 32 complete and well-annotated chloroplast genomes. We split each genome into 312 fragments based on their annotations, where each fragment represents a single exonic or non-exonic region. We then produced 312 individual alignments of the 32 sequences (one alignment for each fragment) using Clustal Omega v1.2.4 [89]. We then manually examined the 312 alignments, deleted regions of each alignment that could not be confidently aligned, and manually adjusted the remaining regions. Subsequently, we concatenated the 312 alignments to create a single alignment with 32 sequences and 156,245 bases. We estimated a phylogeny from this alignment using IQ-TREE v1.5.5 [90], and ModelFinder [91] with default settings, and 1000 ultrafast bootstrap replicates [92]. Scripts for this analysis are available in the phylogenetic_analysis folder of the github repository.

Additional files

Additional file 1: Table S1. Other known Eucalyptus chloroplast genomes. (XLSX 9 kb)

Additional file 2: Figure S1. The summary of long-read-only assemblies. Long_5x indicates the 5x coverage of long-read was used in the assembly. Hinge failed to assemble genome with <20x coverage. None, Racon, Nanopolish and Racon+Nanopolish mean the different genome polishing pipeline. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). **C.** The mapping rate of validation reads to the assemblies after manual curation. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. (PNG 1920 kb)

Additional file 3: Figure S2. The summary of short-read-only and hybrid assemblies with Karect short-read correction. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). Numbers marked with red contained a single contig covering the whole chloroplast genome. The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). Numbers marked with red denote assemblies with the expected length, in the range 155,938 bp–155,945 bp. **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1500 kb)

Additional file 4: Figure S3. The summary of short-read-only and hybrid assemblies with SPAdes short-read correction. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). Numbers marked with red contained a single contig covering the whole chloroplast genome. The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). Numbers marked with red denote assemblies with the expected length, in the range 155,938 bp–155,945 bp. **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1550 kb)

Additional file 5: Figure S4. The summary of short-read-only and hybrid assemblies with Karect+SPAdes short-read correction. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). Numbers marked with red contained a single contig covering the whole chloroplast genome. The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). Numbers marked with red denote assemblies with the expected length, in the range 155,938 bp–155,945 bp. **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1500 kb)

Additional file 6: Figure S5. The summary of short-read-only and hybrid assemblies without any short-read correction. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). Numbers marked with red contained a single contig covering the whole chloroplast genome. The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). Numbers marked with red denote assemblies with the expected length, in the range 155,938 bp–155,945 bp. **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1540 kb)

Additional file 7: Figure S6. The summary of assemblies with Karect-corrected short-reads and ≤ 10 kb long-reads. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1210 kb)

Additional file 8: Figure S7. The summary of assemblies with Karect corrected short-reads and 10–20 kb long-reads. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1190 kb)

Additional file 9: Figure S8. The summary of assemblies with Karect corrected short-reads and 20–30 kb long-reads. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). Numbers marked with red contained a single contig covering the whole chloroplast genome. The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). Numbers marked with red denote assemblies with the expected length, in the range 155,938 bp–155,945 bp. **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1330 kb)

Additional file 10: Figure S9. The summary of assemblies with Karect corrected short-reads and 30–40 kb long-reads. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). Numbers marked with red contained a single contig covering the whole chloroplast genome. The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). Numbers marked with red denote assemblies with the expected length, in the range 155,938 bp–155,945 bp. **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1190 kb)

Additional file 11: Figure S10. The summary of assemblies with Karect corrected short-reads and 40–50 kb long-reads. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). Numbers marked with red contained a single contig covering the whole chloroplast genome. The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). Numbers marked with red denote assemblies with the expected length, in the range 155,938 bp–155,945 bp. **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1210 kb)

Additional file 12: Figure S11. The summary of assemblies with Karect corrected short-reads and ≥ 50 kb long-reads. Long and Short indicate the coverage of this assembly. **A.** The total coverage of the chloroplast genome across all contigs output by the assembler. The number is the number of contigs of each assembly, whereas the heatmap is the genome coverage (it could be over 100% if some duplications exist). Numbers marked with red contained a single contig covering the whole chloroplast genome. The heatmap color is reversed compared to the Fig. 1 to make the color in all figure panel A show consistence. **B.** The assembly length of different assemblies after manual curation (e.g. removing duplicate regions). Numbers marked with red denote assemblies with the expected length, in the range 155,938 bp–155,945 bp. **C.** The mapping rate of validation reads to the assemblies after manual curation. Assemblies with highest mapping rate (99.43%) are marked with red. **D.** The average per-base error rate of validation reads mapped to each manually-curated genome assembly. Assemblies with the lowest error rate (0.0007) are marked with red. (PNG 1180 kb)

Additional file 13: Figure S12. The long−/short-read mapping coverage in the possible heteroplastic site 51,961 (IGV view). The purple "I" in short-reads indicated the T insertion, whereas the black lines in long-reads indicate the deletion during that region. (PNG 978 kb)

Additional file 14: Supplementary result. (DOCX 13 kb)

Additional file 15: Table S2.. The result of NOVOPlast assembly (XLSX 9 kb)

Acknowledgements
We thank Eleanor Beavan for help with tissue collection and DNA extractions.

Funding
This research is supported by the Australian Research Council Future Fellowship, FT140100843.

Availability of data and materials
The datasets supporting the results of this article are available on NCBI, https://www.ncbi.nlm.nih.gov/ (accession number: MG921592), and github, https://github.com/asdcid/Chroloplast-genome-assembly.

Authors' contributions
AS and RL carried out sample collection and AS performed DNA extraction for Illumina sequencing. RL and MS carried out sample collection for MinION sequencing. MS and BS performed DNA extraction, library preparation, and MinION sequencing. WW, DK, RL, MS, and BS were involved in data QC and discussion. WW and RL planned all analyses. WW conducted all data analyses and drafted the original manuscript. RL and WW finished the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Leaves were collected a single *E. pauciflora* individual in Thredbo, Kosciuszko National Park, New South Wales, Australia (Latitude − 36.49433, Longitude 148.282983) with written permission from the Scientific Licensing office of the Office of Environment and Heritage for New South Wales: www.licence.nsw.gov.au, in accordance with national guidelines in Australia. Tissues were not deposited as voucher specimens.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Research School of Biology, Australian National University, Canberra, Australia. [2]Institute of Applied Genetics and Cell Biology, University of Natural Resources and Life Sciences, Vienna, Austria. [3]Department of Biological Sciences, Macquarie University, Sydney, Australia.

Received: 15 May 2018 Accepted: 3 December 2018
Published online: 29 December 2018

References

1. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution and applications in genetic engineering. Genome Biol. 2016;17(1):134.
2. Gibbs AK, Udovicic F, Drinnan AN, Ladiges PY. Phylogeny and classification of Eucalyptus subgenus Eudesmia (Myrtaceae) based on nuclear ribosomal DNA chloroplast DNA and morphology. Aus Syst Botany. 2009;22(3):158–79.
3. Parks M, Cronn R, Liston A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 2009;7:84.
4. Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JM, et al. Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. Am J Bot. 2012;99(2):320–9.

5. Bayly MJ, Rigault P, Spokevicius A, Ladiges PY, Ades PK, Anderson C, et al. Chloroplast genome analysis of Australian eucalypts--Eucalyptus, Corymbia, angophora, Allosyncarpia and Stockwellia (Myrtaceae). Mol Phylogenet Evol. 2013;69(3):704–16.
6. Zheng YH, Alverson AJ, Wang QF, Palmer JD. Chloroplast phylogeny of Cucurbita: evolution of the domesticated and wild species. J Syst Evol. 2013; 51(3):326–34.
7. Li B, Cantino PD, Olmstead RG, Bramley GL, Xiang CL, Ma ZH, et al. A large-scale chloroplast phylogeny of the Lamiaceae sheds new light on its subfamilial classification. Sci Rep. 2016;6:34343.
8. Wei SJ, Lu YB, Ye QQ, Tang SQ. Population genetic structure and Phylogeography of Camellia flavida (Theaceae) based on Chloroplast and nuclear DNA sequences. Front Plant Sci. 2017;8:718.
9. Wicke S, Schneeweiss GM, dePamphilis CW, Muller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order gene function. Plant Mol Biol. 2011;76(3–5):273–97.
10. Zheng XM, Wang JR, Feng L, Liu S, Pang BO, Qi L, et al. Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. Sci Rep. 2017;7(1): 1555.
11. Oldenburg DJ, Bendich AJ. Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. J Mol Biol. 2004; 335(4):953–70.
12. Oldenburg DJ, Bendich AJ. The linear plastid chromosomes of maize: terminal sequences, structures and implications for DNA replication. Curr Genet. 2016;62(2):431–42.
13. Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB, Daniell H. Complete plastid genome sequence of the chickpea (Cicer arietinum) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). Mol Phylogenet Evol. 2008;48(3):1204–17.
14. Hirao T, Watanabe A, Kurita M, Kondo T, Takata K. Complete nucleotide sequence of the Cryptomeria japonica D. Don. Chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. BMC Plant Biol. 2008;8:70.
15. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, et al. The complete chloroplast genome sequence of Pelargonium x hortorum: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol. 2006;23(11):2175–90.
16. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 1986;5(9):2043–9.
17. Kunnimalaiyaan M, Nielsen BL. Fine mapping of replication origins (ori a and ori B) in Nicotiana tabacum chloroplast DNA. Nucleic Acids Res. 1997; 25(18):3681–6.
18. Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH. The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol. 2004;21(7):1445–54.
19. Nie X, Lv S, Zhang Y, Du X, Wang L, Biradar SS, et al. Complete chloroplast genome sequence of a major invasive species, Crofton weed (Ageratina adenophora). PLoS One. 2012;7(5):e36869.
20. Yang JB, Tang M, Li HT, Zhang ZR, Li DZ. Complete chloroplast genome of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evol Biol. 2013;13:84.
21. Lee J, Kang Y, Shin SC, Park H, Lee H. Combined analysis of the chloroplast genome and transcriptome of the Antarctic vascular plant Deschampsia antarctica Desv. PLoS One. 2014;9(3):e92501.
22. Giordano F, Aigrain L, Quail MA, Coupland P, Bonfield JK, Davies RM, et al. De novo yeast genome assemblies from MinION PacBio and MiSeq platforms. Sci Rep. 2017;7(1):3935.
23. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19(6): 1117–23.
24. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20(2):265–72.
25. Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, et al. De novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. Gigascience. 2017;6(2):1–13.
26. Oikonomopoulos S, Wang YC, Djambazian H, Badescu D, Ragoussis J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Sci Rep. 2016;6:31602.
27. de Lannoy C, de Ridder D, Risse J. The long reads ahead: de novo genome assembly using the MinION. F1000Research. 2017;6:1083.
28. Vaser R, Sovic I, Nagarajan N, Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–46.
29. Ferrarini M, Moretto M, Ward JA, Surbanovski N, Stevanovic V, Giongo L, et al. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics. 2013;14:670.
30. Wu Z, Gui S, Quan Z, Pan L, Wang S, Ke W, et al. A precise chloroplast genome of Nelumbo nucifera (Nelumbonaceae) evaluated with sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. BMC Plant Biol. 2014;14:289.
31. Redwan RM, Saidin A, Kumar SV. Complete chloroplast genome sequence of MD-2 pineapple and its comparative analysis among nine other plants from the subclass Commelinidae. BMC Plant Biol. 2015;15:196.
32. Chaney L, Mangelson R, Ramaraj T, Jellen EN, Maughan PJ. The complete chloroplast genome sequences for four Amaranthus species (Amaranthaceae). Appl Plant Sci. 2016;4(9):1600063.
33. Wang L, Wuyun T-n DH, Wang D, Cao D. Complete chloroplast genome sequences of Eucommia ulmoides: genome structure and evolution. Tree Genet Genomes. 2016;12(1):12.
34. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, et al. Comprehensive comparison of Pacific biosciences and Oxford Nanopore technologies and their applications to transcriptome analysis. F1000Res. 2017;6:100.
35. Xu Z, Xin T, Bartels D, Li Y, Gu W, Yao H, et al. Genome analysis of the ancient Tracheophyte Selaginella tamariscina reveals evolutionary features relevant to the Acquisition of Desiccation Tolerance. Mol Plant. 2018;11(7): 983–94.
36. Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, Martin W. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 2005;138(3): 1723–33.
37. Chen H, Yu Y, Chen X, Zhang Z, Gong C, Li J, et al. Plastid DNA insertions in plant nuclear genomes: the sites, abundance and ages and a predicted promoter analysis. Funct Integr Genomics. 2015;15(2):131–9.
38. Ma PF, Zhang YX, Guo ZH, Li DZ. Evidence for horizontal transfer of mitochondrial DNA to the plastid genome in a bamboo genus. Sci Rep. 2015;5:11608.
39. Spooner DM, Ruess H, Iorizzo M, Senalik D, Simon P. Entire plastid phylogeny of the carrot genus (Daucus, Apiaceae): concordance with nuclear data and mitochondrial and nuclear DNA insertions to the plastid. Am J Bot. 2017;104(2):296–312.
40. Hoang NV, Furtado A, McQualter RB, Henry RJ. Next generation sequencing of total DNA from sugarcane provides no evidence for chloroplast heteroplasmy. New Neg Plant Sci. 2015;1-2:33–45.
41. Department of Agriculture and Water Resources Australian forest profiles Eucalypt. 2016.
42. Williams JE. Biogeographic patterns of three sub-alpine eucalypts in south-East Australia with special reference to Eucalyptus pauciflora Sieb. Ex Spreng J Biogeography. 1991;18(2):223–30.
43. Boland DJ, Brooker MIH, Chippendale GM, Hall N, Hyland BPM, R.D. J, et al. Forest trees of Australia. Canberra: CSIRO; 2002.
44. Gauli A, Vaillancourt RE, Bailey TG, Steane DA, Potts BM. Evidence for local climate adaptation in early-life traits of Tasmanian populations of Eucalyptus pauciflora. Tree Genet Genomes. 2015;11:104–15.
45. Cochrane PM, Slatyer RO. Water relations of Eucalyptus pauciflora near the alpine tree line in winter. Tree Physiol. 1988;4(1):45–52.
46. Evans JR, Vogelmann TC. Photosynthesis within isobilateral Eucalyptus pauciflora leaves. New Phytol. 2006;171(4):771–82.
47. Warren CR. Uptake of inorganic and amino acid nitrogen from soil by Eucalyptus regnans and Eucalyptus pauciflora seedlings. Tree Physiol. 2009; 29(3):401–9.
48. Buckley TN, Turnbull TL, Pfautsch S, Adams MA. Nocturnal water loss in mature subalpine Eucalyptus delegatensis tall open forests and adjacent *E pauciflora* woodlands. Ecol Evol. 2011;1(3):435–50.
49. Martorell S, Diaz-Espejo A, Medrano H, Ball MC, Choat B. Rapid hydraulic recovery in Eucalyptus pauciflora after drought: linkages between stem hydraulics and leaf gas exchange. Plant Cell Environ. 2014;37(3):617–26.

50. Way DA, Holly C, Bruhn D, Ball MC, Atkin OK. Diurnal and seasonal variation in light and dark respiration in field-grown Eucalyptus pauciflora. Tree Physiol. 2015;35(8):840–9.
51. Wick RR, Judd LM, Gorrie CL, Unicycler HKE. Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017; 13(6):e1005595.
52. Kamath GM, Shomorony I, Xia F, Courtade TA, Tse DNHINGE. Long-read assembly achieves optimal repeat resolution. Genome Res. 2017;27(5):747–56.
53. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
54. Steane DA, Nicolle D, Sansaloni CP, Petroli CD, Carling J, Kilian A, et al. Population genetic analysis and phylogeny reconstruction in Eucalyptus (Myrtaceae) using high-throughput, genome-wide genotyping. Mol Phylogenet Evol. 2011;59(1):206–24.
55. Huang H, Shi C, Liu Y, Mao SY, Gao LZ. Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evol Biol. 2014;14:151.
56. Wang M, Cui L, Feng K, Deng P, Du X, Wan F, et al. Comparative analysis of Asteraceae Chloroplast genomes: structural organization RNA Editing and Evolution. Plant Mol Biol Rep. 2015;33(5):1526–38.
57. Chen Z, Feng K, Grover CE, Li P, Liu F, Wang Y, et al. Chloroplast DNA structural variation, phylogeny, and age of divergence among diploid cotton species. PLoS One. 2016;11(6):e0157183.
58. Ni Z, Ye Y, Bai T, Xu M, Complete Chloroplast XLA. Genome of Pinus massoniana (Pinaceae): gene rearrangements, loss of ndh genes, and short inverted repeats contraction Expansion. Molecules. 2017;22(9):1528-41.
59. Rice Chromosome 10 Sequencing C. In-depth view of structure, activity, and evolution of rice chromosome 10. Science. 2003;300(5625):1566–9.
60. Scarcelli N, Mariac C, Couvreur TL, Faye A, Richard D, Sabot F, et al. Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it? Mol Ecol Resour. 2016;16(2):434–45.
61. Diekmann K, Hodkinson TR, Fricke E, Barth S. An optimized chloroplast DNA extraction protocol for grasses (Poaceae) proves suitable for whole plastid genome sequencing and SNP detection. PLoS One. 2008;3(7):e2813.
62. Xiang B, Li X, Qian J, Wang L, Ma L, Tian X, et al. The Complete Chloroplast genome sequence of the medicinal plant Swertia mussotii using the PacBio RS II platform. Molecules. 2016;21(8):1029.
63. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45.
64. Schalamun M, Nagar R, Kainer D, Beavan E, Eccles D, Rathjen JP, et al. Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora. Mol Ecol Resourc. 0(0).
65. Suarez AM, Rutherford S. gDNA Extraction of Eucalypts pauciflora for full genome sequencing. Protocolsio. 2018. https://doi.org/10.17504/protocols.io.j7ecrje.
66. BBMap. http://sourceforge.net/projects/bbmap/. Accessed 16 June 2017.
67. FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 1 June 2017.
68. Schalamun M, Schwessinger B. High molecular weight gDNA extraction after Mayjonade et al. optimised for eucalyptus for nanopore sequencing. Protocolsio 2017. doi: https://doi.org/10.17504/protocols.io.ka2csge.
69. Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, Callot C, et al. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. BioTechniques. 2016;61(4):203–5.
70. Porechop. https://github.com/rrwick/Porechop. Accessed 13 July 2017.
71. NanoFit. https://pypi.python.org/pypi/NanoFilt. Accessed 13 July 2017.
72. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9.
73. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13:238.
74. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011;11(5):759–69.
75. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
76. Allam A, Kalnis P, Solovyev V. Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data. Bioinformatics. 2015;31(21):3421–8.
77. Heydari M, Miclotte G, Demeester P, Van de Peer Y, Fostier J. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. BMC Bioinformatics. 2017;18(1):374.
78. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017;45(4):e18.
79. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
80. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963.
81. Loman NJ, Quick J. Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12(8): 733–5.
82. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292–4.
83. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods. 2018;15(6):461-8.
84. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.
85. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
86. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76.
87. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45(W1):W6–W11.
88. Laslett D, Canback BARAGORN. A program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32(1):11–6.
89. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. Mol Syst Biol. 2011;7:539.
90. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016;44(W1):W232–5.
91. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–9.
92. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35(2): 518–22.
93. Lohse M, Drechsel O, Bock R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet. 2007;52(5–6):267–74.