Authors

Teng Li¹, David Kainer², William J. Foley¹, Allen Rodrigo¹, Carsten Külheim^{3*}

Title

The draft genome sequence of *Eucalyptus polybractea* based on hybrid assembly with short- and long-reads reads

Affiliations

1. Research School of Biology, The Australian National University, Canberra 2600, Australia

2. Center for BioEnergy Innovation, Bioscience Division, Oak Ridge National Laboratories, Oak Ridge TN 37831 USA

3. College of Forest Resources and Environmental Science, Michigan Technological University, Houghton, 49931 MI USA

* Author for correspondence: Carsten Külheim, ckulheim@mtu.edu

Orcid ID's: TL: 0000-0002-3224-0477; DK: 0000-0001-7271-4676; WJF: 0000-0001-8587-1814; AR: 0000-0002-8327-7317; CK: 0000-0002-0798-3324

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the US Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doepublic-access-plan>).

Acknowledgements

The authors would like to acknowledge Ms Miriam Schalamun and Dr Benjamin Schwessinger for aiding in the development of the native DNA extraction protocol.

Abstract

Eucalyptus polybractea is a small, multi-stemmed tree, which is widely cultivated in Australia for the production of Eucalyptus oil. We report the hybrid assembly of the *E. polybractea* genome utilizing both short- and long-read technology. We generated 44 Gb of Illumina HiSeq short reads and 8 Gb of Nanopore long reads, representing approximately 83× and 15× genome coverage, respectively. The hybrid-assembled genome, after polishing, contained 24,864 scaffolds with an accumulated length of 523 Mb (N50 = 40.3 kb; BUSCO-calculated genome completeness of 94.3%). The genome contained 35,385 predicted protein-coding genes detected by combining homology-based and de novo approaches. We have provided the first assembled genome based on hybrid sequences from the highly diverse *Eucalyptus* subgenus *Symphyomyrtus*, and revealed the value of including long-reads from Nanopore technology for enhancing the contiguity of the assembled genome, as well as for improving its completeness. We anticipate that the *E. polybractea* genome will be an invaluable resource supporting a range of studies in genetics, population genomics and evolution of related species in *Eucalyptus*.

Key words

Hybrid genome assembly, Eucalyptus, eucalyptus oil, nanopore long reads, genome

Declarations

Funding

Funding provided by The Center for Bioenergy Innovation a U.S. Department of Energy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science to DK. TL and AR received University internal funding from the Australian National University and WJF and CK received funding from the Australian Research Council grant DP14101755.

Conflicts of interest/Competing interests

The authors declare no conflict of interest or competing interests.

Consent for publication

All authors have read and approved the manuscript for submission to Tree Genetics and Genomes.

Availability of data and material

Raw Illumina and Nanopore reads sequenced in this study are available in the Short Read Archive, together with the Whole Genome Shotgun project have been deposited at NCBI under BioProject PRJNA525896. The assembled genome is available at GenBank under the accession number SMSX000000000. All supplementary figures and tables are provided in additional file 1.

Code availability

All tools used in this study were properly described in the Methods section together with the versions used.

Authors' contributions

The study was conceived by CK, WJF and AR. DNA for short-read sequencing was extracted by DK, DNA for long-read sequencing was extracted by CK and sequenced by CK. TL did the bioinformatics analyses of this study and wrote the draft of the MS with DK and CK. All authors read and approved the final version of this manuscript.

Introduction

Eucalypts (*Eucalyptus*, *Angophora* and *Corymbia* spp.) are economically important as the world's most widely planted hardwood genera (Myburg et al. 2014). Globally, eucalypts are planted for pulp and paper production, timber, firewood and 1,8-cineole-rich essential oils ("Eucalyptus oil"), which are extracted from their leaves (Bauhus et al. 2010). They are also the dominant forest trees in Australia and thus of great ecological value. These genera are highly diverse with more than 800 species (Thornhill et al. 2019), ranging from the tallest flowering plant in the world (*E. regnans* over 100 m tall) to small shrubs (e.g. *E. vernicosa*, *E. yaltensis* and *E. surgens* < 2 m). Eucalypts have adapted to a wide range of climates in Australia and nearby islands from tropical to arid and temperate zones. They thrive in tropical monsoon (Köppen-Geiger classification Am), tropical rainforest (Af) and tropical

savannah (Aw) regions with species such as *E. melanophloia*, *E. tereticornis* and *E. tetradonta*, respectively. Eucalypts are common in the arid zones which dominate Australia's interior, with species such as *E. camaldulensis* (Bustos-Segura et al. 2017) surviving in the hot desert climate (BWh). In contrast, *Eucalyptus pauciflora* grows from coastal New South Wales and Victoria (Cfb) to the treeline of the Australian Alps (Wang et al. 2020; Woldendorp et al. 2008), which have a subpolar oceanic climate (Cfc). While some species occupy very narrow geographic and climate space (e.g. *E. polybractea*), others, such as *E. camaldulensis* (Bustos-Segura et al. 2017) and *E. tereticornis* (Bourne et al. 2017) cover vastly different climate space across the Australian continent.

Since the sequencing and annotation of the first *Eucalyptus* genome, *E. grandis* (Myburg et al. 2014), the focus of genetic and genomic research has been biomass accumulation (Cappa et al. 2013; Müller et al. 2019; Resende et al. 2012), wood properties (Cappa et al. 2013; Resende et al. 2017) and pest and pathogen resistance (Mangwanda et al. 2016; Oates et al. 2015; Resende et al. 2017), which are all important traits for the pulp and paper industry. Several studies have investigated species that are exclusively grown for the production of Eucalyptus oil by using genomic information based either on candidate genes in the terpene biosynthetic pathway (Padovan et al. 2017) or utilizing the related *E. grandis* genome sequence for reference mapping of reads (Kainer et al. 2019; Kainer et al. 2018). Due to high levels of synteny across the genus (Hudson et al. 2012), as well as shared ancestral genomic variants (Külheim et al. 2009), mapping of reads obtained from one species to another has been possible. However, regions that are unique to the non-reference species will not align, or align erroneously, and the genetic variants within those regions go undetected. Mapping quality has also improved with several iterations of editing fixed alleles in two eucalypt studies (Kainer et al. 2019; Orr et al. 2020).

Sequencing, assembly and annotation of *E. polybractea* will allow better short-read mapping of this and other closely-related species that are used for the production of essential oils. Having genome sequences from multiple species of eucalypts will also allow comparative studies, including variation of form (tall trees versus shrub), adaptation to climate, or the quantitative and qualitative variation of

essential oils which is a well described trait in this genus (Padovan et al. 2014). Here we report the hybrid assembly of the genome of *E. polybractea* with high-coverage Illumina short reads (~ 83×) and low-coverage (~15×) Nanopore long reads using the MaSuRCA assembler. A schematic workflow of this study is shown in Fig. S1. We anticipate that the assembled and annotated genome of *E. polybractea* will be a catalyst for a range of genetics, population genomics, and evolution studies of this highly diverse genus and other related species.

Material and Methods

Sampling, Library Construction, and Sequencing

Leaf samples were collected from a *Eucalyptus polybractea* R.T. Baker progeny trial on the property of GR Davis Pty Ltd in West Wyalong NSW, Australia (33°58'S, 147°03'E). For Illumina short-read sequencing, leaves from 12 related (half-sib) individuals were flash frozen in liquid nitrogen and stored in -80°C until DNA extraction using the Qiagen DNeasy Plant kit (Qiagen, Valencia, CA, USA) as described in (Kainer et al. 2018). Short-read sequencing libraries were prepared using a modified protocol based on Rohland and Reich (2012) and were sequenced with 125 bp paired-end reads on the Illumina HiSeq 2500 platform at Macrogen (Republic of Korea). A total of 43.4 Gb of paired-end reads were generated from HiSeq data.

For long-read sequencing, branches were cut and placed in sealed bags for transport to Canberra, Australia, where they were placed in water at 4°C. Leaf sections were placed in a 2 ml Eppendorf tube with 5 mm steel beads, frozen in liquid nitrogen and ground in a Qiagen bead mill. DNA was extracted following the protocol of (Schalamun et al. 2018). DNA was quantified using a Qubit fluorimeter and quality was assessed by separation of the DNA on a 0.7% agarose gel. Due to high foliar concentrations of plant secondary metabolites, particularly tannins and essential oils, purification steps by either SPRI beads or precipitation with chloroform/isoamylalcohol were repeated where necessary. Oxford Nanopore long reads were generated by using the Ligation Sequencing 1D Kit (SQK-LSK108) with approximately 5 µg high quality DNA, which was extracted from four individuals due to varying success

of high-quality native DNA extraction. Initial efforts to extract native DNA from the same individuals as for short-read sequencing failed, likely due to high foliar concentrations of eucalyptus oil. We then selected individuals that had the lowest concentration of foliar eucalyptus oil (Kainer et al. 2019). Four libraries were prepared and sequenced on four different flowcells using the MinION device (Oxford Nanopore, Oxford, UK), Nanopore reads were basecalled from their raw FAST5 files using Albacore v2.3.1. Applying a minimum length cutoff of 500 bp and discarding reads with quality <7, this study produced a total of 7.96 Gbp (N50: 16.1 kb, longest read: 145,007 bp).

Genome Characteristics

Jellyfish v. 2.2.6 (Marcais and Kingsford 2011) was employed to generate the frequency distribution of 17-, 21-, and 25-mers of the clean short reads obtained from paired-end Illumina sequencing libraries, and these distributions were subsequently processed by GenomeScope (Vurture et al. 2017). The analysis shows that the genome size of *Eucalyptus polybractea* is between 490 and 543 Mb with a population heterozygosity level of 2% (Fig. S2). The genome sizes estimated by kmer-based statistical approach are within the range of sizes listed for other *Eucalyptus* species (381 Mb - 724 Mb) as reported on the Plant C-values Database (<http://data.kew.org/cvalues>).

Genome assembly

Eucalyptus polybractea genome was assembled *de novo* using Maryland Super-Read Celera Assembler v.3.2.6 (MaSuRCA) (Zimin et al. 2013) with short reads only, followed by a hybrid (Illumina + Nanopore data) assembly method. The short reads used for both assemblies derived from paired-end libraries which were trimmed for adapters by BBDuk(BBMap), but not trimmed for any aspects of quality, cleaning or error correction. The assembler MaSuRCA constructs long and accurate mega-reads from the combination of long and short read data, essentially converting each Nanopore long read into one or more very long, highly accurate reads (Zimin et al. 2017). The short-reads-only and hybrid assemblies yielded total assembly length of 1,063 Mb and 523 Mb, respectively (Table 1). The guanylic and cytidylic acid (GC) content of the *E. polybractea* hybrid assembly genome (mean = 39.0%) is similar

to the *E. grandis* genome (39.3%). The hybrid assembly was subsequently polished by Pilon v.1.22 (Walker et al. 2014), which uses information in short reads to correct bases, fix mis-assemblies, and fill gaps. Pilon was applied for multiple iterations until the polished genome remained unchanged. Compared to the short read only assembly, the inclusion of approximately 15× genome coverage of Nanopore long reads for a hybrid assembly led to a 95% decrease in the number of scaffolds from 502,024 to 24,864 scaffolds and an 11-fold increase in the scaffold N50 length from 3,393 bp to 40,281 bp (Table 1).

Repeat content analysis

Repetitive sequences in the *E. polybractea* genome were identified and classified using a combination of *de novo* and homology-based approaches. Initially, a *de novo* repeat library of *E. polybractea* was built by RepeatModeler v. 1.0.4 (Smit and Hubley 2008-2015) with default parameters based on the hybrid assembly. RepeatMasker v. open-4.0.7 (Smith et al. 2013-2015) was then employed to detect repeats in the *E. polybractea* genome, by aligning sequences from the hybrid assembly to the RepeatMasker combined database (Dfam_consensus (Hubley et al. 2016) and Repbase (Jurka et al. 2005)) as well as to the *de novo* repeat library that we built. Repeat sequences were estimated to account for 40.1% (210 Mb) of the *E. polybractea* genome (Table S1), with long terminal repeats (LTRs) being the most pervasive class (17.6%). However, DNA transposons encompass only 4.1% of the genome. In addition, the percentage of repetitive elements in *E. polybractea* is less than that in *E. grandis* (50.1%), which likely accounts for the smaller genome size of *E. polybractea* (523 Mb) compared to *E. grandis* (640 Mb).

Genome annotation

The gene annotation for the *E. polybractea* genome was performed by combining homology-based and *de novo* approaches. For the homology-based method, the hybrid assembly of the genome was aligned to the protein-coding genes predicted in *E. grandis* (Myburg et al. 2014) using TBLASTN v. 2.2.31 (Camacho et al. 2009). The homologous genome sequences were then aligned against the

matching protein sequences by GeneWise v. 2.4.1 (Birney et al. 2004) to obtain accurate spliced alignments. The *de novo* prediction was estimated by the *ab initio* gene predictors Augustus v. 3.2.1 (Stanke et al. 2006) and GenScan (Burge and Karlin 1997), followed by filtering partial sequences and genes with length less than 100 bp. The final consensus protein-coding gene set of the *E. polybractea* genome was generated by EVIDENCEModeler (EVM) v. 1.1.1 (Haas et al. 2008), with the combined gene annotation results of *de novo* and homology-based predictions. Our results showed that 35,385 protein-coding genes were identified in the *E. polybractea* genome, with an average coding length of 1,373 bp (Table 2).

Results and Discussion

Assessment of the genome assembly

To evaluate the completeness of the genome assembly, Benchmarking Universal Single-Copy Orthologs v.3.0.2 (BUSCO) (Simao et al. 2015) was employed to search against the Eudicotyledons database of 2,121 orthologs. The results showed that the genome completeness was substantially improved in the hybrid assembly compared to the short reads only assembly, from 67.2% to 94.3% of the complete sequences detected by BUSCO. In addition, 1,765 single-copy and 236 duplicated orthologs were observed in the hybrid assembly with only 21 (1%) fragmented genes (Table 1).

In order to validate our genome assembly by the hybrid approach, comparison was performed between the *E. polybractea* genome and *E. grandis* genome. The protein-coding genes predicted in *E. grandis* genome (Myburg et al. 2014) were mapped to the hybrid assembly of *E. polybractea* by BLAT algorithm with default parameters (Kent 2002). Statistical analysis was performed at different levels of percentage of sequence homology and percentage of coverage, which demonstrated that the hybrid assembly of *E. polybractea* covered approximately 98% of the protein-coding genes in *E. grandis* (Table S2). Further, we mapped the clean short reads from Illumina HiSeq sequencing of 12 related (half-sib) individuals to the hybrid assembly of the *E. polybractea* genome using the Burrows-Wheeler Aligner v0.7.17-r1188 (BWA) (Li 2013), and extracted the mapping rate and coverage rate from the result by

Qualimap v2.2.1 (Okonechnikov et al. 2016). The results showed that our hybrid assembly had high mapping rate of the short reads (97.5%) with a mean depth of 49.3x, while 91.6% of the assembled genome had a mean depth over 10x (Fig. S3), ensuring a high level of mapping accuracy at the nucleotide level. Additionally, we also performed variant calling with the BWA mappings of the 12 half-sib individuals using the Genome Analysis Toolkit v3.5 (DePristo et al. 2011), resulting in the detection of 12,983,296 variants. The mapping was followed by filtering of adjacent indels within 5 bp and low quality (quality score less than 30, sequencing depth either less than one third of mean depth or three times bigger than the mean depth) single nucleotide variants (SNVs) using BCFtools (Li 2011), resulting in the detection of 10,740,778 SNVs that passed filtering. The results revealed that the population heterozygosity level of the *E. polybractea* genome was approximately 2.05%. In summary, all of these analyses indicated that our draft genome sequence has high contiguity, accuracy, and, more importantly, a high degree of gene space completeness.

Data availability

Raw Illumina and Nanopore reads sequenced in this study are available in the Short Read Archive, together with the Whole Genome Shotgun project have been deposited at NCBI under BioProject PRJNA525896. The assembled genome is available at GenBank under the accession number SMSX000000000. All supplementary figures and tables are provided in additional file 1.

References

- Bauhus Jr, Meer Pvd, Kanninen M (2010) Ecosystem goods and services from plantation forests. The Earthscan forest library. Earthscan, London ; Washington, D.C.
- BBMap <http://sourceforge.net/projects/bbmap/>
- Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise Genome Res 14:988-995 doi:10.1101/gr.1865504
- Bourne AE, Creek D, Peters JMR, Ellsworth DS, Choat B (2017) Species climate range influences hydraulic and stomatal traits in Eucalyptus species Annals of Botany 120:123-133 doi:10.1093/aob/mcx020
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA J Mol Biol 268:78-94 doi:10.1006/jmbi.1997.0951
- Bustos-Segura C, Dillon S, Keszei A, Foley WJ, Külheim C (2017) Intraspecific diversity of terpenes of *Eucalyptus camaldulensis* (Myrtaceae) at a continental scale Australian Journal of Botany 65:257-269 doi:<https://doi.org/10.1071/BT16183>
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications BMC Bioinformatics 10:421 doi:10.1186/1471-2105-10-421
- Cappa EP, El-Kassaby YA, Garcia MN, Acuña C, Borralho NMG, Grattapaglia D, Marcucci Poltri SN (2013) Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in *Eucalyptus globulus* PloS one 8:e81267 doi:10.1371/journal.pone.0081267
- DePristo MA et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data Nature Genetics 43:491-+ doi:10.1038/ng.806
- Haas BJ et al. (2008) Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments Genome Biol 9:R7 doi:10.1186/gb-2008-9-1-r7
- Hubley R et al. (2016) The Dfam database of repetitive DNA families Nucleic Acids Res 44:D81-89 doi:10.1093/nar/gkv1272
- Hudson CJ et al. (2012) High synteny and colinearity among Eucalyptus genomes revealed by high-density comparative genetic mapping Tree Genetics & Genomes 8:339-352 doi:10.1007/s11295-011-0444-9
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements Cytogenet Genome Res 110:462-467 doi:10.1159/000084979
- Kainer D, Padovan A, Degenhardt J, Krause S, Mondal P, Foley WJ, Külheim C (2019) High marker density GWAS provides novel insights into the genomic architecture of terpene oil yield in Eucalyptus New Phytologist 223:1489-1504 doi:10.1111/nph.15887
- Kainer D, Stone EA, Padovan A, Foley WJ, Külheim C (2018) Accuracy of Genomic Prediction for Foliar Terpene Traits in Eucalyptus polybractea G3 (Bethesda) 8:2573-2583 doi:10.1534/g3.118.200443
- Kent WJ (2002) BLAT--the BLAST-like alignment tool Genome Res 12:656-664 doi:10.1101/gr.229202
- Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways BMC Genomics 10:11 doi:452 10.1186/1471-2164-10-452
- Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data Bioinformatics 27:2987-2993 doi:10.1093/bioinformatics/btr509
- Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM arXiv preprint arXiv:13033997

- Mangwanda R, Zwart L, van der Merwe N, Moleleki L, Berger D, Myburg A, Naidoo S (2016) Localization and transcriptional responses of *Chrysosporthe austroafricana* in *Eucalyptus grandis* identify putative pathogenicity factors *Frontiers in Microbiology* 7 doi:10.3389/fmicb.2016.01953
- Marcais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers *Bioinformatics* 27:764-770 doi:10.1093/bioinformatics/btr011
- Müller BSF et al. (2019) Independent and Joint-GWAS for growth traits in *Eucalyptus* by assembling genome-wide data for 3373 individuals across four breeding populations *New Phytologist* 221:818-833 doi:doi:10.1111/nph.15449
- Myburg AA et al. (2014) The genome of *Eucalyptus grandis* *Nature* 510:356-362 doi:10.1038/nature13308
- Oates CN, Külheim C, Myburg AA, Slippers B, Naidoo S (2015) The Transcriptome And Terpene Profile Of *Eucalyptus grandis* Reveals Mechanisms Of Defence Against The Insect Pest, *Leptocybe invasa* Plant and Cell Physiology 56:1418-1428 doi:10.1093/pcp/pcv064
- Okonechnikov K, Conesa A, Garcia-Alcalde F (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data *Bioinformatics* 32:292-294 doi:10.1093/bioinformatics/btv566
- Orr AJ et al. (2020) A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant *Proceedings of the Royal Society B: Biological Sciences* 287:20192364 doi:doi:10.1098/rspb.2019.2364
- Padovan A, Keszei A, Külheim C, Foley W (2014) The evolution of foliar terpene diversity in Myrtaceae *Phytochemistry Reviews* 13:695-716 doi:10.1007/s11101-013-9331-3
- Padovan A, Webb H, Mazanec R, Grayling P, Bartle J, Foley WJ, Külheim C (2017) Association genetics of essential oil traits in *Eucalyptus loxophleba*: explaining variation in oil yield *Molecular Breeding* 37:73 doi:10.1007/s11032-017-0667-z
- Resende MDV et al. (2012) Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees *New Phytologist* 194:116-128 doi:10.1111/j.1469-8137.2011.04038.x M4 - Citavi
- Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, Grattapaglia D (2017) Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus* *New Phytologist* 213:1287-1300 doi:10.1111/nph.14266 M4 - Citavi
- Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture *Genome Res* 22:939-946 doi:10.1101/gr.128124.111
- Schalamun M et al. (2018) Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from *Eucalyptus pauciflora* *Mol Ecol Resour* doi:10.1111/1755-0998.12938
- Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs *Bioinformatics* 31:3210-3212 doi:10.1093/bioinformatics/btv351
- Smit A, Hubley R (2008-2015) RepeatModeler Open-1.0
- Smith A, Hubley R, Green P (2013-2015) RepeatMasker Open-4.0.
- Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts *Nucleic Acids Res* 34:W435-439 doi:10.1093/nar/gkl200
- Thornhill AH, Crisp MD, Külheim C, Lam KE, Nelson LA, Yeates DK, Miller JT (2019) A dated molecular perspective of eucalypt taxonomy, evolution and diversification *Australian Systematic Botany* 32:29-48 doi:<https://doi.org/10.1071/SB18015>
- Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC (2017) GenomeScope: fast reference-free genome profiling from short reads *Bioinformatics* 33:2202-2204 doi:10.1093/bioinformatics/btx153
- Walker BJ et al. (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement *Plos One* 9 doi:10.1371/journal.pone.0112963

- 315 Wang W, Das A, Kainer D, Schalamun M, Morales-Suarez A, Schwessinger B, Lanfear R (2020) The draft
316 nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing de novo
317 assemblies GigaScience 9 doi:10.1093/gigascience/giz160
- 318 Woldendorp G, Hill MJ, Doran R, Ball MC (2008) Frost in a future climate: modelling interactive effects
319 of warmer temperatures and rising atmospheric [CO₂] on the incidence and severity of frost
320 damage in a temperate evergreen (*Eucalyptus pauciflora*) Global Change Biology 14:294-308
321 doi:10.1111/j.1365-2486.2007.01499.x
- 322 Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler
323 Bioinformatics 29:2669-2677 doi:10.1093/bioinformatics/btt476
- 324 Zimin AV et al. (2017) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*,
325 a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm Genome Res 27:787-
326 792 doi:10.1101/gr.213405.116

328

Table 1: Summary of genome assembly and annotation of *Eucalyptus polybractea*.

	Illumina + Nanopore	Illumina only
Genome assembly		
Number of contigs	24 864	502 024
Total length	523 113 982 bp	1063 617 135 bp
N50 size	40 281 bp	3393 bp
Longest	580 368 bp	73 252 bp
GC (%)	39.0	38.9
BUSCO genome completeness		
Complete	2001 (94.3%)	1426 (67.2%)
Complete and single copy	1765 (83.2%)	1177 (55.5%)
Complete and duplicated	236 (11.1%)	249 (11.7%)
Fragmented	21 (1.0%)	408 (19.2%)
Missing	99 (4.7%)	287 (13.6%)
Genome annotation		
Number of protein-coding genes	35 385	
Mean transcript length	2878.6 bp	
Mean coding sequence length	1373.3 bp	
Mean exon length	313.6 bp	
Mean intron length	445.5 bp	

Table 2. Genome annotation of the hybrid assembly *E. polybractea* genome.

Method	Gene number	Mean transcript length (bp)	Mean CDS length (bp)	Mean exon per gene	Mean exon length (bp)	Mean intron length (bp)
Homology to <i>E. grandis</i>	35,490	2702.3	1330.4	4.2	313.2	422.4
<i>De novo</i> by AUGUSTUS	28,157	2888.5	1231.1	5.0	244.4	410.5
<i>De novo</i> by GENESCAN	26,098	2774.7	1455.4	3.4	423.0	540.6
EVidenceModeler (EVM)	35,385	2878.6	1373.3	4.4	313.6	445.5