Chromosome-level *de novo* genome assembly of *Telopea speciosissima* (New South Wales waratah) using long-reads, linked-reads and Hi-C

Stephanie H Chen[1,2], Maurizio Rossetto[2,3], Marlien van der Merwe[2], Patricia Lu-Irving[2], Jia-Yee S Yap[2,3], Hervé Sauquet[4,5], Greg Bourke[6], Jason G Bragg[2,5], Richard J Edwards[1,]

[1]School of Biotechnology and Biomolecular Sciences, UNSW Sydney, High St, Kensington, NSW 2052, Australia

stephanie.h.chen@unsw.edu.au, richard.edwards@unsw.edu.au

[2]Research Centre for Ecosystem Resilience, Australian Institute of Botanical Science, The Royal Botanic Garden Sydney, Mrs Macquaries Rd, Sydney, NSW 2000, Australia

maurizio.rossetto@rbgsyd.nsw.gov.au, marlien.vandermerwe@rbgsyd.nsw.gov.au, patricia.lu-irving@bgcp.nsw.gov.au, samantha.yap@rbgsyd.nsw.gov.au, jason.bragg@rbgsyd.nsw.gov.au

[3]Queensland Alliance of Agriculture and Food Innovation, University of Queensland, St Lucia 4072, Australia

[4]National Herbarium of New South Wales, Royal Botanic Gardens and Domain Trust, Mrs Macquaries Rd, Sydney, NSW 2000, Australia

herve.sauquet@rbgsyd.nsw.gov.au

[5]School of Biological, Earth and Environmental Sciences, UNSW Sydney, High St, Kensington, NSW 2052, Australia

[6]Blue Mountains Botanic Garden, Bells Line of Road, Mount Tomah, NSW 2758, Australia

greg.bourke@bgcp.nsw.gov.au

Corresponding author

ORCID iD

SHC 0000-0001-8844-6864

MR 0000-0002-4878-9114

MVDM 0000-0003-1307-5143

PL-I 0000-0003-1116-9402

JSY 0000-0002-9141-6006

HS 0000-0001-8305-3236

JGB 0000-0002-7621-7295

RJE 0000-0002-3645-5539

ABSTRACT

Background

Telopea speciosissima, the New South Wales waratah, is Australian endemic woody shrub in the family Proteaceae. Waratahs have great potential as a model clade to better understand processes of speciation, introgression and adaptation, and are significant from a horticultural perspective.

Findings

Here, we report the first chromosome-level reference genome for *T. speciosissima*. Combining Oxford Nanopore long-reads, 10x Genomics Chromium linked-reads and Hi-C data, the assembly spans 823 Mb (scaffold N50 of 69.0 Mb) with 91.2 % of Embryophyta BUSCOs complete. We introduce a new method in Diploidocus (https://github.com/slimsuite/diploidocus) for classifying, curating and QC-filtering assembly scaffolds. We also present a new tool, DepthSizer (https://github.com/slimsuite/depthsizer) , for genome size estimation from the read depth of single copy orthologues and find that the assembly is 93.9 % of the estimated genome size. The largest 11 scaffolds contained 94.1 % of the assembly, conforming to the expected number of chromosomes (2*n* = 22). Genome annotation predicted 40,158 protein-coding genes, 351 rRNAs and 728 tRNAs. Our results indicate that the waratah genome is highly repetitive, with a repeat content of 62.3 %.

Conclusions

The *T. speciosissima* genome (Tspe_v1) will accelerate waratah evolutionary genomics and facilitate marker assisted approaches for breeding. Broadly, it represents an important new genomic resource of Proteaceae to support the conservation of flora in Australia and further afield.

Keywords: Telopea, waratah, genome assembly, reference genome, long-read sequencing, Hi-C

INTRODUCTION

Telopea R.Br. is an eastern Australian genus of five species of large, long-lived shrubs in the flowering plant family Proteaceae. The New South Wales waratah, *Telopea speciosissima* (Sm.) R.Br., is a striking and iconic member of the Australian flora, characterised by large, terminal inflorescences of red flowers (Figure 1) and has been the state floral emblem of New South Wales since 1962 and one of the first Australian plant species collected for cultivation in Europe [1]. The species is endemic to the state of New South Wales, occurring on sandstone ridges in the Sydney region. Previous studies have investigated variation among *Telopea* populations by phenetic analysis of morphology [2] and evolutionary relationships using cladistics [3]. Population structure and patterns of divergence and introgression between *T. speciosissima* populations have been characterised using several loci [4]. Further, microsatellite data and modelling suggest a history of allopatric speciation followed by secondary contact and hybridization among *Telopea* species [5]. These studies point to the great potential of *Telopea* as a model clade for understanding processes of divergence, environmental adaptation and speciation. Our understanding of these processes can be greatly enhanced by a genome-wide perspective, enabled by a reference genome [6–10].

Figure 1. New South Wales waratah (*Telopea speciosissima*). Photo taken by SH Chen.

Genome sequencing efforts have traditionally focused on model species, crops and their wild relatives, resulting in a highly uneven species distribution of reference genomes across the plant tree of life [11]. Despite Proteaceae occurring across several continents and encompassing 81 genera and ca. 1700 species [12,13], the only publicly available reference genome in the family is a widely-grown cultivar of the most economically important crop in the family, *Macadamia integrifolia* (macadamia nut) HAES 74 [14,15]. Waratahs are significant to the horticultural and cut flower industries, with blooms cultivated for the domestic and international markets; a reference genome will accelerate efforts in breeding for traits such as resistance to pests and diseases (e.g. *Phytophthora* root rot) as well as desirable floral characteristics [16]. More reference genomes in the Proteaceae family will also facilitate research into the molecular evolution of the group.

Here, we provide a high quality long-read *de novo* assembly of the *Telopea speciosissima* genome, using Oxford Nanopore long-reads, 10x Genomics Chromium linked-reads and Hi-C, which will serve as an important platform for evolutionary genomics and the conservation of the Australian flora.

DNA EXTRACTION AND SEQUENCING

Sampling and DNA extraction

Young leaves (approx. 8 g) were sampled from the reference genome individual (NCBI BioSample SAMN18238110) where it grows naturally along the Tomah Spur Fire Trail (-33.53° S, 150.42° E) on land belonging to the Blue Mountains Botanic Garden, Mount Tomah in New South Wales, Australia. Leaves were immediately frozen in liquid nitrogen and stored at -80° C prior to extraction.

High-molecular-weight (HMW) genomic DNA (gDNA) was obtained using a sorbitol pre-wash step prior to a CTAB extraction adapted from Inglis et al. [17]. The gDNA was then purified with AMPure XP beads (Beckman Coulter, Brea, CA, USA) using a protocol based on Schalamun et al. [18] (details available on protocols.io [19]). The quality of the DNA was assessed using Qubit, NanoDrop and TapeStation 2200 System (Agilent, Santa Clara, CA, USA).

ONT PromethION sequencing

We performed an in-house sequencing test on the MinION (MinION, RRID:SCR_017985) using a FLO-MINSP6 (R9.4.1) flow cell with a library prepared with the ligation kit (SQK-LSK109). The remaining purified genomic DNA was sent to the Australian Genome Research Facility (AGRF) where size selection was performed to remove small DNA fragments using the BluePippin High Pass

Plus Cassette on the BluePippin (Sage Science, Beverly, MA, USA). Briefly, 10 µg of DNA was split into 4 aliquots (2.5 µg) and diluted to 60 µL in TE buffer. Then, 20 µL of RT equilibrated loading buffer was added to each aliquot and mixed by pipetting. Samples were loaded on the cassette by removing 80 µL of buffer from each well and adding 80 µL of sample or external marker. The cassette was run with the 15 kb High Pass Plus Marker U1 cassette definition. Size selected fractions (approximately 80 µL) were collected from the elution module following a 30 min electrophoresis run. The library was prepared with the ligation sequencing kit (SQK-LSK109). The sequencing was performed using MinKNOW v.19.12.2 (MinION) and v12.12.8 (PromethION) and MinKNOW Core v3.6.7 (in-house test), v3.6.8 (AGRF MinION) and v3.6.7 (AGRF PromethION). A pilot run was first performed on the MinION using the FLO-MIN106 (R9.4.1) flow cell followed by two FLO-PRO002 flow cells (R9.4) on the PromethION (PromethION, RRID:SCR_017987).

Basecalling was performed after sequencing with GPU-enabled Guppy v3.4.4 using the high-accuracy flip-flop models, resulting in 54x coverage. The output from all ONT basecalling was pooled for adapter removal using Porechop (Porechop, RRID:SCR_016967) v.0.2.4 [20] and quality filtering (removal of reads less than 500 bp in length and Q lower than 7) with NanoFilt (NanoFilt, RRID:SCR_016966) v2.6.0 [21] followed by assessment using FastQC (FastQC, RRID:SCR_014583) v0.11.8 [22].

10x Genomics Chromium sequencing

High-molecular-weight gDNA was sent to AGRF for 10x Genomics Chromium sequencing. Size selection was performed to remove DNA fragments <40 kb using the BluePippin 0.75 % Agarose Gel Cassette, Dye Free on the BluePippin (Sage Science, Beverly, MA, USA). Briefly, 5 µg of DNA was diluted to 30 µL in TE buffer and 10 µL of RT equilibrated loading buffer was added to each aliquot

and mixed by pipetting. Samples were loaded on the cassette by removing 40 µL of buffer from each well and adding 40 µL of sample or external marker. The cassette was run with the 0.75 % DF Marker U1 high-pass 30-40 kb v3 cassette definition. Size selected fractions (approximately 40 µL) were collected following the 30 min electrophoresis run. The library was prepared using the Chromium Genome Library Kit & Gel Bead Kit and sequenced (2 x 150 bp paired-end) on the NovaSeq 6000 (Illumina NovaSeq 6000 Sequencing System, RRID:SCR_016387) with NovaSeq 6000 SP Reagent Kit (300 cycles) and NovaSeq XP 2-Lane Kit for individual lane loading.

Hi-C sequencing

Hi-C library preparation and sequencing was conducted at the Ramaciotti Centre for Genomics at the University of New South Wales using the Phase Genomics Plant kit v3.0. The library was assessed using Qubit and the Agilent 2200 TapeStation system (Agilent Technologies, Mulgrave, VIC, Australia). A pilot run on an Illumina iSeq 100 with 2 x 150 bp paired end sequencing run was performed for QC using hic_qc v1.0 [23] with i1 300 cycle chemistry. This was followed by sequencing on the Illumina NextSeq 500 (Illumina NextSeq 500, RRID:SCR_014983) with 2 x 150 bp paired-end high output run and NextSeq High Output 300 cycle kit v2.5 chemistry.

The ONT, 10x and Hi-C sequencing yielded a total of 48.3, 123.4 and 25.0 Gb of sequence, respectively (Table 1).

Table 1. Library information of *Telopea speciosissima* reference genome (Tspe_v1).

Sequencing platform	Library	Median insert size (bp)	Mean read length (bp)	No. of reads	Sequence bases (Gb)
Oxford Nanopore Technologies*	Ligation (SQK-LSK109)	-	13,449	3,595,148	48.3
Illumina NovaSeq 6000	Paired-end 10x Chromium	336	2 x 150	822,558,750	123.4
Total gDNA	-	-	-	**826,153,898**	**171.7**
Illumina NextSeq 500^	Phase Genomics Proximo Hi-C (Plant)	174	2 x 151	165,573,702	25.0

* Two PromethION flow cells and two partial flow cells from a MinION pilot run

^ Includes a pilot iSeq run used to QC the library

GENOME ASSEMBLY AND VALIDATION

Our assembly workflow consisted of assembling a draft long-read assembly, polishing the assembly with Illumina reads and scaffolding the assembly into chromosomes using Hi-C data (Figure 2).

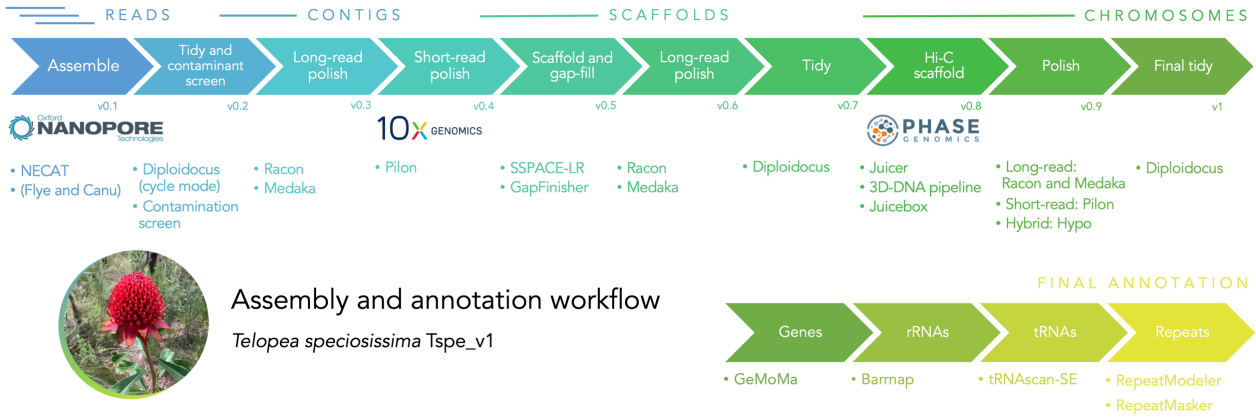


Figure 2. Assembly and annotation workflow for the *Telopea speciosissima* reference genome Tspe_v1. Logos reproduced with permission. Waratah photo by SH Chen.

172 **Draft long-read and 10x assemblies**

The first stage of our assembly approach involved comparing three long-read assemblers using the
174 ONT data as input: NECAT v0.01 [24], Flye (Flye, RRID:SCR_017016) v2.6 [25] and Canu (Canu, RRID:SCR_015880) v1.9 [26]. The genome size parameter used for the assemblers was 1134 Mb, as
176 previously reported for *Telopea truncata* [27]. We later refined genome size estimates for *T. speciosissima* (see 'DepthSizer: genome size estimation using single-copy orthologue sequencing
178 depths' section below).

180 The best draft genome assembly was assessed on three metrics: contiguity (N50), BUSCO completeness, and proximity to the estimated genome size. NECAT resulted in the most contiguous
182 assembly, at 365 contigs and the highest BUSCO completeness at 81.2 %. This was followed by Flye at 2,484 contigs and 81.0 % complete, then Canu at 3,983 contigs at 78.4 % complete.

184

As a comparison to the long-read assemblies, the 10x data were assembled with Supernova
186 (Supernova assembler, RRID:SCR_016756) v2.1.1 [28] with 332 Mb reads as input, aiming for 56x raw coverage. We generated pseudohaploid output of the assembly for comparison (pseudohap2
188 '.1' fasta). The BUSCO score was higher than each of the long-read assemblies at 91.8 %. However, the 10x assembly had much lower contiguity at 43,951 contigs, as expected (Table 2).

190

Assembly completeness and accuracy

192 Completeness was evaluated by BUSCO (BUSCO, RRID:SCR_015008) v3.0.2b [29], implementing BLAST+ v2.2.31 [30], Hmmer (Hmmer, RRID:SCR_005305) v3.2.1 [31], Augustus (Augustus,
194 RRID:SCR_008417) v3.3.2 [32] and EMBOSS (EMBOSS, RRID:SCR_008493) v6.6.0 [33]) against the embryophyta_odb9 dataset (n = 1,440; Table S1). BUSCO results were collated using BUSCOMP

(BUSCO Compilation and Comparison Tool; RRID:SCR_021233) v0.11.0 [34] to better evaluate the gains and losses in completeness between different assembly stages (Figure 3, Additional file 1). Notably, polishing markedly improved the BUSCO score – long-read polishing increased complete BUSCOs from 1,167 to 1,308 and short-read polishing further increased this to 1,333. We recovered a maximal non-redundant set of 1,386 complete single copy BUSCOs across the set of assemblies. Assembly quality (QV) was also estimated using k-mer analysis of trimmed and filtered 10x linked-read data by Merqury v1.0 with k = 20 [35]. First, 30 bp from the 5' end of read 1 and 10 bp from the 5' end of read 2 were trimmed using BBmap (BBmap, RRID:SCR_016965) v38.51 [36]. In addition, reads were trimmed to Q20, then those shorter than 100 bp were discarded.

216 Table 2. *Telopea speciosissima* assembly summaries.

Assembly	Supernova pseudohap2	v0.1 Canu	v0.1 Flye	v0.1 NECAT
Total length (bp)	**850,975,689**	**981,953,849**	**857,703,641**	**842,143,239**
No. of scaffolds	**27,610**	**3,983**	**2,445**	**365**
N50 (bp)	874,466	1,848,137	2,271,126	10,701,597
L50	247	132	94	24
No. of contigs	**43,951**	**3,983**	**2,484**	**365**
N50 (bp)	72,725	1,848,137	2,199,532	10,701,597
L50	3,268	132	101	24
N bases	18,076,790	0	3,900	0
GC (%)	40.1	39.95	40.47	40.15
BUSCO complete (genome; n = 1440)	**91.8 % (1,323)**	**78.4 % (1,129)**	**81.0 % (1,167)**	**81.2 % (1,169)**
Single copy	81.2 % (1,170)	68.6 % (988)	74.9 % (1,079)	75.2 % (1,083)
Duplicated	10.6 % (153)	9.8 % (141)	6.1 % (88)	6.0 % (86)
BUSCO fragmented	3.1 % (44)	5.2 % (75)	4.2 % (60)	4.3 % (62)
BUSCO missing	5.1 % (73)	16.4 % (236)	14.8 % (213)	14.5 % (209)
Merqury completeness (%)	**89.84**	**74.86**	**76.06**	**74.90**
Solid k-mers in the assembly	531,141,929	442,555,507	449,643,538	442,822,843
Total solid k-mers in read set	591,186,146	591,186,146	591,186,146	591,186,146
Merqury QV	**46.75**	**19.90**	**20.45**	**20.21**
k-mers unique to assembly	351,669	182,400,749	141,959,491	146,903,850
k-mers in both assembly and read-set	832,063,849	981,878,172	857,652,545	842,136,304
Error rate	0.000021	0.010223	0.009007	0.009539

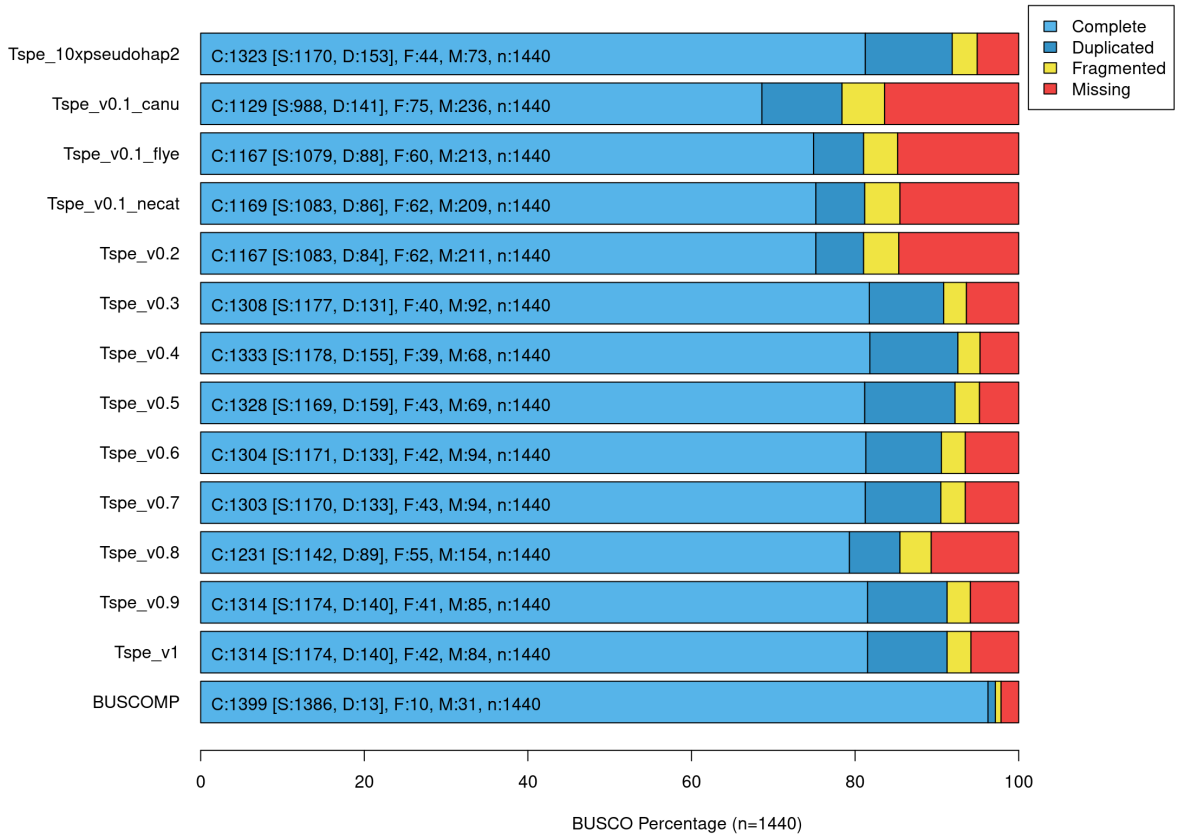


Figure 3. BUSCOMP summary of BUSCO completeness rating compiled over different stages (see Figure 2) of the *Telopea speciosissima* genome assembly. The final BUSCOMP rating uses the best rating per BUSCO gene across any of the assemblies.

DepthSizer: genome size estimation using single-copy orthologue sequencing depths

Telopea speciosissima has been reported as a diploid (2*n* = 22) [37,38]. We confirmed the individual's diploid status using Smudgeplot v0.2.1 [39] (Figure S1a). The 1C-value of *T. truncata* (Tasmanian waratah) has been estimated at 1.16 pg (1.13 Gb) using flow cytometry [27]. Supernova v2.1.1 predicted a genome size of 953 Mb from the assembly of the 10x linked-reads whilst GenomeScope (GenomeScope, RRID:SCR_017014) v1.0 [40] predicted a smaller genome of 794 Mb from the same data (Figure S1b).

We sought to refine the genome size estimate of *T. speciosissima* using the ONT data and draft genome assemblies, implementing a new tool, DepthSizer (https://github.com/slimsuite/depthsizer, RRID:SCR_021232). ONT reads were mapped onto each draft genome using Minimap2 (Minimap2, RRID:SCR_018550) v2.17 [41] (--secondary=no -ax map-ont). The single-copy read depth for each assembly was then calculated as the modal read depth across single copy complete BUSCO genes, which should be reasonably robust to poor-quality and/or repeat regions within these genes [42].

By definition, sequencing depth (X) is the volume of sequencing divided by the genome size. Given a known volume of sequencing, it is therefore possible to estimate the genome size by estimating the achieved sequencing depth. DepthSizer works on the principle that the modal read depth across single copy BUSCO genes provides a good estimate of the true depth of coverage. This assumes that genuine single copy depth regions will tend towards the same, true, single copy read depth. In contrast, assembly errors or collapsed repeats within those genes, or incorrectly-assigned single copy genes, will give inconsistent read depth deviations from the true single copy depth. (The exception is regions of the genome only found on one haplotig – half-depth alternative haplotypes for regions also found in the main assembly – such as heterogametic sex chromosomes [42], but these are unlikely to outnumber genes present in single copy on both homologous chromosomes.) As a consequence, the dominant (i.e. modal) depth across these regions should represent single copy sequencing depth. First, the distribution of read depth for all single copy genes is generated using Samtools (Samtools, RRID:SCR_002105) v0.11 [43] mpileup, and the modal peak calculated using the 'density' function of R (R Project for Statistical Computing, RRID:SCR_001905) v3.5.3 [44] (allowing a non-integer estimation). Genome size, *G*, was then estimated from the modal peak single-copy depth, X_{sc}, and the total volume of sequencing data, *T*, using the formula:

$G = T / X_{SC}$

This estimate does not account for any non-nuclear (or contamination) read data, nor any biases/inconsistences in read mapping and/or raw read insertion/deletion error profiles. As a consequence, this will tend to be an overestimate. We also calculated a second genome size estimate, adjusting for read mapping and imbalanced insertion:deletion ratios. Here, samtools coverage was used to estimate the total number of bases mapped onto the assembly (assembly bases with coverage x average depth) and Samtools fasta to extract all the mapped reads. The ratio of the mapped read bases, *M*, to the summed length of mapped reads, *L*, is then calculated and used to adjust *T*:

$T_{adj} = T.M / L$

$G_{adj} = T_{adj} / X_{SC}$

DepthSizer also outputs genome size predictions based on the integer modal read depth across single-copy complete BUSCO genes, and the mode of modal read depths across single-copy complete BUSCO genes.

We used genome size estimates to assess long-read assembly completeness to guide decisions in the first stage of the assembly progress. DepthSizer analysis of the three draft genome assemblies estimated the genome size of *T. speciosissima* to range from 806 Mb (Flye density mode with mpileup adjusted) to 926 Mb (Canu mode of modes with mpileup; Table S2), which falls between the Supernova and GenomeScope estimates. The mean estimate across the three genomes with six methods each (mode method: BUSCO mode/mode of modes/density mode x depth method:

mpileup/mpileup-adjusted) was 874 Mb. The mpileup-adjusted depth method resulted in slightly smaller estimated genome sizes compared to the non-adjusted depth method, indicating a slight bias towards insertion versus deletion errors in the raw ONT reads. We report an estimated genome size of 876.4 Mb, based on Tspe_v1 using the density and mpileup-adjusted parameters, which is theoretically the most robust method.

Assembly tidying and contamination screening

According to the three chosen metrics, we moved forward with the NECAT assembly (Table S1). Whilst Supernova had a higher BUSCO completeness (91.8% versus 81.2%), NECAT was orders of magnitude better in terms of contiguity (10.7 Mb N50 on 365 contigs vs 874 kb N50 on 27,610 scaffolds). The draft genome was screened and filtered to remove contamination, low-quality contigs and putative haplotigs using more rigorous refinement of the approach taken for the Canfam_GSD (German Shepherd) and CanFam_Bas (Basenji) dog reference genomes [42,45], implemented in Diploidocus v0.9.6 (https://github.com/slimsuite/diploidocus, RRID:SCR_021231).

First, the assembly was screened against the NCBI UniVec database (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/, downloaded 05/08/2019) to identify and remove contaminants. Hits are first scored using rules derived from NCBI Vecscreen (https://www.ncbi.nlm.nih.gov/tools/vecscreen/) and regions marked as 'Terminal' (within 25 bp of a sequence end), 'Proximal' (within 25 bases of another match) or 'Internal' (>25 bp from sequence end or vecsreen match). Then, any segment of fewer than 50 bases between two vector matches or between a match and a sequence end are marked as 'Suspect'. In our experience, default Vecscreen parameters appear prone to excessive false positives in large genomes (data not shown), and so Diploidocus features two additional contaminant identification filters. First, the 'Expected

False Discovery Rate' (eFDR) is calculated for each contaminant sequence. This is simply the BLAST+ Expect value for that hit, divided by the total number of hits at that Expect value threshold. Any hits with an eFDR value exceeding the default threshold of 1.0 were filtered from the vecscreen results. Short matches in long-read assemblies are unlikely to be real contamination and a second filter was applied, restricting contaminant screening to a minimum hit length of 50 bp. Finally, the percentage coverage per scaffold is calculated from the filtered hits. This is performed first for each contaminant individually, before being collapsed into total non-redundant contamination coverage per query. Diploidocus then removes any scaffolds with at least 50% contamination, trims off any vector hits within 1 kb of the scaffold end, and masks any remaining vector contamination of at least 900 bp. This masking replaces every other base with an N to avoid an assembly gap being inserted: masked regions should be manually fragmented if required. Diploidocus can also report the number of mapped long reads that completely span regions flagged as contamination.

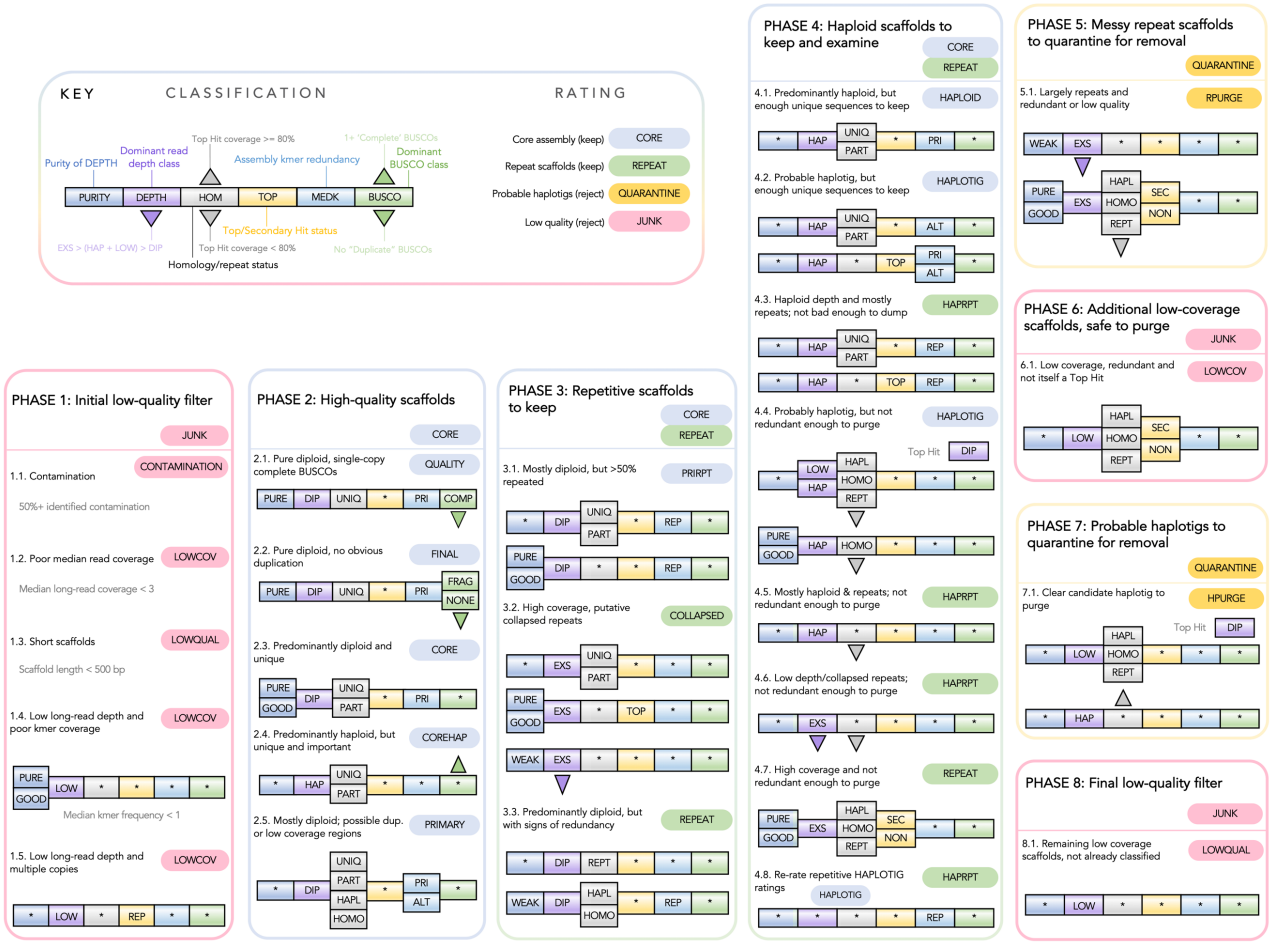
After contamination screening, a sorted BAM file of ONT reads mapped to the filtered assembly is generated using Minimap2 v2.17 (–ax map-ont --secondary = no) [41]. BUSCO Complete genes (see above) were used to estimate a single-copy read depth of 54X. This was used to set low-, mid- and high-depth thresholds for Purge Haplotigs (Purge_haplotigs, RRID:SCR_017616) v20190612 [46] (implementing Perl v5.28.0, BEDTools (BEDTools, RRID:SCR_006646) v2.27.1 [47], R v3.5.3 [44], and SAMTools v1.9 [43]) of 13X, 40X and 108X. Purge Haplotigs coverage bins were adjusted to incorporate zero-coverage bases, excluding assembly gaps (defined as 10+ Ns). Counts of Complete, Duplicate and Fragmented BUSCO genes were also generated for each sequence. General read depth statistics for each sequence were calculated with BBMap v38.51 pileup.sh [36]. The sect function of KAT (KAT, RRID:SCR_016741) v2.4.2 [48] was used to calculate k-mer frequencies for the 10x linked reads (first 16 bp trimmed from read 1), and the assembly itself. Telomeres were

predicted using a method adapted from https://github.com/JanaSperschneider/FindTelomeres, searching each sequence for 5' occurrences of a forward telomere regular expression sequence, C{2,4}T{1,2}A{1,3}, and 3' occurrences of a reverse regular expression, T{1,3}A{1,2}G{2,4}. Telomeres were marked if at least 50% of the terminal 50 bp matches the appropriate sequence.

Diploidocus combines read depth, KAT k-mer frequencies, Purge Haplotigs depth bins, Purge Haplotigs best sequence hits, BUSCO gene predictions, telomere prediction and vector contamination into a single seven-part (PURITY|DEPTH|HOM|TOP|MEDK|BUSCO+EXTRA) classification (Table 3). Diploidocus then performs a hierarchical rating of scaffolds, based on their classifications and compiled data (Table 4, Figure 4). Based on these ratings, sequences are divided into sets (Table 4):

1. Core. Predominantly diploid scaffolds and unique haploid scaffolds with insufficient evidence for removal.
2. Repeats. Unique haploid scaffolds with insufficient evidence for removal but dominated by repetitive sequences. High coverage scaffolds representing putative collapsed repeats.
3. Quarantine. Messy repetitive sequences and strong candidates for alternative haplotigs.
4. Junk. Low coverage, short and/or high-contaminated sequences.

If any sequences are marked as 'Quarantine' or 'Junk', sequences in the 'Core' and 'Repeat' sets are retained and used as input for another round of classification and filtering. Convergence was reached after three cycles with 148 core sequences and 62 repeat sequences retained (see Table S3 for summary of cycles and Table S4 for full output).



350 Figure 4. Diploidocus scaffold rating process based on a six-part classification. Asterisks indicate any
class value is accepted. Phases are executed in order. Consequently, rules for later phases appear
352 less restrictive than the full set of criteria required to receive that rating. See main text, Table 3 and
Table 4 for details of the six-part classification and final ratings.

354

356

358

360

Table 3. Diploidocus sequence classification.

Criterion	Description
PURITY	Purity of dominant read depth class • PURE = At least 80% of sequence in that depth bin • GOOD = At least 50% of sequence in that depth bin • WEAK = Under 50% of sequence in that depth bin
DEPTH	Dominant read depth class based on BBMap and Purge Haplotigs (PH) • LOWX = Median read depth below 3 • LOW = PH Low read depth bin has highest percentage coverage (ties assigned to other class) • HAP = PH Hap read depth bin has highest percentage coverage (non-DIP ties assigned to HAP) • DIP = PH Dip read depth bin has highest percentage coverage (ties assigned to DIP) • EXS = PH High read depth bin has highest percentage coverage
HOM	Homology/repeat status based on Purge Haplotigs Top and Secondary hits • UNIQ = No Top Hit • PART = Partial (<50%) coverage of Top Hit • HAPL = 50%+ Top Hit coverage but no Secondary Hit • HOMO = Top Hit and Secondary Hit but combined coverage < 250% • REPT = Top Hit and Secondary Hit and 250%+ combined coverage
TOP	Top/Secondary Hit status for sequence • TOP = Sequence is a Top Hit for at least one other sequence • SEC = Sequence is a Secondary Hit but not a Top Hit for at least one other sequence • NON = Neither a Top Hit nor a Secondary Hit for any other sequence
MEDK	Assembly redundancy based on KAT assembly kmers • PRI = Over 50% unique kmers (KAT median assembly kmer frequency = 1) • ALT = KAT median assembly kmer frequency of two • REP = KAT median assembly kmer frequency exceeds two
BUSCO	Dominant BUSCO class • COMP = 1+ Complete BUSCO genes and more Complete than Duplicated • DUPL = 1+ Duplicated BUSCO genes and more Duplicated than Complete • FRAG = 1+ Fragmented BUSCO genes and no Complete or Duplicated • NONE = No Complete, Duplicated or Fragmented BUSCO genes
EXTRA	+TEL: If any telomeres are detected, +TEL is added +VEC: If any contamination is detected, +VEC is added

362

Table 4. Diploidocus sequence rating.

Rating	Description	Set
COLLAPSED	High coverage scaffolds representing putative collapsed repeats	Repeat
CONTAMINATION	50%+ identified contamination.	Junk
CORE	Predominantly non-repetitive, diploid depth sequences with <50% covered by Purge Haplotigs Top Hit.	Core
COREHAP	Predominantly haploid read depth but less than 50% covered by Purge Haplotigs Top Hit and at least 1 Complete BUSCO. Probable haploid-depth region of genome.	Core
FINAL	High quality scaffolds with dominant diploid depth	Core
HAPLOID	Predominantly haploid coverage but enough unique sequence to keep - might represent very heterozygous alternative haplotigs.	Core
HAPLOTIG	Predominantly haploid coverage but enough unique sequence to keep - possible alternative haplotig. Or low/haploid coverage scaffold with insufficient coverage of another scaffold to purge.	Core*
HAPRPT	As HAPLOTIG but with evidence for dominant repetitive sequences (high kmer frequencies and/or read depth regions).	Repeat
HPURGE	Clear candidate haplotig to purge.	Quarantine
LOWCOV	Very low read depth; low read depth with additional kmer signatures of poor raw data coverage; low read depth and assembly kmer signature of repetitive sequence	Junk
LOWQUAL	Short scaffolds failing to meet minimum length criterion.	Junk
PRIMARY	Putative primary scaffold but with possible alternative scaffolds still in assembly and/or low-quality regions	Core
PRIRPT	Putative primary scaffold but >50% repeated	Core
QUALITY	Highest quality scaffolds: diploid depth with Complete BUSCOs and no Duplicated BUSCOs.	Core
REPEAT	Predominantly Diploid scaffolds that have major signs of redundancy, probably due to presence of alternative contigs	Repeat
RPURGE	Messy scaffolds that are largely repeats and are sufficiently redundant/low quality to purge	Quarantine

* Sequences rated HAPLOTIG should be reviewed for possible manual exclusion.

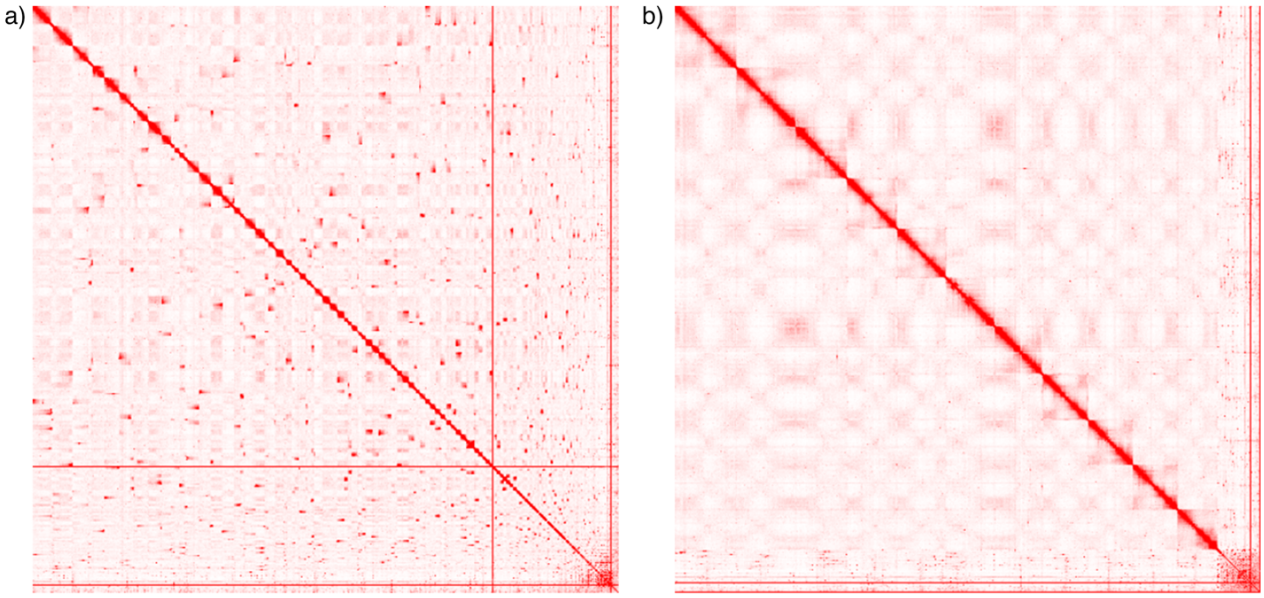
368 **Assembly polishing and gap-filling**

The assembly was first long-read polished with Racon (Racon, RRID:SCR_017642) v1.4.5 [49] and
370 medaka v1.0.2 [50]. Then, the 10x reads were incorporated by short-read polishing using Pilon (Pilon, RRID:SCR_014731) v1.23 [51] with reads mapped using Minimap2 v2.12 [41] and correcting
372 for indels only; we found correcting for indels only resulted in a higher BUSCO score than correcting for indels and SNPs following the steps described in this section. We scaffolded using SSPACE-
374 LongRead v1.1 [52] followed by gap-filling using gapFinisher v20190917 [53]. The assembly was scaffolded from 209 contigs into 138 scaffolds, however, no gaps were filled. After another round
376 of long-read polishing with Racon v1.4.5 [49] and medaka v1.0.2 [50], we moved forward with a second round of tidying in Diploidocus v0.9.6 (default mode). Here, 128 scaffolds out of the 138
378 were retained and consisted of 87 core, 41 repeat, 10 quarantine and 0 junk scaffolds.

380 **Hi-C scaffolding**

Hi-C data were aligned to the draft genome assembly using the Juicer (Juicer, RRID:SCR_017226)
382 pipeline v1.6 [54] then scaffolds were ordered and orientated using the 3D *de novo* assembly pipeline (3D de novo assembly, RRID:SCR_017227) v180922 [55]. The contact map was visualised
384 using Juicebox Assembly Tools v1.11.08 [56] and errors over 3 review rounds were corrected manually to resolve 11 chromosomes (Figure 5). Although the assembly was in 2,357 scaffolds
386 following incorporation of Hi-C data; the N50 increased by over 4-fold to 68.9 Mb. Surprisingly, the contig number increased considerably from 148 to 3,537, suggesting that the Hi-C data and NECAT
388 assembly were in conflict, with possibilities of incorrectly joined sequences in the initial long-read assembly or the Hi-C data causing the draft assembly to split into an unnecessarily large number of
390 fragments. The resulting assembly was tidied again using Diploidocus v10.0.6 (default mode) and 1643 scaffolds (824,534,974 bp) were retained out of 2,357 (833,952,765 bp; 1,347 core, 296

392 repeat, 548 quarantine and 166 junk scaffolds). The fact that Diploidocus removed a high percentage of sequences, together with the assembly statistics from the widely-used long-read
394 assemblers Canu and Flye (Table S1), suggests that NECAT is the cause of the unexpected jump in contig number following Hi-C scaffolding. However, the quality of the Hi-C library was not optimal
396 to start with, so this may also have contributed to the high degree of fragmentation.



398

Figure 5. Hi-C contact matrices visualised in Juicebox.js in balanced normalisation mode a) before
400 and b) after correction.

402 **Final polishing and assembly clean-up**

A further round of long-read polishing with Racon v1.4.5 [49] and medaka v1.0.2 [50] was
404 performed as described above. The assembly contiguity improved and there were 1,399 scaffolds and 1,595 contigs. We then short-read polished using Pilon v1.23 [51]. Two Pilon strategies were
406 applied: (1) indel-only correction; (2) indel and SNP correction. We retained the indel and SNP corrected assembly as it resulted in a marginally higher BUSCO score compared to indel only
408 correction (1311 vs 1310 complete BUSCOs); there was no change to contig nor scaffold numbers. A

final hybrid polish was performed using Hypo v1.0.3 [57]. The number of scaffolds remained as 1,399, however, the BUSCO score improved slightly by 0.1 % to 91.2 % complete. Notably, Hypo polishing improved the Merqury QV score from 29.8 to 33.9. The assembly was concluded with a final tidy with Diploidocus v0.14.1 (default mode). For the final assembly 1,289 scaffolds were retained from the 1,399 scaffolds (1,084 core, 205 repeat, 72 quarantine and 38 junk). All gaps in the assembly were then standardised to 100 bp.

Tspe_v1 reference genome

Assembly of 48.3 Gb of Nanopore long-read data and 123.4 Gb of Illumina short-read data (10x linked-reads) followed by scaffolding with Hi-C data produced a 823.3 Mb haploid genome, representing 93.9 % of the DepthSizer estimated genome size. The final assembly contained 1,289 scaffolds with an N50 of 69.0 Mb and L50 of 6 (Table 5). The Hi-C data facilitated scaffolding into 11 chromosomes, conforming to previous cytological studies [37], and the anchored proportion of Tspe_v1 spanned 94.2 % of the final assembly; the chromosomes were numbered by descending length (Table S5) as this is the first instance *Telopea* chromosomes have been studied in detail.

From a core set of 1,440 single-copy orthologues from the Embryophyta lineage, 91.4 % were complete in the assembly (81.8 % as single-copy, 9.5 % as duplicates), 2.7% were fragmented and only 5.9 % were not found, suggesting that the assembly includes most of the waratah gene space. The Tspe_v1 assembly is comparable in completeness to the *Macadamia integrifolia* (SCU_Mint_v3) assembly [14], which also combined long-read and Illumina sequences (90.6 % vs 80.0 % complete BUSCOs, respectively, in the anchored portion of the assembly). BUSCOMP analysis revealed that only 2.2% genes were not found by BUSCO in any version of the assembly.

The Merqury v1.0 [35] QV score of the assembly was 34.03, indicating a base-level accuracy of >99.99 % (Figure S2).

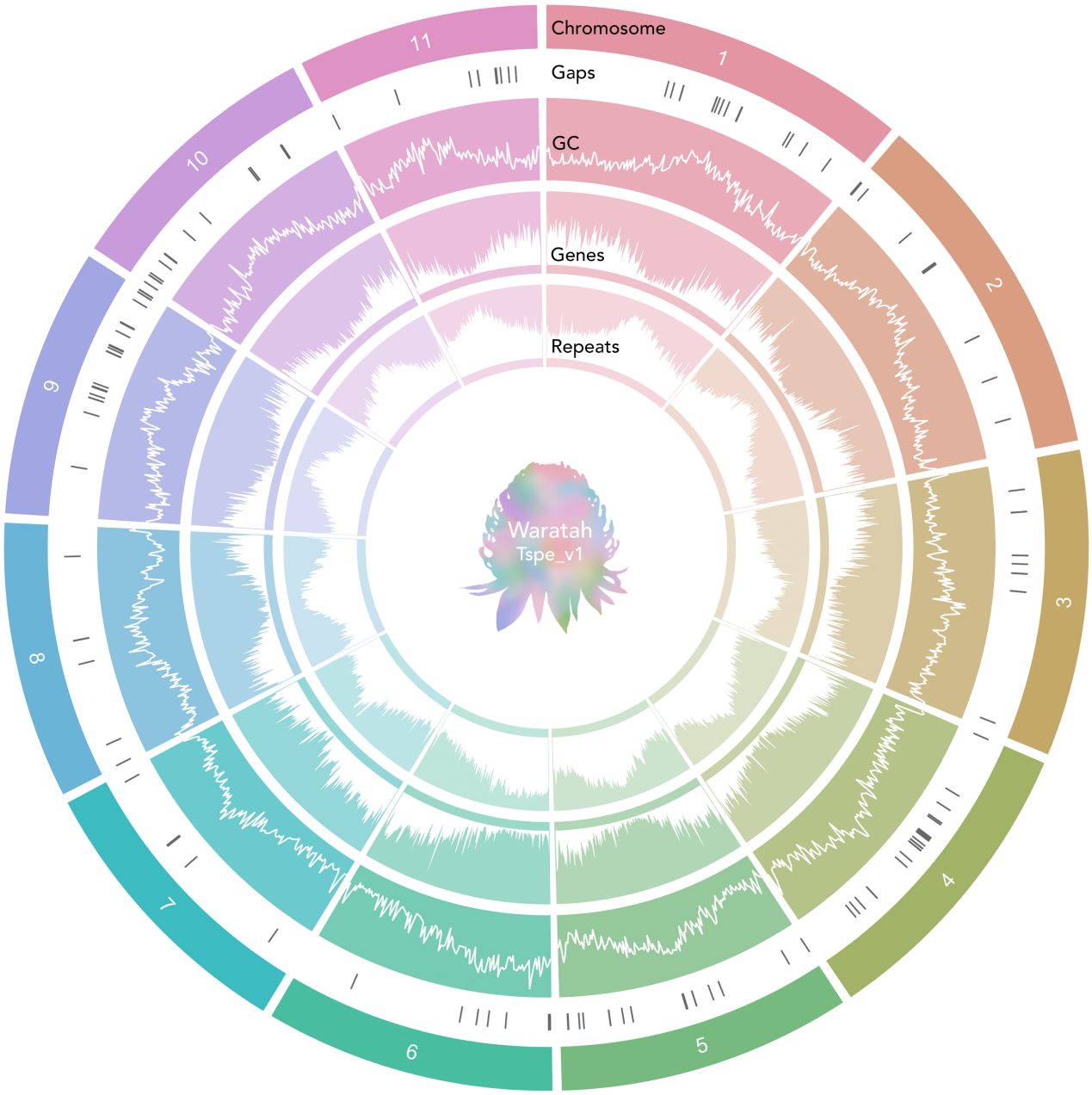
Genome visualisation

Features (gaps, GC content, gene density and repeat density) of the main nuclear chromosome scaffolds of the Tspe_v1 assembly were visualised as a circular diagram using the R package circlize v0.4.12 [58] (Figure 6). GC content was calculated in sliding windows of 50 kb using BEDTools v2.27.1 [47]. There were 147 gaps of unknown length across the 11 chromosomes, represented as 100 bp gaps in the assembly. An inverse pattern in the incidence of genes and repeats was observed across all chromosomes, with repeat content generally peaking towards the centre of each chromosome, suggesting predominantly metacentric and submetacentric chromosomes. This pattern may represent enriched repeat content and reduced coding content in pericentromeric regions, although further study is required to identify the centromeres [59–61].

454 Table 5. Genome assembly and annotation statistics for the *Telopea speciosissima* reference genome.

Statistic	**Tspe_v1**
Total length (bp)	**823,061,212**
No. of scaffolds	**1,289**
N50 (bp)	69,013,595
L50	6
No. of contigs	**1,452**
N50 (bp)	12,206,888
L50	21
N bases	18,174
GC (%)	40.11
BUSCO complete (genome; n = 1440)	**91.2 % (1314)**
Single copy (genome)	81.5 % (1174)
Duplicated (genome)	9.7 % (140)
BUSCO fragmented (genome)	2.9 % (42)
BUSCO missing (genome)	5.9 % (84)
Protein-coding genes	**40,158**
mRNAs	46,877
rRNAs	351
tRNAs	728
BUSCO complete (proteome; n = 1440)	**94.0 % (1353)**
Single copy (proteome)	79.3 % (1143)
Duplicated (proteome)	14.7 % (211)
BUSCO fragmented (proteome)	3.4 % (49)
BUSCO missing (proteome)	2.6 % (38)

456



458 Figure 6. Features of the 11 chromosomes of the *Telopea speciosissima* reference genome depicted as a circlize diagram. Concentric tracks from the outside inward represent: chromosomes, gaps
460 (gaps of unknown length appear as 100 bp in the assembly), GC content, gene density and repeat density. The latter three tracks denote values in 500 kb sliding windows. Density was defined as the
462 fraction of a genomic window that is covered by genomic regions. Plots are white on a solid background coloured by chromosome.

GENOME ANNOTATION

Heterozygosity and repetitive elements

Genome-wide heterozygosity was estimated to be 0.756 % using trimmed 10x reads with GenomeScope [40] from the k-mer 20 histogram computed using Jellyfish (Jellyfish, RRID:SCR_005491) v2.2.10 [62] (Figure S1b).

We identified and quantified repeats using RepeatModeler (RepeatModeler, RRID:SCR_015027) v2.0.1 and RepeatMasker (RepeatMasker, RRID:SCR_012954) v4.1.0 [63] and showed that the *T. speciossisima* genome is highly repetitive, with repeats accounting for 62.3 % of sequences (Table S6). Class I transposable elements (TEs) or retrotransposons were the most pervasive classified repeat class (20.3 % of the genome), and were dominated by long terminal repeat (LTR) retrotransposons (18.1 %). Class II TEs (DNA transposons) only accounted for 0.03 % of the genome.

Gene prediction

The genome was annotated using the homology-based gene prediction program GeMoMa (GeMoMa, RRID:SCR_017646) v1.7.1 [64] with four reference genomes downloaded from NCBI: *Macadamia integrifolia* (SCU_Mint_v3, GCA_013358625.1), *Nelumbo nucifera* (Chinese Lotus 1.1, GCA_000365185.2), *Arabidopsis thaliana* (TAIR10.1, GCA_000001735.2) and *Rosa chinensis* (RchiOBHm-V2, GCA_002994745.2). The annotation files for *M. integrifolia* were downloaded from the Southern Cross University data repository (doi.org/10.25918/5e320fd1e5f06).

Genome annotation predicted 40,126 protein-coding genes and 46,842 mRNAs in the *T. speciosissima* assembly, which fits the expectation for plant genomes [65]. Of these genes, 40,158

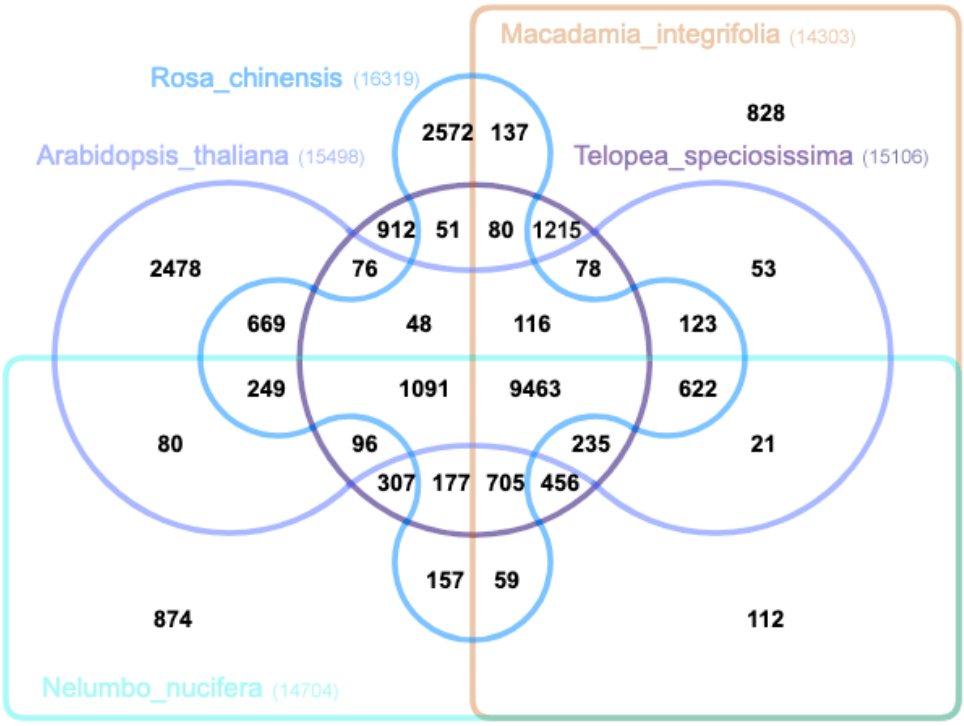
appeared in the 11 chromosomes (Table S5). Of 1,440 Embryophyta orthologous proteins, 94.0 % were complete in the annotation (79.3 % as single-copy, 14.7 % as duplicates), 3.4 % were fragmented and 2.6 % were missing.

Additionally, 351 ribosomal RNA genes were predicted with Barrnap (Barrnap, RRID:SCR_015995) v0.9 [66] and a set of 728 high-confidence transfer RNAs (tRNAs) was predicted with tRNAscan-SE (tRNAscan-SE, RRID:SCR_010835) v2.05 [67], implementing Infernal (Infernal, RRID:SCR_011809) v1.1.2 [68]. A set of 2,419 tRNAs was initially predicted and filtered to 760 using the recommended protocol for eukaryotes. Then, 22 tRNAs with mismatched isotype and 10 with unexpected anticodon were removed to form the high-confidence set.

Orthologous clusters

The protein sequences of Tspe_v1 and the four species used in the GeMoMa annotation were clustered into orthologous groups using OrthoVenn2 [69]. The five species formed 24,140 clusters: 23,031 orthologous clusters (containing at least 2 species) and 1,109 single-copy gene clusters. There were 9,463 orthologous families common to all of the species. The three members of the order Proteales (*T. speciosissima*, *M. integrifolia* and *N. nucifera*) shared 456 families (Figure 7 and Figure S3).

Tests for gene ontology (GO) enrichment of 912 waratah-specific clusters identified 12 significant terms (Table S7). The most enriched GO terms were DNA recombination (GO:0006310, $P = 1.8 \times 10^{-27}$), retrotransposon nucleocapsid (GO:0000943, $P = 3.5 \times 10^{-12}$) and DNA integration (GO:0015074, $P = 4.1 \times 10^{-11}$).



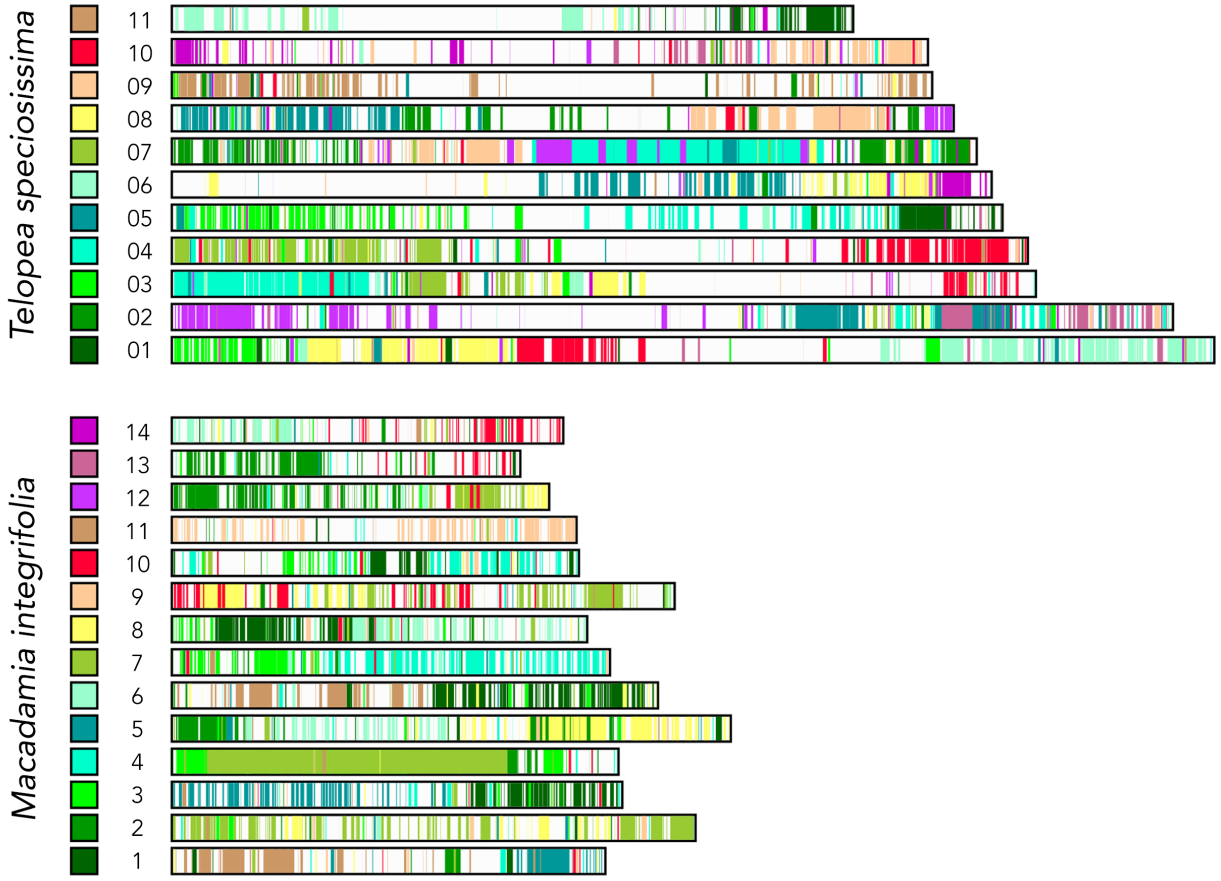
512 Figure 7. Orthologous gene clusters shared among the three members of the order Proteales – *Telopea speciosissima*, *Macadamia integrifolia* and *Nelumbo nucifera* – and the core eudicots – 514 *Arabidopsis thaliana* (Brassicales) and *Rosa chinensis* (Rosales).

516 **Synteny between *Telopea* and *Macadamia***

Synteny between the *Telopea* (Tspe_v1) and *Macadamia* (SCU_Mint_v3) genomes was explored 518 with satsuma2 version untagged-2c08e401140c1ed03e0f with parameters -l 3000 -do_refine 1 -min_matches 40 -cutoff 2 -min_seed_length 48 and visualised with the ChromosomePaint function 520 [70] and MizBee v1.0 [71]. The *Macadamia* genome (2*n* = 28) has six more chromosomes than the *Telopea* genome (2*n* = 22), but the two species have similar estimated genome sizes – 896 Mb [14] 522 compared to 874 Mb. It is thought that the ancestral Proteaceae had a chromosome number of *x* = 7 [72–75], although the occurrence of paleo-polyploidy in family has been debated [76]. Overall, 524 synteny analyses reveal an abundance of interchromosomal rearrangements between the *Telopea* and *Macadamia* genomes, reflecting the long time since their divergence (73-83 Ma [77]).

526 However, a number of regions exhibit substantial collinearity, for example, *Telopea* chromosome 09 and *Macadamia* chromosome 11 (Figure 8 and Figure S4).

528



530 Figure 8. Synteny between *Telopea speciosissima* (2*n* = 22) and *Macadamia integrifolia* (2*n* = 28).

CONCLUSIONS

532

We present a high-quality annotated chromosome-level reference genome of *Telopea* 534 *speciosissima* assembled from Oxford Nanopore long-reads, 10x Genomics Chromium linked-reads and Hi-C (823 Mb in length, N50 of 69.9 Mb and BUSCO completeness of 91.2 %): the first for a 536 waratah, and only the second publicly available Proteaceae reference genome. We envisage these

data will be a platform to underpin evolutionary genomics, gene discovery, breeding and the conservation of Proteaceae and the Australian flora.

DATA AVAILABILITY

The Tspe_v1 genome was deposited to NCBI under BioProject PRJNA712988 and BioSample SAMN18238110 along with the raw data (ONT, 10x and Hi-C) to SRA as SRR14018636, SRR14018635 and SRR14018634. Supporting data are available in the GigaScience database (GigaDB, RRID:SCR_004002) [TBC].

Data for species used for genome annotation are available at the following repositories:

Macadamia integrifolia

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/013/358/625/GCA_013358625.1_SCU_Mint_v3/

doi.org/10.25918/5e320fd1e5f06

Arabidopsis thaliana

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/735/GCF_000001735.4_TAIR10.1/

Rosa chinensis

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/002/994/745/GCA_002994745.2_RchiOBHm-V2/

Nelumbo nucifera

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/365/185/GCF_000365185.1_Chinese_Lotus_1.1

LIST OF ABBREVIATIONS

BLAST: Basic Local Alignment Search Tool

bp: base pairs

BUSCO: Benchmarking Universal Single-Copy Orthologs

CTAB: cetyl trimethylammonium bromide

Gb: gigabase pairs

GC: guanine-cytosine

Hi-C: high-throughput chromosome conformation capture

HMW: high molecular weight

kb: kilobase pairs

LINE: long interspersed nuclear element

LTR: long terminal repeat

Mb: megabase pairs

mRNA: messenger RNA

NCBI: National Centre for Biotechnology Information

ONT: Oxford Nanopore Technologies

PE: paired-end

QV: Merqury consensus quality value

rRNA: ribosomal RNA

SINE: short interspersed nuclear element

SNP: single-nucleotide polymorphism

TE: transposable element

CONSENT FOR PUBLICATION

Not applicable.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

We would like to acknowledge the contribution of the Genomics for Australian Plants Framework Initiative consortium (https://www.genomicsforaustralianplants.com/consortium/) in the generation of data used in this publication. The Initiative is supported by funding from Bioplatforms Australia (enabled by NCRIS), the Ian Potter Foundation, Royal Botanic Gardens Foundation (Victoria), Royal Botanic Gardens Victoria, the Royal Botanic Gardens and Domain Trust, the Council of Heads of Australasian Herbaria, CSIRO, Centre for Australian National Biodiversity Research and the Department of Biodiversity, Conservation and Attractions, Western Australia. SHC was supported through an Australian Government Research Training Program Scholarship. RJE was funded by the Australian Research Council (LP160100610 and LP18010072).

AUTHORS' CONTRIBUTIONS

JGB coordinated the project. MR, MvdM, PL-I, HS, GB, JGB and RJE designed the study and funded the project. GB provided the samples. PL-I and J-YSY performed optimised DNA extraction protocols and performed extractions. SHC performed the genome assembly, scaffolding and annotation. RJE conceptualised and developed Diploidocus and DepthSizer. SHC wrote the manuscript. All authors edited and approved the final manuscript.

ACKNOWLEDGEMENTS

We thank Stuart Allan for providing access to the sequenced plant and assistance with sample collection at Blue Mountains Botanic Garden. We thank the members of UNSW Research Technology Services, particularly Duncan Smith, for help with software installation on the high-performance computing cluster Katana. We acknowledge Mabel Lum for assistance with the Bioplatforms Australia data portal. ONT and 10x sequencing were conducted at the Australian Genome Research Facility (AGRF). Hi-C library prep and sequencing was conducted at the Ramaciotti Centre for Genomics at the University of New South Wales.

REFERENCES

1. Nixon P. The Waratah. Kenthurst, NSW: Kangaroo Press; 1987.

2. Crisp MD, Weston PH. Geographic and ontogenetic variation in morphology of Australian waratahs (*Telopea*: Proteaceae). *Syst Biol*. 1993; doi: 10.2307/2992556.

3. Weston PH, Crisp MD. Cladistic biogeography of waratahs (Proteaceae, Embothrieae) and their allies across the pacific. *Aust Syst Bot*. 1994; doi: 10.1071/sb9940225.

4. Rossetto M, Thurlby KA, Offord CA, Allen CB, Weston PH. The impact of distance and a shifting temperature gradient on genetic connectivity across a heterogeneous landscape. *BMC Evol Biol*. 2011; doi: 10.1186/1471-2148-11-126.

5. Rossetto M, Allen CB, Thurlby KAG, Weston PH, Milner ML. Genetic structure and bio-climatic modeling support allopatric over parapatric speciation along a latitudinal gradient. *BMC Evol Biol*. 2012; doi: 10.1186/1471-2148-12-149.

6. Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, et al.. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*. Nature Publishing Group; 2012; doi: 10.1038/nature11584.

7. Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, et al.. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am Nat*. 2016; doi: 10.1086/688018.

8. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al.. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci*. 2018; doi: 10.1073/pnas.1720115115.

9. Radwan J, Babik W. The genomics of adaptation. *Proc R Soc B Biol Sci*. 2012; doi: 10.1098/rspb.2012.2322.

10. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al.. Genomics and the origin of species. *Nat Rev Genet*. Nature Publishing Group; 2014; doi: 10.1038/nrg3644.

11. Royal Botanic Gardens, Kew. State of the World's Plants 2017. Royal Botanic Gardens, Kew; 2017. Report No.: 978-1-84246-647–6.

12. Mast AR, Willis CL, Jones EH, Downs KM, Weston PH. A smaller *Macadamia* from a more vagile tribe: inference of phylogenetic relationships, divergence times, and diaspore evolution in *Macadamia* and relatives (tribe Macadamieae; Proteaceae). *Am J Bot*. Wiley Online Library; 2008; 95:843–70.

13. Weston PH. Proteaceae. In: Kubitzki K, editor. The Families and Genera of Vascular Plants. Volume IX. Berlin: Springer-Verlag; 2006. p. 364–404.

14. Nock CJ, Baten A, Mauleon R, Langdon KS, Topp B, Hardner C, et al.. Chromosome-scale assembly and annotation of the macadamia genome (*Macadamia integrifolia* HAES 741). *G3 Genes Genomes Genet*. 2020; doi: 10.1534/g3.120.401326.

15. Nock CJ, Baten A, Barkla BJ, Furtado A, Henry RJ, King GJ. Genome and transcriptome sequencing characterises the gene space of *Macadamia integrifolia* (Proteaceae). *BMC Genomics*. 2016; doi: 10.1186/s12864-016-3272-3.

16. Offord CA. Analysis of characters and germplasm of significance to improvement of Australian native waratahs (*Telopea* spp., family Proteaceae) for cut flower production. *Genet Resour Crop Evol*. 2006; doi: 10.1007/s10722-005-3487-7.

17. Inglis PW, Pappas M de CR, Resende LV, Grattapaglia D. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PLOS ONE*. 2018; doi: 10.1371/journal.pone.0206085.

18. 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 Resour*. 2019; doi: https://doi.org/10.1111/1755-0998.12938.

19. Lu-Irving P, Rutherford S: High molecular weight DNA extraction from leaf tissue. dx.doi.org/10.17504/protocols.io.bu9ynz7w (2021).

20. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genomics*. 2017; doi: 10.1099/mgen.0.000132.

668 21. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. 2018; doi: 10.1093/bioinformatics/bty149.

670 22. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010; www.bioinformatics.babraha-m.ac.uk/projects/fastqc/.

672 23. Phase Genomics. hic_qc. 2019; https://github.com/phasegenomics/hic_qc.

24. Chen Y, Nie F, Xie S-Q, Zheng Y-F, Dai Q, Bray T, et al.. Efficient assembly of nanopore reads via highly
674 accurate and intact error correction. *Nat Commun*. Nature Publishing Group; 2021; doi: 10.1038/s41467-020-20236-7.

676 25. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. Nature Publishing Group; 2019; doi: 10.1038/s41587-019-0072-8.

678 26. 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; doi:
680 10.1101/gr.215087.116.

27. Jordan GJ, Carpenter RJ, Koutoulis A, Price A, Brodribb TJ. Environmental adaptation in stomatal size
682 independent of the effects of genome size. *New Phytol*. 2015; doi: 10.1111/nph.13076.

28. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome
684 sequences. *Genome Res*. 2017; doi: 10.1101/gr.214874.116.

29. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome
686 assembly and annotation completeness with single-copy orthologs. *Bioinforma Oxf Engl*. 2015; doi: 10.1093/bioinformatics/btv351.

688 30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. Elsevier; 1990; 215:403–10.

690 31. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. Oxford University Press; 2011; 39:W29–37.

692 32. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*. Oxford University Press; 2005; 33:W465–7.

694 33. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet TIG*. 2000; doi: 10.1016/s0168-9525(00)02024-2.

696 34. Stuart KC, Edwards RJ, Cheng Y, Warren WC, Burt DW, Sherwin WB, et al.. Transcript- and annotation-guided genome assembly of the European starling. *bioRxiv*. Cold Spring Harbor Laboratory; 2021; doi:
698 10.1101/2021.04.07.438753.

35. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing
700 assessment for genome assemblies. *Genome Biol*. BioMed Central; 2020; 21:1–27.

36. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. 2014;
702 https://sourceforge.net/projects/bbmap/.

37. Darlington CD, Wylie AP. Chromosome atlas of flowering plants. London, UK: George Allen and Unwin
704 Ltd.; 1956.

38. Ramsay H. Chromosome numbers in the Proteaceae. *Aust J Bot*. 1963; 11:1–20.

39. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplots: Reference-free profiling of polyploid genomes. *bioRxiv*. 2019; doi: 10.1101/747568.

40. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 2017; doi: 10.1093/bioinformatics/btx153.

41. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018; doi: 10.1093/bioinformatics/bty191.

42. Edwards RJ, Field MA, Ferguson JM, Dudchenko O, Keilwagen J, Rosen BD, et al.. Chromosome-length genome assembly and structural variations of the primal Basenji dog (*Canis lupus familiaris*) genome. *BMC Genomics*. 2021; doi: 10.1186/s12864-021-07493-6.

43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al.. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; doi: 10.1093/bioinformatics/btp352.

44. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; https://www.R-project.org/

45. Field MA, Rosen BD, Dudchenko O, Chan EKF, Minoche AE, Edwards RJ, et al.. Canfam_GSD: De novo chromosome-length genome assembly of the German Shepherd Dog (*Canis lupus familiaris*) using a combination of long reads, optical mapping, and Hi-C. *GigaScience*. 2020; doi: 10.1093/gigascience/giaa027.

46. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*. 2018; doi: 10.1186/s12859-018-2485-7.

47. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010; doi: 10.1093/bioinformatics/btq033.

48. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*. 2017; doi: 10.1093/bioinformatics/btw663.

49. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 2017; doi: 10.1101/gr.214270.116.

50. Oxford Nanopore Technologies Ltd.. medaka. 2020; https://github.com/nanoporetech/medaka.

51. 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; doi: 10.1371/journal.pone.0112963.

52. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*. 2014; doi: 10.1186/1471-2105-15-211.

53. Kammonen JI, Smolander O-P, Paulin L, Pereira PAB, Laine P, Koskinen P, et al.. GapFinisher: A reliable gap filling pipeline for SSPACE-LongRead scaffolder output. *PLOS ONE*. 2019; doi: 10.1371/journal.pone.0216885.

54. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al.. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*. 2016; doi: 10.1016/j.cels.2016.07.002.

742 55. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al.. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017; doi:
744 10.1126/science.aal3327.

56. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, et al.. The Juicebox Assembly
746 Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv*. Cold Spring Harbor Laboratory; 2018; doi: 10.1101/254797.

748 57. Kundu R, Casey J, Sung W-K. HyPo: Super fast and accurate polisher for long read genome assemblies. 2019; doi: 10.1101/2019.12.19.882506.

750 58. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. *Bioinformatics*. 2014; doi: 10.1093/bioinformatics/btu393.

752 59. Jiang J, Birchler JA, Parrott WA, Kelly Dawe R. A molecular view of plant centromeres. *Trends Plant Sci*. 2003; doi: 10.1016/j.tplants.2003.10.011.

754 60. Oliveira LC, Torres GA. Plant centromeres: genetics, epigenetics and evolution. *Mol Biol Rep*. Springer Science and Business Media LLC; 2018; doi: 10.1007/s11033-018-4284-7.

756 61. Simon L, Voisin M, Tatout C, Probst AV. Structure and function of centromeric and pericentromeric heterochromatin in *Arabidopsis thaliana*. *Front Plant Sci*. 2015; doi: 10.3389/fpls.2015.01049.

758 62. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011; doi: 10.1093/bioinformatics/btr011.

760 63. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinforma*. 2009; doi: 10.1002/0471250953.bi0410s25.

762 64. Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol Biol Clifton NJ*. 2019; doi: 10.1007/978-1-4939-9173-0_9.

764 65. Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van de Peer Y. How many genes are there in plants (… and why are they there)? *Curr Opin Plant Biol*. 2007; doi: 10.1016/j.pbi.2007.01.004.

766 66. Seemann T. barrnap. 2018; https://github.com/tseemann/barrnap.

67. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA
768 genes. *Nucleic Acids Res*. 2016; doi: 10.1093/nar/gkw413.

68. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 2013; doi:
770 10.1093/bioinformatics/btt509.

69. Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, et al.. OrthoVenn2: a web server for whole-genome
772 comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res*. 2019; doi: 10.1093/nar/gkz333.

774 70. Grabherr MG, Russell P, Meyer M, Mauceli E, Alföldi J, Di Palma F, et al.. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*. 2010; doi: 10.1093/bioinformatics/btq102.

776 71. Meyer M, Munzner T, Pfister H. MizBee: a multiscale synteny browser. *IEEE Trans Vis Comput Graph*. 2009; doi: 10.1109/TVCG.2009.167.

778 72. Carta A, Bedini G, Peruzzi L. A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytol*. 2020; doi: https://doi.org/10.1111/nph.16668.

780 73. Johnson L, Briggs B. Evolution in the Proteaceae. *Aust J Bot*. 1963; 11:21–61.

74. Johnson LAS, Briggs BG. On the Proteaceae—the evolution and classification of a southern family. *Bot J*
782 *Linn Soc*. 1975; doi: 10.1111/j.1095-8339.1975.tb01644.x.

75. Murat F, Armero A, Pont C, Klopp C, Salse J. Reconstructing the genome of the most recent common
784 ancestor of flowering plants. *Nat Genet*. 2017; doi: 10.1038/ng.3813.

76. Stace HM, Douglas AW, Sampson JF. Did 'Paleo-polyploidy' Really occur in Proteaceae? *Aust Syst Bot*.
786 CSIRO PUBLISHING; 1998; doi: 10.1071/sb98013.

77. Sauquet H, Weston PH, Anderson CL, Barker NP, Cantrill DJ, Mast AR, et al.. Contrasted patterns of
788 hyperdiversification in Mediterranean hotspots. *Proc Natl Acad Sci*. 2009; doi: 10.1073/pnas.0805607106.