The first genome for the Cape Primrose Streptocarpus rexii (Gesneriaceae), a model plant for studying meristem-driven shoot diversity

Citation for published version:
Nishii, K, Hart, M, Kelso, N, Barber, S, Chen, Y-Y, Thomson, M, Trivedi, U, Twyford, AD & Möller, M 2022, 'The first genome for the Cape Primrose Streptocarpus rexii (Gesneriaceae), a model plant for studying meristem-driven shoot diversity', *Plant Direct*, vol. 6, no. 4, e388. https://doi.org/10.1002/pld3.388

Digital Object Identifier (DOI):
10.1002/pld3.388

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Plant Direct

Received: 11 October 2021 | Revised: 26 January 2022 | Accepted: 30 January 2022

DOI: 10.1002/pld3.388

ORIGINAL RESEARCH

American Society of Plant Biologists SEB SOCIETY FOR EXPERIMENTAL BIOLOGY WILEY

The first genome for the Cape Primrose *Streptocarpus rexii* (Gesneriaceae), a model plant for studying meristem-driven shoot diversity

Kanae Nishii[1,2] | Michelle Hart[1] | Nathan Kelso[1] | Sadie Barber[1] | Yun-Yu Chen[1,3] | Marian Thomson[4] | Urmi Trivedi[4] | Alex D. Twyford[1,5] | Michael Möller[1]

[1]Royal Botanic Garden Edinburgh, Edinburgh, UK

[2]Kanagawa University, Hiratsuka, Japan

[3]Institute of Molecular Plant Sciences, The University of Edinburgh, Edinburgh, UK

[4]Edinburgh Genomics, Ashworth Laboratories, The University of Edinburgh, Edinburgh, UK

[5]Institute of Evolutionary Biology, Ashworth Laboratories, The University of Edinburgh, Edinburgh, UK

Correspondence
Kanae Nishii, Kanagawa University, 2946 Tsuchiya, Hiratsuka, Kanagawa 259-1293, Japan.
Email: kanaenishii@gmail.com

Michael Möller, Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, Scotland, UK.
Email: mmoeller@rbge.org.uk

Funding information
Sumitomo Foundation, Grant/Award Number: 170204; Japan Society for the Promotion of Science, Grant/Award Numbers: 15K18593, 18K06375; Edinburgh Botanic Garden (Sibbald) Trust, Grant/Award Number: 2018#18

Abstract

Cape Primroses (*Streptocarpus*, Gesneriaceae) are an ideal study system for investigating the genetics underlying species diversity in angiosperms. *Streptocarpus rexii* has served as a model species for plant developmental research for over five decades due to its unusual extended meristem activity present in the leaves. In this study, we sequenced and assembled the complete nuclear, chloroplast, and mitochondrial genomes of *S. rexii* using Oxford Nanopore Technologies long read sequencing. Two flow cells of PromethION sequencing resulted in 32 billion reads and were sufficient to generate a draft assembly including the chloroplast, mitochondrial and nuclear genomes, spanning 776 Mbp. The final nuclear genome assembly contained 5,855 contigs, spanning 766 Mbp of the 929-Mbp haploid genome with an N50 of 3.7 Mbp and an L50 of 57 contigs. Over 70% of the draft genome was identified as repeats. A genome repeat library of Gesneriaceae was generated and used for genome annotation, with a total of 45,045 genes annotated in the *S. rexii* genome. *Ks* plots of the paranomes suggested a recent whole genome duplication event, shared between *S. rexii* and *Primulina huaijiensis*. A new chloroplast and mitochondrial genome assembly method, based on contig coverage and identification, was developed, and successfully used to assemble both organellar genomes of *S. rexii*. This method was developed into a pipeline and proved widely applicable. The nuclear genome of *S. rexii* and other datasets generated and reported here will be invaluable resources for further research to aid in the identification of genes involved in morphological variation underpinning plant diversification.

KEYWORDS
genome assembly, Gesneriaceae, high-molecular weight DNA, Oxford Nanopore Technologies, PLCL pipeline, *Streptocarpus rexii*

The affiliations of Yun-Yu Chen and Marian Thomson indicate where the work had been carried out.

1 | INTRODUCTION

Biodiversity greatly reflects the genetic properties of organisms (e.g. Supple & Shapiro, 2018). Unraveling the genes underlying this diversity is an ongoing challenge in evolutionary research. Recent advances in sequencing technologies made the acquisition of whole genome data rapid and affordable even for non-model plants (e.g., Dumschott et al., 2020; Liu et al., 2019). In this study, we use Oxford Nanopore Technologies (ONT) long read sequencing (Rang et al., 2018) to assemble the nuclear, chloroplast, and mitochondrial genomes of the emerging model plant species, the Cape Primrose *Streptocarpus rexii* (Bowie ex Hook.) Lindl. The plants' haploid nuclear genome size was previously estimated with flow cytometry to be around 929 Mb, falling within the average range of angiosperm genome sizes (Dodsworth et al., 2015). It has a karyotype of $n = x = 16$, $2n = 32$ chromosomes (Möller, 2018) (Figure 1).

Streptocarpus rexii belongs to a genus of the Old World (OW) members of the plant family Gesneriaceae (Weber et al., 2013). *Streptocarpus* includes 185 species (Hilliard & Burtt, 1971; Nishii et al., 2015; GRC, 2021 onwards, https://padme.rbge.org.uk/grc), and some of these are popular ornamental plants. *Streptocarpus* species are particularly notable for their vegetative and floral diversity (e.g., Hilliard & Burtt, 1971; Nishii et al., 2015). The flowers are zygomorphic and possess a variety of shapes and colors and markings representing different pollination syndromes (Möller et al., 2019).

Besides their ornamental value, *Streptocarpus* are historically of great interest to evolutionary biologists (Crocker, 1861), because of their unique vegetative forms, which develop from uniquely evolved and behaving leaf and shoot apical meristems (Jong, 1970; Jong & Burtt, 1975; Steeves & Sussex, 1989). The roles of shoot apical meristems have partly been transferred to leaf meristems positioned at the proximal end of the leaf. These leaf meristems first appear in the cotyledons and their differential growth leads to asymmetric laminal growth, forming a macrocotyledon and microcotyledon (Jong & Burtt, 1975; Imaichi et al., 2000; Nishii et al., 2004). Although caulescent species retain the shoot apical meristem and a short-lived leaf meristem for lamina expansion, unifoliate species only retain the macrocotyledon enlarging from a persistent basal leaf meristem and are the sole photosynthesizing foliar organ. They produce inflorescences from the base of the lamina of the macrocotyledon and are monocarpic. Some species produce additional leaves from the base of the macrocotyledon, which is arranged in a false rosette to produce the rosulate form such as found in *S. rexii* (Jong & Burtt, 1975; Möller & Cronk, 2001; Nishii & Nagata, 2007).

Streptocarpus rexii has been extensively studied as a model species in the genus, in particular from an evolutionary developmental point of view using candidate gene approaches. The roles of a few key meristematic genes characterized in model plants have been investigated in leaf meristems of *S. rexii*. These studies have shown that genes expressed in the shoot apical meristem of model plants, such as class 1 *KNOX* genes, *WUSCHEL*, *ARP*, and *YABBY*, are also expressed in the leaf meristems of *Streptocarpus* (Mantegazza et al., 2007, 2009; Nishii et al., 2010, 2017; Tononi et al., 2010). Thus, it seems that *Streptocarpus* has evolved unique genetic cascades, perhaps through novel genes or pathways to achieve its varied vegetative forms. The genetic differences between these forms appear to have relatively simple mechanisms, as shown in early traditional genetic studies. For example, Beuttel (1939) and Oehlkers (1964) independently studied the vegetative morphology of *Streptocarpus* and reported the involvement of one or two loci differentiating unifoliate and rosulate forms.

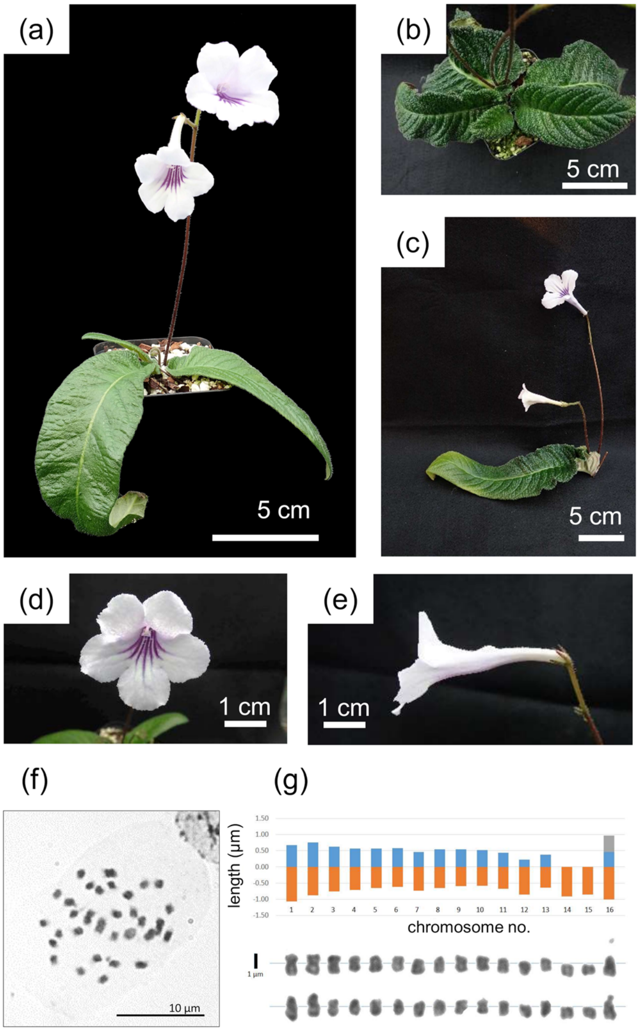

FIGURE 1 *Streptocarpus rexii*. (a) Flowering plant. (b) Top view of the irregular "false rosette" of this rosulate species. (c) An excised individual leaf with inflorescences at its base, similar to the structure of a unifoliate *Streptocarpus* species. (d) Flower in front view. (e) Flower in side view. (f) Root tip mitotic late pro-metaphase showing 32 chromosomes in a cell. (g) Chromosomes, above as schematic diagram showing the 16 unique chromosomes of one genome complement (*n*) aligned along the centromere by decreasing length with the NOR chromosome at the end (NOR in gray), and below as karyotype showing all 32 chromosomes arranged in pairs

Recently, based on restriction site-associated DNA sequencing (RAD-seq), a genetic map was generated for *Streptocarpus* (Chen et al., 2018), and a QTL study on flower morphology reported (Chen et al., 2020). The actual key genes regulating vegetative or floral morphologies are yet to be discovered, and a nuclear genome sequence for *Streptocarpus* would open the door to discover these genes, as well as lineage specific genetic events involved in their evolution.

The unique meristem properties of *S. rexii* mean that newly generated genomic resources may facilitate critical insights into the genetic mechanism underlying meristem establishment in plants. Some genetic resources are already available for the genus, such as a transcriptome (Chiara et al., 2013), and an interspecific genetic map using a cross between the rosulate *S. rexii* and the unifoliate *S. grandis* (Chen et al., 2018). These resources, together with the nuclear genome assembled here would aid studies to pinpoint the genes regulating morphological diversity in *Streptocarpus*. In addition, several methodologies have been successfully applied to investigate gene function in *Streptocarpus*. Most of this work has been carried out on the ornamental *Streptocarpus ionanthus* (formerly *Saintpaulia ionantha*, African Violet) for breeding purposes: here, agrobacterium-mediated transformation was successful (Kushikawa et al., 2001), as well as chemical mutagenesis with ethylmethanesulfonate, colchicine, and physical mutagenesis through exposure to different types of irradiation (da Silva et al., 2017). In *S. rexii*, a successful virus-induced gene-silencing protocol was recently reported (Nishii et al., 2020).

Genetic resources for the family Gesneriaceae are important to annotate the *S. rexii* genome and to understand the establishment of novel meristems and their genetic evolution in *Streptocarpus*. Their evolutionary history could be examined by comparing existing Gesneriaceae genomes, for example, for whole genome duplications (WGDs) known to be a potential source of novel genetic mechanisms (de Bodt et al., 2005). Several nuclear genomes have been published for OW Gesneriaceae, but none for New World (NW) taxa. Nuclear genome assemblies exist for *Dorcoceras hygrometricum* (formerly *Boea hygrometrica*) (Xiao et al., 2015) and *Primulina huaijiensis* (Feng et al., 2020) (Table S1). There are several intriguing developmental and ecological features of these two taxa: for example, *D. hygrometricum* was studied for its desiccation tolerance (Mitra et al., 2013), and a genome-wide DNA methylation study suggested a rapid genome-wide response upon desiccation (Sun et al., 2021). *Primulina*, on the other hand, was studied intensively at a genus level from an ecological and systematic point of view, as well as for the development and diversity of flowers, for which a genetic map had been built (Feng et al., 2016, 2019, 2020).

Several transcriptome studies have been published for both NW and OW Gesneriaceae. In total, transcriptomes are available for 20 species including 11 species of *Primulina* (Table S1). Raw transcriptome reads of *S. ionanthus* are deposited in the data archive of the 1kp project (One Thousand Plant Transcriptomes Initiative, 2019; Table 1). Those genome and transcriptome resources are investigated and incorporated into the annotation pipeline of *S. rexii* here.

In plant cells, DNA is found in the nucleus, the chloroplast, and the mitochondrion. In Gesneriaceae, chloroplast genomes of 18 species have been published at present, including two species and five subspecies of *Streptocarpus* (Table S2), whereas only one mitochondrion sequence is available, for *D. hygrometricum* (Zhang et al., 2012). In the present study, both chloroplast and mitochondrial genome sequences were assembled de novo for *S. rexii* using coverage information from the draft genome assemblies. This plastid assembly method was further developed into a pipeline and tested in *Arabidopsis thaliana* and *Oryza sativa*. This method neither requires a reference nor seed sequences, unlike many other methods (e.g., NOVOPlasty, Dierckxsens et al., 2017; GetOrganelle, Jin et al., 2020). Unlike Organelle_PBA (Soorni et al., 2017), it does not require a scaffolding step. The method is also computationally efficient because it does not deal with all raw reads, but only those from the assembled contigs.

Although the method for whole genome sequencing and assembly are well established for model species (e.g., Dumschott et al., 2020; Michael et al., 2018), there are still technical challenges for successful genome sequencing. These include issues around the extraction of high-molecular weight, high-quality DNA, and the method developed here may prove to be widely applicable. As additional resources for OW Gesneriaceae, we generated a genome annotation pipeline. This study provides the resources and methodological approaches developed and applied here.

TABLE 1 Statistics of *Streptocarpus rexii* genome assembly and annotation

Parameter	Values
Estimated genome size (Mb)[a]	929
Assembled genome size (Mb)	766
Num. of total contigs	5,855
Num. of contigs ≥ 50 Kbp	1,811
Longest contig (bp)	15,643,668
Statistics (≥3,000 bp)	
GC (%)	38.89
N50 (bp)	3,726,467
N75 (bp)	1,476,021
L50	57
L75	135
BUSCO completeness (%)	99.0
Genome repeats (%)	70.97
Num. genes annotated	45,045
Average gene length (bp)	2,609
Num. of exons	213,819
Average num. exons per gene	4.7
Average exon length (bp)	249
BUSCO completeness of annotated gene set (%)	89.6
Chloroplast genome length (bp)	152,571
Mitochondrial genome length (bp)	599,262

[a]Based on Möller (2018).

2 | MATERIALS AND METHODS

2.1 | Plant material

Streptocarpus rexii (RBGE accession numbers: 20180766, descendent of lineage 19870333; K.Jong-Faraway) was in cultivation at the living collection of Royal Botanic Garden Edinburgh (RBGE). RBGE has maintained an inbred lineage of this species (Hughes et al., 2005) by self-fertilization of individuals over more than 10 generations (Chen et al., 2018).

2.2 | Metaphase and karyotype of *S. rexii*

The chromosome preparation followed Jong and Möller (2000). Images were captured with a Zeiss Zen v.3.1 and an AxioCam MRc5 camera mounted on an Axiophot brightfield microscope (Zeiss, Welwyn Garden City, UK). Images were manipulated in Gimp v.2.10.24 (The GIMP Development Team, 2019), chromosomes measured with Zen, and the karyotype created manually in Powerpoint (Microsoft Office).

2.3 | DNA extraction

We established a reliable protocol for extracting high-quality, high-molecular weight plant DNA. To obtain such DNA requires careful consideration of the plant material and extraction methods; for example, the leaf material is best used fresh, not silica-dried; high temperatures should be avoided during extraction to keep damage to the DNA at a minimum. In the preliminary stage, we tested "traditional" methods such as CTAB (Doyle, 1991; Doyle & Doyle, 1987), Qiagen DNeasy kit (Hilden, Germany), and the Invitrogen PureLink™ Genomic Plant DNA kit (Waltham, MA, US), but these did not generate sufficiently high-quality DNA for ONT sequencing (data not shown). Thus, we developed a DNA extraction protocol based on those reported by Souza et al. (2012), Gunter (2015), and PacBio Sample Net (2015), combining a nuclear isolation method, sorbitol buffer wash, and Qiagen Genomic tip columns. The detailed DNA extraction protocol has been deposited at protocol.io (dx.doi.org/10.17504/protocols.io.bempjc5n) and is also available at RBGE's web magazine (https://stories.rbge.org.uk/archives/30792).

2.4 | Library preparation and sequencing

Two libraries (11843TA0001L01 and 11843TA0001L02) were prepared with different shearing settings of 50 and 45 kbp, using a Megaruptor (Diagenode, Denville, NJ, US). Smaller fragments of approximately less than 1 kbp were removed with AMPure XP beads (Beckman Coulter, Brea, CA, US) with a beads: sample ratio of 0.4 : 1 (v : v). For each library, 1 μg of sheared and cleaned DNA was used. Formalin-fixed, paraffin-embedded (FFPE) DNA repair reactions using NEBNext® FFPE DNA Repair Mix (New England Biolabs, Ipswich, MA, US), end repair/A tailing using NEBNext® Ultra End Repair/dA-Tailing Module (NEB) were carried out, and the samples were re-cleaned with AMPure XP beads (beads: sample ratio of 0.4 : 1, v : v). Adapter ligation was carried out using the Ligation Sequencing SQK-LSK109 kit (ONT, Littlemore, Oxford, UK), and the samples were further cleaned with the same ratio of AMPure XP beads to reduce fragment size to <3 kb. The long fragment buffer (LFB, a component of the SQK-LSK109 kit) was used as bead wash buffer. *Escherichia coli* DNA was added as a spiked-in control to verify the quality of the library preparation and sequencing. The two libraries were quantified and were 16.4 and 18.5 fmol, respectively. They were loaded separately on a PromethION Flow cell FLO-PRO002 (ONT). The sequencing run was set for 63 h, with a pore type: R9.4.1, and caller variant set to "fast." Basecalling was done with Guppy v.3.0.5 (ONT).

2.5 | Read quality control

The quality of reads and the read lengths were assessed with NanoPlot v.1.38.0 (De Coster et al., 2018), and the summary text files provided by Guppy were examined. The read length versus read number histogram plots were generated using R ggplot2 (Wickham, 2016) with a bin size of 1,000.

2.6 | Genome assembly

Two fastq read sets generated from the two sequencing libraries were combined into one file and the assembly performed with Canu v.1.8 (Koren et al., 2017). Only reads longer than 10,000 bp were used for genome assembly. The draft assembly was further polished five times with Racon v.1.4.11 (Vaser et al., 2017). The polished assembly was further examined with Medaka v.0.11.5 (https://nanoporetech.github.io/medaka/) to resolve unreliable scaffolds. To check and remove contaminants from non-plant organisms, the assembly was analyzed with BlobTools v.1.1.1 (Laetsch & Blaxter, 2017), and only scaffolds identified as streptophyta or no-hit were kept in the final assembly.

2.7 | Preparation of Gesneriaceae genome annotation materials

To annotate the *S. rexii* genome assembly produced in this study, we generated annotation material using resources available from public repositories (Table S12). The transcriptomes and proteomes of OW Gesneriaceae were obtained and curated for the genome annotation pipeline. The transcriptome of *S. rexii* was downloaded from Angeldust (http://www.beaconlab.it/angeldust) and their CDS predicted using TransDecoder v.5.5.0 (https://github.com/TransDecoder/TransDecoder). Transcriptomes were assembled de novo for *S.* (*Saintpaulia*) *ionantha* and *Haberlea rhodopensis* from raw reads (Table 3). The raw reads were trimmed using Trimmomatic v.0.39

(Bolger et al., 2014) and assembled using Trinity v.2.11.0 (Grabherr et al., 2011). For *P. huaijiensis*, the assembled transcriptome was obtained from Dryad (https://doi.org/10.1111/1755-0998.12333; Ai et al., 2015), and CDS and predicted protein sequences obtained using TransDecoder.

2.8 | Building Gesneriaceae genome repeat libraries

In the family Gesneriaceae, the genome assemblies of *D. hygrometricum* (previously *Boea hygrometrica*; Xiao et al., 2015) and *P. huaijiensis* (Feng et al., 2020) have been published and deposited in public databases. Genome repeat libraries were built from these assemblies and a *S. rexii* assembly generated in this study using the pipeline of Jacques Dainat (https://www.biostars.org/p/411101/). In brief, the de novo repeat libraries were built using RepeatModeler v.2.0 (Flynn et al., 2020). The same species' proteomes were curated with TransposonPSI (http://transposonpsi.sourceforge.net/) to detect protein sequences of transposons, and the detected transposable element (TE) proteins removed from the proteome using the perl script "fasta_removeSeqFromIDlist.pl" from the Genome Assembly Annotation Service (GAAS; https://github.com/NBISweden/GAAS). Blastp searches were carried out using the proteome without TE proteins as query and the generated repeat libraries as reference sequences. The resulting hits (sequences) were removed from the repeat libraries with ProtExcluder (https://github.com/NBISweden/ProtExcluder). This process was carried out for each species separately, with the resulting libraries concatenated and used as a Gesneriaceae genome repeat library for annotation.

2.9 | Genome repeat analyses

The genome repeats (repeated DNA sequences) in the *S. rexii* genome assembly were analyzed using RepeatMasker (Smit et al., 2013–2015; http://www.repeatmasker.org). The Gesneriaceae repeat libraries generated above, Dfam library (release 3.3, curated families) (Storer et al., 2021), and RepBase library (https://www.girinst.org/; RepBaseRepeatMaskerEdition-20181026) were concatenated and used with RepeatMasker.

2.10 | Genome annotation of *S. rexii*

The *S. rexii* genome was annotated with Maker (Holt & Yandell, 2011). The transcriptomes and proteomes of *S. rexii* and the other OW Gesneriaceae (*P. huaijiensis*, *D. hygrometricum*, *S. ionanthus*, *H. rhodopensis*), and the gff3 output of RepeatMasker, were used in the pipeline. After the first round of Maker, gene models were trained ab initio using SNAP (Korf, 2004) and Augustus v.3.3.3 (Stanke et al., 2008) utilizing Busco v.4.0.2 (Simão et al., 2015), and the second round of Maker carried out with the predicted gene models. In addition to the gff3 statistics in the Maker pipeline, Busco analyses were carried out. These, with the CDS sets of the nuclear genome annotated by Maker, were used to evaluate annotation efficiency.

2.11 | WGD analyses

WGD analyses were carried out on three Gesneriaceae species, *S. rexii*, *P. huaijiensis*, and *D. hygrometricum* using the program wgd (Zwaenepoel & van de Peer, 2019). To obtain CDS files mapped on the genome, the genome sequences of *P. huaijiensis* and *D. hygrometricum* were annotated using Maker, as for *S. rexii* described above, and gff3 files obtained. CDS files were generated from the gff3 files. For *D. hygrometricum*, the public CDS file (PRJNA182117, GCA_001598015.1_Boea_hygrometrica.v1_cds_from_genomic.fna.gz) was also analyzed, although not all sequences in the CDS assembly were assigned to a genome location. Orthologous pairs were found within and between species using wgd-mcl (Altschul et al., 1997), and synonymous substitution ratio (*Ks*) values were calculated using wgd-ksd with PAML4 (Yang, 2007) and MAFFT (Katoh & Standley, 2013). Co-linearity analyses were carried out using wgd-syn with i-ADHoRe (Proost et al., 2012). These analyses identified the anchor gene pairs, which are specific orthologous pairs arranged in the same order in a genome before and after a predicted WGD. Bayesian Gaussian mixture model (BGMM) analyses were carried out on the calculated *Ks* values using wgd-mix and visualized using wgd-viz in wgd.

2.12 | Chloroplast and mitochondrial genome assemblies

We tested two approaches for the assembly of chloroplast and mitochondrial genomes. One approach was to screen all raw reads using plastid queries as described in Wang et al. (2018), but this did not produce complete genomes for *S. rexii*; therefore, another approach was tested: The chloroplast and mitochondrial contigs were selected from the final whole genome assembly, the reads for these contigs extracted, and the chloroplast and mitochondrial genomes newly assembled. This strategy worked well for the *S. rexii* chloroplast and mitochondrial genome assemblies.

The draft genome scaffolds were analyzed with blastn (NCBI BLAST+ v.2.10.0; Altschul et al., 1994), against *A. thaliana* and *Nicotiana tabacum* chloroplast and mitochondrial genomes. The coverage of each scaffold was evaluated with BlobTools carried out as described above. The blastn results were compared against read coverage data. The input reads, that is, reads > 10,000 bp, were mapped onto the assemblies using minimap2 v.2.17 (Li, 2018) to generate a bamfile. From the bamfile, the mapped reads of chloroplast and mitochondrial contigs were extracted using samtools v.1.9 (Li et al., 2009).

The extracted chloroplast and mitochondrial reads were assembled using Canu. Several settings were tested, and the final settings used for the chloroplast assembly were minReadLength = 12,000, minOverlapLength = 10,000, corMinCoverage = 8, trimReadsOverlap = 100, and trimReadsCoverage = 100, and for the mitochondrial assembly corOutCoverage = 999, minReadLength = 10,000, minOverlaplength = 500, trimReadsOverlap = 3, and trimReadsCoverage = 3.

The chloroplast and mitochondrial sequences were annotated using GeSeq (Tillich et al., 2017). To compare the chloroplast sequences, first *S. rexii*, *D. hygrometricum*, and *P. huaijiensis* sequences were aligned using MAFFT and manually checked with BioEdit (Hall, 1999). The similarity of the sequences was analyzed in the mVISTA genome browser (Frazer et al., 2004) with the LAGAN aligner (Brudno et al., 2003). To analyze the similarity of the mitochondrial genomes between *S. rexii* and *D. hygrometricum* and also *A. thaliana* and the closely related *Capsella rubella*, the sequences were aligned one to one, and dot plots generated using the D-GENIES web server (Cabanettes & Klopp, 2018). Sequence accession numbers are shown in Table S13. Prior to the final nuclear genome assembly, chloroplast and mitochondrial contigs were removed.

2.13 | PLCL pipeline

The approach used for the assembly of the *S. rexii* chloroplast and mitochondrial genomes was further developed into a pipeline, including perl scripts. The pipeline, named the plant contig clustering-based genome assembly (PLCL) pipeline, was subsequently tested with data from two representative model plants.

The PLCL pipeline was tested with *A. thaliana* KBS-Mac-74 ONT reads (ERR2173373; Michael et al., 2018) and *O. sativa* subsp. *indica* IR64 ONT reads (DRR196880; Tanaka et al., 2020). For each species, the ONT reads were evaluated using NanoPlot, and draft genomes assembled using Wtdbg2 (Ruan & Li, 2020). The resulting genome assembly contigs were blastn searched with chloroplast and mitochondrial sequences of each species as the query. To assess the contig coverage, the ONT reads were re-mapped onto the draft genome assembly using minimap2. The coverage was calculated using samtools. The results of blastn and coverage data were combined and the candidate plastid contigs selected. The contig coverage data were also assessed with the perl program Algorithm-KMeans-2.05 (https://metacpan.org/pod/Algorithm::KMeans) for rough clustering of plastid and nuclear contigs. The coverage and blastn bitscore datasets were manually checked to select chloroplast and mitochondrial contig candidates. The reads were extracted from bamfiles for each contig and combined as the chloroplast or mitochondrial read set for each species, and assembled using Canu with different parameter settings (see Section 3 for details). The scripts of the PLCL pipelines have been deposited in protocol.io (dx.doi.org/10.17504/protocols.io.bx5dpq26).

3 | RESULTS

3.1 | ONT PromethION sequencing

The sequencing of libraries 11843TA0001L01 and 11843TA0001L02 generated 69.55 and 89.14 Gbp of data respectively, with >2,600 channels active in the flow cell during data production (Table S3). In total, we obtained 158.69 Gbp, in >32 billion reads of raw data. The read N50 was 10,281 and 18,877 bp, respectively. The longest read was >1.1 Mbp, in library 11843TA0001L02. This library generally generated longer read lengths (see Table S3) and a higher yield compared with library 11843TA0001L01. The majority of reads for both libraries ranged between 100 and 100,000 bp. The spike observed at around 3,500 bp in both libraries originated from the internal standard *E. coli*.

3.2 | Genome assembly

Only reads longer than 10,000 bp, that is, 4,360,109 raw reads of 89.56 Gbp combined from both libraries, were used for the genome assembly. These provided approximately 119 times coverage based on the estimated genome size of 929 Mb for *S. rexii* for the $x = 16$ chromosomes (Möller, 2018). The Canu analysis initially resulted in 2,488 contigs. While polishing the assembly, Racon reduced the number of contigs to 2,483, whereas Medaka dissolved uncertain contigs and resulted in a total of 6,155 contigs. The longest contig was 15.64 Mbp and the genome N50 was 3.6 Mbp (Table S4). Of the 6,114 contigs (776,293,364 bp) of >100 bp length, 2,035 contigs (96% of total bp) were assigned to streptophyta and 3,788 (2.75%) were no-hits (Figure S2; Table S5). The average coverage was 42.3× for streptophyta, 3.1× for no hit, and 17.2× for all contigs. Among the detected contaminants, the highest by base pairs was from proteobacteria (0.76%). All streptophyta (2,037) and no-hit (3,827) contigs, in total 5,864, were kept. Nine contigs were removed since they were assigned to the chloroplast and/or mitochondrial genomes (see below Section 3.5 for details), resulting in 5,855 nuclear contigs spanning 766 Mbp (Table 1, Table S7). The resulting nuclear genome assembly showed good contiguity, that is, 75% of the genome was covered by 135 contigs of >1.4 Mbp (see N75 and L75 in Table 1). The BUSCO completeness of the genome assembly against viridiplantae_odb10 was, at 99%, very high (Table 1, Table S6).

3.3 | Nuclear genome annotation

In the assembled *S. rexii* nuclear genome, 543,533,959 bp (70.97%) were identified as repeats (Table 1, Table S7). The retrotransposon portion was the highest among repeat elements (37.97%) (Table S8). In particular LTR retrotransposons occupied a large part of the genome space, at 35.29%, in which Gypsy/DIRS1 repeats were most abundant (22.05%), followed by Ty1/Copia (8.81%). In contrast, DNA transposons occupied only 1.15% of the genome. Unclassified repeats made up almost a third (31.05%) of the genome space.

With the available Gesneriaceae annotation resources (Table 2), Maker annotated 45,045 genes with an average gene length of 2,609 bp (Table 1). The BUSCO completeness of the genome annotation derived transcriptome was 89.6% (Table 1, Table S6).

TABLE 2 Gesneriaceae annotation resources used in the Marker annotation pipeline

Type of resource	Taxon	Reference
Genome repeat library	*Dorcoceras hygrometricum*	This study[a]
	Primulina huaijiensis	This study[a]
	Streptocarpus rexii	This study
Transcriptome	*D. hygrometricum*	Xiao et al. (2015)
	Haberlea rhodopensis	This study[b]
	P. huaijiensis	Ai et al. (2015)
	Streptocarpus ionanthus	This study[b]
	S. rexii	Chiara et al. (2013)

[a] Genome sequences obtained from NCBI.

[b] Reads were obtained from SRA archive.

3.4 | WGD analysis

WGD events were estimated for the *S. rexii* nuclear genome in comparison to those of *D. hygrometricum* and *P. huaijiensis* (Table 3). Prior to the WGD analyses, to obtain genome annotated CDS sets, the genomes of *D. hygrometricum* and *P. huaijiensis* were re-annotated in this study using the Maker pipeline, resulting in 24,585 gene models annotated in the *D. hygrometricum* genome and 42,685 in the *P. huaijiensis* genome.

For the paranomes, defined as the complete set of all duplicated genes in a genome (Vandepoele et al., 2004), synonymous substitution (*Ks*) plots were obtained for all candidate gene pairs and those of anchor gene pairs. For *S. rexii*, *P. huaijiensis*, and *D. hygrometricum*, 84,617, 49,048, and 56,877 whole paranome candidate gene pairs and 20,197, 9,919, and 760 anchor gene pairs were obtained respectively, indicating that the number of anchor gene pairs in *D. hygrometricum* was substantially lower than in the other two species (Figure 2; Table 3).

The *Ks* plots for *S. rexii* and *P. huaijiensis* were very similar, with a peak observed at ∼0.19 for both species for all paranomes and for anchor gene pairs (Figure 2). In contrast, the whole paranome of *D. hygrometricum* showed a first peak at ∼0.06 and a second broader peak around ∼1.0, but the first peak was not recognized with the anchor gene pairs. A fitted mixture model (BGMM) analysis also showed the mismatch between the *Ks* plots of whole paranomes and those anchor gene pairs for *D. hygrometricum* (Figure S3).

TABLE 3 Statistics of assembled genomes available for Gesneriaceae

	Streptocarpus rexii	***Primulina huaijiensis***	***Dorcoceras hygrometricum***
Estimated genome size (Mb)	929	511	1,691 (240[a])
Assembled genome size (Mb)	766	478	1,548
Num. of haploid chromosomes	$n = 16$	$n = 18$	$n = 9$[b]
N50 (bp)	3,726,469	23,479,473	110,988
L50	57	9	3,003
Genome repeats (%)	70.95	54.10	75.16
Num. of nuclear scaffolds	5,855	18	520,969
Num. of annotated genes (this study)	45,045	42,685	24,585
Num. of annotated genes (references)	—	31,328	23,250[c] (49,374[d])
Total candidate gene pairs (whole paranome gene pairs)	84,617	49,048	56,877
Anchor gene pairs	20,197	9,919	760
Orthologous gene pairs 1	12,160	12,160	
Orthologous gene pairs 2	12,756		12,756
Orthologous gene pairs 3		10,923	10,923

Note: Statistic of genome assembly and annotation are based on Feng et al. (2020) for *P. huaijiensis* and Xiao et al. (2015) for *D. hygrometricum* unless noted. Orthologous gene pairs 1, *S. rexii* versus *P. huaijiensis*; 2, *S. rexii* versus *D. hygrometricum*; 3, *P. huaijiensis* versus *D. hygrometricum*.

[a] Based on Zhao et al. (2014).

[b] From Kiehn et al. (1998).

[c] Num. of annotated gene models from genome in Xiao et al. (2015).

[d] Num. of predicted gene models from transcriptome in Xiao et al. (2015).

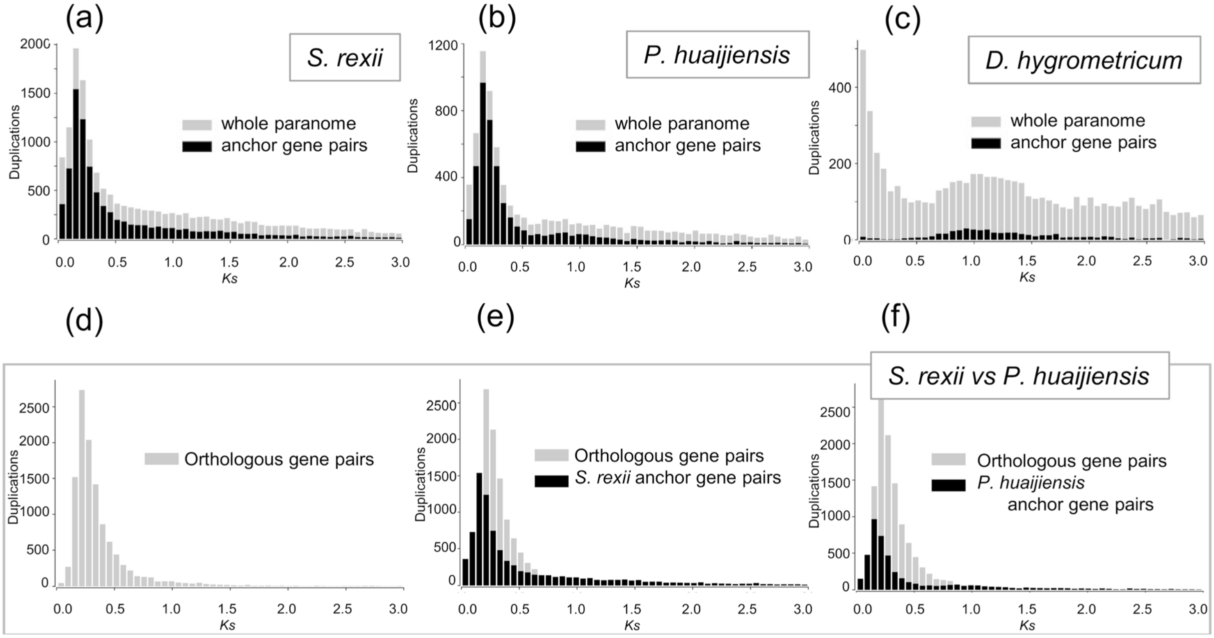

FIGURE 2 Whole genome duplication (WGD) analyses in selected Gesneriaceae. (a–c) Distributions of synonymous substitutions (*Ks*) for the whole paranome and anchor gene pairs for each species. (a) *Streptocarpus rexii*. (b) *Primulina huaijiensis*. (c) *Dorcoceras hygrometricum*. (d–f) Distributions of *Ks* values of one-to-one species' orthologous gene pairs between *S. rexii* and *P. huaijiensis* alone (d), superimposed with *Ks* values for *S. rexii* anchor gene pairs (e), and superimposed with *Ks* values for *P. huaijiensis* anchor gene pairs (f)

The distribution of *Ks* values for orthologous gene pairs between the species showed a peak between *S. rexii* and *P. huaijiensis* of ~0.2 (Figure 2); and peaks between *S. rexii* and *D. hygrometricum* of ~0.25; *P. huaijiensis* and *D. hygrometricum* of ~0.2 (Figure S4).

For *S. rexii* and *P. huaijiensis*, the difference of *Ks* distribution plots between paranomes and orthologous pairs are shown in detail in Figure 2. The *Ks* distribution plot of orthologous gene pair between *S. rexii* and *P. huaijiensis* and those separate paranomes for the two species were similar in shape, but the peaks for individual species were slightly shifted to lower values (Figure 2). The BGMM analyses of these species' *Ks* plots indicated a small peak at $Ks < 0.0625$ with $n > 3$ (Figure S3). This small peak might be caused by tandem duplications rather than WGD.

3.5 | *S. rexii* chloroplast and mitochondrial genome assembly

The chloroplast and mitochondrial genomes of *S. rexii* were assembled from the contigs in the draft genome assembly. In this assembly, six contigs were highly homologous to chloroplast sequences, and three to mitochondrial sequences (Figure 3; Table S9). The chloroplast contigs showed the highest average coverage of 2,694×, and the mitochondrial contigs 213× (Table 4, Table S9). Often, the same contigs were detected by both chloroplast and mitochondrial query searches, such as contig "2179_0," but the bitscore and the contig coverage were distinguishable (Figure 3; Table S9). Based on the selected contigs 24,438 chloroplast raw reads and 2,462 mitochondrial raw reads were extracted for re-assembly of those organelle genomes.

The chloroplast reads were assembled into a circular genome of 152,571 bp, whereas the mitochondrial reads did not produce a circularized assembly and instead are represented by a large linear contig of 599,262 bp (Figure 3). The *S. rexii* chloroplast genome length and gene order were very similar to those of *D. hygrometricum* and *P. huaijiensis*, as well as *A. thaliana* (Figures S5 and S6), although the mitochondrial gene order was highly divergent between *S. rexii*, *D. hygrometricum*, and *A. thaliana* (Figures S7 and S8), and the mitochondrial sequences could not be aligned between species (Figure S8). However, we confirmed that the original mitochondrial contigs aligned well with the final linearised *S. rexii* mitochondrial genome assembly (Figure S9).

3.6 | Benchmark testing of the chloroplast and mitochondrial genome assembly pipeline: PLCL pipeline

The newly developed method for chloroplast and mitochondrial genome assembly used for *S. rexii* was based on an initial long-read assembly of ONT data and identified reads belonging to the nuclear, chloroplast, and mitochondrial genomes from read coverage differences. This method was further developed into the PLCL pipeline and tested with available ONT datasets of the model species *A. thaliana* and *O. sativa* subsp. *indica* (Figure 4). We initially tried to map *A. thaliana* KBS-Mac-74 ONT reads to the published hybrid genome

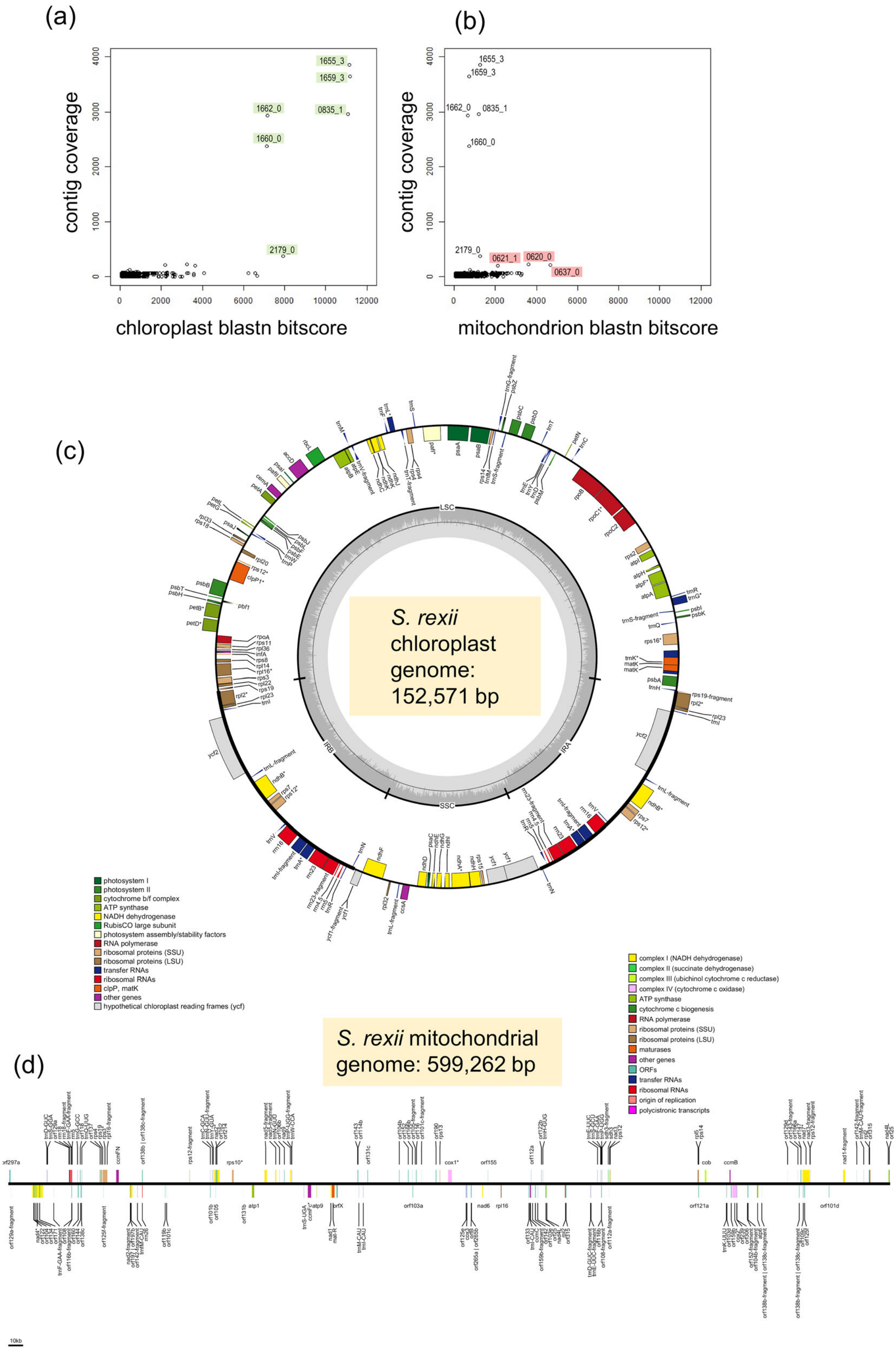

FIGURE 3 *Streptocarpus rexii* chloroplast and mitochondrial genome assemblies. (a and b) Plots of blastn bitscores with *Arabidopsis thaliana* chloroplast or mitochondrion queries to the assembled genome contigs of *S. rexii* (*x*-axis), and the contig coverage (*y*-axis). The ID of contigs is shown for contigs selected as chloroplast or mitochondrial contigs. Green and red shaded markers indicate the selected chloroplast and mitochondrial contigs, respectively, found among the *S. rexii* genome contigs. (c) *S. rexii* circularized chloroplast genome assembly. (d) *S. rexii* mitochondrial genome assembly

TABLE 4 Benchmark statistics of the plant contig clustering-based genome assembly (PLCL) pipeline

Taxon	*Streptocarpus rexii*		*Arabidopsis thaliana* KBS-Mac-74		*Oryza sativa* subsp. *indica* IR64	
Num. total reads	32,242,708		300,071		1,449,788	
Num. total bases (bp)	158,689,714,464		3,421,779,258		9,275,443,298	
Num. total assembled contigs	5,964		353		1,341	
	cp	mt	cp	mt	cp	mt
Num. contigs	6	3	4	2	2	5
Average coverage	2,694	213	2,255	98	3,382	192
Num. reads	24,438	2,462	12,621	8,239	79,822	8,081
Read N50 (bp)	21,036	20,894	19,778	12,157	13,457	12,569
Assembly result						
Whole cp/mt genome	Yes	Unknown	Yes	No	Yes	No
Circular genome	Yes	No	Yes	No	Yes	No

Note: Statistics for chloroplast (cp) and mitochondrial (mt) contigs and reads from the genome assemblies.

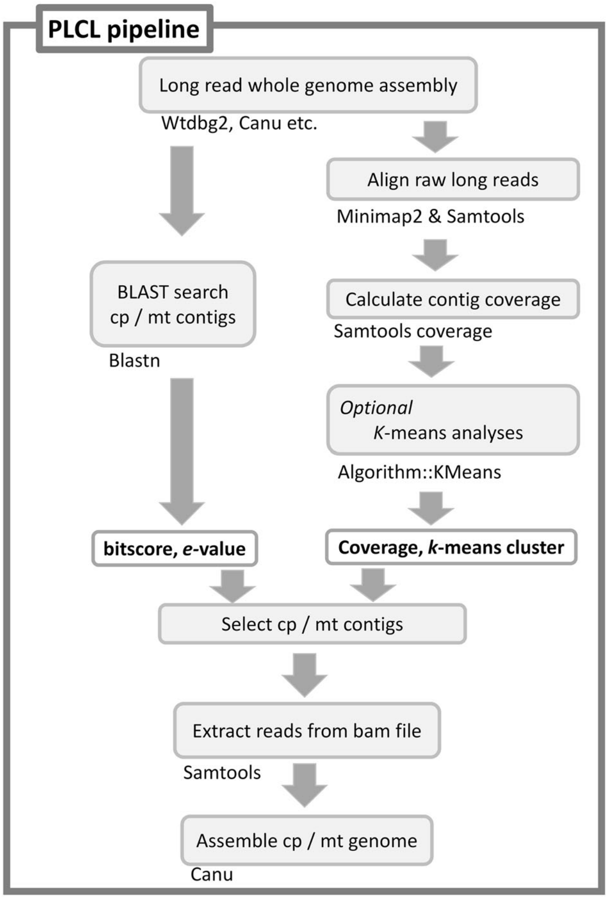

FIGURE 4 Overview of the plant contig clustering-based genome assembly (PLCL) pipeline developed and applied in the present study, with steps, workflow, and programs embedded in the pipeline. cp, chloroplast; mt, mitochondrial

assembly of *A. thaliana* KBS-Mac-74 (NCBI GCA_900303355), but they did not map well, and we were unable to extract plastid contigs and reads (data not shown). Thus, we constructed an assembly solely with ONT reads using Wtdbg2, and obtained 353 contigs with an N50 of 9.8 Mbp and L50 of 5, where N50 is defined as the length of contigs representing half of the bases of the full assembly, and L50 the number of N50 contigs counted from the longest to the shortest (https://www.ncbi.nlm.nih.gov/assembly/help/). In blastn searches of those contigs with the chloroplast sequences as query, 26 contigs were hit, and with the mitochondrial sequence 31 (Figure S11). Several contigs were co-hit by chloroplast and mitochondrial queries, but their blastn bitscores were very different and distinguished them. *K*-means analyses were also used for a quick evaluation of coverage clustering. For the chloroplast, automatic cluster number selection ($k = 0$ in Algorithum::Kmeans) worked well, and four contigs were identified as chloroplast contigs. For the mitochondrion, $k = 0$ resulted in two clusters and the contigs in the c0 cluster ("ctg224," "ctg70," Figure S11c) were also listed as chloroplast contigs (Figure S11a). However, "ctg224" and "ctg70" showed higher blastn bitscores to chloroplast than to mitochondria, and thus these were selected as chloroplast contigs. $K = 4$ seemed to distinguish the mitochondrial contigs (c1 and c2) from nuclear (c0) and chloroplast (c3) contigs. Contig "ctg31" was clustered as c2 by coverage, but it showed a low bitscore to mitochondrion query searches, and thus only "ctg55" and "ctgt36" were selected as the mitochondrial contigs (Figure S11).

In *A. thaliana* KBS-Mac-74, the 12,621 chloroplast reads and 8,239 mitochondrial reads extracted had an estimated coverage of 2,254× and 98×, respectively (Table 4). The chloroplast genome was successfully assembled from the extracted reads using Canu with the parameter sets ps2 and ps3. The resulting chloroplast contigs of *A. thaliana* KBS-Mac-74 were >190,000 and ~148,000 bp when circularized (Table S11). Because the *A. thaliana* KBS-Mac-74 chloroplast sequence has not yet been reported, it was compared with the reference chloroplast sequence of *A. thaliana* Col-0 (154,470 bp). The *A. thaliana* Col-0 chloroplast was longer than the *A. thaliana* KBS-Mac-74 chloroplast in length, and the differences between the two genomes mostly concerned deletions in mononucleotide simple repeat regions. Although it is unknown whether this reflects differences between the lineages or is a result of assembly errors. The assembly quality was shown to be sensitive to the assembler and its parameter settings (Tables S10 and S11; Figure S12). With the Canu parameter setting ps1, the inverted repeat region was not well resolved and was inverted in the final sequence (see Figure S12a,d) but was properly assembled with settings ps2 and ps3 (Figure S12b–d). Ps3 resulted in the best assembly within the tested parameter settings and had fewer deletions/gaps when compared with the *A. thaliana* Col-0 chloroplast genome. Mitochondrial assemblies produced with Canu and Wtdbg2 did not assemble well and the resulting contigs had poorer contiguity (Figure S12f,g) than the initial two original mitochondrial contigs (Figure S12e). When the PLCL pipeline was tested with the *O. sativa* subsp. *indica* IR64 ONT dataset, the chloroplast genome was assembled well but not the mitochondrial genome (Table S11; Figure S12h,i), similar to the *A. thaliana* KBS-Mac-74 ONT dataset.

4 | DISCUSSION

4.1 | A first nuclear genome assembly for the genus *Streptocarpus*

A nuclear genome assembly was successfully generated using two cells of ONT PromethION for *S. rexii*. This was remarkably cost effective compared with other methods, with an estimated sequencing cost of around $3,000 (https://medicine.yale.edu/keck/ycga/services/ontprices/) for the 1-Gb haploid 1 C genome of *S. rexii* (Möller, 2018). This shows the dramatic decline in genome sequencing costs compared with earlier genome sequencing efforts. For example, the *A. thaliana* genome assembly cost around $100 million (reviewed in Li & Harkess, 2018).

Although the assembly is not yet refined to chromosome level, the draft genome is highly contiguous and will be a useful resource for genome-wide studies in *S. rexii* and the genus *Streptocarpus*, as well as in the family Gesneriaceae. The current *S. rexii* genome assembly has an N50 of 3.7 Mbp, and therefore the contigs are much longer than the >1 Mbp threshold recommended for chromosome level scaffolding (Jung et al., 2020). This could be done with a range of technologies such as BioNano optical mapping (Deschamps et al., 2018) or Hi–C chromosome conformation capture (Dudchenko et al., 2017).

The assembled nuclear genome size was 766 Mb, which is smaller than the genome size predicted by flow cytometry of 929 Mb (Möller, 2018). It is known that it is difficult to assemble repetitive elements, where two identical reads belonging to different chromosomal positions are indistinguishable by the assembly programs (Tørresen et al., 2019). These unresolved repeats might be a reason for the smaller assembly size in this study, though further investigations are required here.

To achieve this highly contiguous ONT assembly, it was paramount to obtain high-quality high-molecular weight DNA (see also Mantere et al., 2019). Long read sequencing is known to be very sensitive to any DNA damage caused by drying the leaf tissue in silica gel or high-temperature treatments (e.g., >60°C), as well as the presence of any secondary metabolites such as polysaccharides, polyphenols, phenols, or chemical residues such as guanidinium isothiocyanate, EDTA, or ethanol (https://dnatech.genomecenter.ucdavis.edu/nanopore-sequencing-ont-promethion/). The method developed in this study was a modified nuclear isolation protocol that greatly reduced cytosol components prior to cell lysis, and used proteinase K cell lysis at 50°C to avoid the more conventional high temperature lysis at 60–65°C. Nuclear isolation-based methods seemed to be useful options for obtaining plant DNA suitable for long read sequencing (see also Zerpa-Catanho et al., 2021). We used Qiagen Genomic tip columns, because they were designed to retain 50- to 150-kb DNA fragments, fitting the criteria for ONT long read sequencing. Other more traditional protocols, such as CTAB or the commonly used Qiagen DNeasy kit, tested by us, did not provide sufficient DNA quality or quantity for *S. rexii*. In this plant, polysaccharides appear to represent the main obstacle, and the protocol developed here could well be applicable for other such problematic plants.

4.2 | Gesneriaceae annotation resources

A Maker annotation pipeline was employed to annotate the *S. rexii* genome. To have full function of the pipeline, a vast amount of lineage-specific annotation resources such as transcriptomes and repeat libraries are required (Holt & Yandell, 2011; https://gist.github.com/darencard/bb1001ac1532dd4225b030cf0cd61ce2), and these were generated in the present study from existing Gesneriaceae resources. Maker can utilize multi-species resources and for our purpose the genetic resources of OW Gesneriaceae available in public archives proved very useful. Moreover, all the genome repeat libraries of OW Gesneriaceae generated here have been deposited in public archives for general use. With these Gesneriaceae annotation resources, the Maker pipeline successfully annotated not only the *S. rexii* genome (with 45,045 genes) but also those of the other OW Gesneriaceae, *D. hygrometricum* and *P. huaijiensis*. In their original publications, 23,250 genes were annotated for *D. hygrometricum* (Xiao et al., 2015) and 31,328 for *P. huaijiensis* (Feng et al., 2020), whereas in this study, the number of genes annotated for *D. hygrometricum* (24,585 genes) and *P. huaijiensis* (42,685 genes) was slightly higher, indicating the strength and efficiency of the pipeline (Table 3). This genome annotation pipeline is likely to prove useful for further genomic characterization of the Gesneriaceae.

It is noteworthy that only 55%–60% of genes (24,585) could be annotated in *D. hygrometricum* compared with those annotated in *S. rexii* and *P. huaijiensis*. This might have been due to the relatively low contiguity of the *D. hygrometricum* genome assembly, because the gene models suggested from the transcriptomes was 49,374 genes, a similar number to those in *S. rexii* and *P. huaijiensis*. To annotate the remaining ~47% of genes in the *D. hygrometricum* genome would require an improved genome assembly.

The haploid genome size for *D. hygrometricum* was predicted to be 1,691 Mb by Xiao et al. (2015), which is significantly larger than genome size estimates for the two other study species here: 929 Mb for *S. rexii* (Möller, 2018) and 511 Mb for *P. huaijiensis* (Feng et al., 2020). In these species, the haploid chromosome number is $n = 9$ (*D. hygrometricum*), $n = 18$ (*P. huaijiensis*), and $n = 16$ (*S. rexii*) (see Table 3). The percentage of genome repeats was also the highest in *D. hygrometricum* (75.16%), lowest in *P. huaijiensis* (54.10%), and with *S. rexii* falling towards the higher end (70.95%). Thus, there appears to be a trend that larger genomes in Gesneriaceae retain more genome repeats, with this showing an almost linear relationship to genome size. Long terminal repeat (LTR) retrotransposons are the most abundant genome repeat in *S. rexii* (35.27%) and *P. huaijiensis* (48.4%; Feng et al., 2020) (value not reported for *D. hygrometricum*), with this generally being the case across plants (Kejnovsky et al., 2012). For *D. hygrometricum*, another study reported a haploid genome size estimated with flow cytometry of 240 Mb (Zhao et al., 2014), which is much smaller than the size estimate by Xiao et al. (2015). Neither studies reported the chromosome number of their study material, but this would be essential to establish whether Xiao et al. (2015) worked with polyploid material, or whether the flow cytometry estimation is in error.

Based on molecular data, a WGD analysis by Feng et al. (2020) for *Primulina* suggested two duplications: an ancient Lamiales specific WGD event (*L* event; *Ks* range 0.640–1.407) and a *Primulina* (subtribe Didymocarpinae) specific event (*D* event; *Ks* range 0.050–0.302). Our WGD analyses also suggested a recent WGD event for *S. rexii* (Figure 2), where the distribution of *Ks* values showed a sharp peak at around ∼0.2, in a very similar position to the *D* event of Feng et al. (2020). The peak of *Ks* values of orthologous gene pairs between *S. rexii* and *P. huaijiensis* was also around 0.2, and thus the WGD *D* event might be shared between *Primulina* and *Streptocarpus*, although future studies are required for confirmation. Interestingly, for *D. hygrometricum*, no *D* event was inferred based on the *Ks* values for the anchor gene sets (Feng et al., 2020), as well as the 4DTv analysis (Xiao et al., 2015). However, only half of the gene models were annotated in the currently available genome assembly for *D. hygrometricum*, which might have affected the analysis. The contiguity of the *D. hygrometricum* genome is also low and thus a very small amount of anchor gene pairs were detected when it was compared with the whole paranome of the transcriptome gene set. Thus, a better genome assembly with cytological support might be required to understand whether a WGD event for *D. hygrometricum* exists.

4.3 | The PLCL pipeline

There are a number of pipelines available to assemble organellar genomes, such the short read assemblers NOVOPlasty and GetOrganelle, which rely on seed sequences (Dierckxsens et al., 2017; Jin et al., 2020). With long read sequences, a de novo organellar genome assembly is possible and might be the only option when there are no reliable reference sequences available. For example, for the *Eucalyptus pauciflora* chloroplast genome reads were extracted with the Blasr v.5.1 aligner on the reference sequence, prior to *de novo* assembly (Wang et al., 2018). However, the authors (Wang et al., 2018) reported that it did not assemble the entire chloroplast genome into a single contig, and it required manual curation. Organelle_PBA (Soorni et al., 2017) is a long read sequence assembler for chloroplast and mitochondrial genomes. It also selects the chloroplast or mitochondrial reads at the first step against a reference sequence, and the reads are then assembled. It requires scaffolding of assembled contigs using SSPACE-LongRead (Boetzer & Pirovano, 2014) in the final step. Our method, on the other hand, was able to assemble the chloroplast into a single contig at the assembly step. The different coverage observed when raw reads were remapped to the *S. rexii* genome assembly likely reflects the chloroplast and mitochondrion copy number in the sequencing libraries. For instance, two nuclear genomes exist in a diploid cell, but each chloroplast contains multiple genome copies and each cell contains several chloroplasts, and it was estimated that ∼1,200 copies of the chloroplast genome are contained per mesophyll cell in *A. thaliana* (Sakamoto & Takami, 2018). The mitochondrial genome has a much more complex nature. It varies in structure and can be linear or circular, and often only a subset of mitochondrial genomes are kept in a cell depending on the cell condition in *A. thaliana* (Arimura, 2018). Unlike chloroplasts, mitochondrial genome copy number could be much smaller than the number of mitochondria themselves (Arimura, 2018), though the copy number per cell is still higher than for the nuclear genome. In an *A. thaliana* leaf for example, it was estimated that there may be 50 mitochondrial genome copies per cell (Cai et al., 2015). Thus, in a whole genome assembly, the contig coverage is expected to be "nuclear < mitochondria < chloroplast," and this was demonstrated to be the case in this study. Because our method maps all reads to the whole genome, it has the advantage of eliminating cross-mapped reads between organelles.

The method applies simple blastn searches using a chloroplast or mitochondrial genome as the query sequence, and it was often the case that the same contigs were detected by both chloroplast and mitochondria searches. However, these were distinguishable by differences in the contigs' coverage and/or blastn bitscore values. It might also be an option to use an aligner such as minimap2 to map contigs to the query instead of blastn. The pipeline, which we call the PLCL pipeline, worked well for the chloroplast genome assembly of the plant species *S. rexii*, *A. thaliana*, and *O. sativa*, where perhaps the very high coverage helped.

For the mitochondrial genome assemblies, although the *S. rexii* ONT dataset (>158 Gbp) was assembled into one contig, the *A. thaliana* KBS-Mac-74 ONT dataset (MinION, 3 Gbp; Michael et al., 2018) and *O. sativa* subsp. *indica* IR64 ONT dataset (MinION, 9 Gbp; Tanaka et al., 2020) did not have sufficient coverage. The number of mitochondrial reads was higher in *A. thaliana* KBS-Mac-74 and *O. sativa* subsp. *indica* IR64 than in *S. rexii*, but the read length was longer in *S. rexii* ("Read N50" in Table 4), and thus the resulting coverage was higher in *S.rexii*. In addition, longer reads are known to improve the assembly's contiguity. The pipeline was able to assemble the *S. rexii* mitochondrial genome into one linear scaffold (Table 4). The accuracy of the assembled *S. rexii* mitochondrial genome was supported as well assembled in series with its original mitochondrial contigs (Figure S9). Mitochondrial genomes can exist as linear as well as circular forms (Arimura, 2018), and thus the *S. rexii* linear mitochondrial genome is likely correct and complete, rather than an incomplete assembly. It is difficult to judge the accuracy of the assembly of the mitochondrial genome from comparison with reference genomes, such as *A. thaliana*, because of the lack of conservation of gene order. This is well-known due to frequent mitochondrial fission and fusion events (Arimura, 2018; Kozik et al., 2019) that inhibits the alignment even between closely related genera such as *Arabidopsis* and *Capsella* (Figure S8d).

Thus, the PLCL pipeline appears to be coverage sensitive and requires high-volume input data. In addition, slight differences in the Canu parameter settings lead to differences in indels, and thus users are advised to test several assembler or parameter settings. The assembly, and this is perhaps true for any traditional long read assembly, might not be very accurate for detecting SNPs in mononucleotide repeat regions. In cases where researchers are interested in studying short repeats, it might be useful to polish the assembly with more accurate Illumina short reads. In addition, developments with more accurate long reads and improved algorithms for ONT base calling (https://nanoporetech.com/accuracy), as well as novel high fidelity PacBio HiFi methods, will hopefully greatly reduce sequence error rate in the future. Regardless of future developments, the PLCL pipeline has great potential as an automated chloroplast and mitochondrial genome assembler in projects where many large long read datasets are generated.

5 | CONCLUSIONS

In this study, we assembled the whole genome of *S. rexii*, including the chloroplast, mitochondrial, and nuclear genome. The estimated 929-Mb nuclear genome was assembled solely with the long read sequences into 5,855 scaffolds covering 766 Mb (83%). This provides a valuable example of a cost effective approach to generate a complete genome for a non-model organism. It is now possible to carry out genome-wide studies in *Streptocarpus* to pin down genes regulating its unique meristems. In addition, the resources and methods developed in this study were shown to be more widely applicable. The annotation resources are immediately useful for the genome annotation of Old World Gesneriaceae. The PLCL pipeline provides an effective chloroplast and mitochondrial genome assembly method with wider applications in plants. The DNA extraction method developed here is also likely to prove useful for other polysaccharide-rich plant species. This study provides not only a new genome but also shows the potential of long read sequencing methods to the wider scientific community.

ACKNOWLEDGMENTS

We acknowledge the constructive and positive comments of two anonymous reviewers. The authors greatly thank Edinburgh Genomics and Mark Blaxter for enabling the acquisition of Oxford Nanopore PromethION sequencing data. We especially thank Aureliano Bombarely for useful advice on genome assembly methods. Catherine Kidner, Lucia Campos, and Flávia Pezzini provided general advice on NGS methodologies. The authors received generous support in plant cultivation from the horticulture division at the Royal Botanic Garden Edinburgh (RBGE), and we thank in particular (in alphabetical order) Andrew Ensoll, Louise Galloway, Gunnar Ovstebo, Hazel France, and Stephen Willis. We also thank at the RBGE Science and Technical Services division, Laura Forrest, Ruth Hollands, and Frieda Christie for laboratory support and DTI division and Duncan Reddish for the Linux server support. Pete Hollingsworth, Mark Newman (RBGE) and Akitoshi Iwamoto (Kanagawa University) are thanked for general support and facilitating the research associateships for KN at RBGE and Kanagawa University. KN received support from the Edinburgh Botanic Garden (Sibbald) Trust (2018#18), the Japan Society for the Promotion of Science (JSPS KAKENHI Grant Number 15K18593; 18K06375), and the Sumitomo Foundation (170204). Computation and genome assemblies were performed on the NIG supercomputer at ROIS National Institute of Genetics, Shizuoka, Japan. Iain Milne is thanked for organizing access to the CropDiversity server, James Hutton Institute, Dundee, UK. The authors further thank the NGS community on the web for supporting the many analysis programs running on the servers. RBGE is supported by the Rural and Environment Science and Analytical Services Division (RESAS) of the Scottish Government.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.

DATA AVAILABILITY STATEMENT

The *Streptocarpus rexii* genome assembly and raw reads have been deposited in the European Nucleotide Archive (ENA; Study ID: PRJEB50530, Sample ID: ERS10392821, Genome assembly ID: GCA_929616715, Reads ID: ERR8482532, ERR8484618), and registered in the genome browser of Plant GARDEN (https://plantgarden.jp/en/list/t121488). The transcriptomes are also deposited in ENA (*Streptocarpus ionanthus*: ERZ5128686, *Haberlea rhodopensis*: ERZ5129963. The genome repeat libraries (Data S1) have been deposited in Dfam (https://dfam.org/home) and will be available in a future release. The high-molecular weight (HMW) DNA extraction protocol and PLCL pipeline are deposited in protocol.io (HMW DNA extraction: dx.doi.org/10.17504/protocols.io.bempjc5n, PLCL pipeline: dx.doi.org/10.17504/protocols.io.bx5dpq26).

ORCID

Kanae Nishii https://orcid.org/0000-0001-8141-8908
Michelle Hart https://orcid.org/0000-0001-9503-7786
Alex D. Twyford https://orcid.org/0000-0002-8746-6617
Michael Möller https://orcid.org/0000-0002-2819-0323

REFERENCES

Ai, B., Gao, Y., Zhang, X., Tao, J., Kang, M., & Huang, H. (2015). Comparative transcriptome resources of eleven *Primulina* species, a group of 'stone plants' from a biodiversity hot spot. *Molecular Ecology Resources*, *15*, 619–632. https://doi.org/10.1111/1755-0998.12333

Altschul, S. F., Boguski, M. S., Gish, W., & Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genetics*, *6*, 119–129. https://doi.org/10.1038/ng0294-119

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Arimura, S. I. (2018). Fission and fusion of plant mitochondria, and genome maintenance. *Plant Physiology*, *176*, 152–161. https://doi.org/10.1104/pp.17.01025

Beuttel, E. (1939). Bastardierungsversuche in der Gattung *Streptocarpus* Lindl. II. *Die Heterosis Bei Streptocarpushybriden, Zeitschrift für Botanik*, *35*, 49–91.

Boetzer, M., & Pirovano, W. (2014). SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*, *15*, 211. https://doi.org/10.1186/1471-2105-15-211

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brudno, M., Do, C.-B., Cooper, G. M., Kim, M.-F., Davydov, E., Program, N. C. S., Green, E. D., Sidow, A., & Batzoglou, S. (2003). LAGAN and multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, *13*, 721–731. https://doi.org/10.1101/gr.926603

Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958. https://doi.org/10.7717/peerj.4958

Cai, Q., Guo, L., Shen, Z.-R., Wang, D.-Y., Zhang, Q., & Sodmergen (2015). Elevation of pollen mitochondrial DNA copy number by *WHIRLY2*: Altered respiration and pollen tube growth in *Arabidopsis*. *Plant Physiology*, *169*, 660–673. https://doi.org/10.1104/pp.15.00437

Chen, Y.-Y., Nishii, K., Barber, S., Hackett, C., Kidner, C. A., Gharbi, K., Nagano, A. J., Iwamoto, A., & Möller, M. (2018). A first genetic map in the genus *Streptocarpus* generated with RAD sequencing based SNP markers. *South African Journal of Botany*, *117*, 158–168. https://doi.org/10.1016/j.sajb.2018.05.009

Chen, Y.-Y., Nishii, K., Kidner, C. A., Hackett, C. A., & Möller, M. (2020). QTL dissection of floral traits in *Streptocarpus* (Gesneriaceae). *Euphytica*, *216*, 110. https://doi.org/10.1007/s10681-020-02647-1

Chiara, M., Horner, D. S., & Spada, A. (2013). *De novo* assembly of the transcriptome of the non-model plant *Streptocarpus rexii* employing a novel heuristic to recover locus-specific transcript clusters. *PLoS ONE*, *8*, e80961. https://doi.org/10.1371/journal.pone.0080961

Crocker, C. W. (1861). Notes on the germination of certain species of Cyrtandreae. *Journal of the Proceedings of the Linnean Society, Botany*, *5*, 65–66. https://doi.org/10.1111/j.1095-8312.1860.tb01039.x

da Silva, J. A. T., Dewir, Y. H., Wicaksono, A., Sahijram, L., Kim, H., Zeng, S., Chandler, S. F., & Hosokawa, M. (2017). African violet (*Saintpaulia ionantha* H. Wendl.): Classical breeding and progress in the application of biotechnological techniques. *Folia Horticulture*, *29*, 99–111. https://doi.org/10.1515/fhort-2017-0010

de Bodt, S., Maere, S., & van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*, *20*, 591–597. https://doi.org/10.1016/j.tree.2005.07.008

de Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics*, *34*, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., May, G., & Lin, H. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nature Communications*, *9*, 4844. https://doi.org/10.1038/s41467-018-07271-1

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45*, e18. https://doi.org/10.1093/nar/gkw955

Dodsworth, S., Leitch, A. R., & Leitch, I. J. (2015). Genome size diversity in angiosperms and its influence on gene space. *Current Opinion in Genetics & Development*, *35*, 73–78. https://doi.org/10.1016/j.gde.2015.10.006

Doyle, J. J. (1991). DNA protocols for plants. In G. M. Hewitt, A. W. B. Johnston, & J. P. W. Young (Eds.), *Molecular techniques in taxonomy. NATO ASI series (series H: Cell biology)* (Vol. 57) (pp. 283–293). Springer Verlag.

Doyle, J. J., & Doyle, D. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, *19*, 11–15.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamin, M. S., Machol, I., Lander, E. S., Presser-Aiden, A., & Lieberman-Aiden, E. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-level scaffolds. *Science*, *356*, 92–95. https://doi.org/10.1126/science.aal3327

Dumschott, K., Schmidt, M. H., Chawla, H. S., Snowdon, R., & Usadel, B. (2020). Oxford nanopore sequencing: New opportunities for plant genomics? *Journal of Experimental Botany*, *71*, 5313–5322. https://doi.org/10.1093/jxb/eraa263

Feng, C., Feng, C., & Kang, M. (2016). The first genetic linkage map of *Primulina eburnea* (Gesneriaceae) based on EST-derived SNP markers. *Journal of Genetics*, *95*, 377–382. https://doi.org/10.1007/s12041-016-0650-1

Feng, C., Feng, C., Yang, L., Kang, M., & Rausher, M. D. (2019). Genetic architecture of quantitative flower and leaf traits in a pair of sympatric sister species of *Primulina*. *Heredity*, *122*, 864–876. https://doi.org/10.1038/s41437-018-0170-2

Feng, C., Wang, J., Wu, L.-Q., Kong, H.-H., Yang, L.-H., Feng, C., Wang, K., Rausher, M., & Kang, M. (2020). The genome of a cave plant, *Primulina huaijiensis*, provides insights into adaptation to limestone karst habitats. *The New Phytologist*, *227*, 1249–1263. https://doi.org/10.1111/nph.16588

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 9451–9457. https://doi.org/10.1073/pnas.1921046117

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. *Nucleic Acids Research*, *32*, W273–W279. https://doi.org/10.1093/nar/gkh458

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*, 644–652. https://doi.org/10.1038/nbt.1883

Gunter, L. (2015). Populus nuclear DNA purification using the QIAGEN genomic-tip 100/G. http://1ofdmq2n8tc36m36i46scovo32e.wpengine.netdna-cdn.com/wp-content/uploads/2015/2002/Populus-nuclear-DNA-purification-with-Qiagen-Genomic-tip-2100.pdf

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Hilliard, O. M., & Burtt, B. L. (1971). *Streptocarpus. An African plant study*. Natal University Press.

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491. https://doi.org/10.1186/1471-2105-12-491

Imaichi, R., Nagumo, S., & Kato, M. (2000). Ontogenetic anatomy of *Streptocarpus grandis* (Gesneriaceae) with implications for evolution of monophylly. *Annals of Botany*, *86*, 37–46. https://doi.org/10.1006/anbo.2000.1155

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., & Li, D.-Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*, 241. https://doi.org/10.1186/s13059-020-02154-5

Jong, K. (1970). *Developmental aspects of vegetative morphology of Streptocarpus*. PhD Thesis. University of Edinburgh.

Jong, K., & Burtt, B. L. (1975). The evolution of morphological novelty exemplified in the growth patterns of some Gesneriaceae. *New Phytologist*, *75*, 297–311. https://doi.org/10.1111/j.1469-8137.1975.tb01400.x

Jong, K., & Möller, M. (2000). New chromosome counts in *Streptocarpus* (Gesneriaceae) from Madagascar and the Comoro Islands and their taxonomic significance. *Plant Systematics and Evolution*, *224*, 173–182. https://doi.org/10.1007/BF00986341

Jung, H., Ventura, T., Chung, J.-S., Kim, W.-J., Nam, B.-H., Kong, H.-J., Kim, Y.-O., Jeon, M.-S., & Eyun, S.-I. (2020). Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Computational Biology*, *16*, e1008325. https://doi.org/10.1371/journal.pcbi.1008325

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780. https://doi.org/10.1093/molbev/mst010

Kejnovsky, E., Hawkins, J. S., & Feschotte, C. (2012). Plant transposable elements: Biology and evolution. In J. F. Wendel, J. Greilhuber, I. J. Leitch, & J. Dolezel (Eds.), *Diversity of genomes in plants* (pp. 17–34). Springer Verlag. https://doi.org/10.1007/978-3-7091-1130-7_2

Kiehn, M., Hellmayr, E., & Weber, A. (1998). Chromosome numbers of Malayan and other paleotropical Gesneriaceae. I. Tribe Didymocarpeae. *Beiträge Zur Biologie der Pflanzen*, *70*, 407–444.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research*, *27*, 722–736. https://doi.org/10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59. https://doi.org/10.1186/1471-2105-5-59

Kozik, A., Rowan, B. A., Lavelle, D., Berke, L., Schranz, M. E., Michaelmore, R. W., & Christensen, A. C. (2019). The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genetics*, *15*, e1008373. https://doi.org/10.1371/journal.pgen.1008373

Kushikawa, S., Hoshino, Y., & Mii, M. (2001). *Agrobacterium*-mediated transformation of *Saintpaulia ionantha* Wendl. *Plant Science*, *161*, 953–960. https://doi.org/10.1016/S0168-9452(01)00496-4

Laetsch, D.-R., & Blaxter, M.-L. (2017). BlobTools: Interrogation of genome assemblies. *F1000 Research*, *6*, 1287. https://doi.org/10.12688/f1000research.12232.1

Li, F. W., & Harkess, A. (2018). A guide to sequence your favorite plant genomes. *Applications in Plant Sciences*, *6*, e1030. https://doi.org/10.1002/aps3.1030

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, H., Wei, J.-P., Yang, T., Mu, W.-X., Song, B., Yang, T., Fu, Y., Wang, X.-B., Hu, G.-H., Li, W.-S., Zhou, H.-C., Chang, Y., Chen, X.-L., Chen, H.-Y., Cheng, L., He, X.-F., Cai, H.-C., Cai, X.-C., Wang, M., … Liu, X. (2019). Molecular digitization of a botanical garden: High-depth whole-genome sequencing of 689 vascular plant species from the Ruili botanical garden. *Gigascience*, *8*, 1–9. https://doi.org/10.1093/gigascience/giz007

Mantegazza, R., Möller, M., Harrison, C. J., Fior, S., de Luca, C., & Spada, A. (2007). Anisocotyly and meristem initiation in an unorthodox plant, *Streptocarpus rexii* (Gesneriaceae). *Planta*, *225*, 653–663. https://doi.org/10.1007/s00425-006-0389-7

Mantegazza, R., Tononi, P., Möller, M., & Spada, A. (2009). *WUS* and *STM* homologs are linked to the expression of lateral dominance in the acaulescent *Streptocarpus rexii* (Gesneriaceae). *Planta*, *230*, 529–542. https://doi.org/10.1007/s00425-009-0965-8

Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-read sequencing emerging in medical genetics. *Frontiers in Genetics*, *10*, 426. https://doi.org/10.3389/fgene.2019.00426

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., Loudet, O., Weigel, D., & Ecker, J. R. (2018). High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nature Communications*, *9*, 541. https://doi.org/10.1038/s41467-018-03016-2

Mitra, J., Xu, G., Wang, B., Li, M., & Deng, X. (2013). Understanding desiccation tolerance using the resurrection plant *Boea hygrometrica* as a model system. *Frontiers in Plant Science*, *4*, 446. https://doi.org/10.3389/fpls.2013.00446

Möller, M. (2018). Nuclear DNA C-values are correlated with pollen size at tetraploid but not diploid level and linked to phylogenetic descent in *Streptocarpus* (Gesneriaceae). *South African Journal of Botany*, *114*, 323–344. https://doi.org/10.1016/j.sajb.2017.11.017

Möller, M., Barber, S., Atkins, H. J., & Purvis, D. A. (2019). The living collection at the Royal Botanic Garden Edinburgh illustrates the floral diversity in *Streptocarpus* (Gesneriaceae). *Sibbaldia*, *17*, 155–175. https://doi.org/10.24823/Sibbaldia.2019.272

Möller, M., & Cronk, Q. C. B. (2001). Evolution of morphological novelty: A phylogenetic analysis of growth patterns in *Streptocarpus* (Gesneriaceae). *Evolution*, *55*, 918–929. https://doi.org/10.1554/0014-3820(2001)055[0918:eomnap]2.0.co;2

Nishii, K., Fei, Y., Hudson, A., Möller, M., & Molnar, A. (2020). Virus-induced gene silencing in *Streptocarpus rexii* (Gesneriaceae). *Molecular Biotechnology*, *62*, 317–325. https://doi.org/10.1007/s12033-020-00248-w

Nishii, K., Huang, B.-H., Wang, C.-N., & Möller, M. (2017). From shoot to leaf: Step-wise shifts in meristem and *KNOX1* activity correlate with the evolution of a unifoliate body plan in Gesneriaceae. *Development, Genes and Evolution*, *227*, 41–60. https://doi.org/10.1007/s00427-016-0568-x

Nishii, K., Hughes, M., Briggs, M., Haston, E., Christie, F., DeVilliers, M. J., Hanekom, T., Roos, W. G., Bellstedt, D. U., & Möller, M. (2015). *Streptocarpus* redefined to include all afro-Malagasy Gesneriaceae: Molecular phylogenies prove congruent with geography and cytology and uncovers remarkable morphological homoplasies. *Taxon*, *64*, 1243–1274. https://doi.org/10.12705/646.8

Nishii, K., Kuwabara, A., & Nagata, T. (2004). Characterization of anisocotylous leaf formation in *Streptocarpus wendlandii* (Gesneriaceae): Significance of plant growth regulators. *Annals of Botany*, *94*, 457–467. https://doi.org/10.1093/aob/mch160

Nishii, K., Möller, M., Kidner, C. A., Spada, A., Mantegazza, R., Wang, C.-N., & Nagata, T. (2010). A complex case of simple leaves: Indeterminate leaves co-express *ARP* and *KNOX1* genes. *Development, Genes and Evolution*, *220*, 25–40. https://doi.org/10.1007/s00427-010-0326-4

Nishii, K., & Nagata, T. (2007). Developmental analyses of the phyllomorph formation in the rosulate species *Streptocarpus rexii* (Gesneriaceae). *Plant Systematics and Evolution*, *265*, 135–145. https://doi.org/10.1007/s00606-007-0515-4

Oehlkers, F. (1964). Cytoplasmic inheritance in the genus *Streptocarpus* Lindley. *Advances in Genetics*, *12*, 329–370. https://doi.org/10.1016/S0065-2660(08)60418-6

One Thousand Plant Transcriptomes Initiative. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, *574*, 679–685. https://doi.org/10.1038/s41586-019-1693-2

PacBio Sample Net. (2015). Preparing *Arabidopsis* genomic DNA for size-selected ∼20 kb SMRTbell™ Libraries. Accessed year: 2018. https://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf.

Proost, S., Fostier, J., de Witte, D., Dhoedt, B., Demeester, P., van de Peer, Y., & Vandepoele, K. (2012). i-ADHoRe 3.0-fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research*, *40*, e11. https://doi.org/10.1093/nar/gkr955

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, *19*, 1–11. https://doi.org/10.1186/s13059-018-1462-9

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, *17*, 155–158. https://doi.org/10.1038/s41592-019-0669-3

Sakamoto, W., & Takami, T. (2018). Chloroplast DNA dynamics: Copy number, quality control and degradation. *Plant & Cell Physiology*, *59*, 1120–1127. https://doi.org/10.1093/pcp/pcy084

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit, A., Hubley, R., & Green, P. (2013–2015). RepeatMasker Open-4.0. http://www.repeatmasker.org

Soorni, A., Haak, D., Zaitlin, D., & Bombarely, A. (2017). Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial reads from PacBio DNA sequencing data. *BMC Genomics*, *18*, 49. https://doi.org/10.1186/s12864-016-3412-9

Souza, H., Muller, L., Brandao, R., & Lovato, M. (2012). Isolation of high quality and polysaccharide-free DNA from leaves of *Dimorphandra mollis* (Leguminosae), a tree from the Brazilian Cerrado. *Genetics and Molecular Research*, *11*, 756–764. https://doi.org/10.4238/2012.March.22.6

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*, 637–644. https://doi.org/10.1093/bioinformatics/btn013

Steeves, T. A., & Sussex, I. M. (1989). *Patterns in plant development*. Cambridge University Press. https://doi.org/10.1017/CBO9780511626227

Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., & Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, *12*, 2. https://doi.org/10.1186/s13100-020-00230-y

Sun, R. Z., Liu, J., Wang, Y.-Y., & Deng, X. (2021). DNA methylation-mediated modulation of rapid desiccation tolerance acquisition and dehydration stress memory in the resurrection plant *Boea hygrometrica*. *PLoS Genetics*, *17*, e1009549. https://doi.org/10.1371/journal.pgen.1009549

Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. *Genome Biology*, *19*(1), 1–12. https://doi.org/10.1186/s13059-018-1520-3

Tanaka, T., Nishijima, R., Teramoto, S., Kitomi, Y., Hayashi, T., Uga, Y., & Kawakatsu, T. (2020). De novo genome assembly of the indica rice variety IR64 using linked-read sequencing and nanopore sequencing. *G3: Genes, Genomes, Genetics*, *10*, 1495–1501. https://doi.org/10.1534/g3.119.400871

The GIMP Development Team. (2019). GIMP, available at: https://www.gimp.org

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq—Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*, W6–W11. https://doi.org/10.1093/nar/gkx391

Tononi, P., Möller, M., Bencivenga, S., & Spada, A. (2010). *GRAMINIFOLIA* homolog expression in *Streptocarpus rexii* is associated with the basal meristems in phyllomorphs, a morphological novelty in Gesneriaceae. *Evolution & Development*, *12*, 61–73. https://doi.org/10.1111/j.1525-142X.2009.00391.x

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, *47*, 10994–11006. https://doi.org/10.1093/nar/gkz841

Vandepoele, K., de Vos, W., Taylor, J. S., Meyer, A., & van de Peer, Y. (2004). Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 1638–1643. https://doi.org/10.1073/pnas.0307968100

Vaser, R., Sovic, I., Nagarajan, N., & Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*, 737–746. https://doi.org/10.1101/gr.214270.116

Wang, W., Schalamun, M., Morales-Suarez, A., Kainer, D., Schwessinger, B., & Lanfear, R. (2018). Assembly of chloroplast genomes with long- and short-read data: A comparison of approaches using *Eucalyptus pauciflora* as a test case. *BMC Genomics*, *19*, 977. https://doi.org/10.1186/s12864-018-5348-8

Weber, A., Clark, J. L., & Möller, M. (2013). A new formal classification of Gesneriaceae. *Selbyana*, *31*, 68–94. https://journals.flvc.org/selbyana/article/view/123016

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org, https://doi.org/10.1007/978-3-319-24277-4

Xiao, L., Yang, G., Zhang, L., Yang, X., Zhao, S., Ji, Z., Zhou, Q., Hu, M., Wang, Y., Chen, M., Xu, Y., Jin, H., Xiao, X., Hu, G., Bao, F., Hu, Y., Wan, P., Li, L., Deng, X., … He, Y. (2015). The resurrection genome of *Boea hygrometrica*: A blueprint for survival of dehydration. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 5833–5837. https://doi.org/10.1073/pnas.1505811112

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088

Zerpa-Catanho, D., Zhang, X., Song, J., Hernandez, A. G., & Ming, R. (2021). Ultra-long DNA molecule isolation from plant nuclei for ultra-long read genome sequencing. *STAR Protocols*, *2*, 100343. https://doi.org/10.1016/j.xpro.2021.100343

Zhang, T.-W., Fang, Y.-J., Wang, X.-M., Deng, X., Zhang, X.-W., Hu, S.-N., & Yu, J. (2012). The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: Insights into the evolution of plant organellar genomes. *PLoS ONE*, *7*, e30531. https://doi.org/10.1371/journal.pone.0030531

Zhao, Y., Xu, T., Shen, C.-Y., Xu, G.-H., Chen, S.-X., Song, L.-Z., Li, M.-J., Wang, L.-L., Zhu, Y., Lv, W.-T., Gong, Z.-Z., Liu, C.-M., & Deng, X. (2014). Identification of a retroelement from the resurrection plant *Boea hygrometrica* osmotic and alkaline tolerance in *Arabidopsis thaliana*. *PLoS ONE*, *9*, e98098. https://doi.org/10.1371/journal.pone.0098098

Zwaenepoel, A., & van de Peer, Y. (2019). wgd—Simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics*, *35*, 2153–2155. https://doi.org/10.1093/bioinformatics/bty915

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Nishii, K., Hart, M., Kelso, N., Barber, S., Chen, Y.-Y., Thomson, M., Trivedi, U., Twyford, A. D., & Möller, M. (2022). The first genome for the Cape Primrose *Streptocarpus rexii* (Gesneriaceae), a model plant for studying meristem-driven shoot diversity. *Plant Direct*, *6*(4), e388. https://doi.org/10.1002/pld3.388