Plastid Genome Assembly Using Long-read Data (ptGAUL)

Wenbin Zhou[1*], Carolina E. Armijos[4], Chaehee Lee[5], Ruisen Lu[6], Jeremy Wang[3], Tracey A. Ruhlman[7], Robert K. Jansen[7], Alan M. Jones[1,2], Corbin D. Jones[1,3]

Departments of [1] Biology, [2]Pharmacology, [3]Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [4]Laboratorio de Biotecnología Vegetal, Universidad San Francisco de Quito USFQ, Diego de Robles s/n y Vía Interocéanica, Quito, 170901, Ecuador. [5]Department of Plant Sciences, University of California, Davis, Davis, CA 95616, USA. [6]Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing 210014, China. [7]Department of Integrative Biology, University of Texas at Austin, TX 78712, USA.

*Corresponding authors e-mail address: wenbin.evolution@gmail.com.

Abstract

Although plastid genome (plastome) structure is highly conserved across most seed plants, investigations during the past two decades revealed several disparately related lineages that experienced substantial rearrangements. Most plastomes contain a large, inverted repeat and two single-copy regions and few dispersed repeats, however the plastomes of some taxa harbor long repeat sequences (>300 bp). These long repeats make it difficult to assemble complete plastomes using short-read data leading to misassemblies and consensus sequences that have spurious rearrangements. Single-molecule, long-read sequencing has the potential to overcome these challenges, yet there is no consensus on the most effective method for accurately assembling plastomes using long-read data. We generated a pipeline, ***p***las***t***id ***G***enome ***A***ssembly ***U***sing ***L***ong-read data (ptGAUL), to address the problem of plastome assembly using long-read data from Oxford Nanopore Technologies (ONT) or Pacific Biosciences platforms. We demonstrated the efficacy of the ptGAUL pipeline using 16 published long-read datasets. We showed that ptGAUL produces accurate and unbiased assemblies. Additionally, we employed ptGAUL to assemble four new *Juncus* (Juncaceae) plastomes using ONT long reads. Our results revealed many long repeats and rearrangements in *Juncus* plastomes compared with basal lineages of Poales.

KEY WORDS

Long-read assembly, chloroplast, rearrangement events, *Juncus*, Juncaceae, Poales

1 INTRODUCTION

Plastid genomes (plastomes) are highly conserved, comprising linear, branched or occasionally circular molecules that usually contain a large, inverted repeat (IR) and large and small single-copy regions (LSC and SSC). Due to their conserved structure and low rate of nucleotide substitution, plastome data has made substantial contributions to phylogenetic studies for many plant groups (Jansen & Ruhlman, 2012; Jiang et al., 2022; Liu et al., 2022; Xia et al., 2022; Xu et al., 2022; Yu et al., 2022). Despite the high level of plastome structural conservation in seed plants, rearrangements, including inversions, expansion and contraction of the IR, and IR loss occurred in unrelated lineages of gymnosperms and angiosperms (Ruhlman & Jansen, 2021). Many of these same lineages experienced substantial gene loss with most of these genes functionally transferred to the nuclear genome or substituted by an alternative, nuclear-encoded gene (Ruhlman & Jansen, 2021). Documented transferred/substituted genes include *accD, infA, rpl22, rpl20, rpl32, rpl23, rps7, rps16, ycf1* and *ycf2*.

Genome assembly methods have improved substantially over the past decade (Twyford & Ness, 2017; Freudenthal et al., 2020). NOVOPlasty (Dierckxsens et al., 2017) and GetOrganelle (Jin et al., 2020) are the two most frequently used pipelines for plastome assembly based on Illumina short reads. However these assemblers, which rely on the De Bruijn graph approach (Compeau et al., 2011), do not always yield accurate assembly results when confronted with long repeat regions in plastomes, particularly when those repeats are longer than kmer sizes. In some cases, these tools generate outputs with multiple contigs/scaffolds or hundreds of possible assembly results. The high number of uncertain paths can sometimes be corrected using Bandage (Wick et al., 2015), a software tool that visualizes the depth of read coverage for each contig/scaffold and orders contigs, but the final

arrangement of the contigs is often not well-resolved because the Illumina short reads are insufficient to bridge the repeated sequences and their flanking regions. Using short reads with a typical insert size (300-400 bp) is insufficient to obtain a complete plastome assembly for plant species that have large repeats and may be highly rearranged. So far, few plant systematists have recognized this as an issue likely because most plants possess relatively conservative plastome structures with limited repeated sequences and because their primary interest is the extraction of coding sequences for phylogenetic analyses. Long reads generated by third-generation sequencing methods such as Oxford Nanopore Technologies (ONT) or Pacific Biosciences (Pacbio) platforms may help resolve assembly issues as the longer reads are more likely to span long repeats (Liao et al., 2021).

To date, many tools have been developed to assemble organelle genomes using long-read data and hybrid data (both short- and long-read data), including Organelle_PBA (Soorni et al., 2017), Canu (Koren et al., 2017), Unicycler (Wick et al., 2017), and Flye (Kolmogorov et al., 2019; Syme et al., 2021). However, these pipelines for plastome assembly using long-read or hybrid data have some drawbacks. Organelle_PBA was designed exclusively for PacBio data; the Sprai (Miyamoto et al., 2014) and Celera (Miller et al., 2008) assemblers in Organelle_PBA are no longer maintained limiting its extension to assembly with hybrid datasets. The approach of Syme et al. (2021) requires an extra step to manually filter a subset of raw reads matching the plastome (~250X coverage) and sometimes generates multiple contigs in the assembly result. Canu can generate different results depending on different read coverages (Wang et al., 2018). Unicycler was designed for hybrid data, however it takes an extremely long time to finish as input data is increased. All pipelines can likely assemble the conventional plastome, but are not able to assemble atypical plastome structures accurately.

The angiosperm family Juncaceae contains ~500 species within the seven genera *Juncus* L., *Luzula* DC., *Distichia* Nees and Meyen, *Oxychloe* Philippi, *Patosia* Buchenau, *Marsippospermum* Desv. and *Rostkovia* Desv. (Drábková, 2010). *Juncus* is the largest genus and includes ~300 species (Balslev, 2018) and two major subgenera, *Agathryon* and *Juncus* (Drábková et al., 2006; Drábková, 2010). Although many species of Juncaceae have been included in phylogenetic studies using plastid gene sequences and the internal transcribed spacer region of the nuclear ribosomal repeat (Table S1; Drábková et al., 2006; Drábková, 2010; Brožová et al., 2022), species relationships within *Juncus* remain unresolved. Recently, Brožová et al. (2022) incorporated *rbcL*, *trnL*, *trnL-trnF*, and ITS1-5.8-ITS2 region to reorganize *Juncus* into seven distinct genera: *Juncus*, *Verojuncus*, *Juncinella*, *Alpinojuncus*, *Australojuncus*, *Boreojuncus*, and *Agathryon*.

Not many plastome structures have been reported for *Juncus* (*s.l.*). To avoid the confusion regarding the species names, we did not adopt these latest genera above in our study because a more comprehensive study with more markers is necessary to justify this reranking. So far, the plastome structures are seldom studied in *Juncus* (*s.l.*). Plastomes of just eight *Juncus* species are publicly available in Genbank (Wu et al., 2021; Lu et al., 2021). For example, the focus of Wu et al. (2021) was phylogenetic relationships in the Poales using shared, plastid protein-coding genes and no information was reported on plastome structure. Lu et al., (2021) assembled the plastome of *Juncus effusus* using Velvet (Zerbino, 2010) and NOVOPlasty (Dierckxsens et al., 2017) with GapFiller (Nadalin et al., 2012) without any confirmation by either long range PCR or long-read data leaving the final structure uncertain. Recently, two more *Juncus* (*J. effusus* and *J. inflexus*) nuclear genomes were assembled by Planta et al. (2022), but no plastomes were reported. Adding more complete plastomes of

106 Juncaceae would allow insight into plastome evolution in the family and help gain more
107 phylogenetic insights within Juncaceae and Poales.

108 To assist in assembling potentially complex plastomes and to explore structural
109 variation in *Juncus*, we created a pipeline, ***p***las***t***id ***G***enome ***A***ssembly ***U***sing ***L***ong-read data
110 (ptGAUL), which assembles plastomes using raw ONT long-read sequencing data. The aims
111 of the study were: 1) test the reliability of the ptGAUL pipeline using 16 published plastomes;
112 2) employ the pipeline to assemble plastomes of two *Juncus* species (*J. validus* and *J.*
113 *roemerianus*) sequenced in our study, and assemble two other species (*J. effusus* and *J.*
114 *inflexus*) from the reads of Planta et al. (2022); and 3) compare plastome evolution in *Juncus*
115 to selected members of the Poales.

116

117 **2 MATERIALS AND METHODS**

118 **2.1 *Juncus* sample collection and DNA extraction**

119 Young leaves of *Juncus roemerianus* (voucher number: NCU00441655) and *Juncus validus*
120 (voucher number: NCU00434802) were collected from North Carolina, USA and stored in
121 silica gel. Vouchers were deposited in the herbarium of University of North Carolina at
122 Chapel Hill (NCU). Total genomic DNA extraction of dried leaves was performed using a
123 modified cetyltrimethylammonium bromide (CTAB) protocol described by Cullings (1992)
124 and Xiang et al. (1998). DNA quantity was analyzed with Qubit 2.0 (Life Technologies, USA)
125 and quality was measured using a NanoDrop spectrophotometer 2000 (ThermoFisher
126 Scientific, USA) and 1% w/v agarose gels. Sequencing was performed at the High-
127 Throughput Sequencing Facility (HTSF) at UNC Chapel Hill. For Illumina sequence libraries
128 (Illumina, CA, USA), ~250 ng of total DNA was utilized. Agilent 2100 Bioanalyzer (Agilent
129 Technologies, USA) was used to select ~450 bp fragments for Novaseq 6000, 250 bp paired-

end (PE) sequencing. For the Oxford Nanopore sequencing, ~2000 ng of high-molecular weight DNA was prepared using the ligation sequencing kit (SQK-LSK109) and sequenced on two R9.4.1 flowcells (Oxford Nanopore Technologies, Oxford, UK).

2.2 ptGAUL pipeline and validation

We generated a pipeline to facilitate plastome assembly using long-read data, which can be applied to both PacBio and ONT raw reads. The ptGAUL pipeline (Figure 1) includes three major parts: filtering long reads, setting the depth of coverage, and assembling the filtered plastid data. Step 1: use minimap2 (Li et al., 2018) to find all reads that map partially or completely to the closely related reference plastome, followed by filtering all reads using a customized bash script. Then, use seqkit (Shen et al., 2016) to keep long reads greater than a specified length (default is 3000 bp, "-f" in ptGAUL). Step 2: calculate the coverage by assembly-stats (available in https://github.com/sanger-pathogens/assembly-stats). If the coverage is over 50x, apply seqtk (available in https://github.com/lh3/seqtk) to randomly select a subset of data including about 50x coverage of the plastome (higher coverage might fail in assembly). Step 3: use Flye (Kolmogorov et al., 2019) to assemble the plastome. When only three contigs were detected in the graphical fragment assembly (gfa) file, we used combined_gfa.py, a customized python script, to assemble the plastome into two different paths. Otherwise, the assembly result was checked manually using Bandage. All pipeline code was deposited on Github (https://github.com/Bean061/ptgaul). Step 3: if ptGAUL was different from the assembly coverage setting in Flye (--asm-coverage), ptGAUL implements seqtk to randomly choose a subset of long reads to minimize the bias of read selection. After assembly, if short-read sequencing data are available, the FM-index Long Read Corrector (FMLRC) software (Wang et al., 2018) is recommended to polish and improve the accuracy

of the final assembled sequences because it can generate more accurate assembly result (Mak et al., 2022). All analyses related to ptGAUL were run using 10 CPUs and 40G RAM on the longleaf cluster at UNC Chapel Hill.

Long-read data from 16 published plastomes in NCBI were used to validate the efficacy of ptGAUL for assembly (Table 1). Comparative analyses were conducted including the number of assembled contigs, total genome size (bp) and nucleotide sequence identity between the published results and those obtained with ptGAUL (pairwise identity in alignment) using Geneious v.2022.2 (Kearse et al., 2012).

2.3 Assembly and comparison of four *Juncus* species

The Illumina Novaseq 6000 platform (Illumina, USA) was used to generate 250 bp, paired-end (PE) reads for *Juncus roemerianus* and *J. validus*. Reads were *de novo* assembled using GetOrganelle v1.7.5 (Jin et al., 2020) with default settings. Long-read data were also generated using ONT for *J. roemerianus* and *J. validus*. Long-read data were assembled using ptGAUL with default (3000 bp) filtering parameters ("-f") and using all eight *Juncus* plastomes on GenBank (Table 2) as references for the filtering step. We verified the assembly graph results (gfa file) from Flye using the visualization in Bandage v 0.8.1 (Wick et al., 2015). Then, we conducted FMLRC to polish the final plastomes (an optional step in the ptGAUL pipeline). To examine the assembly result, we mapped all raw Illumina reads and raw ONT reads of each *Juncus* species to our polished assembly and tested the evenness of the coverage at all sites. If every site shares a similar coverage of raw reads without gaps in coverage, this usually indicates a good *de novo* assembly result. We used the samtools v.1.9 (Danecek et al., 2021) depth function to record read depth at every site, followed by a dot plot created by the matplotlib library (Hunter, 2007) in python. We downloaded the raw whole

genome sequencing data of *J. effusus* and *J. inflexus* (both ONT reads: SRR14298760 and SRR14298751 and Illumina reads: SRR14298746 and SRR14298745 from Planta et al., 2022) to assemble the plastomes following the same steps above.

After assembly, we uploaded plastomes of four *Juncus* species (*J. roemerianus*, *J. validus*, *J. effusus*, and *J. inflexus*) to GeSeq online (Tillich et al., 2017) for annotation using Chole (Zhong, 2020), HMMER (Finn et al., 2011) and BLAT (Kent, 2002). We manually checked the start and stop codons of each annotated gene using Geneious v.2022.2. The genes not in frame in each *Juncus* species were either adjusted or removed after a careful comparison with *Typha latifolia* plastid annotation (NC_013823; Guisinger et al., 2010) by mapping the annotations to our *Juncus* assemblies. For the uncertain tRNAs, we confirmed the tRNA secondary structures via RNAfold WebServer (Hofacker, 2003). Linear plastome maps were drawn with OGDRAW v. 1.2 (Lohse et al., 2013). Circular representations were drawn using Circoletto (Darzentas, 2010) to visualize the repeats.

2.4 Examination of repeats and rearrangement events in *Juncus*

We removed one copy of the IR region prior to repeat analyses to avoid counting the repeats from IR region. We implemented BLAST v.2.8.1+ (Altschul et al., 1990) and Tandem Repeats finder v4.09.1 (Benson, 1999) to detect the dispersed repeats and tandem repeats, respectively, following the steps from Lee et al. (2020). We manually checked the result and eliminated duplicated blast hits and recorded the total number of distinct dispersed repeats. We also downloaded complete plastomes of *Eriocaulon decemflorum* (NC_044895; Darshetkar et al., 2019) and two early diverging Poales, *Typha latifolia* (NC_013823; Guisinger et al., 2010) and *Ananas comosus* (NC_026220; Nashima et al., 2015), for comparison. All the plots were drawn using matplotlib (Hunter, 2007) in python.

We focused on the four confirmed assemblies of *Juncus*, i.e., *J. roemerianus*, *J. validus*, *J. effusus*, and *J. inflexus* for characterizing and comparing the rearrangements in *Juncus* plastomes. The other eight publicly available (Lu et al., 2021; Wu et al., 2021) *Juncus* plastomes on GenBank were excluded from the rearrangement analyses (Table 2) because of the uncertain assemblies resulting from short-read data. To eliminate uncertainty in short-read assemblies, we compared them to the *J. effusus* plastome assembled from short-read data (Lu et al., 2021) and to the ptGAUL-assembled plastome of *J. effusus* from long-read data (Planta et al. 2022). To detect rearrangement events within *Juncus*, whole-genome alignments of *J. roemerianus*, *J. validus*, *J. effusus*, and *J. inflexus* were performed to examine the arrangement of locally colinear blocks (LCBs) using progressiveMauve (Darling et al., 2004). One copy of the IR was removed from plastomes prior to Mauve alignment to prevent spurious alignments. *Typha latifolia* was employed as a reference and *Ananas comosus* and *Eriocaulon decemflorum* were also included.

3 RESULTS

3.1 Validation of ptGAUL

Overall, ptGAUL assemblies were successful; assemblies contained either one or three contigs in 11 of 16 the species, with plastome sizes similar to those reported previously (indicated with "S" in Table 1). The assembly graph results (gfa files) indicating plastome structure were visualized and confirmed by Bandage and deposited in Github (https://github.com/Bean061/ptgaul). Assembled plastomes had > 95% nucleotide sequence identity to the references, however the plastome of *Arctostaphylos glauca* was 31,578 bp longer (21% total length) than the published data (Table 1). ptGAUL failed to assemble plastomes of five species (indicated with F in Table 1). The ptGAUL pipeline produced

consistent and reliable results when provided with a dataset of long reads (> 5000 bp N50) with ~50X coverage of the plastome.

Our results indicated that different library preparations affected plastome assembly, regardless of the long-read sequencing platform (PacBio or ONT) employed (Table 1). Plastomes derived from a whole genomic sequencing approach assembled correctly (either one or three contigs), with a reasonable plastome length and structure (by Bandage), while the plastomes using plastid capture approaches (i.e., long range PCR and long fragment target capture) were more fragmented and had a smaller genome size. For example, *Leucanthemum vulgare* had a similar N50 value to *Lepidium sativum* (7900 bp and 7277 bp, respectively), but the *Leucanthemum vulgare* library prepared using long range PCR failed in plastome assembly. All five failed datasets involved the plastid capture approach and most of the raw sequence reads had relatively short length with small N50 values (less than 5000 bp) (Table 1).

3.2 Plastome features of four *Juncus* species

We generated 158,922,322 and 156,712,430 short reads for *Juncus roemerianus* and *J. validus*, respectively along with 427,549 ONT reads from *J. roemerianus* (N50 value: 15,998 bp) and 243,884 ONT reads from *J. validus* (N50 value: 14,365) (Table 2). The data were deposited at NCBI with BioProject accession: PRJNA865266. We also downloaded sequence data (PRJNA723756)of *J. effusus* and *J. inflexus* from Planta et al. (2022) (Table 2). The ptGAUL pipeline produced three contigs each for *J. validus* and *J. roemerianus* (Figure S1a, b) sequenced in this study, and one contig each for *J. effusus* and *J. inflexus* sequenced by Planta et al. (2022) (Figure S1c, d). The final assembled plastomes of *J. validus*, *J.*

roemerianus, *J. effusus*, and *J. inflexus* ranged from 147,183 to 196,852 bp, had similar sized LSCs from, different sizes of the SSC from, and large differences in IR size (Table 2).

The assemblies for the four *Juncus* species were verified by mapping both Illumina and ONT reads back to the assembly. All mapping results showed a high and even coverage of both species (Figure 2 c-f; Figure S2 a,b,d,e). There were no gaps in assemblies regardless of sequencing platform. Annotation of the four *Juncus* plastomes revealed that they contained 114 to 136 genes, 93 to 106 of which were unique. There were 60 to 72 unique protein coding genes (PCGs), 29-30 tRNA genes, and four rRNA genes (Table 2). *Juncus roemerianus* has the greatest gene number at 136, which is similar to the *J. effusus* (133), *J. inflexus* (134), and basal Poales ancestors *Typha latifolia* (133), *Ananas comosus* (132), *Eriocaulon decemflorum* (135) (Table S2). *Juncus effusus* and *J. inflexus* shared a highly similar gene content while *J. validus* lacks 11 *ndh* genes, *rps15* and *trnT-GGU* (Table S2). (Figure S3; Table S2) .

3.3 Verification of published *Juncus effusus* plastome

We compared the *J. effusus* published assembly based on short-read data (Lu et al., 2021, MW366789) with our new assembly using ptGAUL of long-read data of Planta et al. (2022). The result indicated that the short-read assembly generated by Lu et al. was >7.5 kb shorter than our long-read assembly (170,612 bp versus 178,158 bp). The mapping results showed that our assembly was well supported by both long-read and short-read data from Planta et al. (2022) (Figure S2a,b), yet unsupported by the Illumina reads from Lu et al. (2021) with 777 positions with less than 10x coverage, including 295 positions that have no read coverage (Figure S2c). The previous short-read assembly of *J. effusus* (MW366789) was not supported by the long-read data from Planta et al. (2022). Based on this result, we removed the eight

publicly available *Juncus* plastomes assembled with short-read data prior to the comparative analyses of plastomes.

3.4 Repeats in *Juncus* plastomes

Repeat analyses identified many dispersed and tandem repeats in the four *Juncus* plastomes (17.2 – 24.3 % of genome without IRa) in comparison with basal Poales and *Eriocaulon* (1.8 – 3.3 % of genome without IRa) (Table 3). The combined length of both dispersed and tandem repeats in *Juncus* plastomes ranged from 22,577 bp (*J. validus*) to 34,027 bp (*J. roemerianus*), which was far greater than *Typha* (4,436 bp), *Ananas* (3,552 bp) and *Eriocaulon* (2,227 bp) (Table 3). When dispersed repeats were parsed into five different size classes, *Juncus* plastomes contained a greatly increased number of dispersed repeats than basal Poales and *Eriocaulon* (Figure 3 and Table S3). Larger repeats (>201 bp) were found only in *Juncus* (Figure 3 and Table S3). Among four *Juncus* plastomes, *J. effusus* and *J. validus* had more abundant dispersed repeats yet *J. roemerianus* was the only one with a repeat larger than 1 kb. *Juncus* plastomes also experienced substantial accumulation of tandem repeats (Table 3). Tandem repeat accumulation was higher than that of dispersed repeats in *J. inflexus* and *J. roemerianus*. All four *Juncus* plastomes contained exceptionally expanded tandem repeats, ranging from 4.6 – 6.6 kb, some of which contain *clpP* (Table S4).

3.5 Rearrangement of *Juncus* plastomes

Whole-genome alignment using progressiveMauve (Figure 4) detected 27 LCBs from seven complete plastomes (*Typha latifolia*, *Ananas comosus*, *Eriocaulon decemflorum*, *Juncus effusus*, *J. inflexus*, *J. roemerianus*, and *J. validus*). The plastomes of the two basal Poales and *Eriocaulon* were colinear, whereas all *Juncus* species have many breakpoints (BP) relative to

the reference, *T. latifolia* (Figure 4; Table 4). When compared with basal Poales plastomes, the BP and reversal distances were 15 and 19, in *J. effusus* and *J. inflexus,* repspectively. *Juncus roemerianus* has the largest BP (17) and reversal distances (20), and *J. validus* has the smallest BP (14) and reversal distances (17). Among the four *Juncus*, 27 LCBs were identified (Figure S4). While *J. effusus* and *J. inflexus* shared the same gene order, widespread rearrangements were detected in the other two species (*J. roemerianus* and *J. validus*).

4 DISCUSSION

4.1 ptGAUL application and suggestions for sequencing approach

The ptGAUL pipeline generated either one or three contig(s) for 11 publicly available datasets using either PacBio or Oxford Nanopore data (Table 1). However, it failed to assemble the data from five species generating more than three short contigs and predicted much smaller plastome size, which is less than optimal (Table 1). In successful cases, the assemblies were highly similar to the published short-read assemblies with over 96-99% nucleotide sequence identity. The lower percent identity between *Cenchrus americanus* and *Digitaria exilis* and their reference assemblies may be due to different sequencing approaches between the Mariac et al. (2014) combined plastid capture method and Illumina sequencing and our long-read approach. For *Arctostaphylos glauca*, we used the read mapping method to verify that our assembly was more reliable than the result of Huang et al. (2022) as it showed more proportional coverage across the entire plastome (Figure S5). This difference could be caused by the selection of a distantly-related reference genome (*Camellia taliensis*) from another family by Huang et al. (2022).

We found that the five failed samples had some features in common. For example, the sequencing approaches in failed assemblies were different from the whole genomic

sequencing method of those that were successful. In the *Leucanthemum vulgare* study, long-range PCR was implemented to generate amplicons that were then sequenced to produce a set of long reads that had an N50 value of ~ 8000 bp (Scheunert et al., 2020). In the remaining failed assemblies, plastid capture was utilized (Bethune et al., 2019). The PCR processes in both studies can greatly increase the bias among different plastome regions, e.g. AT- and GC-rich regions do not amplify as efficiently as other regions (Quail et al., 2012). This could lead to underrepresentation/unevenness in read coverage of different regions resulting in many fragmented assemblies/contigs. Furthermore, the probes were designed based on the plastome data from distantly related species (Bethune et al., 2019), which may be unable to capture all plastome fragments for the target non-model species due to the divergence between the probe regions and the genome being captured. Additionally, the sequences obtained from PCR methods tend to be much shorter than the reads generated from sequencing total genomic DNA (see N50 values in Table 1). The low N50 values could also result from degraded DNA from poor storage, use of silica dried or herbarium material and/or DNA extraction method. For example, the Qiagen DNEasy Plant kits can generate high quality DNA for short-read sequences because the column shreds the DNA to a maximum of ~25 Kb fragments (Qiagen, 2006). CTAB, SDS or other methods that can produce much higher molecular weight (HMW) DNA are preferred for third generation sequencing (Mayjonade et al., 2016; Jung et al., 2019), emphasizing the importance of sample preparation. Likewise, the assembly approaches, parameter combinations, read coverage, and the presence of nuclear genome and/or mitogenome contaminants could impact the completeness of an assembly (Jung et al., 2019; Scheunert et al., 2020).

Overall, considering the read length and read coverage, ptGAUL performs well for HMW samples using total genomic sequencing resulting in high N50 values. Therefore, we

recommend using HMW DNA extraction methods to isolate highly intact DNA, followed by long-read sequencing and subsequent assembly using ptGAUL.

4.2 Long-read data for plastome assembly

We found that short-read data alone may be insufficient to accurately assemble plastomes in species with many long dispersed repeats. This phenomenon has been seen in several lineages including *Eleocharis* (Lee et al., 2020) and *Monsonia* (Ruhlman et al., 2017). Plastome assembly using GetOrganelle for 11 *Juncus* species (12 accessions) failed using Illumina short reads only, including two samples in this study (Fig. S6). All *Juncus* plastome assemblies indicated either many fragmented contigs or many assembly paths (Figure S6). This is because the many long dispersed repeats in *Juncus* plastomes are longer than the kmer size/length of short reads. Based on our *J. effusus* plastome comparison, the final assembly length and total number of genes based on short read data is much shorter than the ones assembled from long read data (Table 2, Table S2), which might be caused by the random selection one of the paths as the final assembly when using short read data. Other studies demonstrated that this issue can be resolved by a three-step approach: comparing different contigs from short-read assemblers (e.g. SPADES, Velvet), manually checking through the contigs compared with closely related species, and long range PCR to confirm assemblies (Lee et al., 2020; Ruhlman et al., 2017). This approach requires considerable time and effort.

Our ONT data resolved the plastome structure of four *Juncus*, confirming prior work (Lee et al., 2020; Ruhlman et al., 2017) showing that long-read data vastly improves assembly of the plastomes with many long repeats. Based on our study and that of Scheunert et al. (2020) ~50X mapping coverage of long-read data can result in an accurate plastome assembly. In our study, long reads of plastid origin represented 5%-6% of reads generated from total

genomic DNA of *Juncus*. Assuming 5% plastid DNA content from whole-genome HMW extractions, to generate ~ 50X coverage of a 160,000 bp plastome requires only ~160 Mbp reads per sample. Currently one chip of ONT generates ~10 Gbp of sequence data, enabling multiplexing up to 64 samples at a consumables cost of ~$1000 USD (based on the price from HTSF at UNC Chapel Hill).

Although several assembly tools have been developed, several issues persist. Some pipelines/software are no longer maintained (i.e., Sprai, Celera Assembler, Organelle_PBA). The assemblers of Syme et al. (2021); others, Canu, and Hinge (Wang et al., 2018) cannot generate a consistent plastome assembly result with one contig when using different coverages of data. Unicycler (Wick et al., 2017) is computationally intensive and does not produce well resolved assemblies when dealing with complicated plastomes with many long repeats. Compared to current published pipelines for plastome assembly, ptGAUL can help generate an accurate plastome assemblies in less than ~20 minutes (10 CPUs and 40G RAM), making it highly convenient. Thus, ptGAUL should greatly facilitate plastome assembly of long-read data for phylogenetic and molecular evolutionary studies, especially in plastomes with a significant fraction of long repeat regions. Although ptGAUL can expedite plastome assembly, researchers still need to pay close attention to the species with multiple plastid types, such as *Eleocharis* (Lee et al., 2020) and *Monsonia* (Ruhlman et al., 2017).

4.3 *Juncus* plastome organization

While many Poales genera contain plastomes with conserved gene order and content (Jones et al., 2007), including *Typha* (Guisinger et al., 2010), *Ananas* (Redwan et al., 2015) and *Eriocaulon* (Darshetkar et al., 2019), the data from the four *Juncus* examined here suggest that at least some species in this group contain plastome features atypical to most

angiosperms. A limited number of complete plastome sequences are available from *Juncus* or other Juncaceae, but recently assemblies of two *Eleocharis* plastomes, in the sister family Cyperaceae (Hochbach et al., 2018), revealed accumulated duplications, gene losses, gene order rearrangements and intraindividual structural heteroplasmy (Lee et al., 2020). Similar phenomena contributed to size variation in the four *Juncus* plastomes, which ranged from 147,183 bp to 196,852 bp (Table 2). Many long repeats, including an unusually high number of dispersed repeats of 61 – 200 bp and 201 – 1000 bp, were present in the four *Juncus* with the greatest accumulation in *J. effusus*. Repeats >1000 bp were detected only in *J. roemerianus* (Table S3; Figure 3). Accumulation of large repeats may predispose plastome rearrangements in addition to contributing to overall size expansion (Tables 2-3, Figure 4) yet at present it is not clear if repeat accumulation predicated rearrangement or *vice versa* (Lee et al., 2021).

Similar repeat accumulation and plastome rearrangement occur in other taxonomic groups. In the *Trachelium caeruleum*, there are gene-order changes, along with gene duplication, pseudogenization and loss were identified, as well as an abundance of variously sized repeats (Haberle et al., 2008). A relationship between repeat accumulation and rearrangement was suggested (Kim & Lee, 2005); studies of *Pelargonium* (Chumley et al., 2006), *Jasminum*, *Mendora* (Lee et al., 2007) and *Trifolum* (Cai et al., 2008) plastomes show early support for the theory. Many of the repeated sequences, when plotted onto the assembled plastid chromosomes, clustered at rearrangement endpoints. The relationship is also supported by findings in bacterial genomes where repeated sequences lead to gene order rearrangements (Rocha, 2003). Reconfiguration of the ancestral angiosperm plastome through repeat-mediated recombination has now been reported in several groups (Sloan et al., 2014; Weng et al., 2014; Schwarz et al., 2015; Ruhlman et al., 2017; Choi et al., 2019; Choi et al.,

2020). The recombinogenic potential of long repeats identified in the *Juncus* plastomes was likely involved in diversifying gene order.

The observation of slight variations in IR length between *Nicotiana* species was explored in seminal work that focused on the IR/LSC boundary in closely related groups. This work ultimately inferred recombination-mediated gene conversion between poly-A tracts that gave rise to a >12 kb expansion at the *N. acuminata* J_{LB} (IR_B/LSC boundary) placing the new J_{LB} near *clpP* and duplicating the 12 kb sequence now included in the IR (Goulding et al., 1996). Although the details of the mechanism have been clarified and refined over the years, repeat-mediated gene conversion appears to be at the heart of it (Maréchal & Brisson, 2010; Oldenburg & Bendich, 2015; Ruhlman & Jansen, 2021).

Plastomes that contain a large number of long repeats can experience extensive rearrangement of gene order and both loss and gain of plastome sequence, including genes, introns and non-coding sequences alike. Expansion and contraction at both LSC and SSC boundaries contributed to variation in *Juncus* plastome size. Photosynthetic seed plant plastomes and IRs range from ~120-170 kb and 20-30 kb, respectively, however most IR-containing angiosperms sequenced to date display highly similar gene arrangement and plastome size (~150 kb; IR, ~25kb; Ruhlman & Jansen, 2021). Total plastome size in some groups is strongly influenced by IR expansion, yet in other lineages the association is loose at best. For example, a study of five *Cyperus* plastomes revealed the largest plastomes had the smaller IRs (i.e. *C. esculentus*; 186,255 kb/37,438 kb) and the smaller plastomes contained the larger IRs (i.e. *C. difformis*; 167,974 kb/38,427 kb) (Ren et al., 2021).

While total plastome size scaled with IR size (Table 2) and total repeat content (Table 3) in the four *Juncus*, the myriad events that altered each plastome relative to a shared ancestor with more conserved structure remain elusive. The smallest of the four plastomes, in

J. validus would seem like a typical plastome based on the overall plastome and IR size (~147 kb and ~29 kb). However, the assembly and annotation show that it is not always size that matters. This plastome has likely experienced/is experiecing an ongoing series of IR boundary migrations resulting in a novel organization relative to the other taxa evaluated here. The near total elimination of the NDH gene suite, predominantly situated in the SSC in typical angiosperm plastomes, was unique to *J. validus* and suggests that IR boundary migration into the SSC played a role it their eventual loss. Although retained by the three other taxa, NDH sequences appear in alternate loci and several have been duplicated by IR inclusion (Figure 2; Figure S3) suggesting migration at the SSC boundaries. Indeed, the gene order arrangement proximal to IR/LSC boundaries display little rearrangement across all four *Juncus* (Figures 2, 3; Figure S3).

Complete ablation of the plastid-encoded NDH (NADH dehydrogenase-like) gene suite was reported for several unrelated seed plant lineages (Ruhlman et al., 2015). The NDH complex of plant and algal plastids participates in cyclic electron flow (CEF) (Shikanai et al., 1998) and comprises a multisubunit, plastid-localized complex that incorporates imported nuclear-encoded factors. The plastid genes encoding the NDH complex are highly conserved across Streptophyta (Hori et al., 2014) suggesting an essential function in photosynthesis (Ifuku et al., 2011). Using plastome sequencing and nuclear transcriptomics revealed that taxa lacking the plastid genes encoding constituents of NDH concomitantly lacked the relevant nuclear-encoded factors. Probing nuclear transcriptomes revealed that regardless of the state of the plastid NDH gene suite, genes encoding the alternate PGR5-dependent CEF pathway (Shikanai, 2014) were present in the nucleus of all examined taxa (Ruhlman et al., 2015). The loss of the NDH suite from the *J. validus* plastome is unique among examined Poales

463 plastomes and suggest that an active PGR5-dependent pathway accounts for CEF in this
464 species.

465 Apart from the loss of NDH genes, gene losses were shared by all four *Juncus*
466 examined and included other genes that were lost from plastomes of diverse lineages
467 (Ruhlman & Jansen, 2018). The plastid-localized Acetyl-coenzyme A carboxylase (ACCase;
468 prokaryotic) is another multisubunit protein complex that incorporates nuclear-encoded
469 polypeptides and participates in fatty acid metabolism (Ohlrogge & Browse, 1995). The
470 plastid *accD* encodes one subunit of the four-unit complex and was lost in numerous taxa,
471 often those that experienced other gene loss and pseudogenization events (Ruhlman & Jansen,
472 2018). Because plastid ACCase activity was thought an essential function (Kode et al., 2005)
473 *accD* loss in several groups suggested that it may be expressed from a functional transfer to
474 the nucleus or substituted by a redundant, nuclear-encoded enzyme (Konishi et al., 1996). In
475 *Trifolium*, which lacks plastid *accD*, a functional transfer to the nucleus was uncovered
476 (Magee et al., 2010). Further investigation failed to detect any remnant of the *accD* sequence
477 in the plastomes of *T. repens* or *T. pratense* while mutated copies were identified in *T.*
478 *aureum* and *T. grandiflorum* (Sabir et al., 2014). The 15-amino acid C-terminal catalytic
479 domain of the ACCD protein, which is minimally required for prokaryotic ACCase function
480 (Lee et al., 2004), was identified in the mutated copies and may indicate functionality.
481 Probing nuclear transcriptomes from *T. repens* and *T. pratense* revealed that, as in *T.*
482 *subterraneum* (Magee et al., 2010), a putatively functional ACCD protein was being
483 expressed from a fusion sequence that included the ACCD catalytic domain (~270 aa) fused
484 to the plastid target peptide from nuclear-encoded, plastid-targeted LPD1 (493 aa). Probing
485 transcriptomes of related legumes that contained intact plastid *accD* was able to detect high
486 identity copies of the ACCD core sequence suggesting that incorporation at nuclear loci

predated the degradation of plastid *accD* (Sabir et al., 2014). Functional redundancy was demonstrated for prokaryotic ACCase (Babiychuk et al., 2011; Rousseau-Gueutin et al., 2013) and other gene products through transfer or substitution in different lineages (Ueda et al., 2007, 2008).

The fate of *accD* sequences and both the prokaryotic and the single-polypeptide eukaryotic ACCase in Poales has been a matter of investigation for some time. Morton & Clegg (1993) identified a recombination hotspot in seven Poaceae plastomes in the region between *rbcL* and *psaI* (i.e. the locus containing *accD* sequences in non-Poaceae plastomes (Harris et al., 2013). Exploiting the fact that both the eukaryotic and prokaryotic ACCases contain biotinylated polypeptides, Konishi et al. (1996) were able to identify which form of the enzyme was active in plastids from across the diversity of the green plant lineage, including two non-photosynthetic representatives. Differentiating the two enzymes by molecular weight revealed that only one group examined did not contain the 35 kDa peptide that represented the prokaryotic holoenzyme: Poaceae. Closer examination of Poales using PCR product sequencing combined with Southern blots probed with plastid *accD* from Commelinaceae taxa demonstrated pseudogenization or deletion in representatives of three families, Restionaceae, Joinvilleaceae and Poaceae (Harris et al., 2013). Extending the loss of *accD* to include the Cyperaceae (Cyperus, Ren et al., 2021; Elocharis, Lee et al., 2020) and now Juncaceae suggests either extreme lability of the coding sequence in Poales or that this gene was transferred or substituted by a nuclear encoded activity in a common ancestor. Differential nuclear retention, expression and transport of the gene product back to plastids among the various lineages could result in relaxed selection on the plastid gene (Ueda et al., 2007; Park et al., 2017).

510 The opportunity to sample deeply across and within lineages is revealing that the
511 unusual variation identified by early Southern blots and more recent plastome sequencing
512 suggests that these 'unusual' structural changes are not unique. The suite of plastid genes that
513 are susceptible to pseudogenization or loss appears consistent across photosynthetic seed
514 plants. Understanding phylogeny, inherent to evolutionary studies, requires deep sampling,
515 high-quality sequencing, assembly and alignment to infer relationships. As next generation
516 sequencing and single-molecule long-read sequencing platforms expand and become more
517 accessible, reads will be generated for many diverse taxa. Where long sequence repeats
518 exceed insert sizes in next gen systems, long reads will be able to 'bridge the gap'. The ability
519 to translate raw sequence reads into usable data for evolutionary and functional inquiries
520 depends on advanced computational tools that provide fast, flexible platforms without vast
521 computational demand. Facilitating this effort, the ptGAUL pipeline provides a fast and easy
522 tool for assembling plastomes from long-read data, which will enable the characterization of
523 repeat-rich, highly rearranged plastomes.
524

525 **ACKNOWLEDGEMENTS**

526 We thank the UNC greenhouse staff for maintaining living *Juncus* materials and the UNC
527 herbarium staff for storing our voucher specimens. We are very grateful for the UNC longleaf
528 high-performance cluster for computational resources. This work was supported by National
529 Science Foundation IOS-2034929 to A.M.J and C.D.J.
530

531 **REFERENCES**

532 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local
533 alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.
534 Babiychuk, E., Vandepoele, K., Wissing, J., Garcia-Diaz, M., De Rycke, R., Akbari, H.,
535 Joubès, J., Beeckman, T., Jänsch, L., & Frentzen, M. (2011). Plastid gene expression

and plant development require a plastidic protein of the mitochondrial transcription termination factor family. *Proceedings of the National Academy of Sciences*, *108*(16), 6674–6679.

Balslev, H. (2018). Two new species of *Juncus* (Juncaceae) from South America. *Phytotaxa*, *376*(2), 97-102-97–102. https://doi.org/10.11646/phytotaxa.376.2.3

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580.

Bethune, K., Mariac, C., Couderc, M., Scarcelli, N., Santoni, S., Ardisson, M., Martin, J., Montúfar, R., Klein, V., & Sabot, F. (2019). Long-fragment targeted capture for long-read sequencing of plastomes. *Applications in Plant Sciences*, *7*(5), e1243.

Brožová, V., Proćków, J., & Drábková, L. Z. (2022). Toward finally unraveling the phylogenetic relationships of Juncaceae with respect to another cyperid family, Cyperaceae. *Molecular Phylogenetics and Evolution*, *177*, 107588.

Cai, Z., Guisinger, M., Kim, H. G., Ruck, E., Blazier, J. C., McMurtry, V., Kuehl, J. V., Boore, J., & Jansen, R. K. (2008). Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *Journal of Molecular Evolution*, *67*(6), 696–704.

Choi, I., Jansen, R., & Ruhlman, T. (2019). Lost and found: Return of the inverted repeat in the legume clade defined by its absence. *Genome Biology and Evolution*, *11*(4), 1321–1333.

Choi, I., Jansen, R., & Ruhlman, T. (2020). Caught in the act: Variation in plastid genome inverted repeat expansion within and between populations of Medicago minima. *Ecology and Evolution*, *10*(21), 12129–12137.

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., & Jansen, R. K. (2006). The complete chloroplast genome sequence of *Pelargonium*× *hortorum:* Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Molecular Biology and Evolution*, *23*(11), 2175–2190.

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991. https://doi.org/10.1038/nbt.2023

Cullings, K. W. (1992). Design and testing of a plant-specific PCR primer for ecological and evolutionary studies. *Molecular Ecology*, *1*(4), 233–240. https://doi.org/10.1111/j.1365-294x.1992.tb00182.x

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research*, *14*(7), 1394. https://doi.org/10.1101/gr.2289704

Darshetkar, A. M., Datar, M. N., Tamhankar, S., Li, P., & Choudhary, R. K. (2019). Understanding evolution in Poales: Insights from Eriocaulaceae plastome. *PLoS ONE*, *14*(8). https://doi.org/10.1371/journal.pone.0221423

Darzentas, N. (2010). Circoletto: Visualizing sequence similarity with Circos. *Bioinformatics*, *26*(20).

582 Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: De novo assembly of
583 organelle genomes from whole genome data. *Nucleic Acids Research*, *45*(4), e18–e18.
584 https://doi.org/10.1093/nar/gkw955
585 Drábková, L. (2010). Phylogenetic relationships within Juncaceae: Evidence from all three
586 genomic compartments with notes to the morphology. In *Seberg, O., Petersen, G.,*
587 *Barford and Davis: Diversity, Phylogeny, and Evolution in the Monocotyledons* (pp.
588 389–416) Aarhus University Press.
589 Drábková, L., Kirschner, J., & Vlček, Č. (2006). Phylogenetic relationships within *Luzula* DC.
590 and *Juncus* L.(Juncaceae): A comparison of phylogenetic signals of trnL-trnF
591 intergenic spacer, trnL intron and rbcL plastome sequence data. *Cladistics*, *22*(2),
592 132–143.
593 Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola,
594 R., Cavalieri, D., Velasco, R., & Cestaro, A. (2013). An evaluation of the PacBio RS
595 platform for sequencing and de novo assembly of a chloroplast genome. *BMC*
596 *Genomics*, *14*(1), 1–12.
597 Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence
598 similarity searching. *Nucleic Acids Research*, *39*, W29–W37.
599 https://doi.org/10.1093/nar/gkr367
600 Freudenthal, J. A., Pfaff, S., Terhoeven, N., Korte, A., Ankenbrand, M. J., & Förster, F.
601 (2020). A systematic comparison of chloroplast genome assembly tools. *Genome*
602 *Biology*, *21*(1), 1–21. https://doi.org/10.1186/s13059-020-02153-6
603 Goulding, S. E., Wolfe, K., Olmstead, R., & Morden, C. (1996). Ebb and flow of the
604 chloroplast inverted repeat. *Molecular and General Genetics*, *252*(1), 195–206.
605 Guisinger, M. M., Chumley, T. W., Kuehl, J. V., Boore, J. L., & Jansen, R. K. (2010).
606 Implications of the plastid genome sequence of *Typha* (Typhaceae, Poales) for
607 understanding genome evolution in Poaceae. *Journal of Molecular Evolution*, *70*(2),
608 149–166.
609 Haberle, R. C., Fourcade, H. M., Boore, J. L., & Jansen, R. K. (2008). Extensive
610 rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated
611 with repeats and tRNA genes. *Journal of Molecular Evolution*, *66*(4), 350–361.
612 Harris, M. E., Meyer, G., Vandergon, T., & Vandergon, V. O. (2013). Loss of the acetyl-CoA
613 carboxylase (accD) gene in Poales. *Plant Molecular Biology Reporter*, *31*(1), 21–31.
614 Hochbach, A., Linder, H. P., & Röser, M. (2018). Nuclear genes, matK and the phylogeny of
615 the Poales. *Taxon*, *67*(3), 521–536.
616 Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic acids*
617 *research*, 31(13), 3429-3431.
618 Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada,
619 T., Mori, H., & Tajima, N. (2014). *Klebsormidium flaccidum* genome reveals primary
620 factors for plant terrestrial adaptation. *Nature Communications*, *5*(1), 1–9.
621 Huang, Y., Escalona, M., Morrison, G., Marimuthu, M. P., Nguyen, O., Toffelmier, E.,
622 Shaffer, H. B., & Litt, A. (2022). Reference genome assembly of the big berry
623 Manzanita (*Arctostaphylos glauca*). *Journal of Heredity*, *113*(2), 188–196.
624 Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and*
625 *Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55
626 Ifuku, K., Endo, T., Shikanai, T., & Aro, E.-M. (2011). Structure of the chloroplast NADH
627 dehydrogenase-like complex: Nomenclature for nuclear-encoded subunits. *Plant and*
628 *Cell Physiology*, *52*(9), 1560–1568.

Jansen, R. K., & Ruhlman, T. A. (2012). Plastid Genomes of Seed Plants. In *Genomics of chloroplasts and mitochondria* (pp. 103–126). Springer. https://doi.org/10.1007/978-94-007-2920-9_5

Jiang, H., Tian, J., Yang, J., Dong, X., Zhong, Z., Mwachala, G., Zhang, C., Hu, G., & Wang, Q. (2022). Comparative and phylogenetic analyses of six *Kenya Polystachya* (Orchidaceae) species based on the complete chloroplast genome sequences. *BMC Plant Biology*, *22*(1), 1–21. https://doi.org/10.1186/s12870-022-03529-5

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*(1), 1–31. https://doi.org/10.1186/s13059-020-02154-5

Jones, E., Simpson, D. A., Hodkinson, T. R., Chase, M. W., & Parnell, J. A. (2007). The Juncaceae-Cyperaceae interface: A combined plastid sequence analysis. *Aliso: A Journal of Systematic and Floristic Botany*, *23*(1), 55–61.

Jung, H., Winefield, C., Bombarely, A., Prentis, P., & Waterhouse, P. (2019). Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. *Trends in Plant Science*, *24*(8), 700–724. https://doi.org/10.1016/j.tplants.2019.05.003

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., & Duran, C. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656. https://doi.org/10.1101/gr.229202

Kim, K. J., & Lee, H. L. (2005). Widespread occurrence of small inversions in the chloroplast genomes of land plants. *Molecules & Cells*, *19*(1), 104–113.

Kode, V., Mudd, E. A., Iamtham, S., & Day, A. (2005). The tobacco plastid accD gene is essential and is required for leaf development. *The Plant Journal*, *44*(2), 237–244.

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546.

Konishi, T., Shinohara, K., Yamada, K., & Sasaki, Y. (1996). Acetyl-CoA carboxylase in higher plants: Most plants other than gramineae have both the prokaryotic and the eukaryotic forms of this enzyme. *Plant and Cell Physiology*, *37*(2), 117–122.

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*(5), 722–736.

Lee, C., Ruhlman, T. A., & Jansen, R. K. (2020). Unprecedented Intraindividual Structural Heteroplasmy in *Eleocharis* (Cyperaceae, Poales) Plastomes. *Genome Biology and Evolution*, *12*(5), 641–655. https://doi.org/10.1093/gbe/evaa076

Lee, C., Choi, I., Cardoso, D., de Lima, H. C., de Queiroz, L. P., Wojciechowski, M. F., Jansen, R. K., & Ruhlman, T. A. (2021). The chicken or the egg? Plastome evolution and an independent loss of the inverted repeat in papilionoid legumes. *The Plant Journal*, *107*(3), 861–875.

Lee, H. L., Jansen, R. K., Chumley, T. W., & Kim, K. J. (2007). Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Molecular Biology and Evolution*, *24*(5), 1161–1180.

Lee, S. S., Jeong, W. J., Bae, J. M., Bang, J. W., Liu, J. R., & Harn, C. H. (2004). Characterization of the plastid-encoded carboxyltransferase subunit (accD) gene of potato. *Molecules & Cells*, *17*(3), 442–429.

Li, J., Su, Y., & Wang, T. (2018). The repeat sequences and elevated substitution rates of the chloroplast accD gene in cupressophytes. *Frontiers in Plant Science*, *9*, 533.

Li, Y., & Deng, X. (2021). The complete chloroplast genome of the marine microalgae *Chaetoceros muellerii* (Chaetoceroceae). *Mitochondrial DNA Part B*, *6*(2), 373–375.

Liao, X., Li, M., Hu, K., Wu, F. X., Gao, X., & Wang, J. (2021). A sensitive repeat identification framework based on short and long reads. *Nucleic Acids Research*, *49*(17), e100–e100. https://doi.org/10.1093/nar/gkab563

Liu, H., Ye, H., Zhang, N., Ma, J., Wang, J., Hu, G., Li, M., & Zhao, P. (2022). Comparative Analyses of Chloroplast Genomes Provide Comprehensive Insights into the Adaptive Evolution of *Paphiopedilum* (Orchidaceae). *Horticulturae, 8*(5), 391. https://doi.org/10.3390/horticulturae8050391

Lohse, M., Drechsel, O., Kahlau, S., & Bock, R. (2013). OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research*, *41*(W1), W575–W581.

Lu, M., Fang, Z., Sheng, F., Tong, X., & Han, R. (2021). Characterization and phylogenetic analysis of the complete chloroplast genome of *Juncus effusus* L. *Mitochondrial DNA Part B*, *6*(5), 1612–1613.

Magee, A. M., Aspinall, S., Rice, D. W., Cusack, B. P., Sémon, M., Perry, A. S., Stefanović, S., Milbourne, D., Barth, S., & Palmer, J. D. (2010). Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research*, *20*(12), 1700–1710.

Mak, Q. C., Wick, R. R., Holt, J. M., & Wang, J. R. (2022). Polishing de novo nanopore assemblies of bacteria and eukaryotes with FMLRC2. *BioRxiv*.

Maréchal, A., & Brisson, N. (2010). Recombination and the maintenance of plant organelle genome stability. *New Phytologist*, *186*(2), 299–317.

Mariac, C., Scarcelli, N., Pouzadou, J., Barnaud, A., Billot, C., Faye, A., Kougbeadjo, A., Maillol, V., Martin, G., & Sabot, F. (2014). Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular Ecology Resources*, *14*(6), 1103–1113.

Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., Langlade, N., & Muños, S. (2016). Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques*, *61*(4), 203–205.

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., & Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, *24*(24), 2818–2824.

Miyamoto, M., Motooka, D., Gotoh, K., Imai, T., Yoshitake, K., Goto, N., Iida, T., Yasunaga, T., Horii, T., & Arakawa, K. (2014). Performance comparison of second-and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*, *15*(1), 1–9.

Morton, B. R., & Clegg, M. T. (1993). A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near rbcL in the grass family (Poaceae). *Current Genetics*, *24*(4), 357–365.

Nadalin, F., Vezzi, F., & Policriti, A. (2012). GapFiller: A de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics*, *13*(14), 1–16.

Nashima, K., Terakami, S., Nishitani, C., Kunihisa, M., Shoda, M., Takeuchi, M., Urasaki, N., Tarora, K., Yamamoto, T., & Katayama, H. (2015). Complete chloroplast genome sequence of pineapple (*Ananas comosus*). *Tree Genetics & Genomes*, *11*(3), 1–11.

Ohlrogge, J., & Browse, J. (1995). Lipid biosynthesis. *The Plant Cell*, *7*(7), 957.

Oldenburg, D. J., & Bendich, A. J. (2015). DNA maintenance in plastids and mitochondria of plants. *Frontiers in Plant Science*, *6*, 883.

Park, S., Ruhlman, T. A., Weng, M.-L., Hajrah, N. H., Sabir, J. S., & Jansen, R. K. (2017). Contrasting patterns of nucleotide substitution rates provide insight into dynamic evolution of plastid and mitochondrial genomes of Geranium. *Genome Biology and Evolution*, *9*(6), 1766–1780.

Planta, J., Liang, Y.-Y., Xin, H., Chansler, M. T., Prather, L. A., Jiang, N., Jiang, J., & Childs, K. L. (2022). Chromosome-scale genome assemblies and annotations for Poales species *Carex cristatella*, *Carex scoparia*, *Juncus effusus*, and *Juncus inflexus*. *G3*, *12*(10), jkac211.

Quail, M. A., Otto, T. D., Gu, Y., Harris, S. R., Skelly, T. F., McQuillan, J. A., Swerdlow, H. P., & Oyola, S. O. (2012). Optimal enzymes for amplifying sequencing libraries. *Nature Methods*, *9*(1), 10–11.

Redwan, R., Saidin, A., & Kumar, S. (2015). Complete chloroplast genome sequence of MD-2 pineapple and its comparative analysis among nine other plants from the subclass Commelinidae. *BMC Plant Biology*, *15*(1), 1–20.

Ren, W., Guo, D., Xing, G., Yang, C., Zhang, Y., Yang, J., Niu, L., Zhong, X., Zhao, Q., & Cui, Y. (2021). Complete chloroplast genome sequence and comparative and phylogenetic analyses of the cultivated *Cyperus esculentus*. *Diversity*, *13*(9), 405.

Rocha, E. P. (2003). DNA repeats lead to the accelerated loss of gene order in bacteria. *Trends in Genetics*, *19*(11), 600–603.

Rousseau-Gueutin, M., Huang, X., Higginson, E., Ayliffe, M., Day, A., & Timmis, J. N. (2013). Potential functional replacement of the plastidic acetyl-CoA carboxylase subunit (accD) gene by recent transfers to the nucleus in some angiosperm lineages. *Plant Physiology*, *161*(4), 1918–1929.

Ruhlman, T. A., Chang, W. J., Chen, J. J., Huang, Y. T., Chan, M. T., Zhang, J., Liao, D. C., Blazier, J. C., Jin, X., & Shih, M. C. (2015). NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biology*, *15*(1), 1–9.

Ruhlman, T. A., Zhang, J., Blazier, J. C., Sabir, J. S., & Jansen, R. K. (2017). Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. *American Journal of Botany*, *104*(4), 559–572.

Ruhlman, T. A., & Jansen, R. K. (2018). Aberration or analogy? The atypical plastomes of Geraniaceae. In *Advances in botanical research* (pp. 223–262). Elsevier.

Ruhlman, T. A., & Jansen, R. K. (2021). The plastid genomes of flowering plants: Essential principles. In Maliga, P. (Eds), *Chloroplast Biotechnology* (pp.3-27). Humana. https://doi.org/10.1007/978-1-0716-1472-3_1

Sabir, J., Schwarz, E., Ellison, N., Zhang, J., Baeshen, N. A., Mutwakil, M., Jansen, R., & Ruhlman, T. (2014). Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnology Journal*, *12*(6), 743–754.

Scheunert, A., Dorfner, M., Lingl, T., & Oberprieler, C. (2020). Can we use it? On the utility of de novo and reference-based assembly of Nanopore data for plant plastome sequencing. *PLoS One*, *15*(3), e0226234.

Schwarz, E. N., Ruhlman, T. A., Sabir, J. S., Hajrah, N. H., Alharbi, N. S., Al-Malki, A. L., Bailey, C. D., & Jansen, R. K. (2015). Plastid genome sequences of legumes reveal

771 parallel inversions and multiple losses of rps16 in papilionoids. *Journal of Systematics*
772 *and Evolution*, *53*(5), 458–468.
773 Shearman, J. R., Sonthirod, C., Naktang, C., Sangsrakru, D., Yoocha, T., Chatbanyong, R.,
774 Vorakuldumrongchai, S., Chusri, O., Tangphatsornruang, S., & Pootakham, W. (2020).
775 Assembly of the durian chloroplast genome using long PacBio reads. *Scientific*
776 *Reports*, *10*(1), 1–8.
777 Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for
778 FASTA/Q file manipulation. *PloS One*, *11*(10), e0163962.
779 Shikanai, T. (2014). Central role of cyclic electron transport around photosystem I in the
780 regulation of photosynthesis. *Current Opinion in Biotechnology*, *26*, 25–30.
781 Shikanai, T., Endo, T., Hashimoto, T., Yamada, Y., Asada, K., & Yokota, A. (1998). Directed
782 disruption of the tobacco ndhB gene impairs cyclic electron flow around photosystem
783 I. *Proceedings of the National Academy of Sciences*, *95*(16), 9705–9709.
784 Sloan, D. B., Triant, D. A., Forrester, N. J., Bergner, L. M., Wu, M., & Taylor, D. R. (2014).
785 A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe
786 Sileneae (Caryophyllaceae). *Molecular Phylogenetics and Evolution*, *72*, 82–89.
787 Soorni, A., Haak, D., Zaitlin, D., & Bombarely, A. (2017). Organelle_PBA, a pipeline for
788 assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing
789 data. *BMC Genomics*, *18*(1), 1–8.
790 Stadermann, K. B., Weisshaar, B., & Holtgräwe, D. (2015). SMRT sequencing only de novo
791 assembly of the sugar beet (*Beta vulgaris*) chloroplast genome. *Bmc Bioinformatics*,
792 *16*(1), 1–10.
793 Syme, A. E., McLay, T. G. B., Udovicic, F., Cantrill, D. J., Murphy, D. J., McLay, T. G. B.,
794 Udovicic, F., Cantrill, D. J., & Murphy, D. J. (2021). Long-read assemblies reveal
795 structural diversity in genomes of organelles – an example with *Acacia pycnantha*.
796 *Gigabyte*, *2021*, 1–23. https://doi.org/10.46471/gigabyte.36
797 Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner,
798 S. (2017). GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic*
799 *Acids Research*, *45*(W1), W6–W11.
800 Twyford, A. D., & Ness, R. W. (2017). Strategies for complete plastid genome sequencing.
801 *Molecular Ecology Resources*, *17*(5), 858–868. https://doi.org/10.1111/1755-
802 0998.12626
803 Ueda, M., Fujimoto, M., Arimura, S., Murata, J., Tsutsumi, N., & Kadowaki, K. (2007). Loss
804 of the rpl32 gene from the chloroplast genome and subsequent acquisition of a
805 preexisting transit peptide within the nuclear gene in *Populus*. *Gene*, *402*(1–2), 51–56.
806 Ueda, M., Nishikawa, T., Fujimoto, M., Takanashi, H., Arimura, S., Tsutsumi, N., &
807 Kadowaki, K. (2008). Substitution of the gene for chloroplast RPS16 was assisted by
808 generation of a dual targeting signal. *Molecular Biology and Evolution*, *25*(8), 1566–
809 1575.
810 Wang, J. R., Holt, J., McMillan, L., & Jones, C. D. (2018). FMLRC: Hybrid long read error
811 correction using an FM-index. *BMC Bioinformatics*, *19*(1), 1–11.
812 https://doi.org/10.1186/s12859-018-2051-3
813 Wang, W., Schalamun, M., Morales-Suarez, A., Kainer, D., Schwessinger, B., & Lanfear, R.
814 (2018). Assembly of chloroplast genomes with long- and short-read data: A
815 comparison of approaches using *Eucalyptus pauciflora* as a test case. *BMC Genomics*,
816 *19*(1), 1–15. https://doi.org/10.1186/s12864-018-5348-8
817 Weng, M. L., Blazier, J. C., Govindu, M., & Jansen, R. K. (2014). Reconstruction of the
818 ancestral plastid genome in Geraniaceae reveals a correlation between genome

819 rearrangements, repeats, and nucleotide substitution rates. *Molecular Biology and*
820 *Evolution*, *31*(3), 645–659.
821 Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial
822 genome assemblies from short and long sequencing reads. *PLoS Computational*
823 *Biology*, *13*(6), e1005595.
824 Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive
825 visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350–3352.
826 Wu, H., Yang, J.-B., Liu, J.-X., Li, D.-Z., & Ma, P.-F. (2021). Organelle Phylogenomics and
827 Extensive Conflicting Phylogenetic Signals in the Monocot Order Poales. *Frontiers in*
828 *Plant Science*, *12*.
829 Xia, C., Wang, M., Guan, Y., Li, Y., & Li, J. (2022). Comparative analysis of complete
830 chloroplast genome of ethnodrug *Aconitum episcopale* and insight into its
831 phylogenetic relationships. *Scientific Reports*, *12*(1), 1–13.
832 https://doi.org/10.1038/s41598-022-13524-3
833 Xiang, Q. Y., Crawford, D. J., Wolfe, A. D., Tang, Y. C., & DePamphilis, C. W. (1998).
834 Origin and biogeography of *Aesculus* L. (Hippocastanaceae): a molecular
835 phylogenetic perspective. *Evolution*, *52*(4), 988–997. https://doi.org/10.1111/j.1558-
836 5646.1998.tb01828.x
837 Xu, K., Lin, C., Lee, S. Y., Mao, L., & Meng, K. (2022). Comparative analysis of complete
838 *Ilex* (Aquifoliaceae) chloroplast genomes: Insights into evolutionary dynamics and
839 phylogenetic relationships. *BMC Genomics*, *23*(1), 1–14.
840 https://doi.org/10.1186/s12864-022-08397-9
841 Yu, J., Fu, J., Fang, Y., Xiang, J., & Dong, H. (2022). Complete chloroplast genomes of
842 *Rubus* species (Rosaceae) and comparative analysis within the genus. *BMC Genomics*,
843 *23*(1), 1–14. https://doi.org/10.1186/s12864-021-08225-6
844 Zerbino, D. R. (2010). Using the Velvet de novo assembler for short-read sequencing
845 technologies. *Current Protocols in Bioinformatics*, *31*(1), 11–15.
846 Zhong, X. (2020). *Assembly, annotation and analysis of chloroplast genomes*. [Doctoral
847 Thesis, The University of Western Australia]. https://doi.org/10.26182/5f333d9ac2bee
848 Zhu, B., Gao, Z., Luo, X., Feng, Q., Du, X., Weng, Q., & Cai, M. (2019). The complete
849 chloroplast genome sequence of garden cress (*Lepidium sativum* L.) and its
850 phylogenetic analysis in Brassicaceae family. *Mitochondrial DNA Part B*, *4*(2), 3601–
851 3602.
852

853

854 **CONFLICT OF INTEREST**

855 The authors declare no competing financial interests.

856

857 **DATA ACCESSIBILITY AND BENEFIT-SHARING**

858 Demultiplexed sequence data of short-read and long-read data are available for download

859 from the NCBI Sequence Read Archive (SRA) (BioProject PRJNA865266). The accession

numbers of *J. roemerianus* and *J. validus* are OP235509 and OP235510, respectively. Information related to ptGAUL can be fetched in GitHub (https://github.com/Bean061/ptgaul).

AUTHOR CONTRIBUTIONS

WZ developed the ptGAUL pipeline assembled the *Juncus* plastome and prepared most of the first draft of the manuscript. CEAF assembled downloaded, publicly available reads using ptGAUL and made modification to the script. Chaehee provided *Eleocharis dulcis* long-read data and confirmed the analyses on the rearrangement events of plastome and helped annotate the *Juncus* plastomes. Ruisen Lu helped analyze the long repeats and SSR numbers in *Juncus* and polished the annotation result for NCBI. Jeremy Wang helped modified the ptGAUL script. Robert Jansen and Tracey Ruhlman helped discuss and write the introduction and discussion on plastome rearrangement events. Alan Jones and Corbin Jones are the senior corresponding authors guiding this project and they polished the prose.

Table 1 ptGAUL performance on 15 published sequence data sets, including the information of assembled plastome from published papers and the information on assembled plastomes from ptGAUL. NA means low nucleotide sequence identity between assembled plastome between published data and our data. S means the samples are well assembled by ptGAUL, while F means the samples failed using ptGAUL. * column includes the references we used for genome assembly in ptGAUL and the bold references were considered as references for comparisons with ptGAUL results.

Species	Library preparation and sequencing methods	Raw read No./N50 (bp)	Reference	Plastid size from ptGAUL (bp) (% nucleotide sequence identity to references)	Number of assembled plastid contigs from ptGAUL	Plastome reference used for ptGAUL (reference length from original studies)*
Arctostaphylos glauca	WGS/PacBio	1814591/15245	Huang et al., 2022	150241 (NA)	3 (S)	NC_035584.1/NC_042713.1/NC_047438.1/**JAHSPW020000272.1 (118663 bp)**
Lepidium sativum	WGS/PacBio	400322/7277	Zhu et al., 2019	153666 (99.9%)	3 (S)	**NC_047178.1 (154997 bp)**
Chaetoceros muellerii	WGS/PacBio	87313/12921	Li & Deng, 2021	117304 (99.8%)	1 (S)	**MW004650.1 (116284 bp)**
Potentilla micrantha	WGS/PacBio	28638/2464	Ferrarini et al., 2013	159850 (99.8%)	3 (S)	**NC_015206.1 (155691 bp)**
Durio zibethinus	WGS/PacBio	853182/9670	Shearman et al., 2020	142806 (99.95%)	1 (S)	**MT321069 (163974 bp)**
Beta vulgaris	WGS/PacBio	96874/3980	Stadermann et al., 2015	155383 (99.9%)	3 (S)	**KR230391.1 (149722 bp)**
Eleocharis dulcis	WGS/PacBio	68167/16288	Lee et al., 2020	199919 (99.5%)	3 (S)	**NC_047447.1 (199561 bp)**
Eucalyptus pauciflora	WGS/ONT	705554/24988	Wang et al., 2018	158561 (99.0%)	1 (S)	MZ670598.1/HM347959.1/NC_014570.1/AY780259.1/ **NC_039597.1 (159942 bp)**
Leucanthemum vulgare	Long range PCR/ONT	18031/7900	Scheunert et al., 2020	119593 (NA)	5 (F)	**NC_047460.1 (150191 bp)**
Oryza glaberrima	Plastid capture/ONT	81363/4319	Bethune et al., 2019	124133 (NA)	4 (F)	**NC_024175.1 (132629 bp)**
Cenchrus americanus	Plastid capture /ONT	105760/5580	Bethune et al., 2019	143162 (96.6%)	3 (S)	**NC_024171.1 (140718 bp)**
Digitaria exilis	Plastid capture /ONT	141250/4028	Bethune et al., 2019	136650 (96.0%)	3 (S)	**NC_024176.1 (140908 bp)**
Podococcus acaulis	Plastid capture /ONT	249417/2621	Bethune et al., 2019	81976 (NA)	2 (F)	**NC_027276.1 (157688 bp)**
Raphia textilis	Plastid capture /ONT	83833/2495	Bethune et al., 2019	60089 (NA)	2 (F)	**NC_020365.1 (157270 bp)**
Phytelephas aequatorialis	Plastid capture /ONT	202925/2551	Bethune et al., 2019	NA	(F)	**NC_029957.1 (159075 bp)**
Picea glauca	WGS/PacBio	563675/4671	Soorni et al., 2017	123476 (98.9%)	1 (S)	**NC_021456.1 (124084 bp)**

878 **Table 2** Summary of features of the plastid genomes of four *Juncus* species, including length,
879 GC content, and gene numbers. The plastome data of *Juncus effusus* were from two different
880 two different sources, this paper and Lu et al. (2021). PCG =protein-coding genes.

Genome Features	***J. effusus***	***J. effusus***	***J. inflexus***	***J. roemerianus***	***J. validus***
Accession No.	NC_059754.1	Present study	Present study	OP235509	OP235510
No. of Illumina read clusters	12443053	96,653,565	83412073	158922322	156712430
No. of ONT reads and N50	0	2960380/21529	2735792/24397	427549/15998	243884/14365
Genome size (bp)	170612	178158	181566	196852	147183
LSC length (bp)	81818	86497	86649	82944	87215
SSC length (bp)	7522	7539	7509	7902	2046
IR length (bp)	40636	42061	43704	53003	28961
Overall GC content, %	36.0	35.9	35.6	32.2	34.7
GC content in LSC, %	33.2	33.2	33.3	33.1	31.6
GC content in SSC, %	26.3	26	26.2	26.5	23
GC content in IR, %	39.7	39.5	38.7	37.5	39.8
Total No. of genes	129	133	134	136	114
No. of unique genes	105	106	106	106	93
No. of unique PCGs	72	72	72	72	60
No. of unique tRNA genes	29	30	30	30	29
No. of unique rRNA genes	4	4	4	4	4

881 **Table 3** Statistics of dispersed and tandem repeats in *Typha*, *Ananas*, *Eriocaulon*, and *Juncus* plastomes

Species	***Typha latifolia***	***Ananas comosus***	***Eriocaulon decemflorum***	***Juncus effusus***	***Juncus inflexus***	***Juncus roemerianus***	***Juncus validus***
Genome size (no IRa)	134,642	132,862	125,164	136,097	137859	143849	118,221
GC %	35.5	36.3	34.2	34.8	34.6	34.3	33.5
Dispersed repeat (DR)							
Length of DRs	1,210	1,495	1,418	15,117	13,229	14,712	14,714
GC %	33.7	36.3	33	35	36	35.7	34
GC % without DR	35.5	36.3	34.2	34.7	34.6	34.1	33.3
% of DR in genome	0.9	1.1	1.1	11.1	9.6	10.2	12.4
Tandem repeat (TR)							
Length of TRs	3,270	2,057	859	12,248	15,783	22,978	8,797
GC % of TRs	8.8	18.4	20	34.2	33.4	32.1	32.1
Genome size without TRs	131,372	130,805	124,305	123,849	122,076	120,871	109,424
GC % without TRs	36	36.6	34.3	34.8	34.8	34.8	33.6
% of TRs in genome	2.4	1.5	0.7	9.0	11.4	16.0	7.4
Total repeat							
Length of total repeats	4,436	3,552	2,227	23,345	26,451	35,027	22,577
GC % of total repeats	12.6	21.5	27.5	34.7	34.7	33.6	33.5
GC % without total repeats	36	36.6	34.3	34.8	34.8	34.5	33.4
% of total repeats in genome	3.3	2.7	1.8	17.2	19.2	24.3	19.1

882

883 **Table 4** Summary of breakpoint and reversal distances for plastomes of *Juncus*, *Eriocaulon* and basal Poales.

Species	*Typha latifolia*	*Ananas comosus*	*Eriocaulon decemflorum*	*J. effusus*	*J. inflexus*	*J. roemerianus*	*J. validus*
Typha latifolia	—						
Ananas comosus	0/0	—					
Eriocaulon decemflorum	0/0	0/0	—				
J. effusus	15/19	15/19	15/19	—			
J. inflexus	15/19	15/19	15/19	—	—		
J. roemerianus	17/20	17/20	17/20	7/9	7/9	—	
J. validus	14/17	14/17	14/17	8/10	8/10	11/13	—

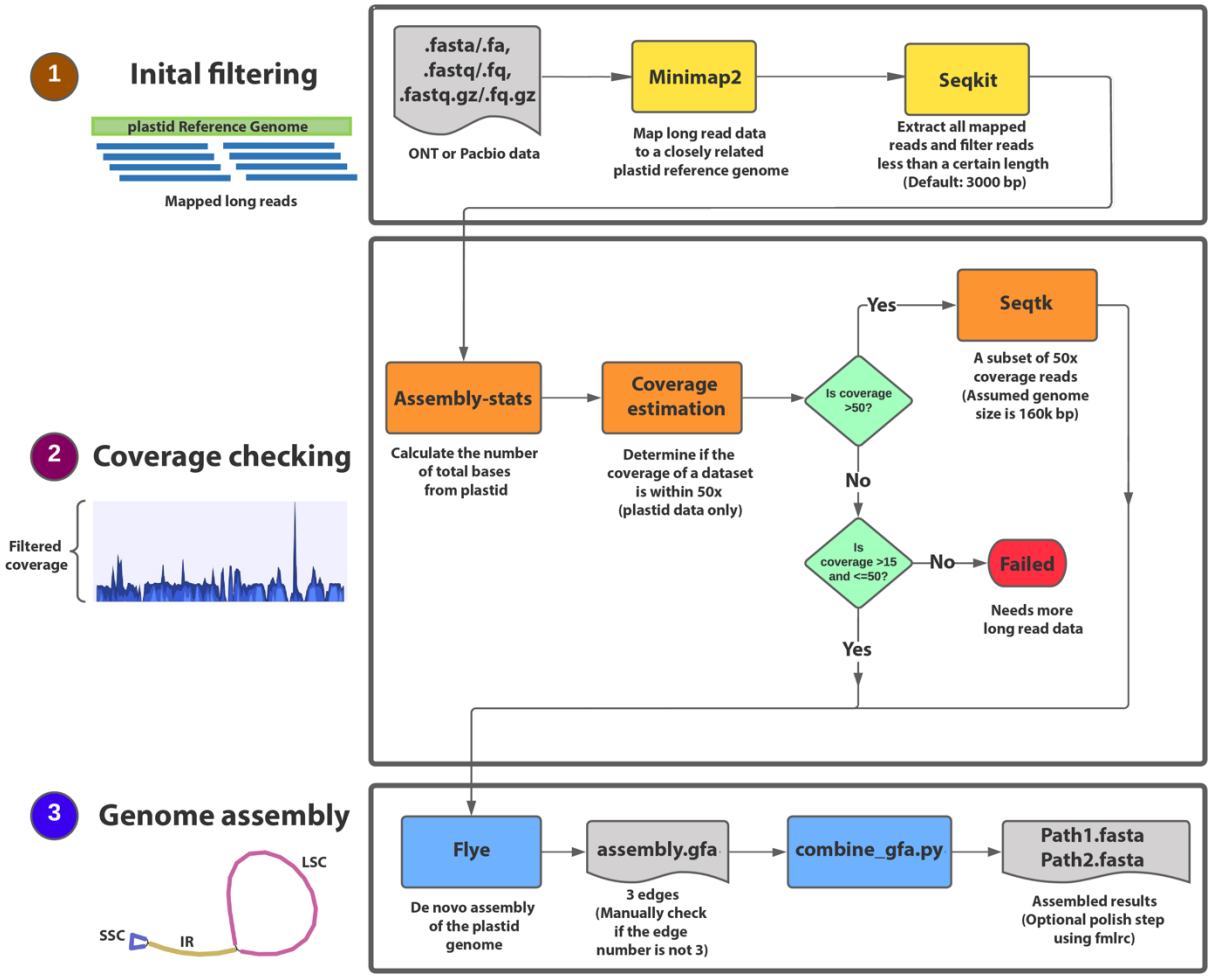

Figure 1 ptGAUL workflow. The program starts with an initial filtering step to filter the long reads of the target species using at least one closely related reference plastome (1). Subsequently, the coverage for those filtered long reads is calculated and filtered to make sure it is about 50x (2). Finally, two paths of plastomes were obtained through Flye and a customized python script, combine_gfa.py (3).

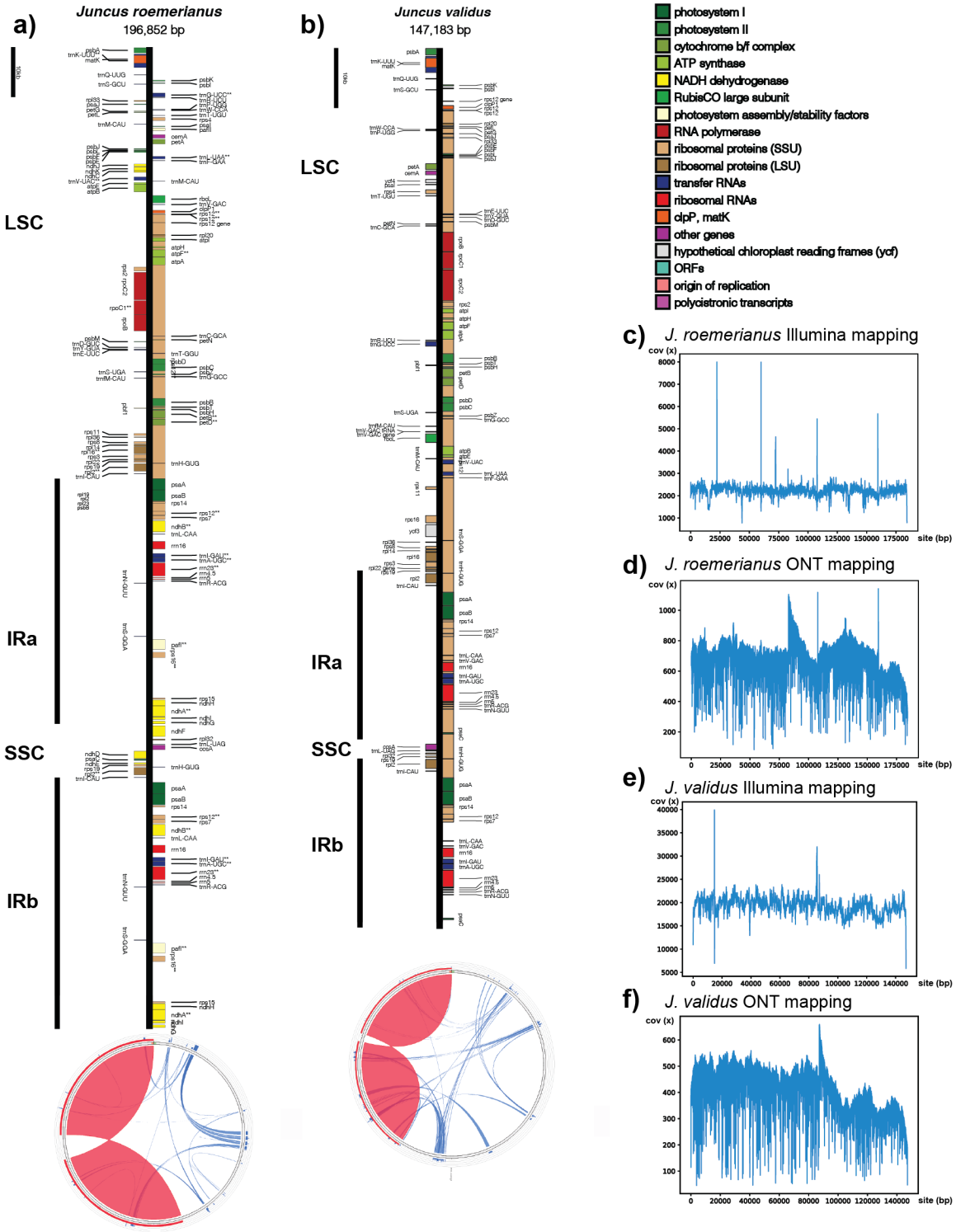

Figure 2 Plastome structural maps and read coverage graphs of *J. roemerianus* and *J. validus*. a) and b) Linear maps of *J. roemerianus* and *J. validus* plastome, respectively, were drawn by OGDRAW (Lohse et al., 2013). Genes that belong to different functional groups are color-coded. Small single copy (SSC), large single copy (LSC), and inverted repeats (IRa, IRb) are indicated for both plastomes. Circular representatioins of the two *Juncus* plasstomes were used to shot locations of repetitive DNA using Circoletto (Darzentas, 2010). The blue lines represent dispersed repeats in the plastome, while red regions represent the IR regions. c) - f) Read coverage plots of *J. roemerianus* and *J. validus* using Illumina reads and ONT reads,

respectively, showing the good quality of the assemblies. The *x* axis represents the position in the plastome, while *y* axis represents the coverage.

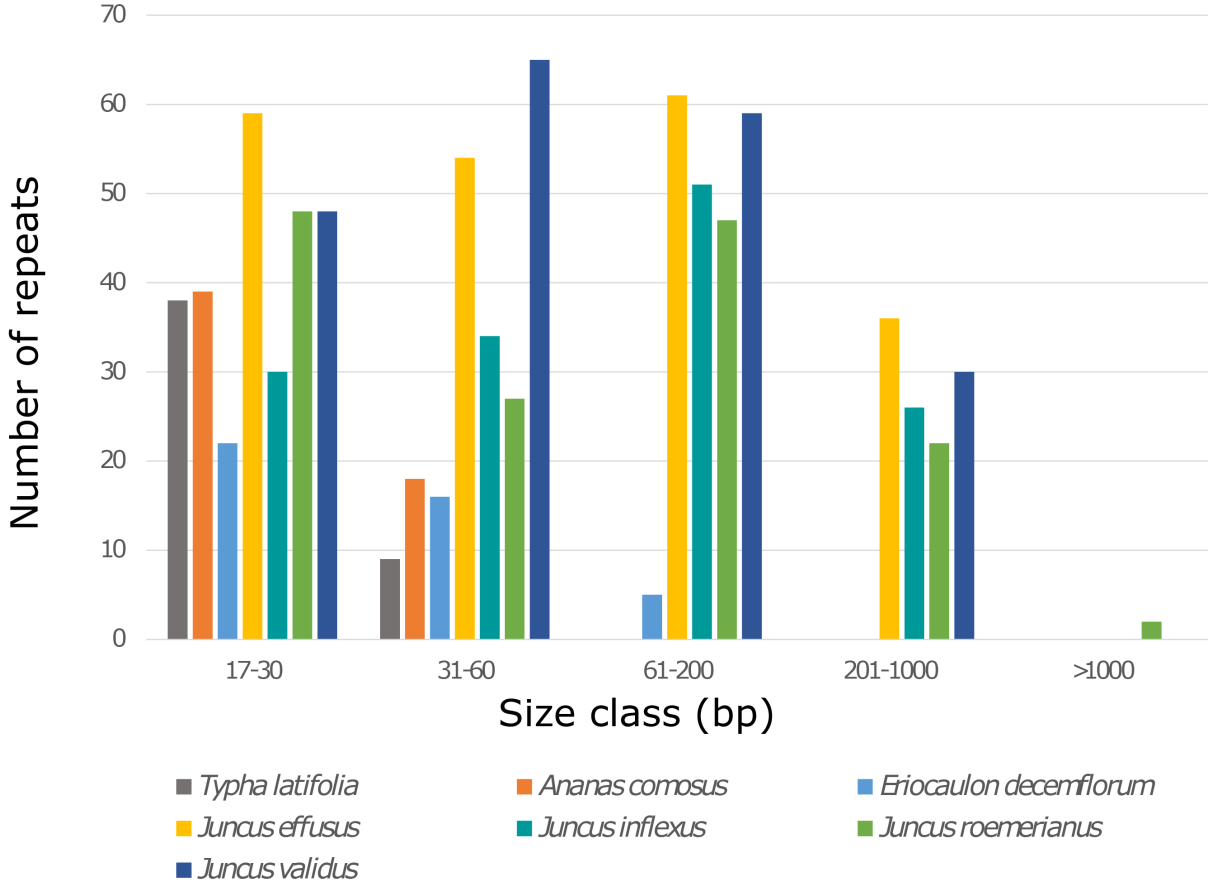

Figure 3 Bar plot of dispersed repeats in plastomes from seven Poales species, including four newly assembled *Juncus* species.

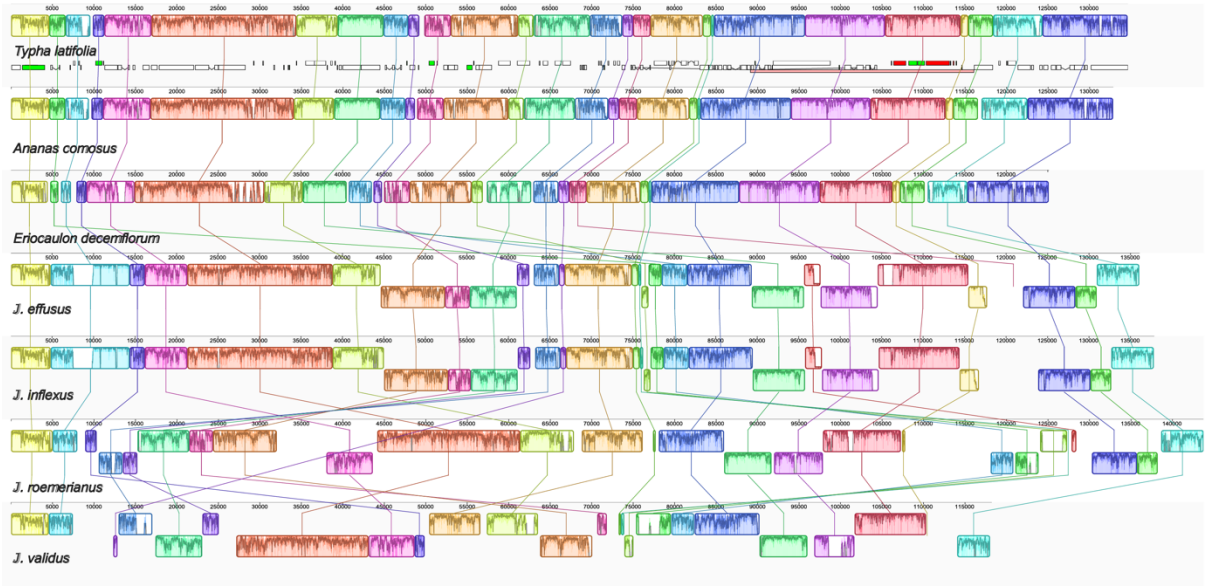

Figure 4 Whole plastome alignment of seven Poales species, including four newly assembly *Juncus* and *Typha latifolia*, *Ananas comosus,* and *Eriocaulon decemflorum*. The local colinear blocks (LCBs) were identified by progressiveMauve with *Typha* plastome as the reference. The corresponding LCBs among seven plastomes are shaded and connected with a line of the same color. LCBs that are flipped indicate inversions. Numbers on the upper *x*-axis are genome map coordinates in basepairs (bp).