Title
Plastid Genome Assembly Using Long-read data.

Permalink
https://escholarship.org/uc/item/8f89x934

Journal
Molecular Ecology Resources, 23(6)

Authors
Zhou, Wenbin
Armijos, Carolina
Lee, Chaehee
et al.

Publication Date
2023-08-01

DOI
10.1111/1755-0998.13787

Peer reviewed

Published in final edited form as:
Mol Ecol Resour. 2023 August ; 23(6): 1442–1457. doi:10.1111/1755-0998.13787.

Plastid Genome Assembly Using Long-read data

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

[1]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[2]Laboratorio de Biotecnología Vegetal, Universidad San Francisco de Quito USFQ, Quito, Ecuador

[3]Department of Plant Sciences, University of California Davis, Davis, California, USA

[4]Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing, China

[5]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[6]Department of Integrative Biology, University of Texas at Austin, Austin, Texas, USA

[7]Department of Pharmacology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Abstract

Although plastid genome (plastome) structure is highly conserved across most seed plants, investigations during the past two decades have revealed several disparately related lineages that experienced substantial rearrangements. Most plastomes contain a large inverted repeat and two single-copy regions, and a few dispersed repeats; however, the plastomes of some taxa harbour long repeat sequences (>300 bp). These long repeats make it challenging to assemble complete plastomes using short-read data, leading to misassemblies and consensus sequences with 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*lastid *G*enome *A*ssembly *U*sing *L*ong-

Correspondence: Wenbin Zhou, Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA., wenbin.evolution@gmail.com.

AUTHOR CONTRIBUTIONS
Wenbin Zhou developed the ptGAUL pipeline, assembled the *Juncus* plastome and prepared most of the first draft of the manuscript. Carolina E. Armijos assembled and 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 analyse 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.

CONFLICT OF INTEREST STATEMENT
The authors declare no competing financial interests.

BENEFIT-SHARING STATEMENT
Benefits Generated: Benefits from this research accrue from the sharing of our data and results on public databases as described above.

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

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 data sets. We showed that ptGAUL quickly produces accurate and unbiased assemblies using only ~50× coverage of plastome data. Additionally, we deployed 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. The ptGAUL pipeline is available on GitHub: https://github.com/Bean061/ptgaul.

Keywords

chloroplast; Juncaceae; *Juncus* ; long-read assembly; Poales; rearrangement events

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 have 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, and expansion and contraction of the IR—and IR loss have 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 (Freudenthal et al., 2020; Twyford & Ness, 2017). 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 either the seed-extend method (Dierckxsens et al., 2017) or 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 larger than kmer sizes. Sometimes 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. However, 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. Short reads with a typical insert size (300–400 bp) are inadequate to obtain a complete plastome assembly for plant species that have large repeats and that may be highly rearranged. So far, few plant systematists have recognized this as an issue, probably 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. This may shift as the cost of sequencing continues to decline and the use of entire plastomes for analysis becomes more common. 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, there are several tools available 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 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 data sets. The approach of Syme et al. (2021) requires an extra step to manually filter a subset of raw reads matching the plastome (~250× coverage) and sometimes generates multiple contigs in the assembly result. CANU can yield different results depending on different read coverages (Wang, Schalamun, et al., 2018). UNICYCLER was designed for hybrid data; however, it takes an extremely long time to finish as input data are increased. All the above pipelines can probably assemble the conventional plastome but cannot assemble atypical plastome structures accurately. Alternatively, a “fishing approach” using either SHASTA (for Nanopore data; Shafin et al., 2020) or HIFIASM (for PacBio data; Cheng et al., 2021; Feng et al., 2022) to assemble the raw reads first, followed by fishing out the plastid contigs using the reference genome can be used. Nevertheless, this assembly process is time- and resource-consuming when the input data are large, according to the guidelines of each software. Additionally, the accuracy of the assembly results may be affected by a considerable number of redundant sequences (Zhang et al., 2022).

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, 2010; Drábková et al., 2006). Although many species of Juncaceae have been included in phylogenetic studies using plastid gene sequences and the internal transcribed spacer (ITS) region of the nuclear ribosomal repeat (Table S1; Brožová et al., 2022; Drábková, 2010; Drábková et al., 2006), species relationships within *Juncus* remain unresolved. Brožová et al. (2022) recently incorporated *rbcL, trnL, trnL-trnF* and ITS1–5.8–ITS2 regions 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 confusion regarding *Juncus* species names, we did not adopt the generic classification of Brožová et al. (2022) in the present investigation because a more comprehensive study with more markers is necessary to justify this reranking. Plastomes of just eight *Juncus* species are publicly available in GenBank (Lu et al., 2021; Wu et al., 2021), but there has been little investigation of how the plastome itself has changed structurally in *Juncus* (*s.l.*). Wu et al. (2021) focused on the phylogenetic relationships in the Poales using shared plastid protein-coding genes, but did not provide any information 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 *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 Juncaceae would allow a deep understanding of plastome evolution inside the family and provide more phylogenetic insights within Juncaceae and Poales.

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

2 | MATERIALS AND METHODS

2.1 | *Juncus* sample collection and DNA extraction

Young leaves of *Juncus roemerianus* (voucher number; NCU00441655) and *Juncus validus* (voucher number: NCU00434802) were collected from North Carolina, USA, and stored in silica gel. Vouchers were deposited in the herbarium of the University of North Carolina at Chapel Hill (NCU). Total genomic DNA extraction of dried leaves was performed using a modified cetyltrimethylammonium bromide (CTAB) protocol described by Cullings (1992) and Xiang et al. (1998). DNA quantity was analysed with Qubit 2.0 (Life Technologies) and quality was measured using a NanoDrop spectrophotometer 2000 (ThermoFisher Scientific) and 1% (w/v) agarose gels. Sequencing was performed at the High-Throughput Sequencing Facility (HTSF) at UNC Chapel Hill. For Illumina sequence libraries (Illumina), ~250 ng of total DNA was utilized. An agilent 2100 Bioanalyzer (Agilent Technologies) 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).

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 entirely 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 at https://github.com/sanger-pathogens/assembly-stats). If the coverage is over 50×, apply SEQTK (available in https://github.com/lh3/seqtk) to randomly select a subset of data, including about 50× coverage of the plastome (excessive coverage will slow down the assembly and may result in a failure to assemble). Step 3: use FLYE (Kolmogorov et al., 2019) to assemble the plastome. When only three contigs are detected in the graphical fragment assembly (.gfa) file, we use "combined_gfa.py," a customized python script, to assemble the plastome into two different paths. Otherwise, the assembly result can be checked manually using BANDAGE. All pipeline code was deposited on GitHub (https://github.com/Bean061/ptgaul). In Steps 2 and 3, we implement SEQTK to randomly choose a subset of long reads to minimize the bias of read selection, which can speed up the assembly process. To make it more user-friendly, we added three parameters that included: the expected plastome size (-g) with default value of 160,000 bp, the expected coverage of plastome data for assembly (-c) with a default value of 50× coverage, and the output directory (-o).

After assembly, if short-read sequencing data are available, the FM-index Long-Read Corrector (FMLRC) software (Wang, Holt, et al., 2018) is recommended to polish and improve the accuracy of the final assembled sequences because it can generate a more accurate assembly result (Mak et al., 2023).

Long-read data from 16 published plastomes in NCBI were used to validate the efficacy of ptGAUL for assembly (Table 1). All analyses were run on the longleaf high-performance compute cluster at UNC Chapel Hill. Comparative analyses were performed that included the number of assembled contigs, total genome size (bp) and nucleotide sequence identity between the published results and those obtained with ptGAUL through a pairwise identity alignment using GENEIOUS version 2022.2 (Kearse et al., 2012). We also compared the performance (memory usage and running time) of ptGAUL to a "fishing approach" by contrasting it against SHASTA version 0.10.0 for ONT data (Shafin et al., 2020) and HIFIASM version 0.16 for PacBio data (Cheng et al., 2021; Feng et al., 2022). Then, we used MINIMAP2 version 2.24 (Li et al., 2018) to map the assembled contigs and a reference sequence (Table S6) to locate the plastome.

2.3 | Assembly and comparison of four *Juncus* species

The Illumina Novaseq 6000 platform (Illumina) was used to generate 250-bpPE reads for *J. roemerianus* and *J. validus.* Reads were de novo assembled using GETORGANELLE version 1.7.5 (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 parameters ("–f"=3000 bp; "–g"=160,000 bp; "–c" = 50×) with all eight *Juncus* plastomes from GenBank (Table 2) as references for the filtering step. We verified the assembly graph results (.gfa file) obtained with FLYE version 2.7 in BANDAGE version 0.8.1 (Wick et al., 2015). Then, we used FMLRC version 1.0.0 to polish the final plastomes (an optional step in the ptGAUL pipeline).

To assess the quality of the assembly result, we mapped all raw Illumina and ONT reads of each *Juncus* species to our polished assembly and tested the evenness of the coverage at all sites. If every site shared a similar coverage of raw reads without gaps in coverage, this usually indicated a good de novo assembly result. We used the SAMTOOLS version 1.9 (Danecek et al., 2021) depth function to record read depth at every site, followed by a dot plot created using python's matplotlib library (Hunter, 2007). We downloaded the raw whole-genome sequencing data of *J. effusus* and *J. inflexus* (both ONT reads: SRR14298760/SRR14298751 and Illumina reads: SRR14298746/SRR14298745 from Planta et al., 2022) to assemble plastomes following the same steps detailed above.

After assembly, we uploaded the 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 version 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 its 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 version 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 the repeat analyses to avoid counting the repeats from this region. We implemented BLAST version 2.8.1+ (Altschul et al., 1990) and TANDEM REPEATS FINDER version 4.09.1 (Benson, 1999) to detect the dispersed repeats and tandem repeats, respectively, following the steps from Lee et al. (2020). We manually checked the result, eliminated duplicated BLAST hits and recorded the total number of distinct dispersed repeats. We also downloaded the complete plastomes of *Eriocaulon decemflorum* (NCJD44895; 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 the MATPLOTLIB library (Hunter, 2007) from Python.

We focused on the four confirmed assemblies of *Juncus,* namely *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 unreliable assemblies resulting from short-read data. To eliminate uncertainty in short-read assemblies, we compared the sequence identity between the *J. effusus* plastome assembled from short-read data (Lu et al., 2021) and 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 arrangements of locally colinear blocks (LCBs) using progressive MAUVE (Darling et al., 2004). One copy of the IR was removed from plastomes prior to MAUVE alignment to prevent spurious alignments. *Typha latifolia* was used 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 species, with plastome sizes similar to those reported previously (indicated with an “S” in Table 1). The assembly graph results (.gfa files) showing plastome structure were visualized and confirmed in BANDAGE and deposited in GitHub (https://github.com/Bean061/ptgaul). Assembled plastomes had >95% nucleotide sequence identity to the references, but 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 an “F” in Table 1). The ptGAUL pipeline produced consistent and reliable results with a data set of long reads (>5000 bp N50) with ~50× coverage of the plastome.

Our results indicated that different library preparation methods 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 (verified in 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 and 7277 bp, respectively), but the *Leucanthemum vulgare* library prepared using long-range PCR failed in plastome assembly. All five failed data sets involved the plastid capture approach and most of the raw sequence reads had relatively short length with small N50 values (<5000 bp) (Table 1).

3.2 | Plastome features of four *Juncus* species

We generated 158,922,322 and 156,712,430 short reads for *J. 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 are accessible at NCBI under the BioProject accession no.: PRJNA865266 (SRR21976089; SRR21976090; SRR21976091; SRR21976092). 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. effuses* and *J. inflexus* ranged from 147,183 to 196,852 bp, had similar sized LSCs, different sizes of the SSC 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 plastome assembly. All mapping results showed a high and even coverage of four species (Figure 2c–f; Figure S2a,b,d,e). There were no gaps in the assemblies regardless of the sequencing platform. Annotation of the four *Juncus* plastomes revealed that they contained 114–136 genes, 93–106 of which were unique. There were 60–72 unique protein coding genes (PCGs), 29–30 tRNA genes and four rRNA genes (Table

2). *J. roemerianus* had the greatest gene number (136), which is similar to *J. effusus* (133), *J. inflexus* (134), and basal Poales ancestors such as *Typha latifolia* (133), *Ananas comosus* (132) and *Eriocaulon decemflorum* (135) (Table S2). *J. effusus* and *J. inflexus* shared a highly similar gene content while *J. validus* lacked 11 *ndh* genes, *rps15* and *trnT-GGU* (Figure S3; Table S2).

3.3 | Verification of published *J. effusus* plastome

We compared the *J. effusus* published assembly based on short-read data (Lu et al., 2021, MW366789) with our new ptGAUL assembly employing the long-read data from 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 vs. 178,158 bp). The mapping results showed that our assembly was well supported by both long- and short-read data from Planta et al. (2022) (Figure S2a,b). Yet, it was unsupported by the Illumina reads from Lu et al. (2021) with 777 positions with less than 10× coverage, including 295 positions that had 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 these results, we removed the eight publicly available *Juncus* plastomes assembled with short-read data prior to the comparative analyses of plastomes as we thought that these would unfairly bias the comparisons in favour of ptGAUL.

3.4 | Repeats in *Juncus* plastomes

Repeat analyses identified many dispersed and tandem repeats in the four *Juncus* plastomes (17.2%–24.3% of the genome without IRa) in comparison with basal Poales and *Eriocaulon* (1.8%–3.3% of the 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 for *Typha* (4436 bp), *Ananas* (3552 bp) and *Eriocaulon* (2227 bp) (Table 3). When dispersed repeats were parsed into five different size classes, *Juncus* plastomes contained more dispersed repeats than basal Poales and *Eriocaulon* (Figure 3; Table S3). Larger repeats (>201 bp) were found only in *Juncus* (Figure 3; 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 the dispersed repeats in *J. inflexus* and *J. roemerianus.* All four *Juncus* plastomes contained exceptionally expanded tandem repeats, ranging from 4.6 to 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 (BPs) 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,* respectively. *J. 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 I DISCUSSION

4.1 | ptGAUL application and suggestions for sequencing approach

The ptGAUL pipeline generated either one or three contig(s) for 11 publicly available data sets using PacBio or ONT data (Table 1). However, it failed to assemble the data from five species generating more than three contigs and predicted a 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 at a basepair level with over 96%–99% nucleotide sequence identity. The lower percentage 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 the failed assemblies were different from the whole-genome sequencing method in 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; for example, 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 nonmodel 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 caused by poor storage, the 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, Crawley, UK). CTAB, SDS or other methods that can produce much higher molecular weight (HMW) DNA are preferred for third-generation sequencing (Jung et al., 2019; Mayjonade et al., 2016), 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 (Figure S6). Overall, the average assembly time for each *Juncus* is about 1 hr (Table S5). All *Juncus* plastome assemblies Indicated many fragmented contigs or assembly paths (Figure S6). This is because the numerous long-dispersed repeats present in the *Juncus* plastomes are longer than the kmer size/length of short reads. Based on our *J. effusus* plastome comparison, the final assembly length and the total number of genes based on short-read data are smaller than those assembled from long-read data (Table 2; Table S2), which might be caused by the random selection of one of the paths as the final assembly when using short-read data. Other studies demonstrated that a three-step approach can resolve this issue: (i) by comparing different contigs from short-read assemblers (e.g., SPADES, VELVET), (ii) by manually checking through the contigs when contrasted with closely related species, and (iii) using 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 previous work (Lee et al., 2020; Ruhlman et al., 2017), showing that long-read data vastly improve the assembly of plastomes with many long repeats. Based on our study and that of Scheunert et al. (2020), ~50× mapping coverage of long-read data can result in an accurate plastome assembly. In our research, long reads of plastid origin represented 5%–6% of reads generated from the total genomic DNA of *Juncus.* Assuming *5%* plastid DNA content from whole-genome HMW extractions, generating ~50× coverage of a 160,000-bp plastome requires only ~160 Mb reads per sample. Currently, one chip of ONT generates ~10 Gb of sequence data, enabling multiplexing up to 64 samples at consumable cost of roughly $1000 USD.

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), CANU and HINGE (Wang, Schalamun, et al., 2018) cannot generate a consistent plastome assembly result with one contig when using different data coverage, UNICYCLER (Wick et al., 2017) is computationally intensive and does not produce well-resolved assemblies when dealing with complicated plastomes with many long repeats. The “fishing approach” associated with the assembly process used in SHASTA and HIFIASM assemblers requires a considerable utilization of either time (HIFIASM) or memory (SHASTA) (Table S6). These “fishing approaches” can find many relatively short plastid contigs, for example 16 contigs in *Eucalyptus pauciflora* with a maximum of 26,790 bp matching to the reference and 13 contigs in *Arctostaphylos glauca* with a maximum of 85,406 bp matching to the reference (Table S6). Compared to currently published pipelines for plastome assembly, ptGAUL can help generate accurate plastome assemblies in less than 10 min with ~16 Gb memory when the raw sequence data are less than 10 Gb (Table S6), making it highly convenient and typically significantly faster than other tools (Tables S5 and S6). 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 species with multiple plastome 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 intra-individual structural heteroplasmy (Lee et al., 2020). Similar phenomena contributed to size variation in the four *Juncus* plastomes, which ranged from 147,183 to 196,852 bp (Table 2). Many long repeats, including an unusually high number of dispersed repeats of 61–200 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 and 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 *Trachelium caeruleum,* 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 (Choi et al., 2019, 2020; Ruhlman et al., 2017; Schwarz et al., 2015; Sloan et al., 2014; Weng et al., 2014). We speculate that the recombinogenic potential of long repeats identified in the *Juncus* plastomes contributed to the diversification of 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 noncoding 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 to 170 kb and 20 to 30 kb, respectively, but most IR-containing angiosperms sequenced to date display highly similar gene arrangement and plastome size (~150 kb; IR, ~25 kb; 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;* 186,255/37,438 kb), and the smaller plastomes contained the larger IRs (i.e., *C. difformis;* 167,974/38,427 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 a more conserved structure remain elusive. The smallest of the four plastomes, in *J. Validus,* would appear to be a typical plastome based on the overall plastome and IR size (~147 and ~29 kb). However, the assembly and annotation show that it is not always size that matters. This plastome has probably experienced/is experiencing 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 and 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 cyclical 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). Loss of the NDH suite from the *J. validus* plastome is unique among examined Poales plastomes and suggest that an active PGR5-dependent pathway accounts for CEF in this species.

Apart from the loss of NDH genes, gene losses were shared by all four *Juncus* examined and included other genes that were lost from plastomes of diverse lineages (Ruhlman & Jansen, 2018). The plastid-localized acetyl-coenzyme A carboxylase (ACCase; prokaryotic) is another multisubunit protein complex that incorporates nuclear-encoded polypeptides and participates in fatty acid metabolism (Ohlrogge & Browse, 1995). The plastid *accD* encodes one subunit of the four-unit complex and was lost in numerous taxa, often those that experienced other gene loss and pseudogenization events (Ruhlman & Jansen, 2018). Because plastid ACCase activity was thought to have an essential function (Kode et al., 2005), *accD* loss in several groups suggested that it may be expressed from a functional transfer to the nucleus or substituted by a redundant, nuclear-encoded enzyme (Konishi et al., 1996). In *Trifolium,* which lacks plastid *accD,* a functional transfer to the nucleus was uncovered (Magee et al., 2010). Further investigation failed to detect any remnant of the *accD* sequence in the plastomes of *T. repens* or *T. pratense,* while mutated copies were identified in *T. aureum* and *T. grandiflorum* (Sabir et al., 2014). The 15-amino-acid (aa) C-terminal catalytic domain of the ACCD protein, which is minimally required for prokaryotic ACCase function (Lee et al., 2004), was identified in the mutated copies and may indicate functionality. Probing nuclear transcriptomes from *T. repens* and *T. pratense* revealed that, as in *T. subterraneum* (Magee et al., 2010), a putatively functional ACCD protein was being expressed from a fusion sequence that included the ACCD catalytic domain (~270 aa) fused to the plastid target peptide from nuclear-encoded, plastid-targeted LPD1 (493 aa). Probing transcriptomes of related legumes that contained intact plastid *accD* was able to detect high-identity copies of the ACCD core sequence, suggesting that incorporation at nuclear loci pre-dated 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 and 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 nonphotosynthetic 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 a relaxed selection of the plastid gene (Park et al., 2017; Ueda et al., 2007).

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank the UNC greenhouse staff for maintaining living *Juncus* materials and the UNC herbarium staff for storing our voucher specimens. We are very grateful to UNC Research Computing and its longleaf high-performance cluster for computational resources. We also appreciate Dr P. F. Ma and Dr H. Wu for kindly sharing the *Juncus* Illumina data with us. This work was supported by National Science Foundation IOS-2034929 to A.M.J. and C.D.J.

DATA AVAILABILITY STATEMENT

Demultiplexed sequence data of short- and long-read data are available for download from the NCBI Sequence Read Archive (SRA) (BioProject PRJNA865266; SRR21976089, SRR21976090, SRR21976091 and SRR21976092). The accession numbers of *J. roemerianus* and *J. validus* are OP235509 and OP235510, respectively. Information related to ptGAUL can be retrieved from GitHub (https://github.com/Bean061/ptgaul).

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. [PubMed: 2231712]

Babiychuk E, Vandepoele K, Wissing J, Garcia-Diaz M, De Rycke R, Akbari H, 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. 10.11646/phytotaxa.376.2.3

Benson G (1999). Tandem repeats finder: A program to analyse DNA sequences. Nucleic Acids Research, 27(2), 573–580. [PubMed: 9862982]

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. [PubMed: 31139509]

Brožová V, Proćków J, & Drábková LZ (2022). Toward finally unravelling the phylogenetic relationships of Juncaceae with respect to another cyperid family. Cyperaceae. Molecular Phylogenetics and Evolution, 177(107), 588.

Cai Z, Guisinger M, Kim HG, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, & Jansen RK (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. [PubMed: 19018585]

Cheng H, Concepcion GT, Feng X, Zhang H, & Li H (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature methods, 18(2), 170–175. [PubMed: 33526886]

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. [PubMed: 31046101]

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. [PubMed: 33209275]

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, & Jansen RK (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. [PubMed: 16916942]

Compeau PEC, Pevzner PA, & Tesler G (2011). How to apply de Bruijn graphs to genome assembly. Nature Biotechnology, 29(11), 987–991. 10.1038/nbt.2023

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

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, & Li H (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2), giab008. 10.1093/gigascience/giab008 [PubMed: 33590861]

Darling ACE, Mau B, Blattner FR, & Perna NT (2004). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Research, 14(7), 1394. 10.1101/gr.2289704 [PubMed: 15231754]

Darshetkar AM, Datar MN, Tamhankar S, Li P, & Choudhary RK (2019). Understanding evolution in Poales: Insights from Eriocaulaceae plastome. PLoS ONE, 14(8), e0221423. 10.1371/journal.pone.0221423 [PubMed: 31430346]

Darzentas N (2010). Circoletto: Visualizing sequence similarity with Circos. Bioinformatics, 26(20), 2620–2621. [PubMed: 20736339]

Dierckxsens N, Mardulyn P, & Smits G (2017). NOVOPIasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Research, 45(4), e18. 10.1093/nar/gkw955 [PubMed: 28204566]

Drábková L (2010). Phylogenetic relationships within Juncaceae: Evidence from all three genomic compartments with notes to the morphology. In Seberg O, Petersen G, Barford AS, & Davis JI (Eds.), Diversity, phytogeny, and evolution in the monocotyledons (pp. 389–416). Aarhus University Press.

Drábková L, Kirschner J, & Vlček Č (2006). Phylogenetic relationships within *Luzula* DC. and *Juncus* L.(Juncaceae): A comparison of phylogenetic signals of trnL-trnF intergenic spacer, trnL intron and rbcL plastome sequence data. Cladistics, 22(2), 132–143. [PubMed: 34892869]

Feng X, Cheng H, Portik D, & Li H (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. Nature Methods, 19(6), 671–674. [PubMed: 35534630]

Ferrarini M, Moretto M, Ward JA, Šurbanovski N, StevanoviŠ V, Giongo L, Viola R, Cavalieri D, Velasco R, & Cestaro A (2013). An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics, 14(1), 1–12. [PubMed: 23323973]

Finn RD, Clements J, & Eddy SR (2011). HMMER web server: Interactive sequence similarity searching. Nucleic Acids Research, 39, W29–W37. 10.1093/nar/gkr367 [PubMed: 21593126]

Freudenthal JA, Pfaff S, Terhoeven N, Korte A, Ankenbrand MJ, & Förster F (2020). A systematic comparison of chloroplast genome assembly tools. Genome Biology, 21(1), 1–21. 10.1186/s13059-020-02153-6

Goulding SE, Wolfe K, Olmstead R, & Morden C (1996). Ebb and flow of the chloroplast inverted repeat. Molecular and General Genetics, 252(1), 195–206. [PubMed: 8804393]

Guisinger MM, Chumley TW, Kuehl JV, Boore JL, & Jansen RK (2010). Implications of the plastid genome sequence of *Typha* (Typhaceae, Poales) for understanding genome evolution in Poaceae. Journal of Molecular Evolution, 70(2), 149–166. [PubMed: 20091301]

Haberle RC, Fourcade HM, Boore JL, & Jansen RK (2008). Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. Journal of Molecular Evolution, 66(4), 350–361. [PubMed: 18330485]

Harris ME, Meyer G, Vandergon T, & Vandergon VO (2013). Loss of the acetyl-CoA carboxylase (accD) gene in Poales. Plant Molecular Biology Reporter, 31(1), 21–31.

Hochbach A, Linder HP, & Röser M (2018). Nuclear genes, matK and the phylogeny of the Poales. Taxon, 67(3), 521–536.

Hofacker IL (2003). Vienna RNA secondary structure server. Nucleic acids research, 31(13), 3429–3431. [PubMed: 12824340]

Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, Sato S, Yamada T, Mori H, & Tajima N (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. Nature Communications, 5(1), 1–9.

Huang Y, Escalona M, Morrison G, Marimuthu MP, Nguyen O, Toffelmier E, Shaffer HB, & Litt A (2022). Reference genome assembly of the big berry Manzanita (*Arctostaphylos glauca*). Journal of Heredity, 113(2), 188–196. [PubMed: 35575079]

Hunter JD (2007). Matplotlib: A 2D graphics environment. Computing in Science and Engineering, 9(3), 90–95. 10.1109/mcse.2007.55

Ifuku K, Endo T, Shikanai T, & Aro E-M (2011). Structure of the chloroplast NADH dehydrogenase-like complex: Nomenclature for nuclear-encoded subunits. Plant and Cell Physiology, 52(9), 1560–1568. [PubMed: 21785130]

Jansen RK, & Ruhlman TA (2012). Plastid Genomes of Seed Plants. In Genomics of chloroplasts and mitochondria (pp. 103–126). Springer. 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. 10.1186/s12870-022-03529-5 [PubMed: 34979920]

Jin JJ, Yu WB, Yang JB, Song Y, Depamphilis CW, Yi TS, & Li DZ (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biology, 21(1), 1–31. 10.1186/s13059-020-02154-5

Jones E, Simpson DA, Hodkinson TR, Chase MW, & Parnell JA (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. 10.1016/j.tplants.2019.05.003 [PubMed: 31208890]

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. [PubMed: 22543367]

Kent WJ (2002). BLAT—The BLAST-like alignment tool. Genome Research, 12(4), 656. 10.1101/gr.229202 [PubMed: 11932250]

Kim KJ, & Lee HL (2005). Widespread occurrence of small inversions in the chloroplast genomes of land plants. Molecules & Cells, 19(1), 104–113. [PubMed: 15750347]

Kode V, Mudd EA, 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. [PubMed: 16212603]

Kolmogorov M, Yuan J, Lin Y, & Pevzner PA (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. [PubMed: 8665091]

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, & Phillippy AM (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research, 27(5), 722–736. [PubMed: 28298431]

Lee C, Choi I, Cardoso D, de Lima HC, de Queiroz LP, Wojciechowski MF, Jansen RK, & Ruhlman TA (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. [PubMed: 34021942]

Lee C, Ruhlman TA, & Jansen RK (2020). Unprecedented intrain-dividual structural heteroplasmy in *Eleocharis* (Cyperaceae, Poales) plastomes. Genome Biology and Evolution, 12(5), 641–655. 10.1093/gbe/evaa076 [PubMed: 32282915]

Lee HL, Jansen RK, Chumley TW, & Kim KJ (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. [PubMed: 17329229]

Lee SS, Jeong WJ, Bae JM, Bang JW, Liu JR, & Harn CH (2004). Characterization of the plastid-encoded carboxyltransferase subunit (accD) gene of potato. Molecules & Cells, 17(3), 422–429. [PubMed: 15232216]

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. [PubMed: 29731764]

Li Y, & Deng X (2021). The complete chloroplast genome of the marine microalgae *Chaetoceros muellerii* (Chaetoceroceae). Mitochondrial DNA Part B, 6(2), 373–375. [PubMed: 33659682]

Liao X, U M., Hu K., Wu FX., Gao X., & Wang J. (2021). A sensitive repeat identification framework based on short and long reads. Nucleic Acids Research, 49(17), e100. 10.1093/nar/gkab563 [PubMed: 34214175]

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. 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. [PubMed: 23609545]

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. [PubMed: 34027070]

Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S, Milbourne D, Barth S, & Palmer JD (2010). Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Research, 20(12), 1700–1710. [PubMed: 20978141]

Mak QC, Wick RR, Holt JM, & Wang JR (2023). Polishing de novo nanopore assemblies of bacteria and eukaryotes with FMLRC2. Molecular Biology and Evolution, 40(3), msad048. [PubMed: 36869750]

Maréchal A, & Brisson N (2010). Recombination and the maintenance of plant organelle genome stability. New Phytologist, 186(2), 299–317. [PubMed: 20180912]

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. [PubMed: 24690362]

Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, Callot C, Langlade N, & Muños S (2016). Extraction of high-molecular-weight genomic DMA for long-read sequencing of single molecules. Biotechniques, 61(4), 203–205. [PubMed: 27712583]

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, & Sutton G (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics, 24(24), 2818–2824. [PubMed: 18952627]

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. [PubMed: 24382143]

Morton BR, & Clegg MT (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. [PubMed: 8252646]

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. [PubMed: 22214541]

Nashima K, Terakami S, Nishitani C, Kunihisa M, Shoda M, Takeuchi M, Urasaki N, Tarara 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. [PubMed: 7640528]

Oldenburg DJ, & Bendich AJ (2015). DNA maintenance in plastids and mitochondria of plants. Frontiers in Plant Science, 6, 883. [PubMed: 26579143]

Park S, Ruhlman TA, Weng M-L, Hajrah NH, Sabir JS, & Jansen RK (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. [PubMed: 28854633]

Planta J, Liang Y-Y,Xin H,Chansler MT, Prather LA, Jiang N, Jiang J, & Childs KL (2022). Chromosome-scale genome assemblies and annotations for Poales species Carex cristatella, *Carex scoparia, Juncus effusus*, and *Juncus inflexus*. G3, 12(10), jkac211. [PubMed: 35976112]

Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, Swerdlow HP, & Oyola SO (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. [PubMed: 25592487]

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 EP (2003). DNA repeats lead to the accelerated loss of gene order in bacteria. Trends in Genetics, 19(11), 600–603. [PubMed: 14585609]

Rousseau-Gueutin M, Huang X, Higginson E, Ayliffe M, Day A, & Timmis JN (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. [PubMed: 23435694]

Ruhlman TA, Chang WJ, Chen JJ, Huang YT, Chan MT, Zhang J, Liao DC, Blazier JC, Jin X, & Shih MC (2015). NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. BMC Plant Biology, 15(1), 1–9. [PubMed: 25592487]

Ruhlman TA, & Jansen RK (2018). Aberration or analogy? The atypical plastomes of Geraniaceae. In Advances in botanical research (pp. 223–262). Elsevier.

Ruhlman TA, & Jansen RK (2021). The plastid genomes of flowering plants: Essential principles. In Maliga P (Ed.), Chloroplast Biotechnology (pp. 3–27). Humana, 10.1007/978-1-0716-1472-3_1

Ruhlman TA, Zhang J, Blazier JC, Sabir JS, & Jansen RK (2017). Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. American Journal of Botany, 104(4), 559–572. [PubMed: 28400415]

Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NA, 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. [PubMed: 24618204]

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. [PubMed: 32208422]

Schwarz EN, Ruhlman TA, Sabir JS, Hajrah NH, Alharbi NS, Al-Malki AL, Bailey CD, & Jansen RK (2015). Plastid genome sequences of legumes reveal parallel inversions and multiple losses of rps16 in papilionoids. Journal of Systematics and Evolution, 53(5), 458–468.

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, & Sedlazeck FJ (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nature biotechnology, 38(9), 1044–1053.

Shearman JR, Sonthirod C, Naktang C, Sangsrakru D, Yoocha T, Chatbanyong R, Vorakuldumrongchai S, Chusri O, Tangphatsornruang S, & Pootakham W (2020). Assembly of the durian chloroplast genome using long PacBio reads. Scientific Reports, 10(1), 1–8. [PubMed: 31913322]

Shen W, Le S, Li Y, & Hu F (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PloS One, 11(10), e0163962. [PubMed: 27706213]

Shikanai T (2014). Central role of cyclic electron transport around photosystem I in the regulation of photosynthesis. Current Opinion in Biotechnology, 26, 25–30. [PubMed: 24679254]

Shikanai T, Endo T, Hashimoto T, Yamada Y, Asada K, & Yokota A (1998). Directed disruption of the tobacco ndhB gene impairs cyclic electron flow around photosystem I. Proceedings of the National Academy of Sciences, 95(16), 9705–9709.

Sloan DB, Triant DA, Forrester NJ, Bergner LM, Wu M, & Taylor DR (2014). A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Molecular Phylogenetics and Evolution, 72, 82–89. [PubMed: 24373909]

Soorni A, Haak D, Zaitlin D, & Bombarely A (2017). Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. BMC Genomics, 18(1), 1–8. [PubMed: 28049423]

Stadermann KB, Weisshaar B, & Holtgräwe D (2015). SMRT sequencing only de novo assembly of the sugar beet (*Beta vulgaris*) chloroplast genome. Bmc Bioinformatics, 16(1), 1–10. [PubMed: 25591917]

Syme AE, McLay TGB, Udovicic F, Cantrill DJ, Murphy DJ, McLay TGB, Udovicic F, Cantrill DJ, & Murphy DJ (2021). Long-read assemblies reveal structural diversity in genomes of organelles – An example with *Acacia pycnantha*. Gigabyte, 2021, 1–23. 10.46471/gigabyte.36

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, & Greiner S (2017). GeSeq–versatile and accurate annotation of organelle genomes. Nucleic Acids Research, 45(W1), W6–W11. [PubMed: 28486635]

Twyford AD, & Ness RW (2017). Strategies for complete plastid genome sequencing. Molecular Ecology Resources, 17(5), 858–868. 10.1111/1755-0998.12626 [PubMed: 27790830]

Ueda M, Fujimoto M, Arimura S, Murata J, Tsutsumi N, & Kadowaki K (2007). Loss of the rpl32 gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in *Populus*. Gene, 402(1–2), 51–56. [PubMed: 17728076]

Ueda M, Nishikawa T, Fujimoto M, Takanashi H, Arimura S, Tsutsumi N, & Kadowaki K (2008). Substitution of the gene for chloroplast RPS16 was assisted by generation of a dual targeting signal. Molecular Biology and Evolution, 25(8), 1566–1575. [PubMed: 18453549]

Wang JR, Holt J, McMillan L, & Jones CD (2018). FMLRC: Hybrid long read error correction using an FM-index. BMC Bioinformatics, 19(1), 1–11. 10.1186/s12859-018-2051-3 [PubMed: 29291722]

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(1), 1–15. 10.1186/s12864-018-5348-8 [PubMed: 29291715]

Weng ML, Blazier JC, Govindu M, & Jansen RK (2014). Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Molecular Biology and Evolution, 31(3), 645–659. [PubMed: 24336877]

Wick RR, Judd LM, Gorrie CL, & Holt KE (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Computational Biology, 13(6), e1005595. [PubMed: 28594827]

Wick RR, Schultz MB, Zobel J, & Holt KE (2015). Bandage: Interactive visualization of de novo genome assemblies. Bioinformatics, 31(20), 3350–3352. [PubMed: 26099265]

Wu H, Yang J-B, Liu J-X, Li D-Z, & Ma P-F (2021). Organelle Phylogenomics and Extensive Conflicting Phylogenetic Signals in the Monocot Order Poales. Frontiers. Plant Science, 12, 3345.

Xia C, Wang M, Guan Y, Li Y, & Li J (2022). Comparative analysis of complete chloroplast genome of ethnodrug *Aconltum episcopate* and insight into its phylogenetic relationships. Scientific Reports, 12(1), 1–13. 10.1038/s41598-022-13,524-3 [PubMed: 34992227]

Xiang QY, Crawford DJ, Wolfe AD, Tang YC, & DePamphilis CW (1998). Origin and biogeography of *Aesculus* L. (Hippocastanaceae): A molecular phylogenetic perspective. Evolution, 52(4), 988–997. 10.1111/j.1558-5646.1998.tb01828.x [PubMed: 28565208]

Xu K, Lin C, Lee SY, Mao L, & Meng K (2022). Comparative analysis of complete *llex* (Aquifoliaceae) chloroplast genomes: Insights into evolutionary dynamics and phylogenetic relationships. BMC Genomics, 23(1), 1–14. 10.1186/s12864-022-08397-9 [PubMed: 34979896]

Yu J, Fu J, Fang Y, Xiang J, & Dong H (2022). Complete chloroplast genomes of *Rubus* species (Rosaceae) and comparative analysis within the genus. BMC Genomics, 23(1), 1–14. 10.1186/s12864-021-08225-6 [PubMed: 34979896]

Zerbino DR (2010). Using the Velvet de novo assembler for short-read sequencing technologies. Current Protocols in Bioinformatics, 31(1), 11–15.

Zhang Z, Xie P, Guo Y, Zhou W, Liu E, & Yu Y (2022). Easy353: A tool to get Angiosperms353 genes for phylogenomic research. Molecular Biology and Evolution, 39(12), msac261. [PubMed: 36458838]

Zhong X (2020). Assembly, annotation and analysis of chloroplast genomes. Doctoral Thesis, The University of Western Australia. 10.26182/5f333d9ac2bee

Zhu B, Gao Z, Luo X, Feng Q, Du X, Weng Q, & Cai M (2019). The complete chloroplast genome sequence of garden cress (*Lepidium sativum* L.) and its phylogenetic analysis in Brassicaceae family. Mitochondrial DNA Part B, 4(2), 3601–3602. [PubMed: 33366103]

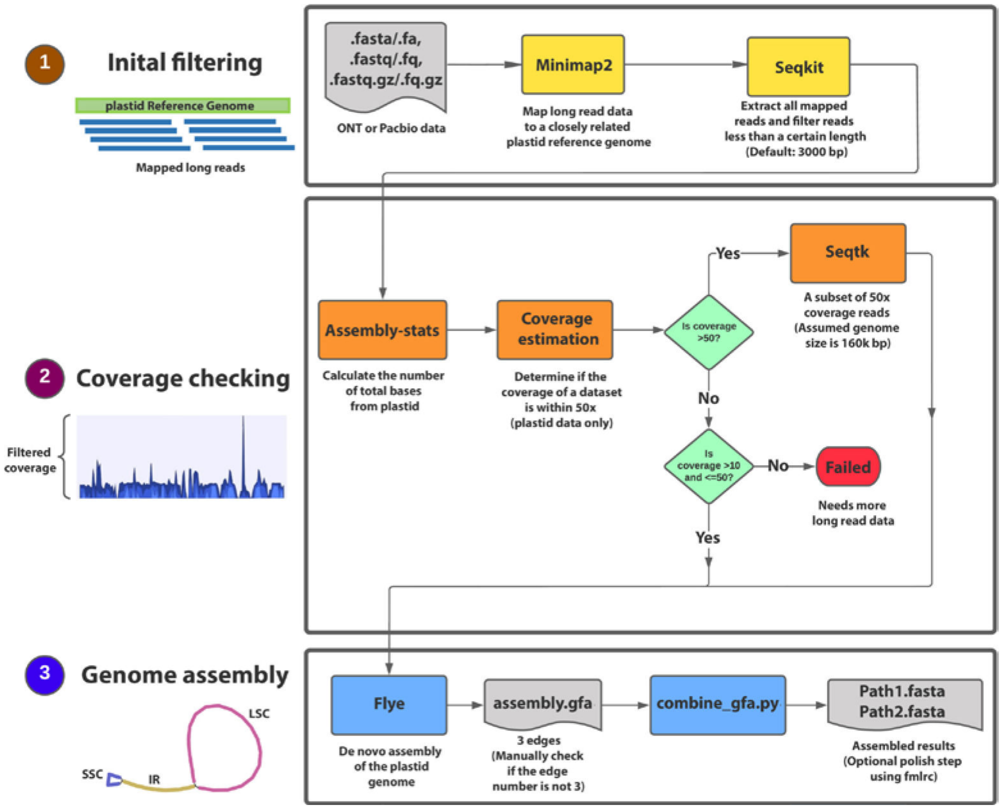

FIGURE 1.

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 50× (2). Finally, two paths of plastomes were obtained through FLYE and a customized Python script, combine_gfa.py (3).

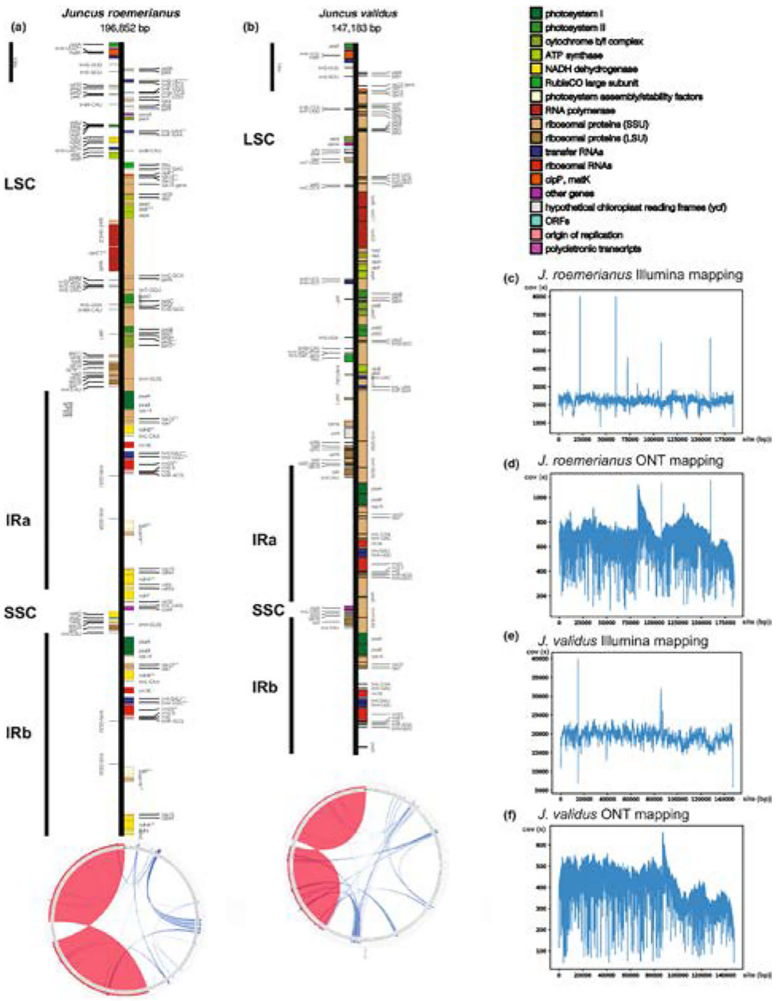

FIGURE 2.

Plastome structural maps and read coverage graphs of *Juncus roemerianus* and *J. validus*. (a,b) Linear maps of the *J. roemerianus* and *J. validus* plastome, respectively, were drawn by OGDRAW (Lohse et al., 2013). Genes that belong to different functional groups are colour-coded. Small single copy (SSC), large single copy (LSC) and inverted repeats (IRa, IRb) are indicated for both plastomes. Circular representations of the two *Juncus* plastomes were used to show 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 the *y*-axis represents the coverage.

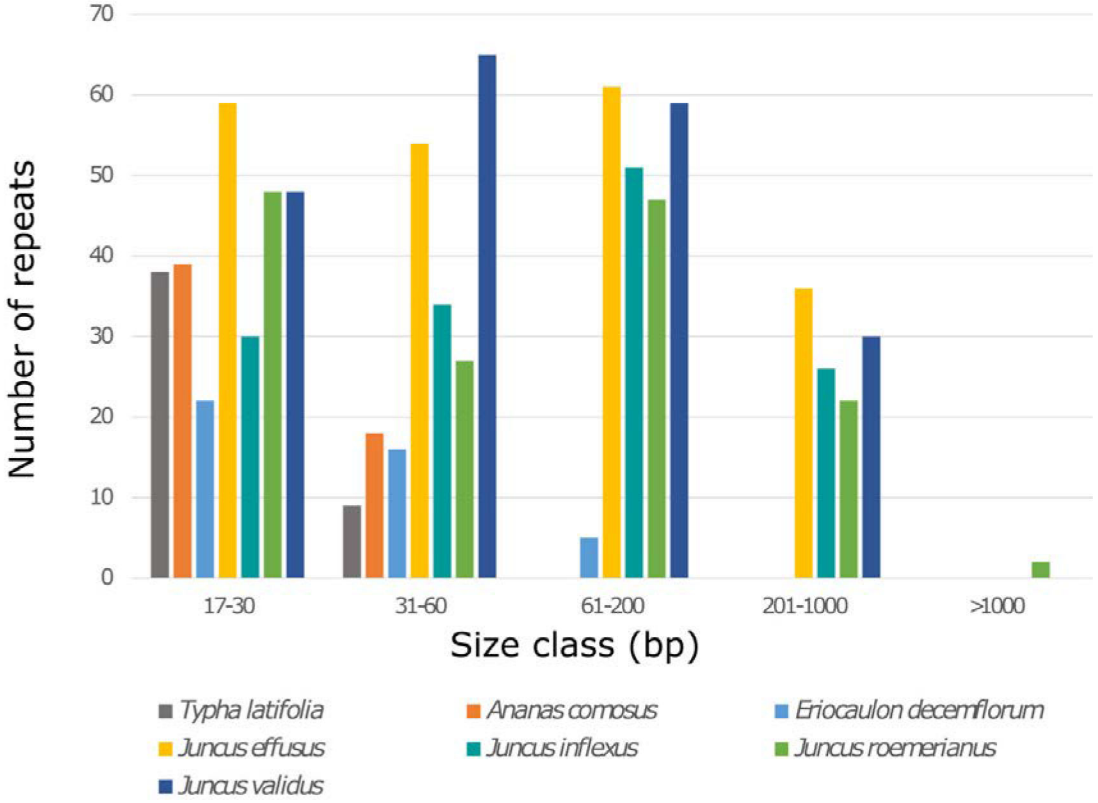

FIGURE 3.

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

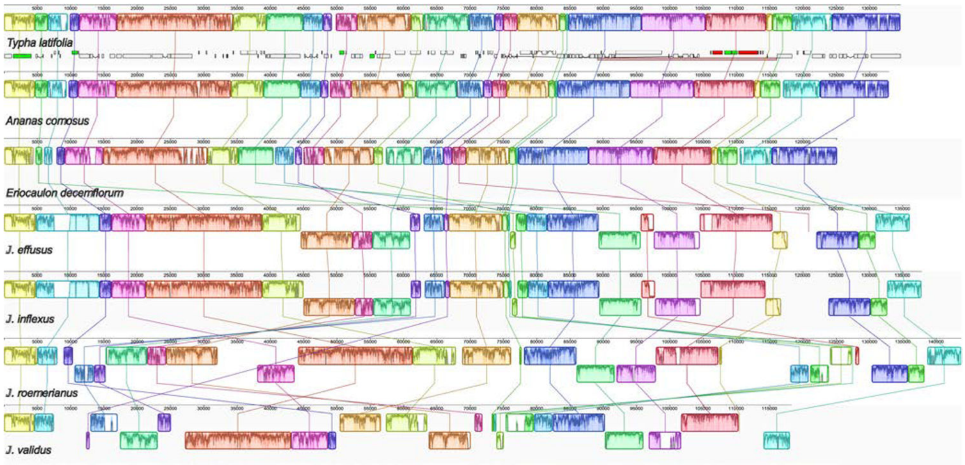

FIGURE 4.

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 the *Typha* plastome as the reference. The corresponding LCBs among seven plastomes are shaded and connected with a line of the same colour. LCBs that are flipped indicate inversions. Numbers on the upper *x*-axis are genome map coordinates in basepairs (bp).

TABLE 1

ptGAUL performance on 16 published sequence data sets, including the information of assembled plastome from published papers and the information on assembled plastomes from ptGAUL.

Species	Library preparation and sequencing methods	Raw read no./N50 (bp)	Reference	Plastid size from ptGAUL (bp) (% nucleotide sequence identity to references)	No. of assembled plastid contigs from ptGAUL	Plastome reference used for ptGAUL (reference length from original studies)[a]
Arctostaphylos glauca	WGS/PacBio	1,814,591/15,245	Huang et al. (2022)	150,241 (NA)	3 (S)	NC_035584.1/NC_042713.1/ NC_047438.1/**JAHSPW020000272.1 (118,663 bp)**
Lepidium sativum	WGS/PacBio	400,322/7277	Zhu et al. (2019)	153,666 (99.9%)	3 (S)	**NC_047178.1 (154,997 bp)**
Chaetoceros muellerii	WGS/PacBio	87,313/12,921	Li and Deng (2021)	117,304 (99.8%)	1 (S)	**MW004650.1 (116,284bp)**
Potentina micrantha	WGS/PacBio	28,638/2464	Ferrarini et al. (2013)	159,850 (99.8%)	3 (S)	**NC_015206.1 (155,691 bp)**
Durio zibethinus	WGS/PacBio	853,182/9670	Shearman et al. (2020)	142,806 (99.95%)	1 (S)	**MT321069 (163,974 bp)**
Beta vulgaris	WGS/PacBio	96,874/3980	Stadermann et al. (2015)	155,383 (99.9%)	3 (S)	**KR230391.1 (149,722 bp)**
Eleocharis dulcis	WGS/PacBio	68,167/16,288	Lee etal. (2020)	199,919 (99.5%)	3 (S)	**NC_047447.1 (199,561 bp)**
Eucalyptus pauciflora	WGS/ONT	705,554/24,988	Wang, Schalamun, et al. (2018)	158,561 (99.0%)	1 (S)	MZ670598.1/HM347959.1/ NC_014570.1/AY780259.1/ **NC_039597.1 (159,942bp)**
Leucanthemum vulgare	Long-range PCR/ONT	18,031/7900	Scheunert et al. (2020)	119,593 (NA)	5(F)	**NC_047460.1 (150,191 bp)**
Oryza glaberrima	Plastid capture/ONT	81,363/4319	Bethune et al. (2019)	124,133 (NA)	4(F)	**NC_024175.1 (132,629 bp)**
Cenchrus americanus	Plastid capture /ONT	105,760/5580	Bethune et al. (2019)	143,162 (96.6%)	3 (S)	**NC_024171.1 (140,718 bp)**
Digitana exilis	Plastid capture /ONT	141,250/4028	Bethune et al. (2019)	136,650(96.0%)	3 (S)	**NC_024176.1 (140,908 bp)**
Podococcus acaulis	Plastid capture /ONT	249,417/2621	Bethune et al. (2019)	81,976 (NA)	2(F)	**NC_027276.1 (157,688 bp)**
Raphia textilis	Plastid capture /ONT	83,833/2495	Bethune et al. (2019)	60,089 (NA)	2(F)	**NC_020365.1 (157,270 bp)**
Phytelephas aequatorialis	Plastid capture /ONT	202,925/2551	Bethune et al. (2019)	NA	(F)	**NC_029957.1 (159,075 bp)**
Picea glauca	WGS/PacBio	563,675/4671	Soorni et al. (2017)	123,476 (98.9%)	1 (S)	**NC_021456.1 (124,084 bp)**

Abbreviations: F, the samples failed using ptGAUL; NA, low nucleotide sequence identity between assembled plastome between published data and our data; S, the samples are well assembled by ptGAUL.

[a] This column includes the references we used for genome assembly in ptGAUL and the references in bold type were considered as references for comparisons with ptGAUL results.

TABLE 2

Summary of features of the plastid genomes of four *Juncus* species, including length, GC content and gene numbers.

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	12,443,053	96,653,565	83,412,073	158,922,322	156,712,430
No. of ONT reads and N50	0	2,960,380/21,529	2,735,792/24,397	427,549/15,998	243,884/14,365
Plastid genome size (bp)	170,612	178,158	181,566	196,852	147,183
LSC length (bp)	81,818	86,497	86,649	82,944	87,215
SSC length (bp)	7522	7539	7509	7902	2046
IR length (bp)	40,636	42,061	43,704	53,003	28,961
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

Note: The plastome data of *Juncus effusus* were from two different two different sources, this paper and Lu et al. (2021). PCG, protein-coding genes.

TABLE 3

Statistics of dispersed and tandem repeats in *Typha, Ananas, Eriocaulon* and *Juncus* plastomes.

	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	137,859	143,849	118,221
GC (%)	35.5	36.3	34.2	34.8	34.6	34.3	33.5
Dispersed repeat (DR)							
Length of DRs	1210	1495	1418	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
Percentage of DR in genome	0.9	1.1	1.1	11.1	9.6	10.2	12.4
Tandem repeat (TR)							
Length of TRs	3270	2057	859	12,248	15,783	22,978	8797
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
Percentage of TRs in genome	2.4	1.5	0.7	9.0	11.4	16.0	7.4
Total repeat							
Length of total repeats	4436	3552	2227	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
Percentage of total repeats in genome	3.3	2.7	1.8	17.2	19.2	24.3	19.1

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	-