Journal of Plant Biology
https://doi.org/10.1007/s12374-025-09475-7

RESEARCH ARTICLE

POLAP: A New Pipeline for Plant Mitochondrial Genome Assembly Using Oxford Nanopore Data Through Reference Generation

Sang Chul Choi[1] · **Sangtae Kim**[1]

Received: 5 March 2025 / Revised: 20 May 2025 / Accepted: 9 June 2025

Abstract
Plant mitochondrial genomes often contain polymorphic fragments, posing a significant challenge to complete genome assembly. Methods for assembling organelle genomes without external reference genomes use either PacBio HiFi data or error-corrected Oxford Nanopore Technologies (ONT) long-read data prior to assembly. In this study, we present a pipeline for assembling a draft mitochondrial genome using uncorrected ONT long-read data and for polishing the assembly using high-quality short-read data. This approach generates references from whole-genome assembly guided by organelle gene annotation. It facilitates users in identifying mitochondrial-derived seed contigs from a plant whole-genome assembly, extracting the raw reads that comprise them, and assembling mitochondrial genome sequences. To determine seed contigs of mitochondrial origin, a whole-genome assembly is evaluated for three criteria: 1) the mitochondrial and plastid gene densities, 2) the number of read coverage, and 3) the connectivity of contigs in the genome assembly graph. We evaluated the effectiveness of our approach by analyzing 11 publicly available plant genome sequencing datasets. We have implemented the approach as the Plant Organelle Long-read Assembly Pipeline (POLAP v0.3.7.3; https://github.com/goshng/polap).

Keywords PtGAUL · Flye · Minimap2 · Long-read sequencing

Introduction

Plant Nuclear, Plastid, and Mitochondrial Genomes

Plant mitochondrial genomes consist of a mixture of DNA molecules with heterogeneous structures, including recombinationally polymorphic circular and branched linear structures (Gaulberto et al. 2014). They are typically represented by a circular "master" chromosome, which tends to be challenging to determine. As reviewed by Wu et al. (2022) and Wang et al. (2024a), plant mitochondrial genomes can vary widely in size, e.g., from approximately 66 kb in *Viscum scurruloideum* Barlow (Liu et al. 2014) to approximately 12 Mb in *Larix sibirica* Ledeb. (Sloan et al. 2014). Very recently, a super-large mitochondrial genome of 18.99 Mb was also reported in Pinaceae (Huang et al. 2024). The complexity of mitochondrial genomes is further increased by alternative structural configurations, including linear and branched forms (Kozik et al. 2019) and its high substitution rate (Wang et al. 2024b). They could have long repeated subsequences spanning hundreds to thousands of base pairs, as in the examples of *Ginkgo biloba* L. (Guo et al. 2016) and *Hemerocallis citrina* Baroni (Zhang et al. 2022), and the long repeats would allow for genomic recombination and rearrangement resulting in multiple structural forms of mitochondrial DNA molecules (Mower et al. 2012; Kozik et al. 2019). Genome structural polymorphism can pose significant challenges in assembling large plant mitochondrial genomes. Because organelle genomes are much smaller than nuclear ones, assembling organelle genomes would have been far easier if DNA sequencing was performed selectively and separately for each of the three types of genomes in plants: nuclear, plastid, and mitochondrial DNA molecules. Unfortunately, one of the major challenges in plant mitochondrial genome assemblies arises from intracellular gene transfer events, which could leave remnants of the three types of genomes within other types in the same plant cells (Wang et al. 2024a). This makes it difficult to identify partial or complete mitochondrial-derived sequences from

✉ Sangtae Kim
amborella@sungshin.ac.kr

Sang Chul Choi
sangchulchoi@sungshin.ac.kr

[1] Department of Biotechnology, Sungshin Women's University, Seoul 01133, Republic of Korea

whole-genome assemblies involving heterogeneous genomic sources.

Plant Mitochondrial Genome Assembly

Organelle-genome assembly pipelines developed so far consider distinctive genomic features in organelles. Genome assemblers such as SPAdes (Bankevich et al. 2012), originally designed for bacterial genome assembly, are frequently employed for organelle-genome assembly because the pipelines for organelles genome assemblies have been developed with small genome sizes in mind. Novoplasty extends the seed sequence into adjacent regions, ultimately leading to the construction of an organelle genome, as organelle genomes are relatively small compared to nuclear genomes (Dierckxsens et al. 2017). GetOrganelle (Jin et al. 2020) uses the reference organelle genome of a species closely related to the target species and selects short reads sequenced only from the organelle genome. It selectively removes reads from a nuclear genome, subsequently using the remaining reads to construct an organelle genome using SPAdes (Bankevich et al. 2012). Unicycler (Wick et al. 2017) employs a short-read genome assembler for an initial genome assembly, which is refined later. Al-Nakeeb et al. (2017) extract mitochondrial DNA sequences from a whole-genome assembly by taking advantage of differences in sequencing depths between organelle and nuclear genomes. The ptGAUL pipeline (Zhou et al. 2023) employs Minimap2 (Li 2018), a long-read alignment tool, to map long-read sequencing data against a reference genome and select well-mapped reads for further plastid genome assembly using Flye (Kolmogorov et al. 2019), a long-read assembler. He et al. (2023) take a graph-based approach to mitochondrial genome assembly, using sequencing data to construct a'mitochondrial master graph'that facilitates obtaining dominant mitochondrial genomes. Xian et al. (2025) used deep learning to identify chloroplast-derived reads for chloroplast genome assembly and to filter out these reads for mitochondrial genome assembly. However, the study of Xian et al. (2025) focused on PacBio HiFi data, which typically produces reads of 10–20 kbp in length—shorter than those generated by Oxford Nanopore Technologies (ONT) long-read sequencing. Despite the advantages of ONT long-read sequencing, its potential has been largely overlooked due to its relatively lower read quality. On the other hand, Bi et al. (2024) introduced a pipeline, PMAT, which could be used for PacBio CLR/HiFi and ONT data in combination with the Newbler genome assembler (Margulies et al. 2005) as a plant mitochondrial genome assembler.

These pipelines have the potential to facilitate practical applications in plant organelle-genome assembly, and they suggest the possibility of a novel approach to plant organelle-genome assembly. Unicycler involves creating references to facilitate the process of genome assembly. Novoplasty's extending-by-seeding approach implies that seed sequences play a role in constructing such a small-sized mitochondrial genome sequence. Al-Nakeeb et al. (2017) suggest selecting mitochondrial-derived sequences from a genome assembly by leveraging the differences in sequencing depths between organelle and nuclear genomes. These developments indicate that an initial genome assembly derived from input sequencing data could potentially provide seed sequences that would then be the basis for assembling organelle genomes (e.g., Bi et al. 2024).

Seed Sequence Identification Based on Organelle-Gene Annotation

Using seed contigs to construct a target genome has been a common approach for the last decade (Hahn et al. 2013). The analysis of plant organelle-genome assemblies has extensively utilized the presence and absence of organelle genes in the selection of organelle-genome sequences from a whole-genome assembly of sequencing data obtained from plant genomic samples (e.g., Xuan et al. 2022). It seems reasonable to posit that seed contigs containing organelle genes can be selected to assemble mitochondrial genomes (Bi et al. 2024). If this is the case, such seed contigs can be a reference for selecting mitochondrial-derived reads to be assembled into a mitochondrial genome sequence. PMAT (Bi et al. 2024) uses a seed-based strategy for organelle-genome assembly. It can applied to long-read sequencing data from both HiFi PacBio and ONT. In the case of lower-quality ONT reads, error correction is performed using either Canu (Koren et al. 2017) or NextDenovo (Hu et al. 2024) prior to organelle-genome assembly with PMAT.

In this study, we present a pipeline that assembles a draft plant mitochondrial genome from low-quality ONT reads without error correction using the Flye assembler, followed by polishing with high-quality short-read data. The method involves the generation of reference sequences that can be used as a basis for the selection of mitochondrial-derived reads. Our approach uses organelle gene annotation and sequencing depth to identify mitochondrial sequence fragments within a whole-genome assembly graph of contig sequences. We then use long-read sequences mapped to the potentially mitochondrial-derived fragments to assemble the organelle genome of a plant species. The following section presents a step-by-step description of the method, illustrating its effectiveness in assembling organelle genomes using publicly available sequencing data. Afterward, we discuss the limitations of our approach, which would warrant further research.

Materials and Methods

Seven Steps of Data Processing in POLAP

Our approach, the Plant Organelle-Genome Long-Read Assembly Pipeline (POLAP), was initially inspired by the ptGAUL pipeline previously developed by Zhou et al. (2023). By following the approach used by Bi et al. (2024), we utilize plant organelle gene annotations to select seed contigs for the organelle-genome assembly. It enables the pipeline to assemble the mitochondrial genome of a target plant species without the necessity of reference genomes from the species that are evolutionarily closely related to the target species.

Our approach to mitochondrial genome assembly consists of seven steps (Fig. 1). The pipeline requires a long-read dataset from the target plant species to assemble a draft mitochondrial genome and a short-read dataset to estimate genome size and later polish the draft mitochondrial genome. The first two steps are long-read subsampling and whole-genome assembly. Next comes organelle gene annotation, followed by organelle DNA seed contig selection and organelle DNA long-read data collection. The final two steps are the assembly of the organelle genome and the polishing of the assembled sequence.

Step 1. Long-Read Subsampling

We filter out long reads shorter than 3 kbp and subsample the data with 50× genome coverage using Seqkit (Shen et al. 2016). Genome size was estimated from the short-read dataset using Jellyfish (Marçais and Kingsford 2011).

Step 2. Whole-Genome Assembly

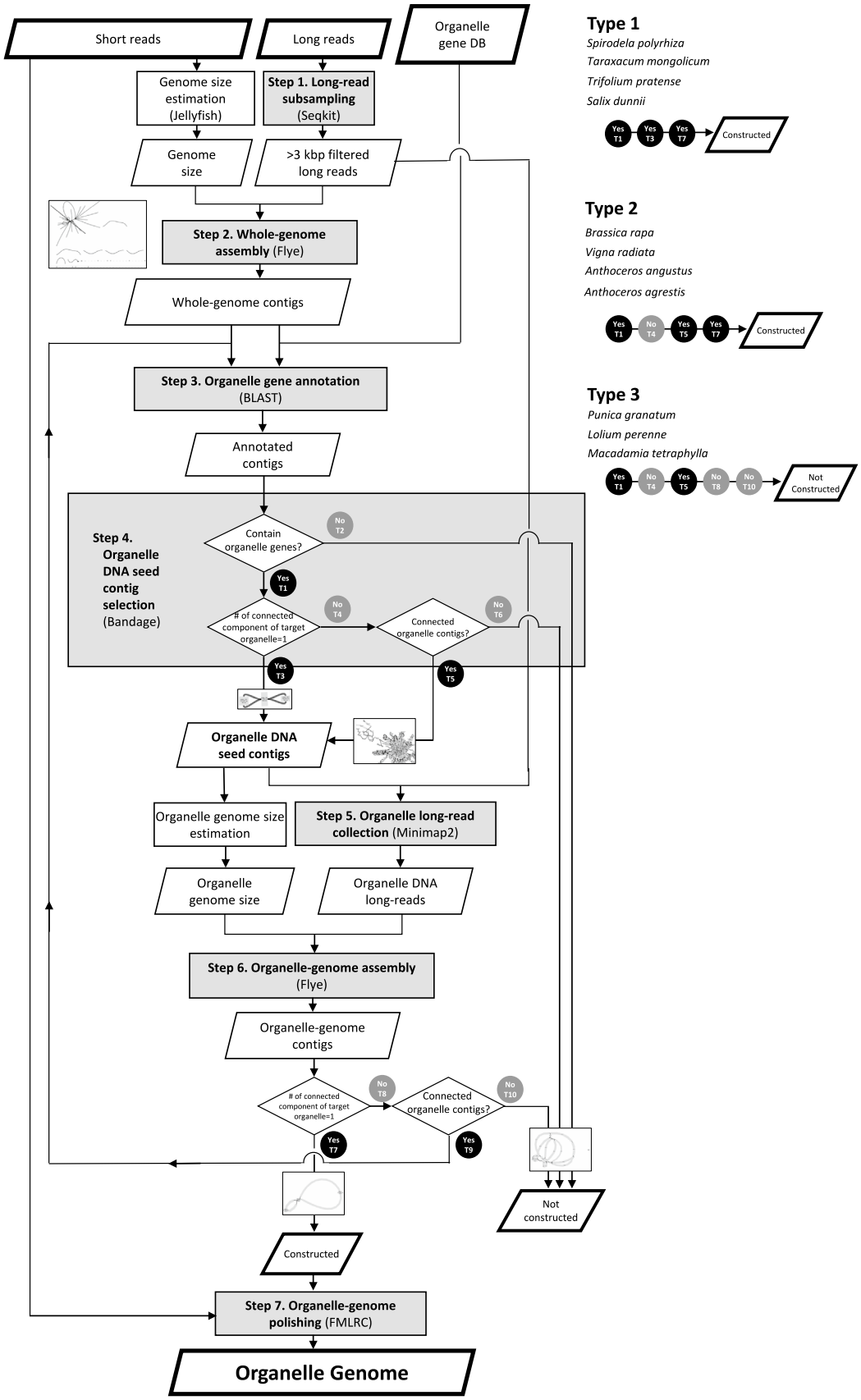
We perform a whole-genome assembly using Flye (Kolmogorov et al. 2019) on the filtered long-read data, specifying an assembly coverage of 30× and using the estimated genome size. We use the term 'Flye whole-genome assembly' to denote the Flye execution in this step, whereas 'Flye organelle-genome assembly' refers to the Flye execution following the organelle long-read collection step. Note that we execute the whole-genome assembly only up to Flye's "contigger" stage because the long-read polishing step may not be necessary for the whole-genome assembly. Once a tangible draft organelle genome is obtained, we proceed with the long-read Flye polishing in the organelle-genome assembly.

Step 3. Organelle Gene Annotation

BLAST (Basic Local Alignment Search Tool) is used to annotate contigs with organelle genes from the whole-genome assembly (Altschul et al. 1997). The contig sequences from a whole-genome assembly are compared with plant organelle genes to estimate the number of mitochondrial and plastid genes in the contigs; however, the organelle gene annotation is intended to guide the selection of mitochondrial-derived contigs rather than a determination of genes in the contigs. As we describe later, the plant organelle protein-coding DNA sequence data sets were prepared to alleviate the high redundancy in the protein-coding genes. We use each protein sequence from the preprocessed mitochondrial and plastid datasets as a query and contig sequences as a BLAST database for TBLASTN search (Altschul et al. 1997). The BEDtools merge function enhances the accuracy of genic region identification and prevents the multiple counting of a gene (Quinlan and Hall 2010). We append the mitochondrial and plastid gene counts to the contig table provided by a Flye whole-genome assembly, referred to as the contig annotation table of the whole-genome assembly. We examine the contig annotation table to choose contigs that may have mitochondrial origins, as elaborated in Step 4.

As mentioned in the previous paragraph, the pipeline needs to prepare organelle amino acid sequences, which we describe as follows. For a collection of plant mitochondrial genes, we retrieved a comprehensive collection of mitochondrial protein sequences from the RefSeq release 218, which was available at https://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/mitochondrion.1.protein.faa.gz (accessed May 5th, 2023). Due to the presence of highly redundant amino acid sequences in the collection, we would overestimate the number of mitochondrial genes on the contigs in the POLAP's annotation step. Three categories of preprocessing were carried out to prepare a sequence data set containing as non-redundant mitochondrial protein sequences as possible. First, we removed short sequences with lengths under 50 amino acids. Second, we purged the mitochondrial reference sequence data of any potentially irrelevant amino acid sequences functionally associated with plastids by eliminating sequences with names such as "orf," "plast," "chloro," "photo," and "hypothetical." Third, we applied the CD-HIT clustering algorithm to choose representatives from groups containing at least three members with a minimum sequence identity of 0.7 (Li and Godzik 2006). Preprocessing the comprehensive collection resulted in 6,743 mitochondrial amino acid sequences.

Similarly, a plastid reference dataset was accessed at https://ftp.ncbi.nlm.nih.gov/refseq/release/plastid/plastid.2.protein.faa.gz (accessed May 5th, 2023). We performed similar steps for plastid protein sequences except for eliminating irrelevant sequences based on their names. The plastid dataset, after preprocessing, comprised 3,764 amino acid sequences.


Short reads
Long reads
Organelle gene DB
Genome size estimation (Jellyfish)
Step 1. Long-read subsampling (Seqkit)
Genome size
>3 kbp filtered long reads
Step 2. Whole-genome assembly (Flye)
Whole-genome contigs
Step 3. Organelle gene annotation (BLAST)
Annotated contigs
Step 4. Organelle DNA seed contig selection (Bandage)
Contain organelle genes?
No T2
Yes T1
# of connected component of target organelle=1
No T4
Connected organelle contigs?
No T6
Yes T3
Yes T5
Organelle DNA seed contigs
Organelle genome size estimation
Step 5. Organelle long-read collection (Minimap2)
Organelle genome size
Organelle DNA long-reads
Step 6. Organelle-genome assembly (Flye)
Organelle-genome contigs
# of connected component of target organelle=1
No T8
Connected organelle contigs?
No T10
Yes T7
Yes T9
Not constructed
Constructed
Step 7. Organelle-genome polishing (FMLRC)
Organelle Genome
Type 1
Spirodela polyrhiza
Taraxacum mongolicum
Trifolium pratense
Salix dunnii
Yes T1
Yes T3
Yes T7
Constructed
Type 2
Brassica rapa
Vigna radiata
Anthoceros angustus
Anthoceros agrestis
Yes T1
No T4
Yes T5
Yes T7
Constructed
Type 3
Punica granatum
Lolium perenne
Macadamia tetraphylla
Yes T1
No T4
Yes T5
No T8
No T10
Not Constructed

◂**Fig. 1** Data processing pipeline of POLAP and three types of the user decision flow. The parallelograms are input or output data, and the gray rectangles are step-by-step processes. In parentheses, the program executes each process. Diamond boxes of decision steps indicate user decisions in the organelle DNA seed contig selection step (upper three) or the result of the organelle-genome assembly step (lower two). Black and gray circles indicate transitions (T) based on Yes and No decisions. A sequence of transitions in each organelle-genome assembly in various land plants is shown under one of three types of user decision flow

Step 4. Organelle DNA Seed Contig Selection

We identify potential mitochondrial contigs in a whole-genome assembly by considering the following three factors: 1) the organelle gene densities annotated on the contigs or the number of genes considering the contig lengths, 2) the number of read coverage in the contigs, and 3) the connectivity of contigs in the genome assembly graph. For "the number of read coverage," we use the following terms interchangeably: multiplicity values, copy number, and sequencing depth. The copy numbers used in our contig annotation tables are originally named "multiplicity values" in Flye's manual. We use the contig annotation table and genome assembly graph from a whole-genome assembly to determine seed contigs for later organelle-genome assembly (see evaluation method below). Using these three factors to select contigs does not require any order. We initiate the seed contig selection process by utilizing the Bandage (Wick et al. 2015) to visually display a genome assembly graph and locate contigs that may have originated from plant mitochondrial genomes.

A contig annotation table listing contigs with at least one organelle gene annotation may facilitate seed contig selection in a genome assembly graph. We categorize the contigs annotated with at least one organelle gene into two groups: a mitochondrial contig group, which consists of contigs containing more mitochondrial genes than plastid genes, and a plastid contig group, comprising the remaining contigs. We evaluate the copy numbers of contigs in a mitochondrial contig group and determine the copy number range of each type of nuclear, mitochondrial, and plastid genome. In general, one would expect mitochondrial-origin contigs to have much higher copy numbers than nuclear-origin contigs and lower than plastid-origin contigs. If the copy numbers of contigs in a mitochondrial contig group fall outside the predetermined range for the mitochondrial genome type, we eliminate those contigs from the mitochondrial contig group. If the copy numbers of contigs in the plastid contig group fall within the predetermined range for the mitochondrial genome type, we add the contigs to the mitochondrial contig group. This is because mitochondrial-derived contigs may have been incorrectly assigned to the plastid contig group. Due to the sparse gene distribution in a long mitochondrial genome, shorter contigs derived from mitochondria may lack annotated genes. To tackle the problem of no gene annotation on mitochondrial-derived short contigs, we use the genome assembly graph to recognize mitochondrial-derived contigs linked to those in the preselected mitochondrial contig group. To ensure accurate identification of contigs in a genome assembly graph, it is important to carefully compare the copy numbers of contigs in the mitochondrial contig group and those in the genome assembly graph. The copy numbers of contigs preselected based on organelle gene annotation should be comparable to those without annotation but linked to the preselected contigs.

Step 5. Organelle DNA Long-Read Collection

We use mitochondrial-derived seed contigs as a reference for selecting long reads before a Flye organelle-genome assembly by using Minimap2 (Li 2018) to map long-read sequences onto the preselected seed contigs. First, we borrow the read selection approach due to the ptGAUL. Second, we devise POLAP's read selection to increase the specificity of long-read selection because a final mitochondrial contig group often contains multiple contigs. POLAP's read selection consists of inter-contig and intra-contig mappings. In inter-contig mapping, we select long reads that align to two contigs and bridge the gap between them. We intend to select reads that are completely covered by a single contig in the intra-contig mapping. To concisely describe inter- and intra-contig mappings, we define variables for some columns of our pipeline's Minimap2 pairwise mapping file format (PAF). The variables are denoted by V followed by a column number in the PAF table: $V2$, the length of a read sequence; $V3$, the start position of the mapping in the read; $V4$, the end position of the mapping in the read; $V7$, seed contig sequence length; $V10$, number of matching bases in the mapping; $V11$, number of all bases in the mapping.

The inter-contig mapping rule requires that a read must meet two conditions to be considered mapped on one of two contigs: (1) $V10/V11 > 0.7$, and (2) $V11 > \omega$ (e.g., $\omega =$ 3,000 bp). Based on the Minimap2 manual, $V10/V11$ can be used as a proxy for the origin of a read from a contig since it is described as a BLAST-like alignment identity. The two conditions are due to Zhou et al. (2023). We assess whether each read meets the inter-contig mapping criteria for a pair of contigs in the final mitochondria contig group. For the intra-contig mapping, a read and its corresponding contig must meet three specific conditions: (1) $(V4 - V3)/V2 > 0.7$, (2) $V10/V11 > 0.7$, and (3) $V11 > \omega$. The intra-contig mapping ensures that a long-read sequence aligns entirely within a single contig, thereby gathering reads mapped within chosen contigs. Without the first condition, the other two conditions are due to Zhou et al. (2023).

In addition to the ptGAUL's read selection approach, we use the POLAP's read selection with the intra- and inter-contig mapping criteria to evaluate whether long-read sequences align within or between two contigs, thereby identifying long reads originating from seed contigs and potential gaps. In the remaining text, ω is referred to as the read-selection option. We subsample the reads so that the long-read data used for organelle genome assembly do not exceed 50 × coverage of the total size of the seed contigs.

Step 6. Organelle-Genome Assembly

We use the ptGAUL process in the last two steps, using the previously selected long-read sequences to perform a subsequent genome assembly with a fixed assembly coverage of 30 × and a genome size estimate equal to the total number of bases in the selected seed contigs from Step 4. Upon completing an organelle-genome assembly, we either proceed to Step 7 or return to Step 3. If we return to Step 3, we iterate through Steps 3, 4, and 5 using the organelle-genome assembly result instead of the whole-genome assembly result. As mentioned in Step 2, we perform Flye genome assemblies up to the "contigger" stage.

Step 7. Organelle-Genome Polishing

We employ the methodology outlined by Zhou et al. (2023) to refine a draft mitochondrial genome using the sequence polishing tool FMLRC (Wang et al 2018; Mak et al. 2023), using the high-quality short-read data we already utilized in our genome size estimation in Step 1. To ensure more accurate mitochondrial DNA sequencing, we complete the Flye organelle-genome assembly process to its final polishing stage before utilizing the Bandage software (Wick et al. 2015) to export DNA sequences from a genome assembly graph. One could examine whether the long-read dataset's mapping onto the mitochondrial genome sequence showed a consistent sequencing depth using SAMtools (Li et al. 2009). By employing the seven-step mitochondrial DNA assembly pipeline, we assembled mitochondrial genomes for 11 plant species, as described in the following subsection.

Evaluation of POLAP With Mitochondrial Genome Sequencing Data

We evaluated the performance of POLAP by assembling mitochondrial DNAs from 11 plant species using publicly available plant sequencing data (Table 1). These plant species are distributed across the various land plant orders. We compared the mitochondrial genome assembled in our study with a reported genome of the same species in the NCBI Genbank database. The user decision flow in the organelle DNA seed contigs and the process right after the organelle-genome assembly (see Fig. 1) allows the results of the 11 test data analyses to be categorized into three POLAP analysis types. We refer to transitions in the user decision flow as "T" followed by a number.

Type 1 POLAP Analysis

This type of POLAP analysis can be completed by sequentially following transitions T1, T3, and T7 in the user decision flow in POLAP (Fig. 1). A feedback transition (T8, T9, and back to T1) can be involved in any user decision flow. An organelle-genome assembly can also serve as a reference seed for another organelle-genome assembly, just as the whole-genome assembly. A whole-genome assembly generates a set of contiguous sequences, including a well-formed and easily identifiable single connected component for the target plant mitochondrial genome sequence. The well-formed graph serves as seed contigs for a subsequent organelle-genome assembly. Both whole-genome and organelle-genome assemblies yield similar mitochondrial genome graph structures in the genome assembly graphs. One could skip the organelle DNA seed contig selection, organelle long-read collection, and organelle-genome assembly steps of the POLAP pipeline and go directly from Steps 2 to 7. However, in this study, we go through all seven steps.

Type 2 POLAP Analysis

Type 2 POLAP data analysis follows a sequence of transitions T1, T4, and T5 before transitioning T7 in the user decision flow (see Fig. 1). A whole-genome assembly produces multiple connected components of a target organelle, but it is challenging to identify potential mitochondrial-derived contigs or their connections. If we cannot identify a seed contig set, we could perform an organelle-genome assembly using any contigs that have more mitochondrial than plastid genes annotated. After seed contig selection and organelle long-read collection, an organelle-genome assembly yields a well-structured mitochondrial genome. This type of data analysis best demonstrates the utility of POLAP.

Type 3 POLAP Analysis

Type 3 POLAP data analysis ends with one of the transitions T2, T6, or T10, regardless of the previous transitions taken in the user decision flow in POLAP (Fig. 1). An organelle-genome assembly fails to make a circular contig path for a mitochondrial genome, likely due to the multiple forms of the mitochondrial genome. We cannot perform a pairwise sequence alignment between our mitochondrial genome and the publicly available one because no "master" complete circular genome sequence is found.

Table 1 Assembly results of POLAP for 11 taxa and comparison of assembly performance between POLAP and PMAT

Higher category	Order	Family	Species	Reported mt genome		POLAP assembly							PMAT assembly[#]				
				GenBank accession number	L (bp)	Type	W	L (bp)	D	AC	M (GB)	T (hrs)	L (bp)	D	AC	M (GB)	T (hrs)
Bryophytes	Anthocerotales	Anthocerotaceae	*Anthoceros agrestis*	NC_049004	228,021	2	G3	228,019	2	0.99	37	< 1	NA	NA	NA	11	< 1
			Anthoceros angustus	NC_037476	242,410	2	G3	240,016	2,394	0.98	52	3	240,490	1,920	0.99	33	4
Monocot	Poales	Poaceae	*Lolium perenne*	JX999996	678,580	3	P11	NA	NA	NA	> 165*	81	NA	NA	NA	114	11
	Alismatales	Araceae	*Spirodela polyrhiza*	NC_017840	228,493	1	G3	229,383	−890	0.99	64	1	228,762	−269	0.99	12	< 1
	Asparagales	Asparagaceae	*Trifolium pratense*	NC_048499	301,823	1	G3	301,810	13	0.94	88	11	300,775	1,048	0.99	126	5
Eudicot	Proteales	Proteaceae	*Macadamia tetraphylla*	MW566572	682,795	3	P3	NA	NA	NA	77	13	NA	NA	NA	41	4
	Malpighiales	Salicaceae	*Salix dunnii*	CP161459	711,424	1	P3	711,423	1	0.99	75	5	NA	NA	NA	80	< 1
	Fabales	Fabaceae	*Vigna radiata*	NC_015121	401,262	2	P3	401,259	3	0.99	77	3	NA	NA	NA	110	< 1
	Myrtales	Lythraceae	*Punica granatum*	NC_071229	404,807	3	P3	NA	NA	NA	103	6	NA	NA	NA	103	6
	Brassicales	Brassicaceae	*Brassica rapa*	NC_049892	219,736	2	G3	219,756	−20	0.99	70	12	219,568	168	0.99	190	3
	Asterales	Asteraceae	*Taraxacum mongolicum*	NC_067879	304,467	1	G3	304,463	4	0.99	56	3	303,635	832	0.99	75	< 1

Type: Type 1 assembles sequences in a circular shape, almost identical to the whole-genome and organelle-genome assemblies run by POLAP; Type 2 assembles a circular mitochondrial genome sequence, but the visual structure of the two assemblies is different; Type 3 does not produce a circular shape of the final mitochondrial genome sequence

W: Read selection method and omega value (kbp); G for ptGAUL and P for POLAP read selection method, respectively; e.g., G3 represents the ptGAUL read selection method with 3 kbp

L: Length of mitochondrial genome sequence. NA indicates that the method used was unable to assemble the circular mitochondrial genome

D: Length difference between the previously reported mt genome and an assembled genome

AC: Sequence alignment coverage using progressiveMauve from pairwise alignment of the POLAP assembled sequence and the GenBank mitochondrial genome sequence. N/A means no such pairwise alignment

M: Peak memory used in the assembly process

T: Running time for mitochondrial genome assembly, not including short-read polishing, on a computer with dual Intel(R) Xeon(R) CPU E5-2690 v4 2.60 GHz with 128 GB of RAM

*Out of memory on the computer used in this study

#Long reads were error-corrected using the NextDenovo assembler (Hu et al. 2024) incorporated in the PMAT. PMAT assemblies were adjusted using the subsampling option -fc, with a value of 0.1 for most datasets, except for *Taraxacum mongolicum* (0.2), *Salix dunnii* (0.02), and *Anthoceros angustus* (1.0)

Results

The POLAP data analyses with the 11 datasets were categorized into three types based on how the plant mitochondrial genomes were assembled (Table 1; see Materials and Methods). We had four cases for either Type 1 or 2 and three for Type 3.

Type 1 POLAP Analysis

Spirodela polyrhiza

Six contigs in the annotation table had more mitochondrial genes than plastid genes annotated; four had a copy number of 0 or 1, while the remaining two had multiple copies (Table S1). A shorter inverted-repeat contig with no organelle gene annotations linked the remaining two contigs and had approximately twice the sequencing depth of either contig (Fig. S1). We selected these two and the inverted-repeat contig as a set of mitochondrial seed contigs for an organelle-genome assembly. These three contigs formed a single connected component in the whole-genome assembly graph. The whole-genome assembly graph played a role in identifying the three seed contigs, while an organelle gene annotation helped pinpoint two of them among numerous contigs in the whole-genome assembly (Fig. S1). The organelle-genome assembly graph was the same as the mitochondrial part of the whole-genome assembly graph, which led to the POLAP data analysis of Type 1 in the user decision flow in POLAP.

Taraxacum Mongolicum

Most contigs in the annotation table had a single copy, whereas two had multiple copies and many annotated mitochondrial genes (Table S2). One inverted repeat contig with no organelle gene annotations bridged the two contigs, forming a figure of eight (Fig. S2), as in the case of *Spirodela polyrhiza*. We selected these three as a set of mitochondrial seed contigs for an organelle-genome assembly. The seed contigs from the whole-genome and organelle-genome assemblies shared the same structure. Notably, two contigs in the contig annotation table had far more copies than the others, suggesting their plastid genomic origin (Table S2). Apparently, mitochondria and plastids share similar genes, which might make it difficult to distinguish between them.

Trifolium Pratens

Three seed contigs were identified as potentially mitochondrial in origin based on the higher number of annotated mitochondrial genes and higher copy numbers, indicating a mitochondrial origin (Table S3). Most other contigs in the mitochondrial contig annotation table had a single copy. Two (edge_3009 and edge_1395) had two copies, making them candidates for mitochondrial selection because their copy number was greater than one, which is typical for nuclear-origin contigs. Because the copy numbers for these contigs were significantly lower than those of the other three contigs above, they were more likely to have originated from the nuclear genome; therefore, we excluded them from our selection of seed contigs. A contiguous path along the three seed contigs in the whole-genome assembly graph resembled a figure of eight (Fig. S3). In contrast to *Spirodela polyrhiza* and *Taraxacum mongolicum*, all three seed contigs were present in the mitochondrial contig annotation table (Table S3). This case study was a Type 1 POLAP analysis because the mitochondrial genomes derived from the whole-genome and first organelle-genome assemblies were structurally identical (Fig. S3).

Salix Dunnii

A single circular contig from a whole-genome assembly yielded a draft mitochondrial genome sequence with more annotated mitochondrial genes and multiple copies (Table S4). Three contigs diverged from the circular contig (Fig. S4) but had no annotated organelle genes. We followed decision transitions T1, T3, and T7 in the user decision flow in POLAP (Fig. 1).

Type 2 POLAP Analysis

Anthoceros Agrestis

Two contigs from a whole-genome assembly were present in the mitochondrial contig annotation table and had high copy numbers, indicating their mitochondrial origin (Table S5). As they formed a non-circular path between the two contigs, connected by three links (Fig. S5), they differed from the Type 1 cases. Selecting the two as seeds, a Flye organelle-genome assembly produced a circular contig containing more mitochondrial than plastid genes. We followed the transitions T1, T4, T5, and T7 in the decision flowchart, resulting in an organelle-genome assembly.

Anthoceros Angustus

Of the contigs with more mitochondrial genes than plastid genes, seven contigs from a whole-genome assembly had significantly more copies (Table S6). The genome assembly graph did not allow for easy selection of seed contigs (Fig. S6). Instead, we used all contigs from the mitochondrial contig annotation table except those with excessively

long contigs with few mitochondrial genes annotated, adding 14 additional contigs to the seeds. An additional contig was selected based on the genome assembly graph. An organelle-genome assembly using the seed contigs yielded a circular sequence (Fig. S6) characterized by a prevalence of mitochondrial genes compared to plastid genes. We followed the transitions T1, T4, T5, and T7 in the user decision flow in POLAP (Fig. 1)

Brassica Rapa

Twenty contigs from a whole-genome assembly had more mitochondrial genes annotated than plastid genes (Table 2). Relying on the genome assembly graph guided by organelle gene annotation and copy number, we identified 10 of the 20 contigs used as seed contigs to assemble an organelle genome, resulting in a circular assembly graph (Fig. 2C). The circular DNA sequence from this organelle-genome assembly had more mitochondrial genes than plastid genes. In the user decision flow in POLAP, we followed transitions T1, T4, T5, and T7 in this case study (Figs. 1, 2).

Vigna Radiata

Four contigs annotated with more mitochondrial than plastid genes had more copies than three rather long contigs in the mitochondrial contig annotation table (Table S7). Two additional contigs were directly or indirectly related to the four based on the genome assembly graph (Fig. S7). Because these six contigs had similar copy numbers, they served as seed contigs for an organelle-genome assembly that resulted in a circular contig (Fig. S7). Our decision path passed through transitions T1, T4, T5, and T7 in the user decision flow in POLAP (Fig. 1).

Type 3 POLAP Analysis

Macadamia Tetraphylla

Three contigs were selected as seeds based on organelle gene annotations (Table S8) and a whole-genome assembly graph (Fig. S8). The structure of the organelle-genome graph was challenging to interpret as a contiguous circular mitochondrial genome sequence; however, we found a shorter contiguous circular sequence of approximately 480

Table 2 Annotation table of the mitochondrial contigs for Brassica rapa

Contig	Length (bp)	Depth	Copy	MT	PT	Seed
edge_3234	96,904	525	13	21	3	A
edge_3235	50,012	480	12	12	2	A
edge_661	15,811	1	0	5	3	X
edge_1	4,274,825	26	1	4	0	X
edge_3290	14,878	527	13	3	0	A
edge_3372	13,007	447	11	3	0	A
edge_60	740	14,605,556	365,139	2	1	X
edge_1690	72,901	25	1	2	0	X
edge_1709	223,103	19	0	2	0	X
edge_2729	4,258	432	11	2	0	A
edge_2730	11,878	0	0	2	0	X
edge_2731	9,440	433	11	2	0	A
edge_3167	6,609	473	12	2	0	A
edge_10	1,986,105	27	1	1	0	X
edge_485	5,800,128	25	1	1	0	X
edge_551	19,744	1	0	1	0	X
edge_2294	1,494	434	11	1	0	A
edge_2295	3,784	31	1	1	0	X
edge_2760	2,414	990	25	1	0	A
edge_2761	1,042	525	13	1	0	A

Depth: Read depths of a contig

Copy: Copy number of the contig

MT: Number of mitochondrial genes

PT: Number of plastid genes

Seed: Denoted as A if the contig is a seed for an organelle-genome assembly and has more MT genes than PT ones annotated. Denoted as X if it is not selected as a seed

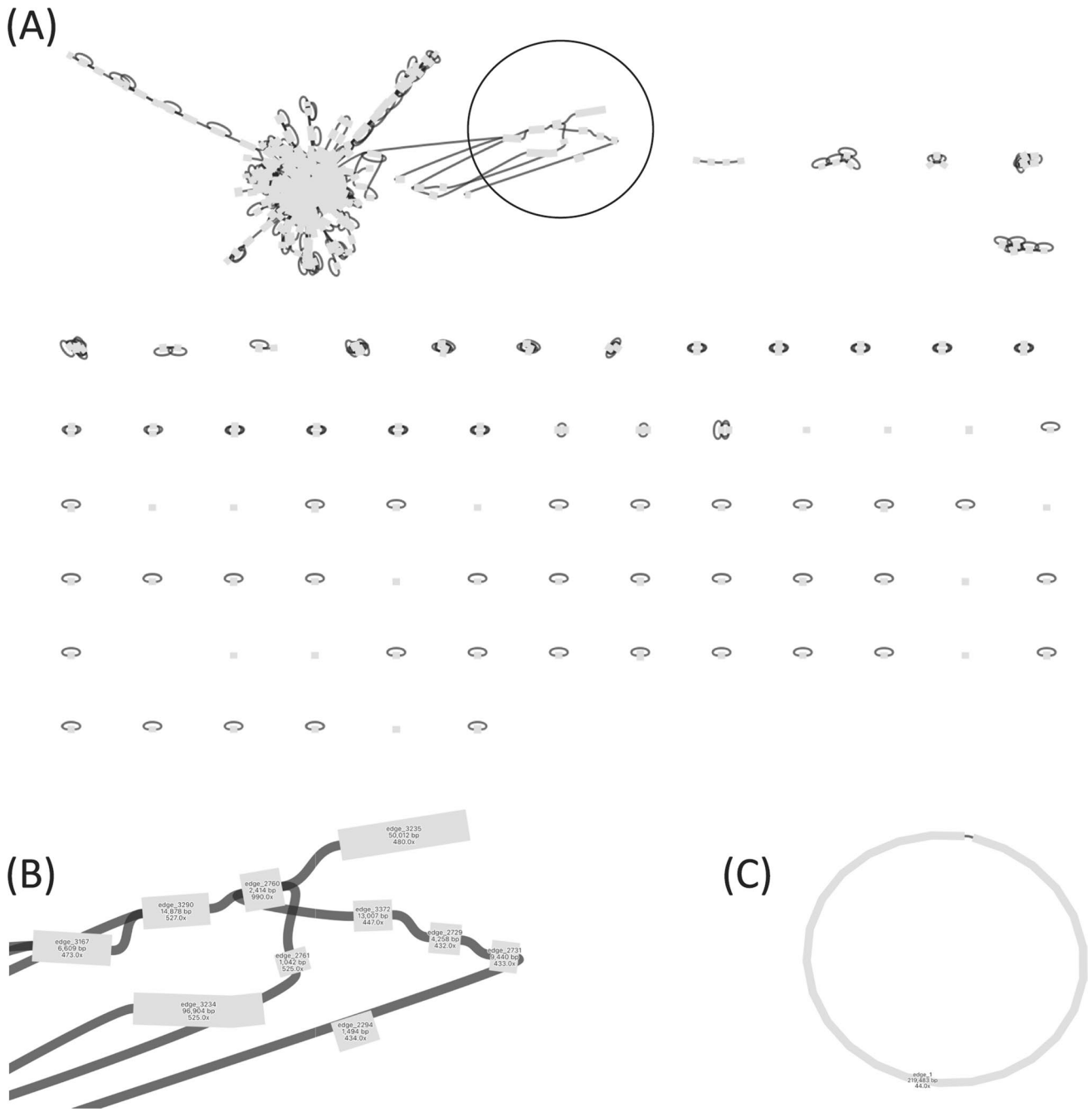


Fig. 2 Bandage graphs from each step of the mitochondrial genome assembly of *Brassica rapa* using POLAP. **A** Graphs of a long-read whole-genome assembly. The genome assembly graph is displayed by depth filtering of a range between 200× and 1,500× to facilitate visual representation. The circle indicates selected seed contigs (see Table 2). **B** The seed contigs for the organelle-genome assembly. **C** The graph of the mitochondrial genome assembly is based on the seed contigs using the read selection option of 3 kbp

kbp, which appears to be one of the potentially multiple forms of the mitochondrial genome. In this case study, we followed transitions T1, T4, T5, T8, and T10 in the user decision flow in POLAP (Fig. 1).

Punica Granatum

We selected 11 seed contigs, 10 of which had more mitochondrial genes than plastid genes (Table S9). One was selected based on the assembly graph (Fig. S9). Complicated

connections of an organelle-genome assembly made it difficult to untangle and generate a circular sequence. We followed transitions T1, T4, T5, T8, and T10 in the user decision flow in POLAP (Fig. 1).

Lolium Perenne

We identified 13 contigs with more mitochondrial genes than plastid genes and more copies than the rest of the mitochondrial contig group (Table S10). In addition, we selected 11 additional contigs that were directly or indirectly related to the 13 contigs based on the genome assembly graph (Fig. S10). Organelle-genome assembly was performed on the selected seed contigs with 11 kbp of the read selection option (Fig. S10). Due to the high interconnectedness of the assembly graph of an organelle-genome assembly, we could not construct a circular mitochondrial genome sequence.

Read-Selection Options

The read selection option from the ptGAUL or POLAP mapping method indicates the affinity of reads to seed contigs. Higher values increase specificity to seed contigs. Using several different read-selection option values (ω), we evaluated organelle genome assembly results (Figs. S11–S32). Metrics included the organelle genome assembly's fragment count, total bases, and read depth. The results suggested that the 3 kbp option value was generally effective for our datasets, with read-selection method preference varying by tested species.

Changing the reading selection resulted in very similar graphs and fragment counts in some cases (e.g., *Spirodela polyrhiza*, *Salix dunnii*, and *Anthoceros agrestis*) and very different graphs and fragment counts in others (e.g., *Taraxacum mongolicum* and *Lolium perenne*) (Figs. S11–S32). This suggests that multiple analyses with varying option values should be performed to select the optimal result when assembling a new genome.

Comparison with PMAT

To demonstrate the characteristics of POLAP, we compared its running performance with that of PMAT, a previously published tool, using the 11 datasets we included in this study (Table 1). Circular mitochondrial genome assemblies were obtained from eight taxa using POLAP and five taxa using PMAT (Table 1 and Fig. S33). PMAT and POLAP both failed to complete circular mitochondrial genome assemblies in *Lolium perenne*, *Macadamia tetraphylla*, and *Punica granatum*. However, PMAT also failed to assemble the mitochondrial genomes of *Anthoceros agrestis*, *Salix dunnii*, and *Vigna radiata*, which were shown as highly fragmented assembly graphs, preventing the extraction of complete DNA sequences. In cases where the circular genome was successfully derived from both tools, the length of the genome assembly was almost identical to that already reported in GenBank. The largest differences among the dataset were 0.79% for PMAT and 0.99% for POLAP in *Anthoceros angustus*. The proportion of aligned nucleotides in the assembly based on progressive Mauve alignment was high in all the cases (0.94 ~ 0.99) (Table 1). When comparing the computer resources used by the two tools, there was no significant difference (Table 1).

Discussion

POLAP implements a method for constructing plant mitochondrial genomes by selecting seed contigs from a whole-genome assembly using a long-read genome assembler, Flye (Kolmogorov et al. 2019), and then selecting reads from the long-read data using a long-read alignment tool, Minimap2 (Li 2018). The selection of organelle DNA seed contigs is based on three features of plant organelle contigs: the mitochondrial or plastid gene densities, the read coverage, and the connectivity of contigs within a genome assembly graph. The 11 case studies using publicly available sequencing data illustrate the utility of POLAP for constructing plant mitochondrial genomes. The POLAP approach of selecting mitochondrial DNA seed contigs and long reads was actually not necessary in the four case studies (i.e., Type 1 POLAP analysis). For three, it failed to generate a circular mitochondrial genome sequence (i.e., Type 3 POLAP analysis). In the other four case studies, we constructed circular mitochondrial genomes even when an initial whole-genome assembly did not appear to produce mitochondrial-derived contigs.

We consider the limitations of POLAP and further development for the seven steps in POLAP (Fig. 1). Steps 2, 6, and 7 will benefit from genome assemblers and polishing tools. In Step 1, one could use long-read sequencing data to estimate genome size (e.g., GCE, genomic character estimator, Liu et al. 2020) if one only wants to construct the organelle-genome structure without short-read polishing. Although Flye can perform genome assembly without specifying the genome size, we found that including the genome size made the assembly more efficient. Additionally, short-read data is required for the polishing step. In Step 3, we could exclude nuclear-derived contigs with single copies or contigs that are too long for mitochondrial genome sequences to improve the efficiency of BLAST search in POLAP. In Step 4, constructing a mitochondrial genome assembly depends on selecting "seed" contigs from a whole-genome assembly based on subjective selection criteria, making it difficult to determine the effect of different contig selection criteria on the resulting mitochondrial genome. Due to the lack of objective criteria for the selection of mitochondrial-derived contigs, it is

not easy to automate the process of mitochondrial genome assembly. It would become even more problematic with the increasing availability of high-throughput sequencing data for plant species. To facilitate the identification of mitochondrial-derived contigs, one could overlay the organelle gene annotation and copy numbers of a genome assembly onto the genome assembly graph, allowing computational traversal algorithms to identify contigs of mitochondrial origin. Further investigation is needed to develop a better methodological approach for selecting mitochondrial-derived contigs.

Although POLAP was originally developed for mitochondrial genome assembly, it can also be adapted for plastid genome assembly. This can be achieved by modifying its key steps (i.e., Steps 3, 4, and 5) to reflect the characteristics of plastid genomes better. For instance, contigs can be selected based on a higher number of plastid gene annotations relative to mitochondrial genes and substantially higher copy numbers than those typically observed in mitochondrial-derived contigs. Since plastid genomes are generally sequenced at a higher depth than mitochondrial genomes, the long-read dataset used for plastid genome assembly can be substantially downsized. As a result, while the plastid genome may be efficiently assembled, nuclear and mitochondrial genomes are likely to remain fragmented due to the reduced data size.

It is useful to provide the appropriate amount of data required for the assembly of an unknown mitochondrial genome. However, it is challenging to provide an appropriate amount because the number of organelle genome copies can vary depending on the taxon and the physiological condition of the plant. Nevertheless, it is reasonable to assume that there are at least ten times more copies of mitochondria and one hundred times more copies of chloroplasts than of the nuclear genome (Cai et al. 2015; Cole 2016; Morley and Nielsen 2016; Sakamoto and Takami 2018; Shen et al. 2019). Based on this assumption, we evaluated the possibility of reducing large data inefficiency while maintaining sufficient coverage using fixed data volumes from the various taxonomic groups employed in this study. Table S11 shows the potential depth of the mitochondrial genome under the assumption of an initial data volume of 10 GB, with ten copies of the mitochondrial genome per cell and a whole-genome size ratio for the taxa used in this study (average 0.116%). Under these conditions, the average depth of coverage of the mitochondrial genome was 369x, ranging from 43 × to 855x. Therefore, we recommend using 10 Gbp of raw sequencing data for an optimal balance between sequencing costs and assembly reliability for mitochondrial genome assembly in most cases.

Our pipeline is designed to prioritize identifying a primary circular mitochondrial structure. However, plant mitochondrial genomes are often structurally complex and polymorphic, potentially including circular and branched linear forms (Gaulberto et al. 2014). Detecting such structural diversity is limited by the three criteria used in our pipeline for seed contig selection during mitochondrial genome assembly. Nevertheless, final assemblies can be evaluated further by annotating organelle genes to determine the presence of sequences likely derived from mitochondrial DNA. Future research aims to develop methods for assessing the likelihood that assembled sequences represent authentic plant mitochondrial genomes.

The three types of POLAP data analysis produce a mitochondrial genome assembly, whether it is circular or not. We consider several possible ways in which circular forms of mitochondrial genomes are not constructed, as follows. We may fail at either a whole or organelle-genome assembly stage. This may be due to insufficient organelle genome sequencing coverage or insufficient sequencing data. A whole-genome assembly may produce contigs with no organelle gene annotations at all. Even after organelle gene annotation, seed contig selection from a whole-genome assembly may fail. We may have difficulty extracting a draft mitochondrial genome from a resulting organelle-genome assembly graph.

Conclusion

POLAP, an open-source tool licensed under the GNU General Public License version 3.0, is developed to facilitate the plant mitochondrial genome assembly pipeline using long-read sequencing data in the absence of reference genomes. It demonstrates its use with 11 publicly available plant sequencing datasets and is freely available at https://github.com/goshng/polap.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12374-025-09475-7.

Acknowledgements We thank Chanhong Kim for his encouragement and support in this research and Jieun Lee for her insightful discussions on mitochondrial DNA analysis. This study was funded by grants from the National Research Foundation of Korea (2021R1F1A1059131).

Authors Contribution S.C.C. and S.K. conceptualized and designed the study, S.C.C. analyzed the data and drafted the initial manuscript, and both authors revised and edited the text. Both authors approved the final version of the manuscript.

Data Availability The POLAP source code, as well as the example dataset used to demonstrate its usefulness, is available at https://github.com/goshng/polap.

References

Al-Nakeeb K, Petersen TN, Sicheritz-Pontén T (2017) Norgal: extraction and de novo assembly of mitochondrial DNA from

whole-genome sequencing data. BMC Bioinf 18:510. https://doi.org/10.1186/s12859-017-1927-y

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021

Bi C, Shen F, Han F, Qu Y, Hou J, Xu K, Xu L-a, He W, Wu Z, Yin T (2024) PMAT: An efficient plant mitogenome assembly toolkit using low-coverage HiFi sequencing data. Hortic Res 11:023. https://doi.org/10.1093/hr/uhae023

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

Cole LW (2016) The evolution of per-cell organelle number. Front Cell Dev Biol 4:85. https://doi.org/10.3389/fcell.2016.00085

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

Gualberto JM, Mileshina D, Wallet C, Niazi AK, Weber-Lotfi F, Dietrich A (2014) The plant mitochondrial genome: dynamics and maintenance. Biochimie 100:107–120. https://doi.org/10.1016/j.biochi.2013.09.016

Guo W, Grewe F, Fan W, Young GJ, Knoop V, Palmer JD, Mower JP (2016) Ginkgo and welwitschia mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution. Mol Biol Evol 33:1448–1460. https://doi.org/10.1093/MOLBEV/MSW024

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res 41:e129. https://doi.org/10.1093/nar/gkt371

He W, Xiang K, Chen C, Wang J, Wu Z (2023) Master graph: an essential integrated assembly model for the plant mitogenome based on a graph-based framework. Briefings Bioinf 24:522. https://doi.org/10.1093/bib/bbac522

Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K, Ruan J, Xiao C-L, Wang D, Wu D-D, Wang S (2024) NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. Genome Biol 25:107. https://doi.org/10.1186/s13059-024-03252-4

Huang K, Xu W, Hu H, Jiang X, Sun L, Zhao W, Long B, Fan S, Zhou Z, Mo P, Jiang X, Tian J, Deng A, Xie P, Wang Y (2024) The mitochondrial genome of *Cathaya argyrophylla* Reaches 18.99 Mb: analysis of super-large mitochondrial genomes in Pinaceae. arXiv. https://doi.org/10.48550/arXiv.2410.07006

Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z (2020) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol 21:241. https://doi.org/10.1186/S13059-020-02154-5

Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546. https://doi.org/10.1038/s41587-019-0072-8

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 Res 27:722–736. https://doi.org/10.1101/gr.215087.116

Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC (2019) The alternative reality of plant mitochondrial DNA: One ring does not rule them all. PLoS Genet 15:e1008373. https://doi.org/10.1371/JOURNAL.PGEN.1008373

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

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

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W (2020) Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv. https://doi.org/10.48550/arXiv.1308.2012

Liu Y, Medina R, Goffinet B (2014) 350 My of mitochondrial genome stasis in mosses, an early land plant lineage. Mol Biol Evol 31:2586–2591. https://doi.org/10.1093/molbev/msu199

Mak QXC, Wick RR, Holt JM, Wang JR (2023) Polishing *de novo* nanopore assemblies of bacteria and eukaryotes with FMLRC2. Mol Biol Evol 40:048. https://doi.org/10.1093/molbev/msad048

Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770. https://doi.org/10.1093/bioinformatics/btr011

Margulies M et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380. https://doi.org/10.1038/nature03959

Morley SA, Nielsen BL (2016) Chloroplast DNA copy number changes during plant development in organelle DNA polymerase mutants. Front Plant Sci 7:57. https://doi.org/10.3389/fpls.2016.00057

Mower JP, Case AL, Floro ER, Willis JH (2012) Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (*Mimulus guttatus*) lineage with cryptic CMS. Genome Biol Evol 4:670–686. https://doi.org/10.1093/gbe/evs042

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033

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

Shen J, Zhang Y, Havey MJ, Shou W (2019) Copy numbers of mitochondrial genes change during melon leaf development and are lower than the numbers of mitochondria. Hortic Res 6:1–9. https://doi.org/10.1038/s41438-019-0177-8

Shen W, Le S, Li Y, Hu F (2016) SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE 11:e0163962. https://doi.org/10.1371/journal.pone.0163962

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10:e1001241. https://doi.org/10.1371/journal.pbio.1001241

Wang J, Kan S, Liao X, Zhou J, Tembrock LR, Daniell H, Jin S, Wu Z (2024a) Plant organellar genomes: much done, much more to do. Trends Plant Sci 29:754–769. https://doi.org/10.1016/j.tplants.2023.12.014

Wang J, Zou Y, Mower P, Reeve W, Wu Z (2024b) Rethinking the mutation hypotheses of plant organellar DNA. Genomics Commun 1:e003

Wang JR, Holt J, McMillan L, Jones CD (2018) FMLRC: Hybrid long read error correction using an FM-index. BMC Bioinf 19:1–11. https://doi.org/10.1186/s12859-018-2051-3

Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wick RR, Schultz MB, Zobel J, Holt KE (2015) Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31:3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Wu Z-Q, Liao X-Z, Zhang X-N, Tembrock LR, Broz A (2022) Genomic architectural variation of plant mitochondria—a review of multichromosomal structuring. J Syst Evol 60:160–168. https://doi.org/10.1111/jse.12655

Xian W, Bezrukov I, Bao Z, Vorbrugg S, Gautam A, Weigel D (2025) TIPPo: a user-friendly tool for *de novo* assembly of organellar genomes with high-fidelity data. Mol Biol Evol 42:247. https://doi.org/10.1093/molbev/msae247

Xuan L, Qi G, Li X, Yan S, Cao Y, Huang C, He L, Zhang T, Shang H, Hu Y (2022) Comparison of mitochondrial genomes between a cytoplasmic male-sterile line and its restorer line for identifying candidate CMS Genes in *Gossypium hirsutum*. Int J Mol Sci 23:9198. https://doi.org/10.3390/ijms23169198

Zhang K, Wang Y, Zhang X, Han Z, Shan X (2022) Deciphering the mitochondrial genome of *Hemerocallis citrina* (Asphodelaceae) using a combined assembly and comparative genomic strategy. Front Plant Sci. https://doi.org/10.3389/FPLS.2022.1051221

Zhou W, Armijos CE, Lee C, Lu R, Wang J, Ruhlman TA, Jansen RK, Jones AM, Jones CD (2023) Plastid genome assembly using long-read data. Mol Ecol Resour 23:1442–1457. https://doi.org/10.1111/1755-0998.13787