TIPP_plastid: A User-Friendly Tool for De Novo Assembly of Organellar Genomes with HiFi Data

Wenfei Xian[1], Ilja Bezrukov[1], Zhigui Bao[1], Sebastian Vorbrugg[1], Anupam Gautam[2,3], Detlef Weigel[1,4]*

[1]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany

[2]Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany

[3]International Max Planck Research School "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany

[4]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany

*corresponding author: weigel@tue.mpg.de (D.W.)

Abstract

Plant cells have two major organelles with their own genomes: chloroplasts and mitochondria. While chloroplast genomes tend to be structurally conserved, the mitochondrial genomes of plants, which are much larger than those of animals, are characterized by complex structural variation. We introduce TIPP_plastid, a user-friendly, reference-free assembly tool that uses PacBio high-fidelity (HiFi) long-read data and that does not rely on genomes from related species or nuclear genome information for the assembly of organellar genomes. TIPP_plastid employs a deep learning model for initial read classification and leverages k-mer counting for further refinement, significantly reducing the impact of nuclear insertions of organellar DNA on the assembly process. We used TIPP_plastid to completely assemble a set of 54 complete chloroplast genomes. No other tool was able to completely assemble this set. TIPP_platiid outperforms PMAT in mitochondrial genome assembly, especially with respect to the completeness of protein coding genes. We also used the assembled organelle genomes to identify instances of nuclear plastid DNA (NUPTs) and nuclear mitochondrial DNA (NUMTs) insertions. The cumulative length of NUPTs/NUMTs positively correlates with the size of the nuclear genome, suggesting that insertions occur stochastically. NUPTs/NUMTs show predominantly C:G to T:A changes, with the mutated cytosines typically found in CG and CHG contexts, suggesting that degradation of NUPT and NUMT sequences is driven by the known elevated mutation rate of methylated cytosines. siRNA loci are enriched in NUPTs and NUMTs, consistent with the RdDM pathway mediating DNA methylation in these sequences.

Introduction

In the cells of green plants, DNA is found in three main locations: chloroplasts or chloroplast-related plastids, mitochondria, and the nucleus. The chloroplast is the primary site of photosynthesis, converting solar energy into chemical energy, while mitochondria are crucial for cellular energy metabolism. Chloroplasts and mitochondria are thought to have originated from ancient endosymbiosis events (Zimorski et al. 2014). Due to secondary and tertiary endosymbiosis, chloroplasts or plastids are present across various kingdoms, collectively referred to as photosynthetic eukaryotes (Yoon et al. 2004).

Chloroplast genomes are structurally conserved across species, and they typically comprise four distinct fragments: one large single copy (LSC), one small single copy (SSC), and two inverted repeats (IRs). In contrast, the genomes of mitochondria, present in all eukaryotic organisms except for the microorganism Monocercomonoides sp. (Karnkowska et al. 2016), vary significantly across kingdoms. The structure of animal mitochondrial genomes is very conserved, presenting as a single small circular DNA with sizes from 11 to 28 kb (Jin et al. 2020). The situation is very different in plants, which have structurally complex mitochondrial genomes with large variation in size, with the largest known mitochondrial genomes reaching up to 11 Mb (Sloan et al. 2012; Putintseva et al. 2020).

Compared to nuclear genomes, much less attention has been paid to the high-quality assembly of organellar genomes. Short read data are useful, with some caveats, for the assembly of the relatively small and conserved mitochondrial genomes of animals and

chloroplast genomes of plants (Dierckxsens et al. 2017; Jin et al. 2020), but their utility is limited for the larger and more complex mitochondrial genomes of plants (Štorchová and Krüger 2024). Long and highly accurate read data have substantially enhanced our ability to assemble nuclear genomes (Wenger et al. 2019; Sereika et al. 2022). With the help of long reads, even highly repetitive regions such as centromeres and telomeres can be assembled (Naish et al. 2021; Nurk et al. 2022; Wlodzimierz et al. 2023), although challenges persist with the assembly of rDNA clusters. Moreover, the typically very high coverage of organellar genomes in data sets of genomic DNA interferes with productive assembly using standard tools, which are optimized for the nuclear genomes (Cheng et al. 2021). In addition, chloroplast and mitochondrial DNA fragments are often transferred to the nucleus, which also interferes with assembly of the true organellar genomes (Uliano-Silva et al. 2023).

Several tools have been developed to enable the specific use of long-read data for organelle genome assembly, primarily focusing on chloroplast genomes, such as Organelle_PBA (Soorni et al. 2017) , ptGAUL (Zhou et al. 2023) and CLAW (Phillips et al. 2024). The general approach begins with extracting chloroplast reads from the data set by aligning long reads to the chloroplast genomes of closely related species. This is straightforward and effective for the chloroplast genome, as there are now over 12,000 published chloroplast genomes available, making it almost always possible to find a sufficiently closely related species for successful extraction of chloroplast reads. However, this approach has limitations for mitochondrial genomes, given the much smaller number of available plant mitochondrial genomes (approximately 500 as of July 2023) and the much lower conversation of mitochondrial genomes, even between closely related species. Recently, an alternative approach has been proposed - PMAT (Bi et al. 2024). It begins with downsampling the initial read data set to an estimated coverage of the organellar genomes that is suitable for standard assembly tools. Next, a normal assembly is performed, and then the contigs that appear to belong to organellar genomes are identified based on the presence of conserved protein coding genes. While useful, this approach may result in incomplete assemblies, especially for species with multichromosomal mitochondrial genomes where some chromosomes lack coding genes (Sanchez-Puerta et al. 2017). Clearly, the preferred approach would be a (largely) reference-free and tool for organelle genome assembly that has similar power for both chloroplast and mitochondrial genomes.

As stated above, organellar DNA can be transferred to the nucleus, and it is common to find organellar sequences in the nuclear genome (Richly and Leister 2004; Hazkani-Covo et al. 2010; Zhang et al. 2020). These sequences are known as nuclear mitochondrial DNA (NUMTs) and nuclear chloroplast DNA (NUPTs). The nuclear genome evolves much faster than mitochondrial genome, typically by an order of magnitude (Wolfe et al. 1987; Drouin et al. 2008). Accordingly, NUPTs and NUMTs tend to diverge from the ancestral organellar genomes quite rapidly. By aligning NUPTs and NUMTs, which should not carry any function, to the corresponding organelle genomes, one can explore presumably neutral processes of sequence change in the integrated organellar DNA (Huang et al. 2005; Rousseau-Gueutin et al. 2011; Yoshida et al. 2014; Fields et al. 2022). Questions of interest are whether NUPTs and NUMTs behave in a similar manner, and how their evolutionary fate compares to that of other large insertions, such as transposons (Wang et al. 2013; Maumus and Quesneville 2014).

We have developed TIPP_plastid, a user-friendly, reference-free assembly tool for plant organelle genomes that integrates TIARA, a deep-learning-based approach for organellar DNA classification (Karlicki et al. 2022), eliminating the need for knowledge of organellar genomes from closely related species genomes or nuclear genome information of the target species. We use k-mer information to optimize TIARA's output, distinguishing NUPTs, NUMTs, and misclassifications caused by repetitive sequences. Using TIPP_plastid, we not only successfully assembled 54 complete chloroplast genomes but also demonstrated superior performance in mitochondrial assembly compared to PMAT, revealing the complex structure of mitochondrial genomes. Additionally, we detailed the insertion patterns of NUPTs and NUMTs, and analyzed nucleotide substitutions in NUPTs and NUMTs.

Approach

We designed and implemented a reference-free, user-friendly tool for assembly of plant organelle genomes called TIPP_plastid from highly accurate PacBio HiFi long reads. It begins with a deep learning model to identify candidate organelle reads, followed by use of a k-mer count approach to filter out the remaining nuclear reads and finishing with assembly of the organellar genomes. **Figure 1** illustrates the entire workflow.

TIPP_plastid uses TIARA to classify the reads. We evaluated the accuracy of TIARA (Karlicki et al. 2022) using *Arabidopsis thaliana* and *Oryza sativa* (**Figure S1**). As described in the original paper, TIARA will classify the NUMTs/NUPTs as organelle reads, and there is also an increased proportion of misclassification in highly repeated regions, such as centromeres and rDNA clusters. Hence, further filtering is necessary.

The assumption for subsequent filtering is that true organellar reads are the largest class in the TIARA output, and that misclassifications are relatively rare. We use KMC3 (Kokot et al. 2017) to generate a k-mer (k=31) count database from the reads identified by TIARA. Next, we perform filtering based on k-mer counts separately for chloroplast and mitochondrial reads. We use readskmercount to obtain the read median k-mer count *rmkc*, which is used as a representative for each read. Reads labeled as plastid are processed first because chloroplast genomes are more conserved than mitochondrial genomes.

After calculating *rkmc* for all input reads, the median kmer count *mkc* of all chloroplast reads, and of all mitochondrial reads after chloroplast assembly, will be used for filtering. To this end, we set the low kmer count threshold *lkc* to 0.3 * *mkc*, and the high kmer count threshold *hkc* to 5 * *mkc*. A read is removed if more than one fifth of its k-mer counts are either lower than *lkc* or higher than *hkc*. Reads with many k-mer counts below the *lkc* threshold likely originate from the nucleus, and possibly correspond to NUPTs or NUMTs. Reads with many k-mer counts above the *hkc* threshold are likely from highly repetitive nuclear regions such as centromeres,

and rDNA clusters. After filtering, flye (Kolmogorov et al. 2019) is used to assemble the chloroplast genome in the first assembly step. The assembly is performed iteratively with a random selection of reads, until the assembly graph matches the typical chloroplast structure. In each assembly round, only 800 reads are used, which is around 100x coverage, since excessive coverage might negatively affect flye results. Following the assembly with flye, the assembly graph is checked for a typical chloroplast structure or a circular DNA when inverted repeats were set as lost. The structural check is aiming to match two isomeric chloroplast genomes that coexist equimolarly, differing only in the orientation of the LSC and SSC, as is the case in most land plants and algae (Palmer 1983; Aldrich et al. 1985; Wang and Lanfear 2019). Once this is achieved, the cycle ends with output of two typical heteroplasmic fasta sequences or one circular sequence.

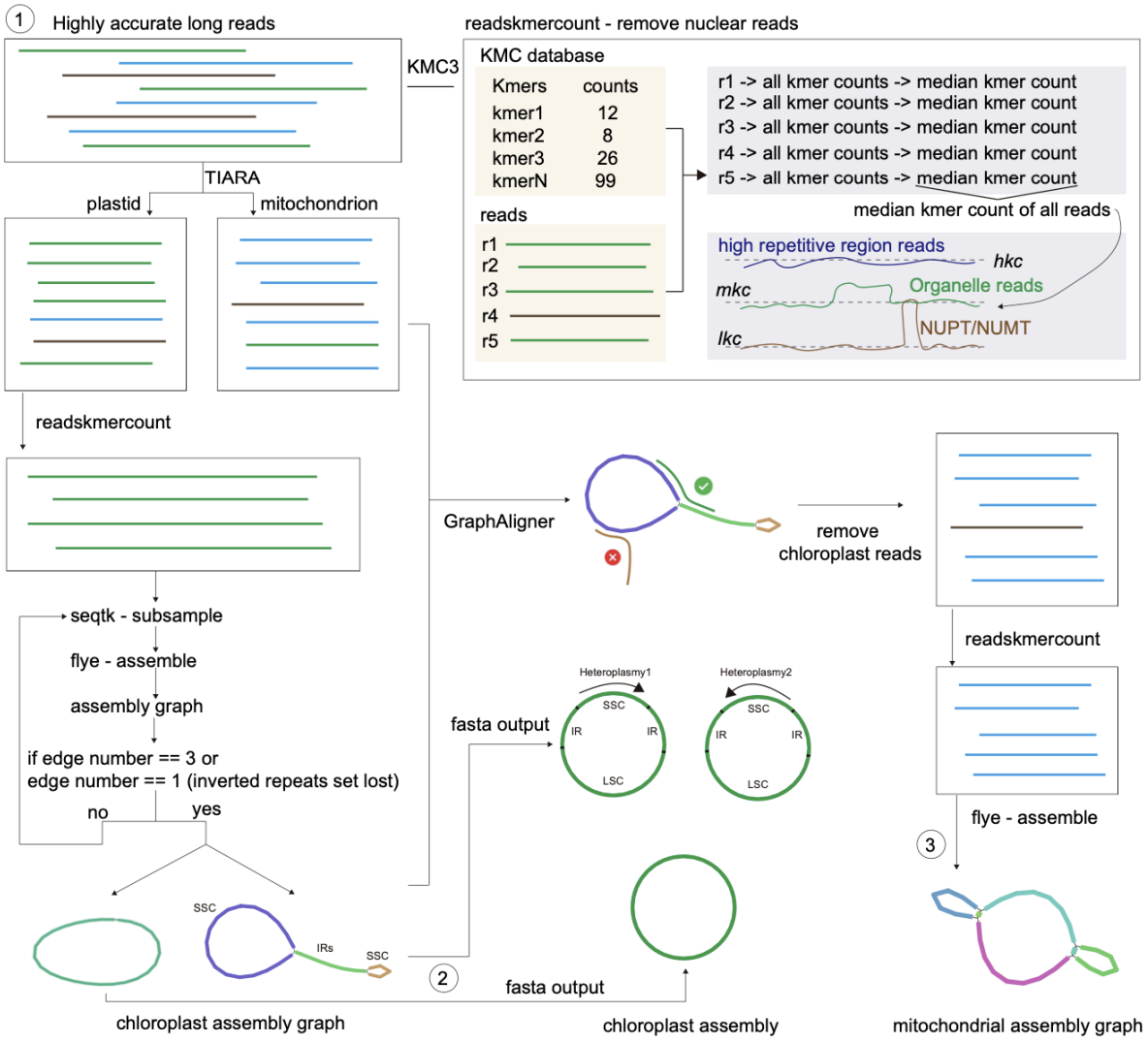

Figure 1. Workflow of TIPP_plastid.

The next step is the assembly of the mitochondrial genome. Considering that some chloroplast reads might be misclassified as mitochondrion by TIARA, GraphAligner

(Rautiainen and Marschall 2020) is used to align all reads labeled as mitochondrion to the chloroplast assembly graph as a further refinement step. If the read alignment is almost end-to-end (left clip length ≤ 100 bp, right clip length ≤ 100 bp and identity > 95%), reads are considered as likely originating from the chloroplast and are removed. It is worth noting that mapping reads directly to an organelle assembly graph is the optimal solution for the organelle genome alignment, since linearized circular DNA combined with heteroplasmy will lead to clipped alignments. CLAW (Phillips et al. 2024) also addresses the alignment issues caused by a linearized circular DNA target by joining the two linear DNA sequences. Although this approach avoids clipping alignment, it introduces the issue of mapping quality of zero.

As a final step in TIPP_plastid, the reads remaining after alignment to the chloroplast assembly graph will be processed by readskmercount to exclude reads originating from the nucleus, as described above. Given that the coverage of mitochondrion is generally lower than that of chloroplast and the genomes size are usually larger, all finally remaining reads serve directly as input to flye for generating the assembly graph.

Results and Discussion

Chloroplast genome assembly

Given the conserved structure of chloroplast genomes, we categorized the assemblies based on the structure on the assembly graph into three classes: 1) containing only the typical chloroplast genome or one circular DNA (complete genome); 2) consisting of the complete genome and other sequences; 3) incomplete assembly (**Figure 2**).

To test the performance of TIPP_plastid, we selected 54 phylogenetically diverse planta and compared the performance with that of ptGUAUL, CLAW and PMAT. Using TIPP_plastid, we successfully assembled all 54 complete chloroplast genomes without any extraneous sequences (**Figure S2**). We obtained two chloroplast genomes (**Figure S2**) for *Acorus gramineus*, suggesting that the sample might contain reads from two genotypes. It was therefore excluded from downstream analysis. ptGAUL assembled 46 complete genomes, produced six assemblies containing complete chloroplast genomes along with other sequences, and was unable to assemble one species (**Figure S3**). CLAW successfully assembled 35 complete chloroplast genomes, 14 assemblies included complete chloroplast genomes as well as other sequences, and it did not assemble four species (**Figure S4**). PMAT assembled 16 complete chloroplast genomes, 26 assemblies featured complete chloroplast genomes with additional sequences, and it failed to assemble 11 species (**Figure S5**).

Whole-genome alignments against published chloroplast genomes indicated high consistency between the published and TIPP_plastid assemblies (**Figure S6**). Typical chloroplast genomes have three distinct regions: SSC, LSC and IR. Public chloroplast genomes are typically presented as linear circular DNA sequences. Thus, in whole-genome alignments, the single IR from the assembly graph aligned to two regions of the linear representations, with one forward and one reverse orientation (**Figure S6**). We also assembled two previously unpublished chloroplast genomes, *Adenosma buchneroides*

(153,640 bp) (**Figure S7**) and *Helichrysum umbraculigerum* (154,011 bp) (**Figure S8**). Comparing the chloroplast genome lengths across 53 species, we observed that those from green algae are larger than those from terrestrial plants, with terrestrial plant chloroplast genomes generally around 150 kb (**Figure 2; Table S4**).

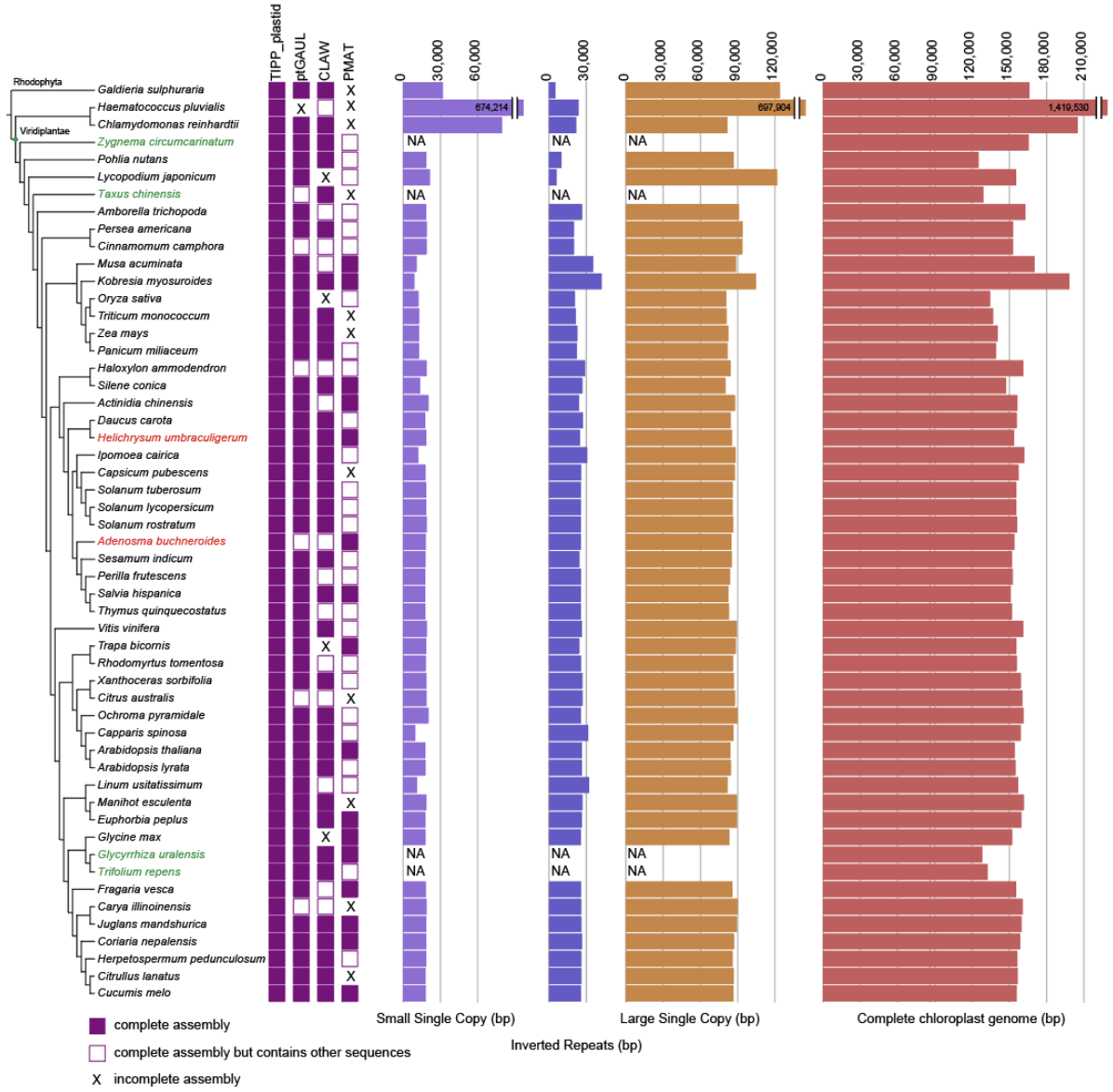

Figure 2. Benchmarking of four chloroplast genome assembly tools and genome statistics. See Methods for phylogenetic tree. Red species names indicate de novo assemblies for the first time. Green species names indicate loss of inverted repeats, and the three topologically defined regions are therefore not measured.

Mitochondrial genome assembly

Only PMAT assembled also mitochondrial genomes, and we therefore compared the ability of TIPP_plastid to assemble mitochondrial genomes with PMAT. Given that PMAT assembled genomes often contain sequences from both organelles, we aligned distinct parts of the mitochondrial assembly graph from both PMAT and TIPP_plastid against the chloroplast genomes assembled by TIPP_plastid. For PMAT, mitochondrial genome

assemblies from 33 out of 53 species contained also chloroplast sequences (**Figure S9**). For *Musa acuminata*, *Adenosma buchneroides*, *Trapa bicornis* (master), and *Fragaria vesca* (master), all parts aligned fully to the chloroplast genome graph, indicating that the assembly of mitochondrial genomes had failed. Since TIPP_plastid removes chloroplast reads first, none of the assemblies contain chloroplast sequences. Thus, in subsequent analyses, we removed the chloroplast sequences from PMAT mitochondrial genome assemblies.

Given the structural diversity of plant mitochondrial genomes, it is challenging to assess the completeness of results from the assembly graph structure as we did with chloroplasts. Inspired by BUSCO (Seppey et al. 2019) for assessing the completeness for nuclear genomes, we use 41 protein-coding genes (PCGs) collected by mitopy (Alverson et al. 2010) to evaluate the completeness of mitochondrial assemblies. Out of the 53 species, 35 mitochondrial genomes had previously been published, which we also included in our evaluation (**Table S5-S8**). Considering that the output of mitochondrial genomes from PMAT and TIPP_plastid is in the form of assembly graphs, where large repetitive fragments are represented only once, we focused on the presence or absence of genes, and did not consider orientation or copy number.

As shown in **Figure 3A**, TIPP_plastid and PMAT are in agreement regarding the completeness of PCGs in the mitochondrial assemblies of 43 species. In eight species, the mitochondrial genomes assembled by TIPP_plastid had higher PCG completeness, while for two species PMAT outperformed TIPP_plastid. For *Musa acuminata*, both TIPP_plastid and PMAT failed to assemble the mitochondrial genome.

The eight species in which TIPP_plastid was superior include one red alga and two Chlorophyta for which PMAT failed to output mitochondrial genomes, with the TIPP_plastid assemblies matching the published assemblies for these three species. Although *Haematococcus lacustris* and *H. pluvialis* belong to the same genus, their mitochondrial genomes exhibit poor synteny (**Figure S10**). In *Kobresia myosuroides*, the PMAT, but not the TIPP_plastid mitochondrial assembly lacked the *rps1* gene. Further whole-genome alignment (**Figure S11**) demonstrated that this was due to a mis-assembly by PMAT. However, the TIPP_plastid assembly includes non-mitochondrial fragments.

In *S. conica*, which has one of the largest mitochondrial genomes (11 Mb), TIPP_plastid assembled a mitochondrial genome that was highly consistent with the published genome (**Figure 3B**). PMAT, in contrast, only assembled parts of the mitochondrial genome. The mitochondrial assembly graph from TIPP_plastid had numerous small circular DNAs (**Figure 3C**), which PMAT failed to identify (**Figure 3D**). A similar issue with missing small circular DNAs in PMAT occurred in *Actinidia chinensis and Linum usitatissimum*. TIPP_plastid assembly of *A. chinensis* matched the published genome, which includes a large circular DNA of 724 kb and a smaller circular DNA of 200 kb, whereas PMAT only generates a linearized sequence of the large circle (**Figure S12**). In *L. usitatissimum*, the PMAT assembly had lost two protein-coding genes, *rpl5* and *rps14*, which are present in a circular DNA sequence assembled by TIPP_plastid. Whole-genome alignment again indicated that PMAT the assembly had lost the circular DNA with these two genes (**Figure S13**). In *Adenosma buchneroides*, PMAT failed to assemble the mitochondrial genome, whereas TIPP_plastid assembled a 346 kb linear DNA sequence containing 38 PCGs (**Figure S14**). Given the

number of PCGs in related species — 39 in *Sesamum*, 35 in *Perilla*, 37 in *Salvia*, and 36 in *Thymus* — this suggests that the linear DNA sequence from TIPP_plastid is largely complete.

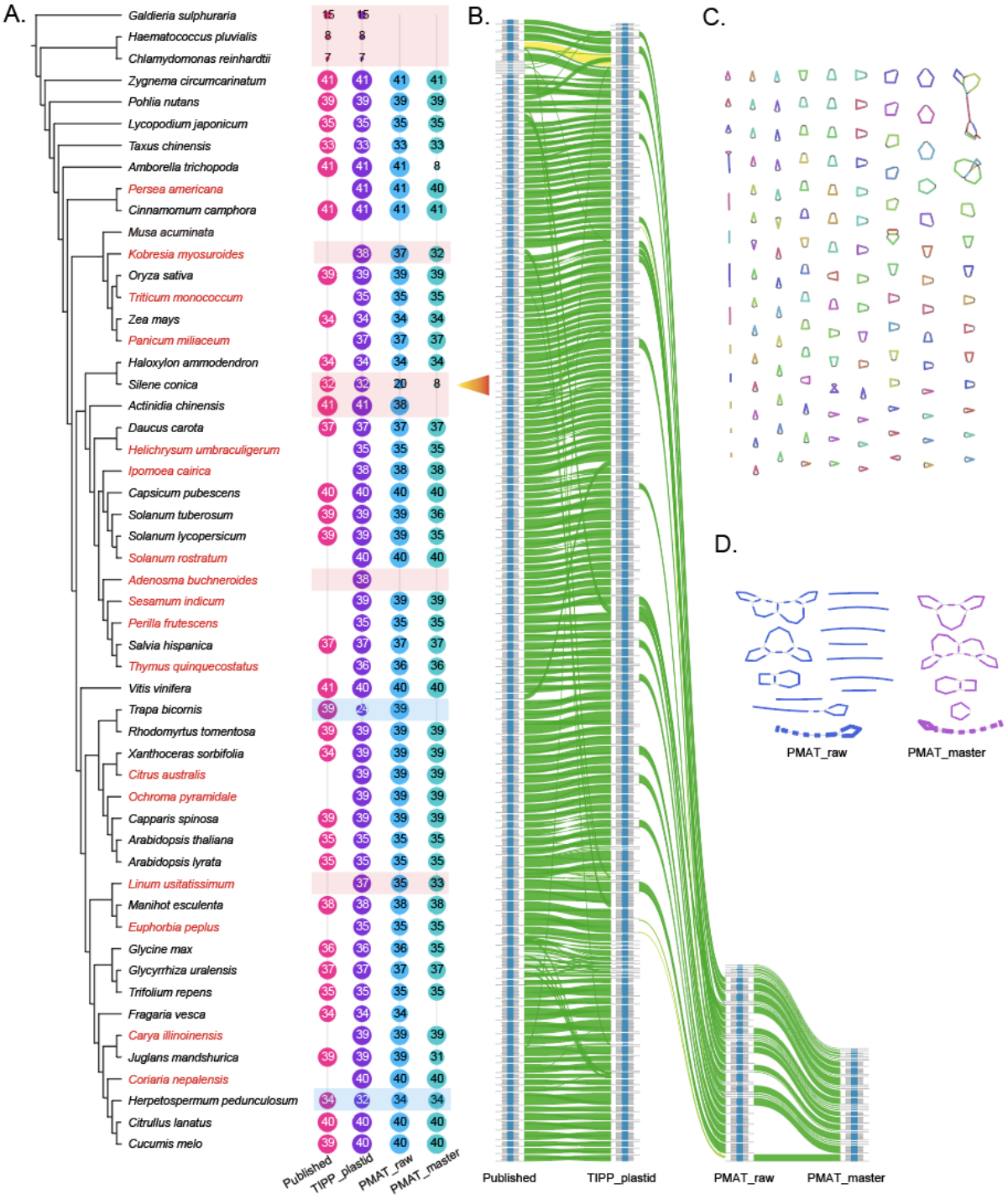

Figure 3. Benchmarking of mitochondrial genome assembly. A. See Methods for phylogenetic tree. Red species names indicate de novo assemblies for the first time. The numbers inside the circles indicate the number of non-redundant Protein Coding Genes (PCGs) in the assembly. Light red and light blue backgrounds indicate superior results with TIPP_plastid or PMAT. B, Whole genome alignment, including the published, TIPP_plastid, and PMAT assemblies (both raw and master), of the *S. conica* mitochondrial genome, visualized with Alitv. C. TIPP_plastid assembly graph of *S. conica* visualized with Bandage. D. PMAT assembly graph of *S. conica* visualized with Bandage.

As mentioned, PMAT outperformed TIPP_plastid for two species. For *Trapa bicornis*, the TIPP_plastid assembly graph comprised only linear DNA fragments, indicating the erroneous identification of a large number of non-mitochondrial reads. Using verkko to construct a whole-genome assembly graph revealed that *Trapa bicornis* possesses a large rDNA cluster that is misidentified by TIARA (**Figure S15**). For *Herpetospermum pedunculosum*, the TIPP_plastid assembly lacked two genes, *nad3* and *atp6*, due to assembly errors, as inferred from whole-genome comparisons. However, the PMAT raw assembly included non-mitochondrial fragments (**Figure S16**).

Computational cost

The most time-consuming step for PMAT is the initial whole genome assembly (Newbler), while for TIPP-plastid it is read classification (TIARA). We therefore evaluated runtime and peak memory usage by using PMAT in the pt module and TIPP_plastid in the chloroplast module. This allowed us to concurrently assess the other two chloroplast assembly tools ptGAUL and CLAW. Using the same data from 53 species, we again performed chloroplast assembly using four different tools. The results indicated that PMAT required the longest time, being 48 times slower than ptGAUL, and 8 times slower than either TIPP_plastid or CLAW, which in turn were only half as fast as ptGAUL (**Figure 4A**).

In terms of peak memory usage, PMAT also consumed the most memory, 27 times that of ptGAUL, 17 times that of CLAW, and 4 times that of TIPP_plastid. Despite the similar runtimes of TIPP_plastid and CLAW, TIPP_plastid memory consumption was three times that of CLAW and five times that of ptGAUL (**Figure 4B**). For detailed time and memory usage, please refer to **Table S9-10**.

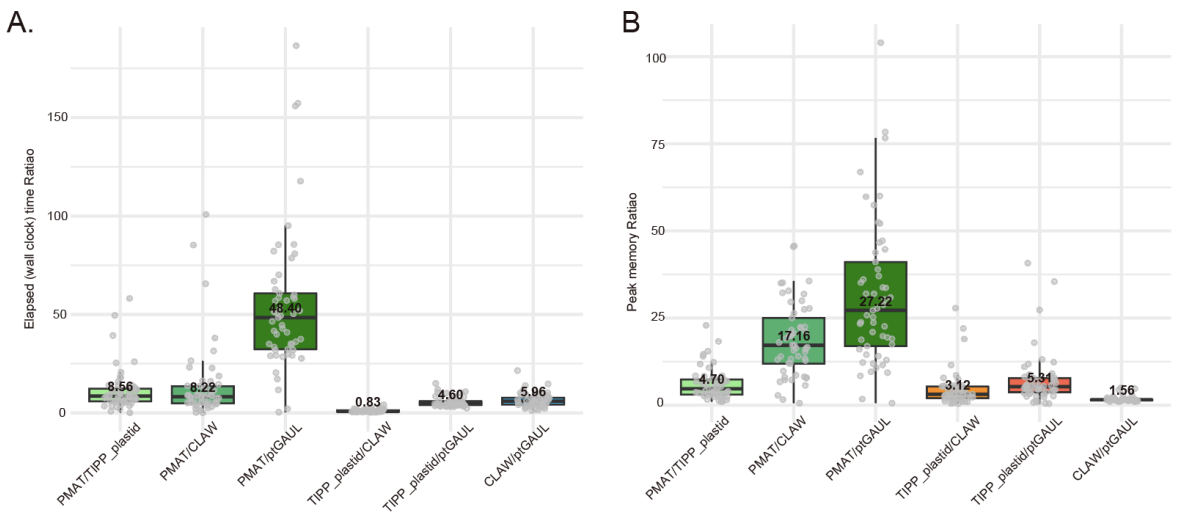

Figure 4. Computational cost. A. Ratio of elapsed times between each pair of the four tools. B. Ratio of peak memory usage between each pair of the four tools. Grey dots indicate different species. The means are shown as horizontal lines, with the upper and lower box indicating the interquartile range (IQR), and the whiskers extending to the most extreme values within 1.5 times the IQR from the first and third quartiles.

Identification of NUPTs/NUMTs

Next, we wanted to know whether we could improve on the accurate identification of NUPTs and NUMTs and the elimination of potential contamination of nuclear assemblies with pieces of organellar genomes. High-quality nuclear genomes assembled from PacBio HiFi data are available for all of the species used in this study except *Lycopodium japonicum*, *Ochroma pyramidale* and *Perilla frutescens*, with the assemblies of the latter two being highly fragmented. Because algal genomes are small and have very few NUPTs and NUMTs (Zhang et al. 2020), we excluded them from further analysis. *Musa acuminata* was not included either, because we had not been able to assemble the mitochondrial genome. For all other 45 nuclear genome assemblies, we retrieved all contigs/scaffolds over 500 kb.

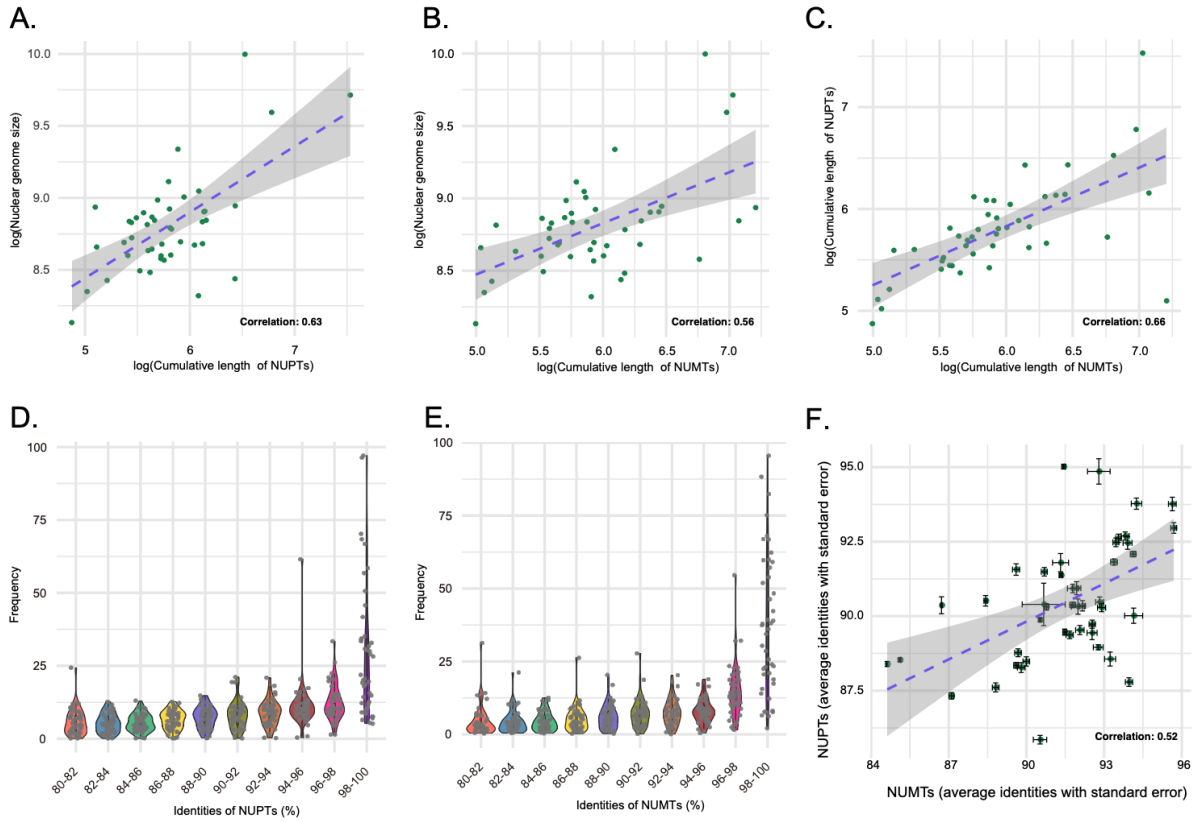

Figure 5. Comparison of NUPT and NUMT sequences and the corresponding organellar genomes. A. Comparison of cumulative lengths of NUPTs and of nuclear genome size. B. Comparison of cumulative lengths of NUMTs and of nuclear genome size. C. Comparison of cumulative lengths of NUPTs and of NUMTs. D. Cumulative length distribution of NUPTs across different identities. E. Cumulative length distribution of NUMTs as a function of sequence identity with the corresponding mitochondrial genome. F. Correlation between NUPT/chloroplast genome identity and NUMT/mitochondrial genome identity. Bars indicate standard errors.

The species with the longest cumulative lengths of NUMTs were *S. conica*, *Amborella trichopoda*, *Triticum monococcum*, *Capsicum pubescens* and *Taxus chinensis*. This might be attributed to *S. conica* and *A. trichopoda* having large mitochondrial genomes (11 and 3.9 Mb) and *T. monococcum*, *C. pubescens*, and *T. chinensis* having large nuclear genomes (5, 3.9 and 10 Gb). The latter three species also had the highest cumulative lengths of NUPTs

(**Table S11**). As observed before (Zhang et al. 2020), both NUPT and NUMT lengths are positively correlated with nuclear genome size (Pearson's correlation coefficients of 0.63 and 0.56) (**Figure 5 A&B**). Since NUPTs and NUMTs are part of the nuclear genome, their lengths are also positively correlated (**Figure 5C**).

NUPTs and NUMTs appear to evolve mostly neutrally, as evidenced by the gradual accumulation of mutations (Huang et al. 2005; Noutsos et al. 2005). Because the substitution rates of plant organellar genomes is typically an order of magnitude lower than that of nuclear genomes (Wolfe et al. 1987; Drouin et al. 2008), the number of differences between NUPT and NUMT sequences and the corresponding organellar genomes reflect the age of nuclear insertions (Richly and Leister 2004; Michalovova et al. 2013; Yoshida et al. 2019). We found that recent insertion events, with sequence identities of 98% to 100%, are most frequent (**Figure 5D, 5E, Table S12**), which is also reflected by the correlation of average sequence identities between NUPTs and NUMTs and their organellar genomes being well correlated (Pearson's correlation coefficient = 0.52) (**Figure 5F**). We conclude that NUPTs and NUMTs tend to be deleted quite quickly from the nuclear genomes, which is consistent with individual NUPTs and NUMTs in *A. thaliana* genomes having low allele frequencies (Igolkina et al. 2024).

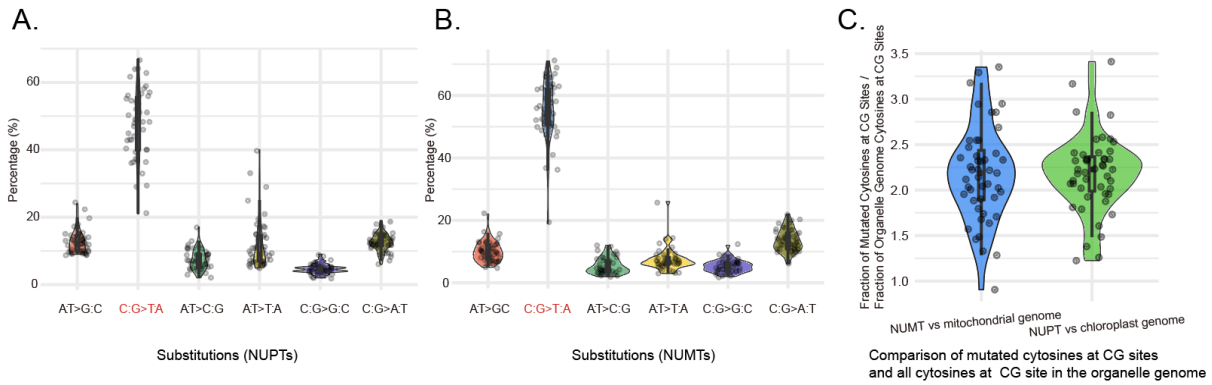

Figure 6. The landscape of substitutions in NUPTs and NUMTs. A. Distribution of nucleotide substitutions in NUPTs, inferred from sequence comparison with the corresponding chloroplast genome. B. Distribution of nucleotide substitutions in NUMTs, inferred from sequence comparison with the corresponding mitochondrial genome. C. Enrichment of cytosine substitutions in NUPTs and NUMTs at CG sites.

Mutation spectra of NUPTs/NUMTs

C:G>T:A substitutions dominate the substitution spectrum in *A. thaliana* mutation accumulation lines, both in the greenhouse and the wild, although not in older natural populations (Ossowski et al. 2010; Cao et al. 2011; Exposito-Alonso et al. 2018; Weng et al. 2019). The excess of C:G>T:A substitutions has been attributed to spontaneous deamination of methylated cytosines (Ossowski et al. 2010), which is found in plants in three contexts, CG, CHG and CHH, with most of it in the CG context (Law and Jacobsen 2010). Previous studies have found that C:G>T:A substitutions to be the most common substitutions in NUPTs and NUMTs (Huang et al. 2005; Rousseau-Gueutin et al. 2011; Fields et al. 2022). We

confirm this phenomenon in our set of 45 species, with the highest substitution rates at CG sites (**Figure 6, Tables S13, S14**).

siRNA targeting NUPTs and NUMTs

The increased substitution rate at CG sites in NUPTs and NUMTs suggested that these are often methylated, which has been directly confirmed in several instances (Yoshida et al. 2014; Fields et al. 2022). The most common type of DNA methylation in plants, RNA-directed DNA methylation (RdDM), is associated with small interfering RNAs (siRNAs) (Sigman and Slotkin 2016), and we therefore tested the hypothesis that NUPTs and NUMTs are enriched for siRNAs. In a previous study, siRNA data were generated for 11 of the 45 species that we investigated (Lunardon et al. 2020), and we annotated siRNA loci by mapping siRNA reads (Axtell 2013).

For all 11 species, the overlap of siRNA loci with NUPT/NUMTs was significantly higher than expected by chance (**Figure 7, Table S15**), demonstrating that siRNAs are indeed enriched in NUPTs and NUMTs.

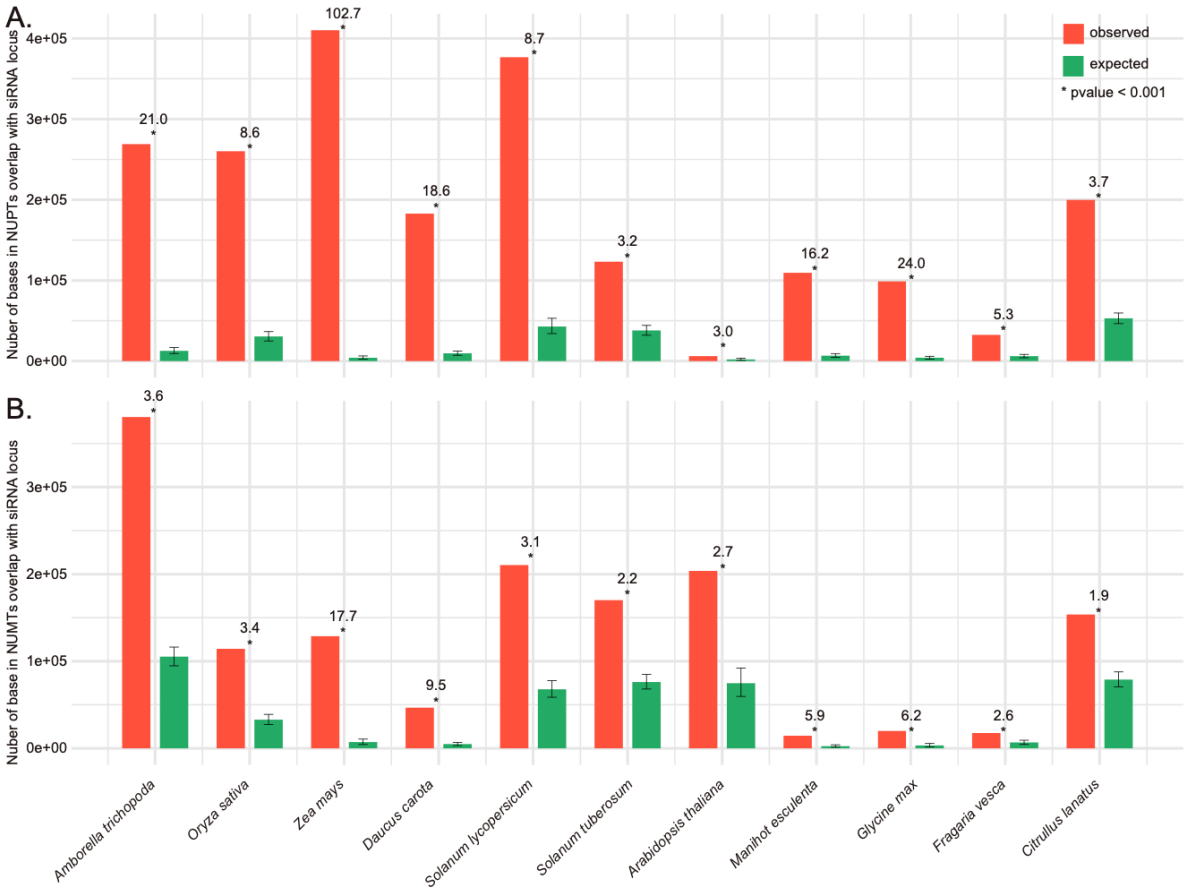

Figure 7. Enrichment of siRNAs in NUPTs and NUMTs. A. Overlap of siRNA loci with NUPTs. B. Overlaps of siRNA loci with NUMTs. Species in A and B annotated at the bottom. The numbers on top of each bar represent the enrichment, and the error bars represent the 95% confidence interval from random sampling of the genome.

Conclusion

We introduce TIPP_plastid, a user-friendly, reference-free approach for assembling plant organellar genomes. TIPP_plastid leverages high-accuracy long reads to provide a streamlined and universal assembly process without the need for external reference genomes. For both chloroplast and mitochondrial genomes, we provide assembly graphs. For chloroplast genomes, we provide in addition information on heteroplasmy.

TIPP_plastid outperforms all other tested assemblers for chloroplast genomes. Compared to chloroplast genomes, assessing the performance for mitochondrial genomes is more difficult due to the diversity of plant mitochondrial genomes. Based on the completeness of protein-coding genes, TIPP_plastid outperforms the second-best tool PMAT (Bi et al. 2024) in eight species, while PMAT was superior for two species, *T. bicornis* and *H. pedunculosum*. A significant factor appears to be the presence of a large rDNA cluster in the nuclear genome of *T. bicornis*, which results in poor classification by Tiara (Karlicki et al. 2022), the initial tool used by TIPP_plastid for selecting input reads for the assembly. Why PMAT outperforms TIPP_plastid for *H. pedunculosum* is unclear.

Materials and Methods

Data sources: HiFi datasets were downloaded from publicly available databases (**Table S1**). The accession numbers for chloroplast and mitochondrial genomes are provided in **Table S2** and **Table S3**. A phylogenetic tree of the 53 species was constructed with rtrees (Li 2023).

Evaluation of Tiara for read classification: First, minimap2 (2.24-r1122) was used to align all HiFi reads to the *A. thaliana* and *O. sativa* reference genomes, retaining only the primary alignments. Next, Tiara (1.0.3) (Karlicki et al. 2022) was used to classify HiFi reads as organellar. A 100 kb sliding window was applied to calculate the proportion of reads classified as organellar by Tiara compared to minimap2 in each window. The results were visualized using ggplot2.

Assembly of organellar genomes: We subsampled PacBio HiFi reads to approximate 4x nuclear genome coverage for each species, except for 2x for *T. chinensis*, which has a particularly large nuclear genome (Xiong et al. 2021). For *S. conica*, with its large mitochondrial genome (Sloan et al. 2012) , we used 10x nuclear genome coverage. We used identical datasets for assembly with the different tools. TIPP_plastid (v2.1) and PMAT (v1.5.3) (Bi et al. 2024) were set to assemble chloroplast and mitochondrial genomes simultaneously. For PMAT, the size of the nuclear genome was provided. For ptGAUL (v1.0.5) (Zhou et al. 2023) and CLAW (Phillips et al. 2024), which only assemble chloroplast genomes, the chloroplast genome sequences of closely related species were provided.

Whole-genome alignment and visualization: To compare genomes assembled from different sources, whole-genome alignments were performed with MiniTV (https://github.com/weigelworld/minitv), which uses minimap2 (Li 2018) for alignment, followed by visualization with AliTV (Ankenbrand et al. 2017).

Removal of chloroplast sequences from mitochondrial assemblies: First, we converted the mitochondrial assembly graphs into fragments. Given that the TIPP_plastid chloroplast assembly results are the cleanest and complete, we aligned the mitochondrial contigs from PMAT (Bi et al. 2024) to the TIPP_plastid chloroplast genome using minimap2 (Li 2018). Contigs that were covered over more than 90% of their length by the chloroplast genome and had greater than 95% similarity to it were labeled as "chloroplast". Using Bandage (Wick et al. 2015), we colored the nodes identified as chloroplast sequences in green and confirmed their identity after visual inspection. We removed the chloroplast sequences from the mitochondrial assemblies.

Assessing assembly completeness: We obtained amino acid sequence files for 41 conserved mitochondrial genes from mitopy (Alverson et al. 2010). We used BLASTX (2.9.0+) (McGinnis and Madden 2004) to align mitochondrial genome assemblies to each of the 41 genes, using a threshold of 1e-3. Considering that the current mitochondrial assembly results are presented in the format of an assembly graph, where long repeats will be collapsed into a single node, we evaluate gene completeness based on the presence or absence of genes, without accounting for their copy number.

Performance benchmarking: All organellar genomes were assembled on an AMD EPYC 7742 processors with 64 cores and 1 TB of RAM. Runtime and peak memory usage were calculated using the /usr/bin/time -v command.

NUMT and NUPT analysis: To identify NUPTs and NUMTs in the nuclear genome, we used BLASTN (2.9.0+) (McGinnis and Madden 2004) with the parameters -evalue 1e-5, -dust no, -penalty -2, -word_size 9, -outfmt 6. We aligned the chloroplast and mitochondrial genomes to their respective nuclear genomes and retained hits with an identity of > 80% and a length > 100 bp. Considering the redundancy in the BLASTN output, we removed all high-scoring segment pairs (HSPs) completely embedded in longer HSPs. We merged overlapping HSPs with bedtools (v2.31.1) (Quinlan and Hall 2010). The identity of the merged interval in the nuclear genome to the organellar genome was calculated as the average of the identities before merging.

To identify substitutions in NUPTs and NUMTs relative to the chloroplast and mitochondrial genomes, we used minimap2 (version 2.24-r1122) (Li 2018) with the parameters --paf-no-hit -ax asm5 --cs -r2k to generate alignment files. Finally, we used htsbox (version r345) (https://github.com/lh3/htsbox) with the parameters pileup -q5 -evcf to call variants.

Annotation of siRNA loci and overlap with NUPTs/NUMTs: For each of the selected 11 species, we downloaded data from two libraries. We used ShortStack4 (Axtell 2013) with default parameters to annotate siRNA loci. In short, reads with one or no mismatch were retained, and multi-mapping reads were assigned to a single location with the U model. GAT (Heger et al. 2013) was used to test whether the siRNA locus overlaps were greater than expected by chance with the parameter -num-samples=1000.

Data availability

Chloroplast and mitochondrial assembly graphs are available on Figshare at https://doi.org/10.6084/m9.figshare.26362141.v1.

Code availability

TIPP_plastid is available at Github (https://github.com/Wenfei-Xian/TIPP). Code to reproduce results from this paper can be found at Github (https://github.com/Wenfei-Xian/Reproducible_for_TIPP_paper).

Acknowledgements

We thank Andrea Movilli, Adrian Contreras, Yueqi Tao, Svitlana Sushko, Li He and Haim Ashkenazy for the discussions. This work was supported by the Max Planck Society and the Novozymes Prize of the Novo Nordisk Foundation (DW).

Author Contributions

W.X. designed the project and conducted the analyses. I.B. set up the computational environment. I.B., Z.B., S.V. and A.G. tested the tool. D.W. supervised the project. W.X. wrote the first draft of the manuscript. W.X. and D.W. prepared the final manuscript with input from all authors.

Competing Interests

DW holds equity in Computomics, which advises plant breeders. DW also consults for KWS SE, a plant breeder and seed producer with activities throughout the world. All other authors declare no conflicts.

References

Aldrich J, Cherney B, Merlin E, Williams C, Mets L. 1985. Recombination within the inverted repeat sequences of the Chlamydomonas reinhardii chloroplast genome produces two orientation isomers. *Curr. Genet.* 9:233–238.

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae). *Mol. Biol. Evol.* 27:1436–1448.

Ankenbrand MJ, Hohlfeld S, Hackl T, Förster F. 2017. AliTV—interactive visualization of whole genome comparisons. *PeerJ Comput. Sci.* 3:e116.

Axtell MJ. 2013. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* 19:740–751.

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:uhae023.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* 43:956–963.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18:170–175.

Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45:e18.

Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol. Phylogenet. Evol.* 49:827–831.

Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, Brachi B, Hagmann J, Grimm DG, Chen J, et al. 2018. The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.* 14:e1007155.

Fields PD, Waneka G, Naish M, Schatz MC, Henderson IR, Sloan DB. 2022. Complete Sequence of a 641-kb Insertion of Mitochondrial DNA in the Arabidopsis thaliana Nuclear Genome. *Genome Biol. Evol.* [Internet] 14. Available from: http://dx.doi.org/10.1093/gbe/evac059

Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.* 6:e1000834.

Heger A, Webber C, Goodson M, Ponting CP, Lunter G. 2013. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* 29:2046–2048.

Huang CY, Grünheit N, Ahmadinejad N, Timmis JN, Martin W. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiol.* 138:1723–1733.

Igolkina A, Vorbrugg S, Rabanal F, Liu H-J, Ashkenazy H, Kornienko A, Fitz J, Collenberg M, Kubica C, Morales AM, et al. 2024. Towards an unbiased characterization of genetic polymorphism. *bioRxiv* [Internet]:2024.05.30.596703. Available from: https://www.biorxiv.org/content/10.1101/2024.05.30.596703v1

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.

Karlicki M, Antonowicz S, Karnkowska A. 2022. Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics* 38:344–350.

Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, Novák L, Žárský V, Barlow LD, Herman EK, et al. 2016. A Eukaryote without a Mitochondrial Organelle. *Curr. Biol.* 26:1274–1284.

Kokot M, Dlugosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33:2759–2761.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37:540–546.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation

patterns in plants and animals. *Nat. Rev. Genet.* 11:204–220.

Li D. 2023. rtrees: an R package to assemble phylogenetic trees from megatrees. *Ecography* [Internet] 2023. Available from: https://onlinelibrary.wiley.com/doi/10.1111/ecog.06643

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100.

Lunardon A, Johnson NR, Hagerott E, Phifer T, Polydore S, Coruh C, Axtell MJ. 2020. Integrated annotations and analyses of small RNA-producing loci from 47 diverse plants. *Genome Res.* 30:497–513.

Maumus F, Quesneville H. 2014. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. *Nat. Commun.* 5:4104.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32:W20–W25.

Michalovova M, Vyskot B, Kejnovsky E. 2013. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity* 111:314–320.

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, Mandáková T, Jamge B, Lambing C, Kuo P, et al. 2021. The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* 374:eabi7489.

Noutsos C, Richly E, Leister D. 2005. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. *Genome Res.* 15:616–628.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* 376:44–53.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* 327:92–94.

Palmer JD. 1983. Chloroplast DNA exists in two orientations. *Nature* 301:92–93.

Phillips AL, Ferguson S, Burton RA, Watson-Haigh NS. 2024. CLAW: An automated Snakemake workflow for the assembly of chloroplast genomes from long-read data. *PLoS Comput. Biol.* 20:e1011870.

Putintseva YA, Bondar EI, Simonov EP, Sharov VV, Oreshkova NV, Kuzmin DA, Konstantinov YM, Shmakov VN, Belkov VI, Sadovsky MG, et al. 2020. Siberian larch (Larix sibirica Ledeb.) mitochondrial genome assembled using both short and long nucleotide sequence reads is currently the largest known mitogenome. *BMC Genomics* 21:654.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Rautiainen M, Marschall T. 2020. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* 21:253.

Richly E, Leister D. 2004. NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Mol. Biol. Evol.* 21:1972–1980.

Rousseau-Gueutin M, Ayliffe MA, Timmis JN. 2011. Conservation of plastid sequences in the plant nuclear genome for millions of years facilitates endosymbiotic evolution. *Plant*

Physiol. 157:2181–2193.

Sanchez-Puerta MV, García LE, Wohlfeiler J, Ceriotti LF. 2017. Unparalleled replacement of native mitochondrial genes by foreign homologs in a holoparasitic plant. *New Phytol.* 214:376–387.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol. Biol.* 1962:227–245.

Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. 2022. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* 19:823–826.

Sigman MJ, Slotkin RK. 2016. The First Rule of Plant Transposable Element Silencing: Location, Location, Location. *Plant Cell* 28:304–313.

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.

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* 18:49.

Štorchová H, Krüger M. 2024. The overview of methods for assembling complex mitochondrial genomes in land plants. *J. Exp. Bot.* [Internet]. Available from: http://dx.doi.org/10.1093/jxb/erae034

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, Darwin Tree of Life Consortium, Formenti G, Abueg L, Torrance J, Myers EW, Durbin R, Blaxter M, et al. 2023. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* 24:288.

Wang W, Lanfear R. 2019. Long-Reads Reveal That the Chloroplast Genome Exists in Two Distinct Versions in Most Plants. *Genome Biol. Evol.* 11:3372–3381.

Wang X, Weigel D, Smith LM. 2013. Transposon variants and their effects on gene expression in Arabidopsis. *PLoS Genet.* 9:e1003255.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37:1155–1162.

Weng M-L, Becker C, Hildebrandt J, Neumann M, Rutter MT, Shaw RG, Weigel D, Fenster CB. 2019. Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in Arabidopsis thaliana. *Genetics* 211:703–714.

Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31:3350–3352.

Wlodzimierz P, Rabanal FA, Burns R, Naish M, Primetis E, Scott A, Mandáková T, Gorringe N, Tock AJ, Holland D, et al. 2023. Cycles of satellite and transposon evolution in Arabidopsis centromeres. *Nature* 618:557–565.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U. S. A.*

84:9054–9058.

Xiong X, Gou J, Liao Q, Li Y, Zhou Q, Bi G, Li C, Du R, Wang X, Sun T, et al. 2021. The Taxus genome provides insights into paclitaxel biosynthesis. *Nat Plants* 7:1026–1036.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.* 21:809–818.

Yoshida T, Furihata HY, Kawabe A. 2014. Patterns of genomic integration of nuclear chloroplast DNA fragments in plant species. *DNA Res.* 21:127–140.

Yoshida T, Furihata HY, To TK, Kakutani T, Kawabe A. 2019. Genome defense against integrated organellar DNA fragments from plastids into plant nuclear genomes through DNA methylation. *Sci. Rep.* 9:2060.

Zhang G-J, Dong R, Lan L-N, Li S-F, Gao W-J, Niu H-X. 2020. Nuclear Integrants of Organellar DNA Contribute to Genome Structure and Evolution in Plants. *Int. J. Mol. Sci.* [Internet] 21. Available from: http://dx.doi.org/10.3390/ijms21030707

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.

Zimorski V, Ku C, Martin WF, Gould SB. 2014. Endosymbiotic theory for organelle origins. *Curr. Opin. Microbiol.* 22:38–48.