TIPP_plastid: A User-Friendly Tool for De Novo Assembly of Plastid Genomes

Wenfei Xian[1], Detlef Weigel[1]*

1 Department of Molecular Biology, Max Planck Institute for Biology, Tuebingen 72076, Germany
*corresponding author: weigel@tue.mpg.de (D.W.)

Abstract

Motivation: The chloroplast is the primary site of photosynthesis in plants, converting solar energy into chemical energy. The chloroplast genome is typically a circular DNA composed of four sections: a large single copy (LSC), a small single copy (SSC), and two intervening repeats (IRs). Long-read sequencing data now allow for assembly of chloroplast genomes along with nuclear genomes, but this requires that reads originating from the chloroplast genome are distinguished from reads that represent nuclear insertions of chloroplast DNA or related sequences from the mitochondrial genomes and that chloroplast heteroplasmy, i.e., the presence of multiple distinct chloroplast genomes is taken into account.

Results: We introduce TIPP_plastid, an efficient tool for the assembly of plastid genomes based on the exclusion of reads originating from the mitochondrial and the nuclear genome, followed by the construction of assembly graphs and extraction of distinct chloroplast genomes. We demonstrate the usefulness of TIPP_plastid by assembling the chloroplast genomes of 45 phylogenetically diverse species.

Availability and Implementation: 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). Chloroplast genomes and assembly graphs of 45 phylogenetically diverse species can be downloaded at figshare (https://doi.org/10.6084/m9.figshare.24715017.v1).

1 Introduction

The currently most widely used chloroplast genome assembly tools, NOVOPlasty (Dierckxsens, Mardulyn and Smits, 2017) and GetOrganelle (Jin *et al.*, 2020) were built for assembling chloroplast genomes from short read data and they cannot make use of third-generation long reads. Organelle_PBA (Soorni *et al.*, 2017) was designed for early versions of long reads of low accuracy and is no longer being updated. MitoHiFi (Uliano-Silva *et al.*, 2023) can use PacBio HiFi data, but it is specifically optimized for assembling the mitochondrial genomes of metazoans and fungi, which are much smaller than those of plant mitochondria. Consequently, there exists a notable gap in the availability of a chloroplast assembly tool that effectively utilizes PacBio HiFi data.

In a plant cell, there are primarily three locations where DNA can be found: the chloroplasts or chloroplast-related plastids, the mitochondria, and the nucleus. Because the transfer of genomic DNA from organelles to the nucleus or other organelles is common and can lead to integration of organellar DNA into one of the other genomes, it is common to detect chloroplast derived sequences in both the nuclear (Martin *et al.*, 2002) and mitochondrial (Wang *et al.*, 2007) genomes. Conventional chloroplast assembly tools typically recognize nuclear integrants of plastid DNA (NUPTs), but often overlook the potential presence of such fragments in the mitochondria. Therefore, accurately identifying reads originating from the chloroplast genome requires the exclusion of sequences deriving from both the nuclear and mitochondrial genomes.

Another phenomenon that is often overlooked is heteroplasmy (Palmer, 1983), the presence of multiple, often two, chloroplast genomes in the same individual, primarily characterized by variations in the orientation of the small single copy (SSC) region of the chloroplast genome. If assembly tools only reconstruct one, the SSC region is frequently misinterpreted as an inversion hotspot in subsequent comparative chloroplast genome analyses (Ibrahim, Azuma and Sakamoto, 2006; Liu *et al.*, 2013; Walker, Zanis and Emery, 2015). This misinterpretation typically stems from neglecting heteroplasmy (Walker *et al.*, 2015). Therefore, to reduce such misrepresentations, an assembly tool should aim to output at least two heteroplasmic genomes (Wang and Lanfear, 2019).

To address the issues outlined above, we developed TIPP_plastid, a tool for assembling chloroplast genomes using PacBio HiFi data. We demonstrate the robustness and effectiveness of TIPP_plastid with the assembly of chloroplast genomes from 45 phylogenetically diverse species.

2 Approach

The overall approach for TIPP_plastid is described in **Figure 1**. To be able to recognize sequences originating from mitochondria, we first downloaded a comprehensive set of chloroplast and mitochondrial genomes from NCBI. To reduce data redundancy, we selected only one genome per genus, resulting in a chloroplast database comprising 4,452 genera and a mitochondrial database including data from 485 genera.

We use Minimap2 (Li, 2018) to align HiFi reads from lineages with chloroplasts to our chloroplast and mitochondrial DNA databases. Since a specific read can originate from only one source, we adopted a 'best match' approach to exclude reads representing *bona fide* mitochondrial sequences, assuming that *bona fide* chloroplast sequences will always have a better match in the chloroplast than in the mitochondrial DNA database. This step leads to a collection of reads that originated from either the chloroplast or the nuclear genome of each target species.

Considering that a plant cell typically carries dozens of chloroplasts, sequences originating from chloroplast genomes are expected to have higher sequencing depth than those from the nuclear genome. We employed a reference-free method based on k-mer counts to estimate the depth distribution of each read using KMC3 (Kokot, Dlugosz and Deorowicz, 2017), resulting in a k-mer database for reads from either chloroplast or nucleus.

Utilizing the KMC3 API, we developed a C++ tool named *readskmercount* to obtain the count distribution of each k-mer in each read, thereby determining the dataset's median count. We avoided using the average count due to its susceptibility to outliers. By comparing the k-mer count distribution of each read with the baseline median count, we can identify and exclude reads representing the nuclear genome.

From the final set of chloroplast-derived PacBio HiFi reads, we select 2,000 to construct an assembly graph with Flye (Kolmogorov *et al.*, 2019). This downsampling step is introduced to accelerate the assembly process. At an approximate average length of HiFi reads of 15 kb (Wenger *et al.*, 2019), this provides around 200x coverage of the typical chloroplast genome with a size of close to 150 kb (Lee *et al.*, 2021).

We automatically determine whether the structure of the assembly graph is indicative of the presence of inverted repeats, common in plant chloroplasts, or absence of inverted repeats. If inverted repeats are detected, two heteroplasmic genomes are rebuilt from the assembly graph.

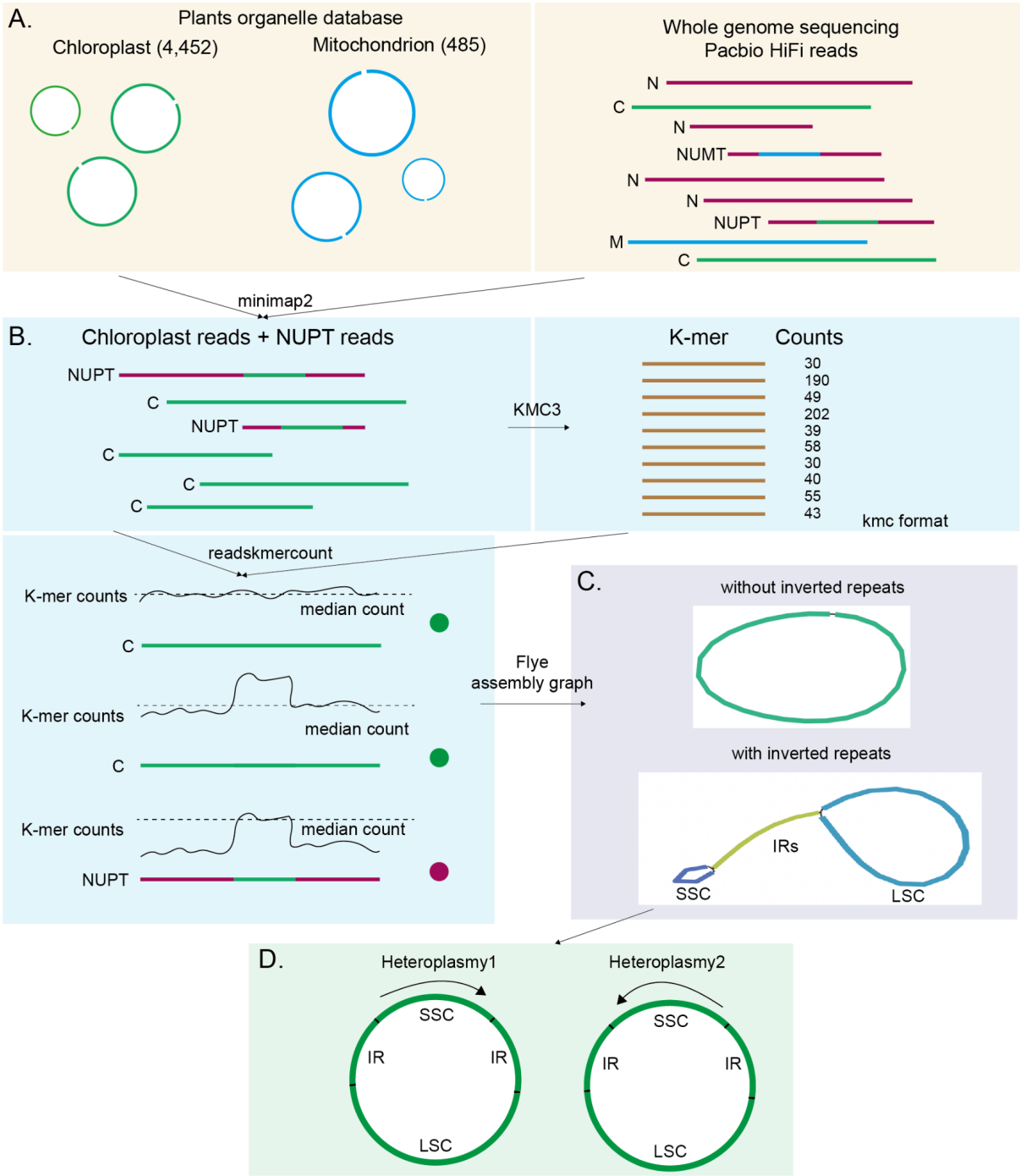

Figure 1: Workflow of TIPP_plastid. The process can be divided into four main stages: A. Exclusion of sequences originating from the mitochondrial genome. B. Exclusion of sequences originating from the nuclear genome. C. Construction of the assembly graph. D. Building of heteroplasmic genomes.

3 Results

To validate the efficacy of our methodology, we used publicly available PacBio HiFi data from 45 species, spanning from Rhodophyta (red algae) to angiosperms (flowering plants). HiFi reads from a Chlorophyta species, *Haematococcus lacustris*, initially failed to generate a

complete assembly under our default parameters. Literature review revealed that the size of its chloroplast genome is an astonishing 1.35 Mb (Bauman *et al.*, 2018) . Our default settings included selecting only a subset of 2,000 chloroplast HiFi reads for assembly, intended to expedite processing speed. In this case, the chloroplast genome coverage, of around 20x, was not sufficient to support complete assembly. By increasing the number of chloroplast HiFi reads to 6,000, we obtained a complete assembly of 1.42Mb for *Haematococcus lacustris* as well.

Among the chloroplast assembly graphs of the 45 species, we identified 42 with inverted repeats, and 3 species without inverted repeats: the gymnosperm *Torreya grandis*, and the two Fabaceae *Trifolium repens* and *Glycyrrhiza uralensis*, in agreement with the literature (Palmer and Thompson, 1982; Wu *et al.*, 2011).

4 Conclusion

We offer TIPP_plastid, an easy-to-use and efficient tool for chloroplast genome assembly, which is particularly suitable for the rapidly increasing number of pangenome projects using PacBio HiFi sequencing data (Tang *et al.*, 2022; Zhou *et al.*, 2022; Wlodzimierz *et al.*, 2023).

Supplementary Information

Supplementary Information can be found at
https://drive.google.com/file/d/1xi0flAchaHePd5YqmuOqKT4i9sh_3TvS/view?usp=drive_link

Conflicts of interest

D.W. holds equity in Computomics, which advises plant breeders. D.W. also consults for KWS SE, a plant breeder and seed producer with activities throughout the world. The other author has declared no conflicts of interest.

Funding

This work was supported by the Max Planck Society.

References

Bauman, N. *et al.* (2018) ‘Next-Generation Sequencing of Haematococcus lacustris Reveals an Extremely Large 1.35-Megabase Chloroplast Genome’, *Genome announcements*, 6(12). Available at: https://doi.org/10.1128/genomeA.00181-18.

Dierckxsens, N., Mardulyn, P. and Smits, G. (2017) ‘NOVOPlasty: de novo assembly of organelle genomes from whole genome data’, *Nucleic acids research*, 45(4), p. e18.

Ibrahim, R.I.H., Azuma, J.-I. and Sakamoto, M. (2006) ‘Complete nucleotide sequence of the cotton (Gossypium barbadense L.) chloroplast genome with a comparative analysis of

sequences among 9 dicot plants', *Genes & genetic systems*, 81(5), pp. 311–321.

Jin, J.-J. *et al.* (2020) 'GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes', *Genome biology*, 21(1), pp. 1–31.

Kokot, M., Dlugosz, M. and Deorowicz, S. (2017) 'KMC 3: counting and manipulating k-mer statistics', *Bioinformatics* , 33(17), pp. 2759–2761.

Kolmogorov, M. *et al.* (2019) 'Assembly of long, error-prone reads using repeat graphs', *Nature biotechnology*, 37(5), pp. 540–546.

Lee, C. *et al.* (2021) 'The chicken or the egg? Plastome evolution and an independent loss of the inverted repeat in papilionoid legumes', *The Plant journal: for cell and molecular biology*, 107(3), pp. 861–875.

Li, H. (2018) 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics* , 34(18), pp. 3094–3100.

Liu, Y. *et al.* (2013) 'Complete chloroplast genome sequences of Mongolia medicine Artemisia frigida and phylogenetic relationships with other plants', *PloS one*, 8(2), p. e57533.

Martin, W. *et al.* (2002) 'Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus', *Proceedings of the National Academy of Sciences of the United States of America*, 99(19). Available at: https://doi.org/10.1073/pnas.182432999.

Palmer, J.D. (1983) 'Chloroplast DNA exists in two orientations', *Nature*, 301(5895), pp. 92–93.

Palmer, J.D. and Thompson, W.F. (1982) 'Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost', *Cell*, 29(2), pp. 537–550.

Soorni, A. *et al.* (2017) 'Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data', *BMC genomics*, 18. Available at: https://doi.org/10.1186/s12864-016-3412-9.

Tang, D. *et al.* (2022) 'Genome evolution and diversity of wild and cultivated potatoes', *Nature*, 606(7914), pp. 535–541.

Uliano-Silva, M. *et al.* (2023) 'MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads', *BMC bioinformatics*, 24(1), pp. 1–13.

Walker, J.F. *et al.* (2015) 'Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes', *American journal of botany*, 102(11), pp. 1751–1752.

Walker, J.F., Zanis, M.J. and Emery, N.C. (2015) 'Correction to "Comparative analysis of complete chloroplast genome sequence and inversion variation in Lasthenia burkei (Madieae, Asteraceae)"', *American journal of botany*, 102(6), p. 1008.

Wang, D. *et al.* (2007) 'Transfer of Chloroplast Genomic DNA to Mitochondrial Genome Occurred At Least 300 MYA', *Molecular biology and evolution*, 24(9), pp. 2040–2048.

Wang, W. and Lanfear, R. (2019) 'Long-Reads Reveal That the Chloroplast Genome Exists in Two Distinct Versions in Most Plants', *Genome biology and evolution*, 11(12), p. 3372.

Wenger, A.M. *et al.* (2019) 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature biotechnology*, 37(10), pp.

1155–1162.

Wlodzimierz, P. *et al.* (2023) ‘Cycles of satellite and transposon evolution in Arabidopsis centromeres’, *Nature*, 618(7965), pp. 557–565.

Wu, C.-S. *et al.* (2011) ‘Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny’, *Genome biology and evolution*, 3, pp. 1284–1295.

Zhou, Y. *et al.* (2022) ‘Graph pangenome captures missing heritability and empowers tomato breeding’, *Nature*, 606(7914), pp. 527–534.