NOVOWrap: An automated solution for plastid genome assembly and structure standardization

Ping Wu[1], Hao Chen[2], Chao Xu[1], Jie Yang[1], Xian-chun Zhang[1], and Shi-Liang Zhou[1]

[1]Institute of Botany Chinese Academy of Sciences
[2]Shaanxi University of Science and Technology

September 11, 2020

Abstract

Plastid genomes are unique to plants and play an important role in genomics and evolutionary biology. Next-generation sequencing has revolutionized plastid genome data acquisition in a way that genome assembly and annotation became bottlenecks for large plastid genome data usage. Here we develop a novel open-source, cross-platform tool, NOVOWrap, with both command-line and graphical user interfaces for plastid genome automatic assembly using personal computers. With minimum inputs and user intervention, NOVOWrap could automatically assemble plastid genomes, validate results and standardize the structure with affordable computer resources. The performance of the software has been successfully benchmarked against eleven plastid genomes of species belonging to lycopods, gymnosperms, and angiosperms. The program is expected to liberate researchers from laborious computer manipulations and create reliable and standard genomic data.

Title:

NOVOWrap: An automated solution for plastid genome assembly and structure standardization

Authors:

Ping Wu[1,2], Hao Chen[3], Chao Xu[1], Jie Yang[1,2], Xianchun Zhang[1,2] and Shiliang Zhou[1,2*]

Affiliations:

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3] Shaanxi University of Science and Technology, Xi'an 710021, China

*To whom correspondence should be addressed.

Abstract

Plastid genomes are unique to plants and play an important role in genomics and evolutionary biology. Next-generation sequencing has revolutionized plastid genome data acquisition in a way that genome assembly and annotation became bottlenecks for large plastid genome data usage. Here we develop a novel open-source, cross-platform tool, NOVOWrap, with both command-line and graphical user interfaces for plastid genome automatic assembly using personal computers. With minimum inputs and user intervention, NOVOWrap could automatically assemble plastid genomes, validate results and standardize the structure with affordable computer resources. The performance of the software has been successfully benchmarked against eleven plastid genomes of species belonging to lycopods, gymnosperms, and angiosperms. The program is expected

to liberate researchers from laborious computer manipulations and create reliable and standard genomic data.

KEYWORDS: plastid genome, assembly, quadripartite structure

Availability

The source code and portable packages for various operating systems are available at https://github.com/wpwupingwp/novowrap. The software is released under the AGPL-3.0 license.

Introduction

Plastids (or chloroplasts in green plants) are considered to originate from a single endosymbiotic event involving an alpha-proteobacterium and cyanobacterium (Keeling 2010). During the process of evolution, the genome of the ancient cyanobacterium shrank and became the plastid genome of approximately 120-160 kb in size (Green 2011).

A typical plastid genome is a circular, double-stranded DNA molecule with a quadripartite structure, which consists of two inverted repeated regions (IR), a small single-copy region (SSC) and a large single-copy region (LSC). The plastid genome usually encodes 110-130 genes with high homology and collinearity among plant species (Wicke *et al.*2011).

Owing to its conserved structure, moderate sequence variability (Smith 2015), and high copy number in a cell, plastid genomes of partial or full-length have been widely used in plant phylogeny, comparative genomics and biotechnology (Tonti-Filippini *et al.* 2017).

The first two plastid genomes were determined in 1986 from*Nicotiana tabacum* (Sugiura *et al.* 1986) and *Marchantia polymorpha*(Brassell *et al.* 1986). To date, there are ˜5,000 complete plastid genomes deposited in GenBank (Sayers *et al.*2019), and the number is soaring since the advent of next-generation sequencing (NGS) technologies (Twyford & Ness 2017). Furthermore, several software for annotating plastid genomes have been developed (Huang & Cronk 2015; Qu *et al.* 2019; Shi *et al.* 2019; Tillich *et al.* 2017). Thus, the biggest obstacle to the data acquisition of plastid genomes seems to be in the assembly step.

The mainstream strategy to determine a plastid genome is by sequencing the total DNA including plastid, mitochondrion and nuclear genome components using NGS technology. Therefore, handling mixed NGS reads is a challenge in plastid genome assembly. One solution is to map all the reads to a close related reference genome allowing reads of plastid to be filtered and collected for assembly. Another solution is to construct contigs by*de novo* assembly, and plastid related contigs are then screened and assembled into a complete genome. Because all reads are used for assembly, this method requires relatively high computing power that personal computers can hardly provide. Both methods require manual adjustment of the IR regions to form a complete plastid genome and a highly similar reference genome is sometimes indispensable (Twyford & Ness 2017).

A novel and increasingly popular strategy is to use a universal seed sequence to bait plastid reads and extend the assemblage cyclically until the full circle is formed. Such method not only overcomes the computing burden of processing all the reads, but also obviates the requirement of complete genomes as a reference (Freudenthal *et al.* 2019). Two widely used implementations of such strategy are NOVOPlasty (Dierckxsens *et al.* 2016) and GetOrganelle (Jin*et al.* 2019). The former hashes all the reads before the extension step and has its own assembly algorithm instead of calling SPAdes (Bankevich *et al.* 2012), which requires less running time.

Unfortunately, all available software only have command-line interfaces and involve complex inputs or settings, which is a challenge for those who have limited computer skills or knowledge of operating systems (Attwood *et al.* 2019). Manual intervention to handle the questionable outputs is also unavoidable. Even NOVOPlasty and GetOrganelle, which usually generate full-length genome sequences, produce multiple outputs with opposite directions, different starting sites, alternate orientation of LSC/SSC, or sometimes mis-assembly. Moreover, although the commonly used *rbcL* for baiting reads usually works well, it may fail

in some cases such as gene transfer events or poor quality. Thus, developing more seeds could be helpful (Lim *et al.* 2018).

Here, we present NOVOWrap, a user-friendly, cross-platform Python package for plastid genome assembly. The program could work effectively on a personal computer and generate reliable assembly results with a standardized structure, with minimal user intervention during the process. By providing a highly automatic solution, the program could help to empower researchers with limited bioinformatics skills or computer resources to more easily determine plastid genomes for phylogeny, genomics, and biotechnology.

Methods

Development of new seeds for assembly

In order to determine the seeds used for assembly, BarcodeFinder (https://github.com/wpwupingwp/BarcodeFinder) was used to screen plastid genome sequences of diverse taxa. Briefly, 5,254 plastid genomes of green plants were downloaded from NCBI RefSeq database (Pruitt *et al.* 2012). In order to remove redundant data and reduce computation time, only one genome of each genus was used and 1,906 genomes were retained. Following this, sequences of each locus were extracted into multiple files according to annotations. After sequences were aligned using MAFFT (Katoh & Standley 2013), sequence polymorphism of each locus was investigated (Nei 1987; Spellerberg & Fedor 2003). Finally, the seeds were determined based on the following five principles:

(1) the sequences of seed are highly conserved; (2) the locus as a seed should exist and not pseudogenize in most of the taxa; (3) the locus as a seed should not transfer to the mitochondrial genome or nuclear genome commonly; (4) the seed candidates should be located far enough from each other in the plastid genome to prevent insufficient coverage of data or notorious repeats influencing the assembly start by taking full use of alternative seeds (for instance, seeds are from different regions of the quadripartite structure); and (5) since *rbcL* has been recommended and widely used, it is kept regardless of other principles.

Eventually, *rbcL* , *psaB* , *psaC* and *rrn23* were selected as candidate seeds (Table 1). The information of other loci is shown in Supplementary material 1.

The Assembly module

For assembly, NOVOWrap accepts both gzip-compressed and uncompressed NGS data (FASTQ format) as input. Because memory consumption is positively correlated to data size, oversize files can hardly be handled on normal personal computers. The Assembly module offers options for extracting partial data from the original data files to reduce memory usage.

With user-provided taxon information (scientific name or others) and the assistance of NCBI Taxonomy database, the module downloads the plastid genome of the most closely related species deposited in NCBI RefSeq database. Alternatively, the program could run offline if the user provides the reference locally. Subsequently, the sequences of seeds were extracted according to the annotations.

With the above discussed inputs (NGS data files and taxonomy) and all parameters that can be automatically detected by the program (for instance, read length, pair-end or not, assembly type, and file path of each input and output), several configuration files using different seeds are generated for NOVOPlasty.

Next, NOVOPlasty is called by the module. The outputs of NOVOPlasty are first checked to verify the completeness of the assembly and then are organized to avoid conflicts between different assembly processes.

To assemble samples, the module could read input information from a comma separated values file to handle all assemblies automatically. The three-column CSV file should contain the file names of “reads 1” data, “reads 2” data (optional, only for pair-end sequencing) and the taxon information of samples. Owing to the limited memory resources of personal computers, the program is executed sequentially to complete batch assembly, although users can also achieve parallelism by opening multiple programs if the hardware resources are abundant.

The Validation module

When the Assembly module finishes running, the Validation module is internally called. In addition, users could invoke this module manually by providing the sequence to be validated and a local reference file or taxon information.

The Validation module uses "Rotate" algorithm to adjust the structure of target and reference sequences and then conducts the collinearity analysis (Figure 1). It takes several steps:

1. Generate a full-length plastid genome. Extend the target sequence by adding a copy to its end. Independently from where the sequence starts, there will be one complete genome started with the beginning of the LSC region instead of a truncated region.
2. Perform self-to-self alignment. Call blastn of BLAST 2.9.0 (Madden*et al.* 2019) to perform a self-to-self pairwise alignment and default options of blastn are used.
3. Locate IR regions. Analyze the BLAST output. Find the longest match that has at least three copies (four if the sequence does not start with one truncated IR region), which should be IR.
4. Determine the other regions. According to the boundary of the IRs, locate the LSC and SSC regions. Extract a complete plastid genome with the order of LSC-IR-SSC-IR ensuring that the starting site of the sequence is on the 5' or 3' (if the LSC region is reverse-complemented) terminal of the LSC region.
5. Adjust the reference. Repeat steps 1 to 4 on the reference sequence.
6. Align target and reference sequences. Use BLAST to perform the pairwise alignment for adjusted target and reference sequences. While the sequence identity threshold of alignment is between 0 and 100%, in this step, the default low threshold (0.7) allows BLAST to ignore the mismatch in the middle of alignment and continue to extend the alignment. Hence, the alignment process prefers to focus on the structure similarity instead of the sequence one.
7. Analyze the alignment result. Because the target and reference sequences have the same starting site (steps 4 and 5) and the boundary of each region is known, regions of two sequences can be easily compared. This may have three possible results. If all the sequences match but they are in different directions, the direction of the whole sequence needs to be altered. If the direction of LSC or SSC is inconsistent with the reference, the orientation of the LSC or SSC needs to be adjusted. Besides the two cases mentioned above, the process treats it as a problematic assembly due to the conservatism of the plastid genome.
8. Output validated sequences with a standardized structure. After adjusting the starting site, direction of the strand, and orientation of four major regions, the verified plastid genome with a standardized structure is generated. In addition, the results of collinearity analysis and unadjusted sequences are also available in case of need.
9. Since both the plus and minus strands are considered when finding the repeated regions, apart from IR regions, theoretically, the program could also recognize direct repeats (DR). For those species that do not have a quadripartite structure, the Validation module may not work as normal, such as species of *Erodium* (Blazier *et al.*2011) and some parasitic plants (Bellot & Renner 2016).

If the assembly failed to pass the validation, the Assembly module was recalled to use another seed for assembly. When at least one validated assembly is found by the Validation module or all seeds are tried, the program stops. If no validated assembly is generated after trying all seeds, the Merge module tries to build a complete plastid genome based on contigs created using different seeds.

The Merge module

The Merge module implemented an edited overlap-Layout-consensus algorithm (Li *et al.* 2012) to merge contigs according to the overlaps between them. The procedures are as follows:

1. Make a copy. Since the input contigs could be in sense or antisense strands, add a reverse-complemented copy of the original input to ensure that at least one copy of contigs are all in sense strands.
2. Self-to-self alignment. Call blastn of BLAST 2.9.0 to do self-to-self alignment with default options.

3. Analyze the BLAST output. Only the forward overlapped match is kept, that is, the query and subject sequence are both in the sense strand, and the match is between the beginning of the downstream sequence and the ending of the upstream sequence. In addition, nested overlap is omitted to remove short redundant contigs.
4. Generate a unidirectional graph. According to the overlap information above, a unidirectional graph is generated. The nodes of the graph represent the contigs and the directed edges represent overlaps between contigs. Since there are two copies of contigs in the opposite pattern, ideally, there might be two major circles in the graph.
5. Cut edges. The transitively inferable edges, non-branching stretches and alternative paths that go through the same node are removed. The edges across the two circles are also removed. All removed edges represent incorrect overlapping relationships, mainly due to repeated sequences, especially IR regions in plastid genomes.
6. Extract the full circle found in the graph. According to the overlap information, the program merges contigs to generate circular sequences.

If the input contigs contain enough head-to-tail overlapping sequences, a whole plastid genome is likely to be formed and the program may generate two circular sequences with opposite orientation. Finally, the Validation module is called to test the output.

Results

Benchmark with sequencing data

Ten species were used to benchmark the usability of the software and the applicability of the seeds. Species representing gymnosperms, basal angiosperms, monocots, and eudicots were sequenced. *Pycnostachys reticulate* and *Nepeta stewartiana* are from newly sequenced genus and another four are newly sequenced species. Additionally, in order to test the ability of the program to recognize DR structure and to expand its application to wider taxonomy, *Selaginella sanguinolenta* , a species of lycopods, was also used.

Purified DNA was acquired from the Plant DNA Bank of China. Sequencing was carried out at the Huada Genomics Institute (BGI, Shenzhen, China). Approximately 50 GB of PE-100 sequencing data per sample were collected. In order to reduce the memory requirement, only 10,000,000 reads of each sample were used for assembly. NOVOWrap v0.95 was run on a PC with Microsoft Windows 10 operating system and 32 GB memory to assemble all data using default options. The reference genomes of each sample were automatically chosen by the program. The output genomes were crossverified based on the results of the collinearity analysis. For further verification, the assemblies of four species, *Cycas debaoensis* , *Brasenia schreberi* , *Oryza alta* and*Firmiana simplex,* were aligned to existing plastid genomes deposited in GenBank using MAFFT v7.409 (Katoh & Standley 2013).

The assembled sequences were then annotated using Plann (Huang & Cronk 2015) and GeSeq (Tillich et al. 2017). The start and stop codon of genes were manually corrected with Unipro UGENE (Okonechnikov *et al.*2012) if necessary. The annotation results were finally visualized with OGDRAW (Greiner *et al.* 2019).

The assembled results are listed in Table 2. All samples were successfully assembled using NOVOWrap v0.95. The memory usage of the program stabilized at ˜12 GB and the time cost ranged between 6 and 20 minutes. For the candidate seeds, two samples failed with the frequently used *rbcL* but succeeded with other seeds. While *rbcL* , *psaB* and *psaC* have similar success rates, *rrn23* only succeeded on *Nepeta stewartiana* , one of the eleven samples. Each seed generated one or two identical results for each sample, with one's structure adjusted if necessary. The vast majority of the samples have identical assembly results with different seeds with the exception of *Pycnostachys reticulate* . In this case, 140 candidate results were generated, with lengths ranging between 152,466 and 152,520 bp (Supplementary material 2). The genome structures of all these candidates are relatively conserved compared to the reference, *Ocimum tenuiflorum* (species in same family) and have the same size of SSC (17,671 bp). Their LSC ranged between 83,397 and 83,423 bp and IRs ranged between 25,695 and 25,713 bp. The variations in size are due to the existence of simple tandem repeats, a difficulty that NGS alone could hardly solve sometimes.

The Validation module successfully recognized the quadripartite regions of the sequences, including the DR structure of the lycopod species. The starting site of each sequence was adjusted according to the respective references. Inverted LSC and SSC regions were detected and adjusted in the subsequent collinearity analysis. Most newly determined genomes displayed high collinearity to the references, except for a few gaps in the alignment. In addition, the IRs of each genome were similar in size to the references with no obvious expansion or contraction (Figure 2).

The annotation results further verified the reliability of the assembly results. No gene loss or pseudogenization occurred in most of the samples. The only exception is *Salvia campanulata* , which has a premature termination in one side of the SSC flanked by IRa causing truncated *ycf1* . Moreover, all the boundaries of the quadripartite regions determined by the Validation module are consistent with the results of OGDRAW (Supplementary material 3).

According to the alignment results, the four resequenced genomes are identical to their references, except for a few one-base insertions or deletions located in mono nucleotide repeats. This is likely due to sequencing errors of NGS platforms or intraspecific differences given that the re-sequenced samples and samples existing in GenBank are from the same species but different individuals. Annotations demonstrate that such mutations have no effect on coding sequences.

The Merge module is designed to merge contigs originating from each seed into one complete plastid genome. For the cases in this study, the Merge module is skipped because all samples are successfully assembled with one of the seeds. To test the Merge module, several artificially generated datasets were used (Supplementary material 4). The result showed that the Merge module generates a complete plastid genome in most of the datasets.

Usability

NOVOWrap v0.95 provides both a command-line interface and a graphical user interface. It is a cross-platform and easy to install. In addition, it offers portable packages for Linux, Microsoft Windows, and MacOS operating systems. For the installation, all dependent third-party tools could be automatically installed if necessary. To further improve user-friendliness, extensive tests have been conducted to improve the robustness of the software. Moreover, preventive measures have been set up for factors that may interrupt the running of the program and cause inconvenience for users, such as Internet accessibility, the validity of input data, and the permission to access the work path. The program requires very few inputs to run and uses universal file formats for both input and output. Finally, NOVOWrap supports remote access of input files on the same local area network for convenience.

NOVOWrap is designed to be run on a normal personal computer. The Validation and Merge modules could finish running in minutes, with negligible consumption of memory and running time. Although the Assembly module requires much more resources, based on the feedback from voluntary testers and our own tests, 16 GB of memory is sufficient to handle approximately 10,000,000 PE100 reads, 8,000,000 PE150 reads, or 3,000,000 PE200 reads, which is sufficient for obtaining complete plastid genomes in general cases.

Discussion

The *rbcL* gene, the most frequently used seed for plastid genome assembly, usually works efficiently, but may fail occasionally. Hence, three novel seeds were introduced in this study. The *psaB* and*psaC* functioned as well as *rbcL* . Although *rrn23*only succeeded in the assembly of *Nepeta stewartiana* , it could be valuable for some extreme cases. Compared with other newly developed seeds, *rrn23* is very conservative in 1,331 evaluated genera. The possible cause of the lower success rate may be the location difference, that *rrn23* is located in the repeated region, while others are distributed in single-copy regions. For assembly, starting with IRs might be harder than LSC/SSC, since the former has extra possible directions to extend the sequences.

In the benchmark, each sample was assembled with at least one seed. Although identical sequences derived from different seeds can be used as supporting evidence for validating the assembly, assembly with one seed is sufficient for practical use, given that more attempts require much more time. According to the above,

the order of using seeds is optimized to *rbcL* first, then *psaB* , *psaC,* and *rrn23* . If none of the seeds work, the whole reference plastid genome is used as a backup seed.

In early time, the plastid genome is considered to have two natural isomers with SSC having the opposite polarity (Palmer 1983). The author considered that the inversion is mediated by the IR structure. However, some recent studies have used inversion as a phylogenetic character which is thought to be divergent in different lineages. (Walker *et al.* 2015). Such misinterpretation has deepened, given the fact that GenBank does not have a mandatory standard for the direction of the SSC. In this study, both isoforms were found in the assembly results of multiple samples. In addition, such structural variations are more likely to be misassembled owing to the existence of IRs (Jin et al. 2019; Wick *et al.* 2015). NOVOWrap automatically adjusts the orientation of SSC and LSC as the default output. In the case of potential needs, unchanged sequences are also generated.

Recently, several studies on the structural variation of plastid genomes, including the abnormal structure of the genome, the expansion and shrinking of IRs, and the recombination of orders of genes and other variations have been conducted (Weng *et al.* 2017; Zhu *et al.* 2016). However, the inconsistent structure of existing data may hamper further studies. Not only the orientation of SSC/LSC, but also other features such as different starting sites, inconsistent order of quadripartite regions and opposite strands of the whole genome, could cause problems in downstream analysis, such as annotation, alignment, collinearity analysis and phylogenetic reconstruction. For instance, the unadjusted sequence of *Salvia campanulata* has been reported to have drastically shortened IRs, with a length of 3,845 bp. However, the adjusted sequence showed a normal length of IRs, with a length of 25,535 bp (Supplementary material 3).

Although manual adjustment may solve the problem, the lack of automatic tools and standards makes it laborious. The Validation module of NOVOWrap ensures the consistency of assembly results automatically, liberating users from tedious error-prone manual manipulation. It is expected to accelerate plastid genome assembly and ease data upload to public databases with all consistent genome structures.

Acknowledgements

Thanks Yixuan Huang for testing and packaging program in MacOS. We are grateful to numerous volunteers who have tested the program with different data and various operating systems on their own personal computers or workstations.

This work is funded by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA23080204 and XDA19050303).

Author Contributions

Ping Wu wrote the program and the manuscript. Hao Chen tested and optimized the code and wrote the user manual. Chao Xu collected the data. Jie Yang analyzed data of lycopods and Xianchun Zhang guided the analysis of DR structure. Shiliang Zhou designed the study and edited the final manuscript.

Data Accessibility

The assembly results have been uploaded to GenBank. The source code of the program and the user manual are available at*https://github.com/wpwupingwp/novowrap*.

References

Attwood TK, Blackford S, Brazas MD, Davies A*, et al.* (2019). A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics* , 20, 398-404.

Bankevich A, Nurk S, Antipov D, Gurevich AA*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology, 19* (5), 455-477. doi:10.1089/cmb.2012.0021

Bellot S, Renner SS. (2016). The plastomes of two species in the endoparasite genus pilostyles (apodanthaceae) each retain just five or six possibly functional genes. *Genome Biol Evol* , 8, 189-201.

Blazier CC, Guisinger MM, Jansen RK. (2011). Recent loss of plastid-encoded ndh genes within Erodium (Geraniaceae). *Plant Molecular Biology* , 76, 263-272.

Brassell SC, Eglinton G, Marlowe IT, Pflaumann U*, et al.* (1986). Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. *Nature* , 320, 129-133.

Dierckxsens N, Mardulyn P, Smits G. (2016). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* , 45.

Freudenthal JA, Pfaff S, Terhoeven N, Korte A*, et al.* (2019) The landscape of chloroplast genome assembly tools. *bioRxiv* . doi:https://doi.org/10.1101/665869

Green BR (2011) Chloroplast genomes of photosynthetic eukaryotes.*Plant Journal, 66* , 34-44. doi:10.1111/j.1365-313X.2011.04541.x

Greiner S, Lehwark P, Bock R (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research, 47* (W1), W59-W64. doi:10.1093/nar/gkz238

Huang DI, Cronk QC (2015) Plann: A command-line application for annotating plastome sequences. *Applications in Plant Sciences, 3* (8). doi:10.3732/apps.1500026

Jin J, Yu W, Yang J, Song Y*, et al.* (2019) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *bioRxiv* . doi:https://doi.org/10.1101/256479

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution, 30* (4), 772-780. doi:10.1093/molbev/mst010

Keeling PJ (2010) The endosymbiotic origin, diversification and fate of plastids. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 365* , 729-748. doi:10.1098/rstb.2009.0103

Li Z, Chen Y, Mu D, Yuan J*, et al.* (2012). Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics* , 11, 25-37.

Lim CE, Kim GB, Ryu SA, Yu HJ*, et al.* (2018). The complete chloroplast genome of Artemisia hallaisanensis Nakai (Asteraceae), an endemic medicinal herb in Korea. *Mitochondrial DNA Part B: Resources* , 3, 359-360.

Madden TL, Busby B, Ye J. (2019). Reply to the paper : Misunderstood parameters of NCBI BLAST impacts the correctness of bioinformatics workflows.1-2.

Nei M. (1987). *Molecular Evolutionary Genetics* : Columbia University Press.

Okonechnikov K, Golosova O, Fursov M, team U (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics, 28* (8), 1166-1167. doi:10.1093/bioinformatics/bts091

Palmer JD. (1983). Chloroplast DNA exists in two orientations.*Nature* , 301, 92-93.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. (2012). NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Research* , 40, 130-135.

Qu XJ, Moore MJ, Li DZ, Yi TS. (2019). PGA: A software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* , 15, 1-12.

Sayers EW, Cavanaugh M, Clark K, Ostell J*, et al.* (2019) GenBank.*Nucleic Acids Research, 47* (D1), D94-D99. doi:10.1093/nar/gky989

Shi L, Chen H, Jiang M, Wang L*, et al.* (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Research* , 47, W65-W73.

Smith DR (2015) Mutation rates in plastid genomes: They are lower than you might think. *Genome Biol Evol, 7* , 1227-1234. doi:10.1093/gbe/evv069

Spellerberg IF, Fedor PJ. (2003). A tribute to Claude-Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' Index. *Global Ecology and Biogeography* , 12, 177-179.

Sugiura M, Shinozaki K, Zaita N, Kusuda M*, et al.* (1986). Clone bank of the tobacco (Nicotiana tabacum) chloroplast genome as a set of overlapping restriction endonuclease fragments: mapping of eleven ribosomal protein genes. *Plant Science* , 44, 211-217.

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES*, et al.*(2017). GeSeq - Versatile and accurate annotation of organelle genomes.*Nucleic Acids Research* , 45, W6-W11.

Tonti-Filippini J, Nevill PG, Dixon K, Small I (2017) What can we do with 1000 plastid genomes? *Plant Journal, 90* , 808-818. doi:10.1111/tpj.13491

Twyford AD, Ness RW. (2017). Strategies for complete plastid genome sequencing. *Molecular Ecology Resources* , 17, 858-868.

Walker JF, Jansen RK, Zanis MJ, Emery NC. (2015). Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes.*American Journal of Botany* , 102, 1751-1752.

Weng ML, Ruhlman TA, Jansen RK. (2017). Expansion of inverted repeat does not decrease substitution rates in Pelargonium plastid genomes.*New Phytologist* , 214, 842-851.

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

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF*, et al.* (2011) The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Molecular Biology, 76* , 273-297. doi:10.1007/s11103-011-9762-4

Zhu A, Guo W, Gupta S, Fan W*, et al.* (2016). Evolutionary dynamics of the plastid inverted repeat: The effects of expansion, contraction, and loss on substitution rates. *New Phytologist* , 209, 1747-1756.

Figure legends

Figure 1. Schematic diagram of the Rotate algorithm. In order to avoid possible interference of truncated region at the starting site, the original sequence was firstly duplicated to form a full-length plastid genome in the middle of the sequence. With the self-to-self comparison with BLAST, the longest repeated region with the highest number should be the IRs. Subsequently, according to the location of IRs, the boundaries of LSC and SSC were determined. Finally, full-length plastid genomes were extracted from the extended sequences and the collinearity analysis was performed.

Figure 2. Collinearity analysis of the assembly results. A) plastid genomes with DR structure; B) assembly has inverted LSC compared with the reference; C) normal comparison result; D) inverted SSC. Red strips represent the collinearity between the reference and the plus strand of the assembly, and green strips represent the minus strand. Blank strips indicate the mismatch of sequences. Crossed strips could be the matches between IRs in both strands, or indicate the presence of inverted region, such as B.

Figure

Figure 1. Schematic diagram of the Rotate algorithm

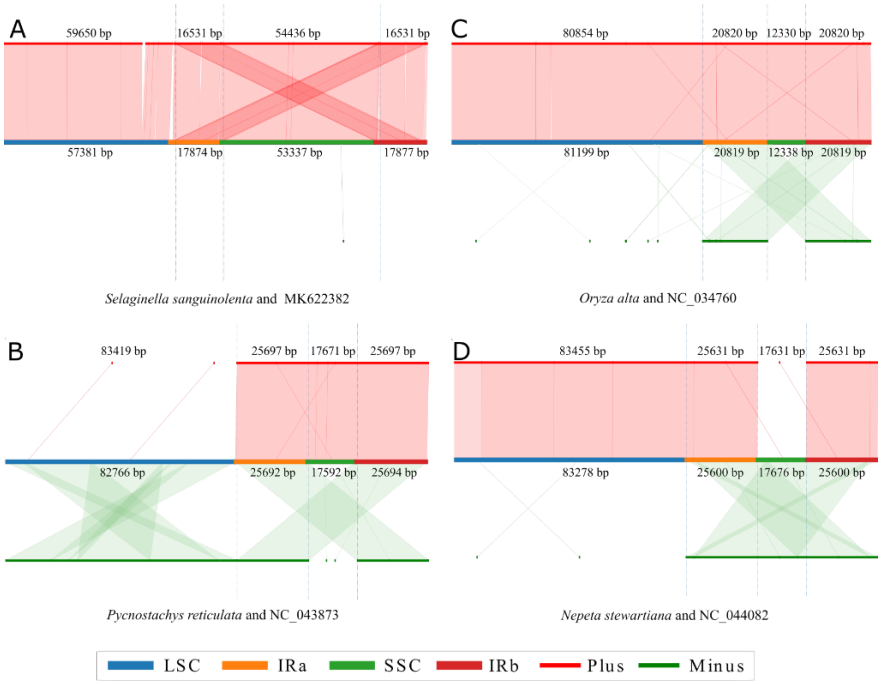

Figure 2. Collinearity analysis of the assembly results

Hosted file

Table 1. The sequence polymorphism of selected seeds.docx available at https://authorea.com/users/308401/articles/479879-novowrap-an-automated-solution-for-plastid-genome-assembly-and-structure-standardization

Hosted file

Table 2. Samples information and the assembly results.docx available at https://authorea.com/users/308401/articles/479879-novowrap-an-automated-solution-for-plastid-genome-assembly-and-structure-standardization