GetOrganelle: a simple and fast pipeline for de novo assembly of a complete circular chloroplast genome using genome skimming data

Jian-Jun Jian[2,4,5], Wen-Bin Yu[1,3,5 *], Jun-Bo Yang[2], Yu Song[1,3], Ting-Shuang Yi[2 *], De-Zhu Li[2*]

1. Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Mengla, Yunnan 666303, China

2. Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China

3. Southeast Asia Biodiversity Research Institute, Chinese Academy of Science, Yezin, Nay Pyi Taw 05282, Myanmar

4. Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, Yunnan 650201, China

5. These authors contributed equally to this work.

* Corresponding author: yuwenbin@xtbg.ac.cn, tingshuangyi@mail.kib.ac.cn, dzl@mail.kib.ac.cn

Abstract

Background: Chloroplast genes and genomes are the most important genomic data for plant phylogeny and DNA barcoding. Since the rapid development of high throughput sequencing technologies, it is cheap to get the low coverage data of whole genome, which is enough to assemble a complete chloroplast genome. To date, there are many assembly processes/pipelines described to assemble a complete chloroplast genome. In this study, we reported a simple and fast procedure to assemble a circular chloroplast genome using GetOrganelle pipeline.

Findings: The GetOrganelle pipeline consists of four steps: 1) recruiting plastid-like reads; 2) de novo assembly using SPAdes; 3) filtering plastid-like contigs; and 4) visualizing and editing de novo assembly graph. Of them, the first three steps can be fulfilled automatically just using a combined command; and the fourth step is to visualize and evaluate the assemblies. Of 57 tested species with public datasets, we directly reassembled the circular chloroplast genome in 47 species. The eight non-circular species having break points, which may be caused by mononucleotide or dinonucleotide repeats, or small reads pool. In addition, we successfully assembled the circular chloroplast genome for the other 903 species of angiosperms using this pipeline, representing 41 families and 358 genera.

Conclusion: The GetOrganelle pipeline is an effective way for land plants to assemble the circular chloroplast genome, without needs for reference-guided scaffolding, gap filling nor start-end point closing. This pipeline can be also applied to assemble mitochondrial genomes and nuclear Ribosomal DNAs using genome skimming data.

Keywords: Assembly; Bandage; Bowtie2; GetOrganelle; Organelle genome; Plastome; SPAdes.

Background

In green plants, organelles contain plastid and mitochondrion genomes. Generally, substitution rate of mitochondrion genes is lower than that of plastid genes, as well as nuclear genes. By contrary, plastid genome maintains a conserved circular and quadripartite structure, with a pair of invert repeat regions that separate large single copy (LSC) and small single copy (SSC) regions (Bock and Knoop 2012). Meanwhile, the plastid genome includes 100-120 unique genes with around 120 - 150 kb in size (Tonti-Filippini et al. 2017). So far, plastid DNA markers are widely used for plant phylogenies (Moore et al. 2010; Soltis et al. 2011; Refulio-Rodriguez and Olmstead 2014; Gitzendanner et al. 2018) and DNA barcoding (CBOL Plant Working Group 2009; China Plant BOL Group 2011; Hollingsworth et al. 2011). Since the rapid development of high throughput sequencing technologies, sequencing cost become cheaper and cheaper. Due to high number of copies of plastid genome in a single cell, it is easy to get enough reads to assemble a whole chloroplast genome from a low coverage of the whole genome sequencing data (Twyford and Ness 2016), or called genome skimming data (Straub et al. 2012).

To date, there are more 3000 chloroplast genomes summited to GenBank (accessed on 31 January, 2018). There are many assembly processes or pipelines described in the publication of the complete chloroplast genome (Figure 1, top). For example, SOAPdenovo2 (Luo et al. 2012) and CLC Genomics Workbench (https://www.qiagenbioinformatics.com/) were widely used to assemble the whole genome sequencing data, then the plastid scaffolds/contigs were filtered out using a reference genome for the further concatenation (Huang and Madan 1999) or post assembly gap filling and closing (Boetzer and Pirovano 2012; Paulino et al. 2015). The pipelines can get a linear plastome with or without gaps. The IOGA (Iterative Organellar Genome Assembly) pipeline (Bakker et al. 2016) incorporate Bowtie2 (Langmead and Salzberg 2012), SOAPdenovo2, SPAdes 3.0 (Bankevich et al. 2012) and other programs for recruiting plastid-like reads and de novo assembly. The plastid scaffolds/contigs need to be finalized by other programs. Moreover, the FASTG files can be visualized by Bandage (Wick et al. 2015), but the Bakker et al. (2016) did not recommend to do that yet. In addition, ORG.asm (Coissac 2017) and NOVOPlasty (Dierckxsens et al. 2017) were reported as fast ways to do de novo assembly for the plastid genome.

In this study, we recommended a simple and fast pipeline, GetOrganelle (https://github.com/Kinggerm/GetOrganelle), for de novo assembly of a complete circular chloroplast genome using the whole genome sequencing data (Figure 1, bottom). This pipeline exploits Bowtie2, BLAST (Camacho et al. 2009) and SPAdes 3.0 as dependencies. It starts with recruitment of initial plastid-like reads by taking any plastome or plastid fragments as references/seeds; the initial plastid-like reads will be treated as “baits” to get more plastid-like reads with multiple extension iterations, which is similar to the MITObim (Hahn et al. 2013) and IOGA (Bakker et al. 2016) pipelines. The core comparison algorithm for extension is hashing method, which cuts reads into substrings with certain length, called “words”, and adds them to a hash table, called “baits pool”. During each extension iteration, the “baits pool” is dynamically increasing as new plastid-like reads being added as “words”, and decreasing by dumping “words” that had been taken as “baits” for a complete iteration. Then, the total plastid-like reads are de novo assembled into a FASTA Graph ("fastg") file using SPAdes. Non-plastid contigs in the FASTG assembly are further automatically identified and trimmed by their connections, coverages and BLAST hit information using a built-in library. The slimmed fastg file can be visualized by Bandage to finalize the complete circular chloroplast genome. If the final plastome is circular, PCR verification for boundaries of four regions is not required.

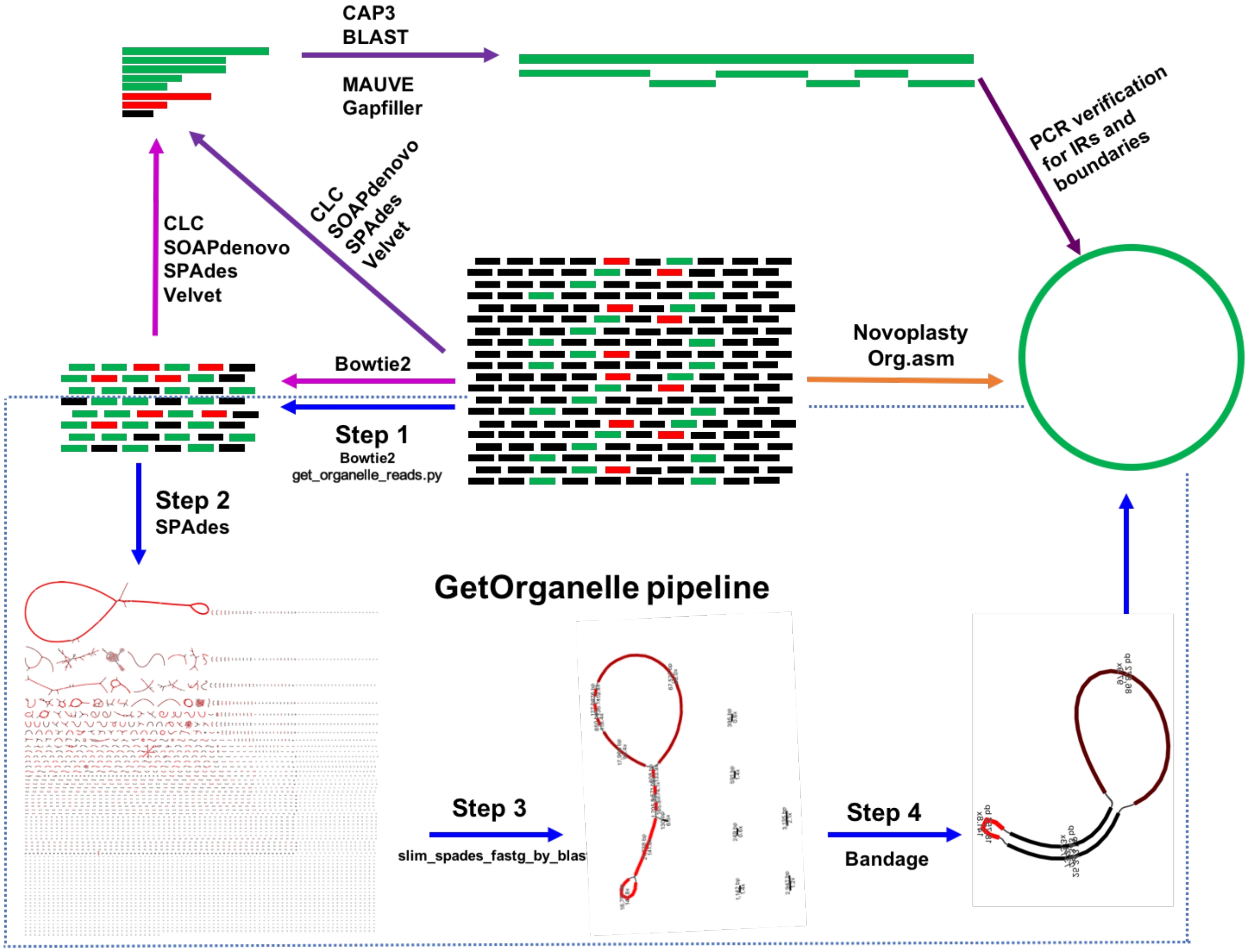

Figure 1. Summary of de novo assembly processes and pipelines for chloroplast genomes. The GetOrganelle pipeline is illustrated at the bottom in a blue dash box.

Methods

Data resources for testing

The data resources were included for assembly tests from the database of Sequence Reads Archive (SRA) (Table 1). More raw data generated by ourselves or collaborators were also tested (Table S1). The examined taxa covered all classes of land plants, with emphasis on the flowering plants. After searching for the published chloroplast genomes, we downloaded the raw reads of 50 species of angiosperms (representing eight major clades, 19 orders and 27 families), of five species of gymnosperms (representing three major clades, five orders and families), and of two species of non-seed plants (one moss and one fern) from the SRA database (Table 1) to reassemble using the GetOrganelle pipeline. We have confirmed that the raw reads of one gymnosperm and of 40 angiosperms from the SRA database corresponded to the published chloroplast genome (Guo et al. 2016; Roquet et al. 2016; Ivanova et al. 2017; Zhang et al. 2017). These data provided a great opportunity to compare the differences between the published plastomes and the newly reassembled plastomes using the GetOrganelle pipeline.

Working pipeline

Prepare the data

The data resources included both paired-end and single-end short reads. The read length varied from 75bp to 300bp (Table 1). We used the reads less than 5,000,00 for each end of paired-end data, or 10,000,000 for single-end data. For paired-end data, if the data was more than 5,000,000 reads of each end, we used "head" command to select the first 5,000,000 reads. For the single-end data, we selected the first 10,000,000 reads if needed. We also tested 1,000,000 reads or more of each end for some samples if a circular plastome was failed to be assembled. A large raw data may consume more time for recruiting the plastid-like reads.

Recruiting plastid-like reads by a Python script

Plastid-like reads were recruited using a new Python script (i.e. get_organelle_reads.py), which was written in version 3.5, and it was compatible with version 2.7. The initial step is using Bowtie2 to map reads to a reference, which can be any plastid genome or contigs in fasta format or Bowtie2 index format. The hitting reads will be treated as "baits" to recruit more plastid-like reads with multiple extension iterations. The feasibility and efficiency of the recruiting process depended on the value of word size, which was similar to the k-mer value in assembly. The best word size changes from data to data and will be affected by read length, read quality, base coverage, organelle DNA content, and other factors. The empirically recommended word size for recruiting was 80%-90% of the average read lengths for the paired-end reads. For the single-end reads, an empirically optional size was about 70% of the average read lengths. To reduce time consuming, number of iterations can be specified. For low memory machines, extracted reads per round can be outputted separately. Additional options can be chosen with special requirements. The raw reads do not need to be trimmed, because SPAdes will do error correction and reduce mismatches during the de novo assembly.

De novo assembly using SPAdes

The recruited plastid-like reads are automatically assembled using SPAdes (Bankevich et al. 2012), if "spades.py" has been added into the environment variables of the machine or server. Both paired and unpaired reads were used in this pipeline. The recommended k-mer value for assembling the chloroplast genome was 75% to 93% of the read length, or less than 127 for some long reads. For example, the 75 bp reads used a combined k-mer, 55, 65 and 71; the 100 bp reads used a combined k-mer, 75, 85, and 93; and the 150 bp or longer reads used a combined k-mer, 85, 95, 105, or even larger when there are longer repeats and coverage is sufficient.

Filtering plastid-like contigs by script

The scaffolding is a default option of SPAdes, so the scaffolds are automatically stored in a "fasta" file. Meanwhile, the SPAdes create an assembly graph, the "fastg" file, which displays the connections of contigs/scaffolds as graph with some allelic polymorphism and assembly uncertainty. The script "slim_fastg_by_blast", which would be also automatically called by main script, can use BLAST and a built-in library to search the plastid-like contigs/scaffolds by contig connections, contig coverages and BLAST hitting table. So the original "fastg" file can be simplified by deleting nuclear and mitochondrial contigs/scaffolds. The annotation of the retained contigs/scaffolds is stored in a concomitant cognominal "csv" file.

Visualizing and editing de novo assembly graph by Bandage

The filtered "fastg" file can be visualized by Bandage (Wick et al. 2015). Meanwhile, the concomitant annotation "csv" file can be imported into the graph, which helps identify plastid-like contigs/scaffolds. The read depth, which correlates to base coverage, is one important criterion to identify the repeats (e.g. inverted repeat, short repeats). For quadripartite plastome, the LSC forms a big circle, and the SSC forms a small circle, then, the two circles were connected by the linear IR region (with roughly doubled read depth) (Figure 2A). Some superfluous branches or connections might be nuclear or mitochondrial fragments (with high similarity of transfer RNA), or assembled from low quality reads of plastid. These fragments need to be removed. For some short repeats, they could be loosened following the process presented in a video (https://youtu.be/cXUV7k-F26w).

Results

Assembling a circular chloroplast genome

Of 57 species listed in in Table 1, we directly reassembled a complete circular chloroplast genome (Figure 2A) from 47 species, including one fern, three gymnosperms and 43 angiosperms. The complete chloroplast genome sequence is free of gap filling, start-end point closing and PCR verification of the LSC-IR-SSC conjunctions. In addition, four species had one break point in the LSC region (Figure 2B), and four species had two break points in LSC/SSC (Figure 2C). Although a whole sequence of the chloroplast genome was assembled in the species with one break point, the break point needs to be closed on the basis of an overlapped region in the both sides by removing some superfluous bases at both ends if needed. For the species with two break points, the two scaffolds were concatenated and circled using the overlapped regions. For the two cases, gap filling may close the break points, and PCR verifications are needed. We were failed to export a whole chloroplast genome sequence of *Picea abies* using Bandage, because five intricate (short inverted) repeats (Figure 2D) together can create 15 possible circular plastomes. Moreover, we successfully assembled a complete circular chloroplast genome for the other 903 species of angiosperms (Table S1: 9 major clades, 22 orders, 41 families and 358 genera) using this pipeline.

Comparisons between the reassembled and published plastomes

Comparative analyses showed that the identity between the reassembled and published plastomes were more than 99.25% in 55 species, except *Dendrobium nobile* was 92.267, and *Phelipanche aegyptiaca* was 97.512%. For 40 species using the identical raw data, the reassembled plastomes of 11 species were identical to the published ones, and of 11 species were fewer than 30 site-differences (Table 1). In addition, the reassembled plastomes of *Amborella trichopoda* and *Nicotiana tabacum* were the same to the published ones. The site-differences in *Laurus nobilis* occurred in the boundary between LSC-IRb and SSC-IRb. *Dendrobium nobile* having the highest differences between the reassembled and published plastomes indicated that one of the sample should be misidentified.

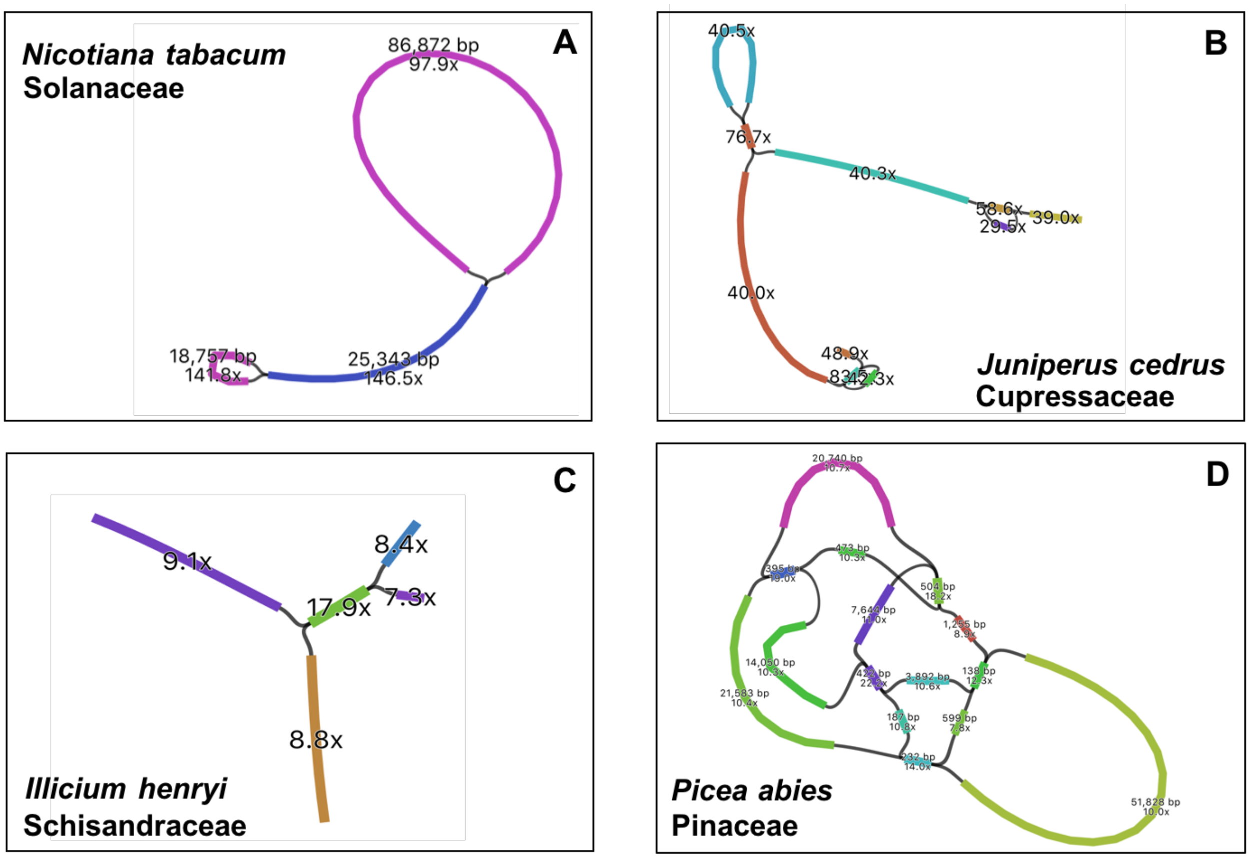

Figure 2. Four types of assembly graphs visualized using Bandage: a complete circular chloroplast genome (A), LSC having one break point (B), both LSC and SSC having one break point (C), and multiple short inverted repeats hard to loosen (D).

Discussion

The GetOrganelle pipeline comprise of four key steps (Figure 1, bottom). Of them the first three steps can be fulfilled automatically just using a combined command; and the fourth step is to visualize and evaluate the assemblies. The filtering step is the most crucial step, also the most time-consuming steps. Like the k-mer, a proper word-size ("-w") may significantly increase both the time and memory efficiency of this step and downstream steps, also ensure the completeness and simplicity of the final assembly graph. The k-mer values of SPAdes need to be adjusted if there are breaks or too many redundant contigs. Some breaks may be caused by mononucleotide or dinucleotide repeats, in some case increasing the k-mer value may avoid the break if there is enough read coverage of chloroplast genome.

According to our tested species listed in Table 1 and Table S1, we have demonstrated that the GetOrganelle pipeline is an effective way for so many land plants to assemble a complete circular chloroplast genome, without needs for gap filling and start-end point closing. Noteworthily, the boundaries between the SC and IR regions can be easily identified from the complete circular chloroplast genome sequence. The reassembled plastomes have less than 0.5% of differences from the published ones if using the same raw data except *Juniperus cedrus*. The most of differences between reassembled and published plastomes are insertions, deletions or some repeats. In the GetOrganlle pipeline, the copies of those repeats can be better identified due to the visualization of assembly graph with read depth information.

It is a challenge to assemble some "weird" chloroplast genomes (e.g., plastome reduction, gene translocations, and IR expansion, or contraction, or loss) using some traditional methodologies. For example, non-photosynthetic plants have reduced plastomes and many gene translocations (Graham et al. 2017; Wicke and Naumann 2018). If *de novo* assembly cannot yield a complete sequence of the plastome, gap filling and PCR verification are needed to close the gap of the contigs/scaffolds (Logacheva et al. 2016; Wicke and Naumann 2018). In our tested species, we successfully assembled a complete circular chloroplast genome for many holoparasites from Balanophoraceae, Cytinaceae, *Cuscuta* spp. (Convolvulaceae), Lennoaceae, and Orobanchaceae, as well as hundreds of hemiparasites (Table 1, S1). For the IR lacking species, like some Fabaceae species, a single circle of the assembly can be displayed by Bandage, which provides the evidence for exempting the PCR verifications.

Of the non-circular species in Table 1, we found that the break points of eight species had mononucleotide or dinonucleotide repeats, and raw data of five species were less than 7,000,000 reads in total. The eight species were failed to assemble a circle plastome which may be caused by single reason or by the combined reasons. The mononucleotide or dinonucleotide repeats causing breaks in our other tested samples, whereas we finally assembled a circular plastome when using more number of reads. In the case of *Picea abies*, we have to acknowledge that a circular plastome cannot be exported by Bandage if the assembled species had multiple (short) inverted repeats occurring in multiple regions. In the published plastomes, so far, we found that plastomes of Ericaceae (Fajardo et al. 2013; Logacheva et al. 2016), Geraniaceae (Weng et al. 2014), and Pinaceae

(Tsumura et al. 2000; Sullivan et al. 2017) had multiple short inverted repeats. In gymnosperms, some studies have revealed that a paired short invert repeats created isomers (Tsumura et al. 2000; Guo et al. 2014; Qu et al. 2017). This phenomenon was also reported in *Tylosema* spp. (Fabaceae) of angiosperms (Wang et al. 2018). However, it is unclear whether the short invert repeats may cause isomeric polymorphism in Ericaceae, Geraniaceae and other groups. To date, there is no effective approaches to assemble this type of plastome using short insert size library, and the concatenated plastome sequence may use relative species as reference or verify using PCR amplification and/or sequencing. As the development of long insert size library or long-read sequencing (e.g. PacBio), in the future, these data can be added for gap closure and repeat resolution when using SPAdes as an assembler to attain a complete sequence of chloroplast genome.

Perspectives of the pipeline

For most of the tested plants, the assembly graph of the plastome is as simple as Figure 2A. A new function that could automatically loosen the graph and export two candidate forms of circular sequences is expected under this situation. Following the procedure of assembling the chloroplast genome, the GetOrganelle pipepine can effectively assemble mitochondrial genome and nuclear Ribosomal DNA (rDNA). Only three changes are: 1) adjusting parameters, basically the word size and k-mer to accommodate the ratio of the coverage of your target sub-genome to the average coverage of total genome (the higher the ratio is, the larger word size and k-mer value should be); 2) using mitochondrial genome or fragments, and rDNA as the reference for recruiting, repectively; and 3) filtering mitochondrion-like and rDNA-like contigs/scaffolds using the built-in mitochondrion and rDNA libraries, respectively.

Availability of supporting source code and requirements

Project name: GetOrganelle pipeline
Project home page: https://github.com/Kinggerm/GetOrganelle
Operating system(s): Linux or Mac OS
Programming language: Python 3.5.1
Other requirements: SPAdes, bowtie2, BLAST
License: Apache License 2.0

Author contributions

W-BY, J-JJ, T-SY, D-ZL conceived the research; J-BY generate the data; J-JJ wrote the script; W-BY, J-JJ, YS analyzed the data; W-BY, J-JJ wrote the draft manuscript; W-BY, J-JJ, J-BY, YS, T-SY and D-ZL revised and approved the manuscript.

References

Bakker, F. T., D. Lei, J. Yu, S. Mohammadin, Z. Wei, S. Kerke, B. Gravendeel, M. Nieuwenhuis, M. Staats, D. E. Alquezar-Planas and R. Holmer. 2016. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society* 117: 33-43.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev and P. A. Pevzner. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455-477.

Bock, R. and V. Knoop. 2012. *Genomics of chloroplasts and mitochondria*. Berlin: Springer.

Boetzer, M. and W. Pirovano. 2012. Toward almost closed genomes with GapFiller. *Genome Biology* 13: R56.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. L. Madden. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* 106: 12794-12797.

China Plant BOL Group. 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the United States of America* 108: 19641-19646.

Coissac, E. 2017. Org.Asm: The ORGanelle ASeMbler. In.

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

Fajardo, D., D. Senalik, M. Ames, H. Zhu, S. A. Steffan, R. Harbut, J. Polashock, N. Vorsa, E. Gillespie, K. Kron and J. E. Zalapa. 2013. Complete plastid genome sequence of Vaccinium macrocarpon: structure, gene content, and rearrangements revealed by next generation sequencing. *Tree Genetics & Genomes* 9: 489-498.

Gitzendanner, M. A., P. S. Soltis, T.-S. Yi, D.-Z. Li and D. E. Soltis. 2018. Chapter Ten - Plastome Phylogenetics: 30 Years of Inferences Into Plant Evolution. Pp. 293-313 in *Advances in Botanical Research*, eds. C. Shu-Miaw and Robert K. J.: Academic Press.

Graham, S. W., V. K. Y. Lam and V. S. F. T. Merckx. 2017. Plastomes on the edge: the evolutionary breakdown of mycoheterotroph plastid genomes. *New Phytologist*: n/a-n/a.

Guo, Q., D. Bianba and W. Zheng. 2016. Characterization of the complete chloroplast genome of *Juniperus cedrus* (Cupressaceae). *Mitochondrial DNA Part A* 27: 4355-4356.

Guo, W., F. Grewe, A. Cobo-Clark, W. Fan, Z. Duan, R. P. Adams, A. E. Schwarzbach and J. P. Mower. 2014. Predominant and Substoichiometric Isomers of the Plastid Genome Coexist within Juniperus Plants and

Have Shifted Multiple Times during Cupressophyte Evolution. *Genome Biology and Evolution* 6: 580-590.

Hahn, C., L. Bachmann and B. Chevreux. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research* 41: e129-e129.

Hollingsworth, P. M., S. W. Graham and D. P. Little. 2011. Choosing and using a plant DNA barcode. *PLOS ONE* 6: e19254.

Huang, X. and A. Madan. 1999. CAP3: A DNA sequence assembly program. *Genome Research* 9: 868-877.

Ivanova, Z., G. Sablok, E. Daskalova, G. Zahmanova, E. Apostolova, G. Yahubyan and V. Baev. 2017. Chloroplast genome analysis of resurrection tertiary relict *Haberlea rhodopensis* highlights genes important for desiccation stress response. *Frontiers in Plant Science* 8: 204.

Langmead, B. and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Meth* 9: 357-359.

Logacheva, M. D., M. I. Schelkunov, V. Y. Shtratnikova, M. V. Matveeva and A. A. Penin. 2016. Comparative analysis of plastid genomes of non-photosynthetic Ericaceae and their photosynthetic relatives. *Scientific Reports* 6: 30042.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam and J. Wang. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1: 1-6.

Moore, M. J., P. S. Soltis, C. D. Bell, J. G. Burleigh and D. E. Soltis. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences* 107: 4623-4628.

Paulino, D., R. L. Warren, B. P. Vandervalk, A. Raymond, S. D. Jackman and I. Birol. 2015. Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics* 16: 230.

Qu, X.-J., C.-S. Wu, S.-M. Chaw and T.-S. Yi. 2017. Insights into the existence of isomeric plastomes in Cupressoideae (Cupressaceae). *Genome Biology and Evolution* 9: 1110-1119.

Refulio-Rodriguez, N. F. and R. G. Olmstead. 2014. Phylogeny of Lamiidae. *American Journal of Botany* 101: 287-299.

Roquet, C., É. Coissac, C. Cruaud, M. Boleda, F. Boyer, A. Alberti, L. Gielly, P. Taberlet, W. Thuiller, J. Van Es and S. Lavergne. 2016. Understanding the evolution of holoparasitic plants: the complete plastid genome of the holoparasite *Cytinus hypocistis* (Cytinaceae). *Annals of Botany* 118: 885-896.

Soltis, D. E., S. A. Smith, N. Cellinese, K. J. Wurdack, D. C. Tank, S. F. Brockington, N. F. Refulio-Rodriguez, J. B. Walker, M. J. Moore, B. S. Carlsward, C. D. Bell, M. Latvis, S. Crawley, C. Black, D. Diouf, Z. Xi, C. A. Rushworth, M. A. Gitzendanner, K. J. Sytsma, Y.-L. Qiu, K. W. Hilu, C. C. Davis, M. J. Sanderson, R. S. Beaman, R. G. Olmstead, W. S. Judd, M. J. Donoghue and P. S. Soltis. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704-730.

Straub, S. C. K., M. Parks, K. Weitemier, M. Fishbein, R. C. Cronn and A. Liston. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349-364.

Sullivan, A. R., B. Schiffthaler, S. L. Thompson, N. R. Street and X.-R. Wang. 2017. Interspecific plastome recombination reflects ancient reticulate evolution in *Picea* (Pinaceae). *Molecular Biology and Evolution*.

Tonti-Filippini, J., P. G. Nevill, K. Dixon and I. Small. 2017. What can we do with 1000 plastid genomes? *Plant J* 90: 808-818.

Tsumura, Y., Y. Suyama and K. Yoshimura. 2000. Chloroplast DNA Inversion Polymorphism in Populations of Abies and Tsuga. *Molecular Biology and Evolution* 17: 1302-1312.

Twyford, A. D. and R. W. Ness. 2016. Strategies for complete plastid genome sequencing. *Molecular Ecology Resources*: n/a-n/a.

Wang, Y.-H., S. Wicke, H. Wang, J.-J. Jin, S.-Y. Chen, S.-D. Zhang, D.-Z. Li and T.-S. Yi. 2018. Plastid genome evolution in the early-diverging legume subfamily Cercidoideae (Fabaceae). *Frontiers in Plant Science* 9.

Weng, M.-L., J. C. Blazier, M. Govindu and R. K. Jansen. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Molecular Biology and Evolution* 31: 645-659.

Wick, R. R., M. B. Schultz, J. Zobel and K. E. Holt. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31: 3350-3352.

Wicke, S. and J. Naumann. 2018. Chapter Eleven - Molecular evolution of plastid genomes in parasitic flowering plants. Pp. 315-347 in *Advances in Botanical Research*, eds. S.-M. Chaw and Jansen R. K.: Academic Press.

Zhang, N., P. Ramachandran, J. Wen, J. A. Duke, H. Metzman, W. McLaughlin, A. R. Ottesen, R. E. Timme and S. M. Handy. 2017. Development of a reference standard library of chloroplast genome sequences, GenomeTrakrCP. *Planta medica* 83: 1420-1430.