Open access • Posted Content • DOI:10.1101/697821

PACVr: Plastome Assembly Coverage Visualization in R — Source link

Michael Gruenstaeudl, Nils Jenke

Institutions: Free University of Berlin

Published on: 11 Jul 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Genome, Sequence assembly, Genomics and Population

BMC Bioinformatics

SOFTWARE Open Access

PACVr: plastome assembly coverage visualization in R

Michael Gruenstaeudl[1*] and Nils Jenke[2]

*Correspondence:
m.gruenstaeudl@fu-berlin.de
[1]Institut für Biologie, Systematische Botanik und Pflanzengeographie, Freie Universität Berlin, 14195 Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Background: Plastid genomes typically display a circular, quadripartite structure with two inverted repeat regions, which challenges automatic assembly procedures. The correct assembly of plastid genomes is a prerequisite for the validity of subsequent analyses on genome structure and evolution. The average coverage depth of a genome assembly is often used as an indicator of assembly quality. Visualizing coverage depth across a draft genome is a critical step, which allows users to inspect the quality of the assembly and, where applicable, identify regions of reduced assembly confidence. Despite the interplay between genome structure and assembly quality, no contemporary, user-friendly software tool can visualize the coverage depth of a plastid genome assembly while taking its quadripartite genome structure into account. A software tool is needed that fills this void.

Results: We introduce 'PACVr', an R package that visualizes the coverage depth of a plastid genome assembly in relation to the circular, quadripartite structure of the genome as well as the individual plastome genes. By using a variable window approach, the tool allows visualizations on different calculation scales. It also confirms sequence equality of, as well as visualizes gene synteny between, the inverted repeat regions of the input genome. As a tool for plastid genomics, PACVr provides the functionality to identify regions of coverage depth above or below user-defined threshold values and helps to identify non-identical IR regions. To allow easy integration into bioinformatic workflows, PACVr can be invoked from a Unix shell, facilitating its use in automated quality control. We illustrate the application of PACVr on four empirical datasets and compare visualizations generated by PACVr with those of alternative software tools.

Conclusions: PACVr provides a user-friendly tool to visualize (a) the coverage depth of a plastid genome assembly on a circular, quadripartite plastome map and in relation to individual plastome genes, and (b) gene synteny across the inverted repeat regions. It contributes to optimizing plastid genome assemblies and increasing the reliability of publicly available plastome sequences. The software, example datasets, technical

(Continued on next page)

(Continued from previous page)
documentation, and a tutorial are available with the package at https://cran.r-project.org/package=PACVr.

Keywords: software, R package, genome assembly, plastid genome, sequencing coverage, visualization

Background

The sequencing and comparison of complete plastid genomes has become a popular method in plant evolutionary research, rendering the precise genome assembly and its quality assessment of high importance. The plastid genomes of most photosynthetically active land plants display a circular, quadripartite structure and comprise two single copy (SC) regions separated by two identical inverted repeats (IR) [1]. A total of four partitions with markedly different lengths can, thus, be defined in typical land plant plastomes: the large single copy (LSC) region of ca. 70-90 kilobases (kb), the small single copy (SSC) region of ca. 15-25 kb, and the two IR regions (IRa and IRb) of ca. 20-25 kb each [2]. The IR regions represent reverse complements of each other and are primarily homogenized through a recombination-mediated replication process [3, 4]. The plastid genomes of most photosynthetically active land plants encode a total of ca. 100-120 proteins, which play a central role in organelle metabolism and photosynthesis [5]. Due to their strong structural conservation, uniparental inheritance, a near absence of recombination, and a high copy number per plant cell, plastid genomes are highly suitable for comparative genomic studies [6]. Numerous investigations have sequenced and compared complete plastid genome sequences over the past decade [7, 8], and the number of publicly available plastid genomes continues to increase dramatically [9]. Recent studies on plastid genome structure and evolution have evaluated polymorphisms across hundreds [10–13] or even thousands [14, 15] of plastid genome sequences, rendering the precise assembly process of plastid genomes and their quality assessment ever more important.

Despite the development of assembly algorithms customized for plastid genomes, the plastome assembly process remains imperfect and often requires the verification, if not manual correction, of the assembly product. Concurrent with the surge in plastid genome sequencing, many new algorithms and pipelines specifically designed for the assembly of plastid genomes have been developed [16–23]. Most of these tools allow a more accurate and targeted assembly of the plastid genome than generic assembly software, but in many cases some form of manual intervention or post-processing of the assembly results remains necessary [18, 21]. The post-processing of automated assembly results often pertains to the correction of the IR length [6, 9], differences in junction boundaries [24], and genome circularization [25]. Common uncertainties and outright errors in plastid genome assemblies include the inequality of the IR regions in length or sequence [26–28], long homopolymer runs [9, 29], and the imperfect duplication of repeats at the junction of the SC regions [24]. The differential orientation of the SSC, by contrast, does not constitute an assembly error but reflects the natural presence of heteroplasmy in organelle genomes [30, 31]. To ensure correctness and reproducibility in plastid genome sequencing and analysis, it is paramount to confirm the validity of plastid genome assemblies [9, 21, 24]. Many of the ambiguities and putative errors recognized in published plastid genome sequences, including those of genome annotations

[9, 27, 28], could potentially be averted by the application of simple quality assessment strategies [21, 29].

Several measures have been used to indicate the quality of plastid genome assemblies, including contiguity metrics and the length and sequence equality of the IR regions, but sequencing coverage remains one of the most popular proxies for assembly quality. In genome research, the length of the shortest among all those contigs that cover at least 50% of a reference genome is often used as an indicator for the quality of a draft genome [32]. The closer this length is to the complete length of the reference genome, the more confidence is placed in the completeness and, by extension, the quality of the assembly [33]. This concept is one of several contiguity metrics used to indicate the quality of a genome assembly (e.g., NG50, [32]; NA50 and NGA50, [34]). However, these contiguity metrics are difficult to apply to so far unsequenced organisms due to the requirement of a known reference sequence. Another, more specific measure for validating the quality of genome assemblies constitutes the degree of gene synteny across draft genomes or subsections thereof [35]. The IRs of a plastid genome, for example, represent recombinogenic isomers and, thus, share the same DNA sequence and gene synteny [3, 36]; exceptions to this rule are very rare [37]. Equality in length, sequence, and gene synteny of the IR regions can, thus, be used as a general indicator for the quality of the plastid genome assembly [27]. The depth of sequencing coverage ('coverage depth' hereafter) represents yet another indicator for the quality of a genome assembly [38]. Average coverage depth is defined as the average number of times each nucleotide of a genome region is represented by aligned reads from a sequence set [39]; it is a unit-less integer. Coverage depth is an important and highly popular indicator for the quality of a genome assembly in biological research [40, 41]. For plastid genomes, coverage depth is reported almost by default in relation to genome assemblies [6, 42] and has been implemented as a quality metric in several plastome assembly pipelines [18–20]. Information on coverage depth is critical for the assessment of large-scale sequence rearrangements or other structural variation of a genome because a greater coverage depth increases the chance that rearrangement endpoints are captured and confirmed by multiple independent reads [39, 43]. Information on coverage depth is also critical for the assembly process itself, as many de novo assembly algorithms operate under the implicit assumption of even coverage depth across the target genome [6, 44, 45]. In the present investigation, coverage depth, as well as gene synteny across the IR regions, are used as specific indicators for plastid genome assembly quality.

Currently available software tools can generate either unpartitioned plots of plastome coverage depth or quadripartite plastome maps, but the simultaneous, user-friendly visualization of both aspects is presently unsupported. When employing currently available software tools, plant biologists must decide if they wish to visualize either plastome sequencing coverage as unpartitioned, often linear plots or, alternatively, the circular, quadripartite structure of a plastid genome. The assembly pipeline FastPlast [19], for example, analyzes coverage depth during run-time and, upon genome assembly, generates a linear coverage plot as part of the pipeline execution (Fig. 1a). Similarly, the assembly pipeline IOGA [46] generates linear coverage plots during run-time, allowing users to evaluate the progress of the assembly process during different pipeline iterations (Fig. 1b). The assembly pipeline ORG.asm [20] also estimates average coverage depth during the assembly process but does not visualize this metric. None of

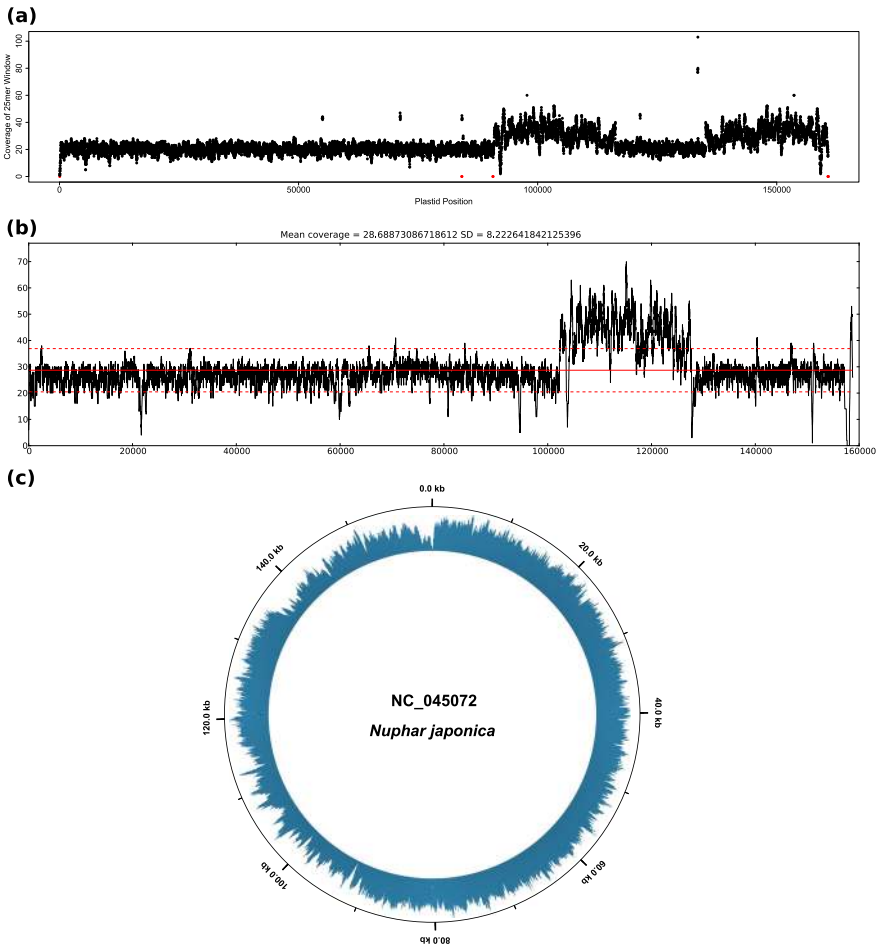

Fig. 1 Visualizations of coverage depth of the plastid genome of *Nuphar japonica* (GenBank accession NC_045072) as generated by the software tools (**a**) FastPlast, (**b**) IOGA, and (**c**) Circleator. For easier viewing, the size of the individual data points was reduced in the visualization of FastPlast, and the tick mark at 160 kb was removed in the visualization of Circleator

these assembly pipelines generate visualizations that account for the circular, quadripartite structure of the plastid genome or for the location of the individual plastome genes. On the other hand, several software tools and web-services exist that visualize complete plastid or bacterial genomes as circular maps. The web-service OrganellarGenomeDraw (OGDRAW; [47, 48]), for example, generates circular maps of plastid and mitochondrial genomes and visualizes gene position and GC content across the genomes. Similarly, the software Circleator [49] generates circular maps of bacterial genomes and can visualize gene position, GC content, and single nucleotide polymorphism locations in comparison to a reference genome. When co-supplied with text-based configuration instructions and a read mapping file, Circleator can also visualize coverage depth on the circular visualizations (Fig. 1c), but the configuration instructions are complex, and unless an intricate, multi-layered visualization procedure is applied, additional genome annotations such as genes are not displayed. The software Circos [50] can also be used to generate elaborate visualizations of circular genomes, including plastomes [51–53], but even more bioinformatics expertise is required to generate

the source code underlying these visualizations, which is typically beyond the ability of a normal user in plant biology. Several older software tools and web-services for generating circular genome maps also exist [54–58], but their application in recent research has been minimal, and some of these services have become inaccessible (e.g., [54, 57, 58]; inaccessible since at least October 2018). To the best of our knowledge, none of the presently available software tools can visualize the coverage depth of a plastid genome assembly on a circular, quadripartite plastome map, as well as gene synteny across the IR regions, while simultaneously displaying the locations of the plastome genes and the location and relative sizes of the SC and the IR regions, especially in a user-friendly fashion.

Given the plethora of complete plastid genomes generated in biological research each year [9], strong demand for a software tool exists that enables a visual quality assessment of plastid genome assemblies. Specifically, it would be desirable to have a tool that allows users to visually explore the coverage depth of a plastid genome assembly as well as the gene synteny across its IR regions [59], as both aspects are indicative for the quality of the genome assembly. To be useful to a wide audience, such a software tool must fulfill four criteria: it must (a) be user-friendly and applicable to users with minimal bioinformatics knowledge; (b) generate publication-ready visualizations that allow the determination if and where a genome assembly displays insufficient coverage depth; (c) allow an easy integration into automated workflows or analysis pipelines; and (d) allow users to set customized window sizes and thresholds for coverage depth calculation. Here, we present such a tool, titled 'PACVr' for 'Plastome Assembly Coverage Visualization in R'. PACVr is a package for the common statistical environment R [60] that visualizes (i) coverage depth of a plastid genome assembly on a circular, quadripartite plastome map, and (ii) gene synteny across the IR regions of the genome assembly. Specifically, PACVr visualizes coverage depth across the entire plastid genome in user-defined window sizes and in relation to the gene annotations, calculates and displays average coverage depth values for each of the four plastome regions, highlights sectors with coverage depth below a user-specified threshold, and visually connects the genes of one IR with their counterparts of the other IR using variable-width connector lines. By applying PACVr upon plastid genome assembly, users can visually inspect coverage depth and IR gene synteny across the input genome and, where applicable, identify regions of potentially reduced assembly confidence. Specifically, users can identify sectors of a plastid genome with low coverage depth or IRs with missing gene synteny and then subject these sectors to re-evaluation or post-processing. Upon presenting the details of the software, we illustrate the application of PACVr on four plastid genome assemblies from different plant lineages. Two of the assemblies represent newly sequenced plastid genomes with a plastome size typical for most angiosperms and a quadripartite genome structure; for these assemblies, we compare the visualizations of PACVr with the output of other software tools for visualizing plastome coverage depth. The other two assemblies represent previously published plastid genomes with plastome sizes that considerably deviate from the typical size range of plastid genomes; one of these assemblies also represents a plastid genome without quadripartite genome structure. Our application of PACVr on empirical data, thus, illustrates the flexibility of the software with regard to plastid genome size and structural configuration.

Implementation

Input and output specifications

The input to PACVr consists of two different files of common file format which contain information on (A) genome sequence and structure, and (B) coverage depth. Information on genome sequence and structure, as well as the genes encoded by the sequence, is supplied via an input file in the GenBank flatfile format. GenBank flatfiles represent the default file type for sequence retrievals from NCBI Nucleotide [61] and contain one or more sequence records, with each record comprising general metadata, an annotation table with the names and locations of genes and other sequence features, and the nucleotide sequence itself [62]. In PACVr, GenBank flatfiles are parsed via the R package genbankr [63] and must, thus, contain only a single sequence record per file, with the locus name no longer than ten alphanumeric characters. Moreover, genbankr requires the location of sequence features that span multiple positions or occur on complementary strands to be specified with the use of only a single invocation of the commands 'join' and 'complement' each, and all sequence features of class 'exon' to be removed. For optimal visualizations, the sequence record of the GenBank file should represent a complete, fully annotated plastid genome and contain feature annotations for each of the two IRs, if these repeats are naturally present in the genome; flatfile qualifiers for the IRs must hereby have the text values 'IRa' and 'IRb', or 'inverted repeat A' and 'inverted repeat B', respectively. For optimal visualizations, the sequence record should have a total sequence length between 50 kb and 250 kb. This preferred size range encompasses all plastid genomes of photoautotrophic land plants currently available on GenBank (sizes of the smallest and largest photoautotrophic embryophyte plastome on GenBank as of 01 January 2020: 59,190 bp [NC_014874] and 242,575 bp [NC_031206], respectively) and is a consequence of the practical limitations of scaling a circular, multi-layered plastome map to overall genome size. The scaling conducted by PACVr particularly aims to balance the visualization of the complete genome with sufficient spacing between adjacent plot layers and a font size large enough for text elements to be legible. Information on coverage depth is supplied via an input file in the binary alignment/map (BAM) format, which stores alignment and mapping information [64] and is typically generated by the mapping of sequence reads to a reference genome with short read alignment packages [65] (such as BWA [66] or Bowtie2 [67] in conjunction with the software samtools [64]). To be suitable for PACVr, the BAM file must also be indexed and, thus, accompanied by an ancillary index file. Generating the BAM file is done prior to, and independent of, the functionality of PACVr and can be conducted under a series of different settings, which may be reflected in the resulting visualization. For example, users may wish to visualize coverage depth calculated only from sequence reads that map to the reference genome as concordant read pairs, which can be beneficial in the identification of assembly errors [68]. Similarly, users may wish to visualize coverage depth calculated only from sequence reads that map to more than one location in the reference genome, which, if applied to plastid genomes, typically highlights the location of the IRs. This autonomy in generating BAM files provides users with considerable flexibility in the application of PACVr , especially as part of a bioinformatic workflow. Several additional input parameters can be specified upon initiation of PACVr, including the window size used for calculating coverage depth, the threshold below which coverage depth is highlighted, and the name of the output file, among other aspects, but these parameters are optional and have well-tested default values set for them.

The output of PACVr is a multi-layered, annotated plastome map with coverage information across the genome. Specifically, PACVr generates a circular, quadripartite map of the plastid genome in which coverage depth values are displayed as histogram bars, with bars below a predefined threshold highlighted in red and partition-wide average coverage values superimposed as horizontal, yellow lines. The map also displays positional information in regular intervals in the form of labeled tick marks as well as the location of all plastome genes, allowing the user to relate areas of low coverage depth to specific genome regions and genes. The map generated by PACVr is saved in PDF format to a user-defined output file.

Coverage calculation and display

Coverage depth is calculated by PACVr via the application of user-defined window sizes with the software mosdepth [69]. Window-based coverage calculations have the flexibility of measuring coverage on customized scales, which can be necessary to account for the variability in read length across different Illumina reagent types or sequencing cycle numbers. Using a sorted BAM file plus its ancillary index file as input, mosdepth rapidly infers the coverage of a particular chromosome by tracking all start and end positions of mapped sequence reads and calculating the cumulative sum of their incremented start positions while decrementing the respective end positions [69]. Based on the results of mosdepth, PACVr infers the average coverage depth for each of the four regions of the plastid genome (i.e., LSC, SSC, IRa and IRb). Two types of coverage depth information are plotted on the plastome map: (i) window-based depth values are displayed in the form of a circular histogram, with the width of each histogram bar equal to the width of the window size, and (ii) partition-wide coverage averages are displayed as horizontal, yellow lines superimposed on the histogram bars as well as numerically in the plastome map legend. PACVr is, thus, different from most other software tools for visualizing coverage depth, which typically display coverage depth as stacked sequence reads [70, 71], line graphs [46], dot graphs [19] or bedGraphs plots [72], and primarily on linear representations of the input genome.

IR equality assessment and display

Equality among the IR regions is evaluated by PACVr both directly and indirectly. The direct evaluation is done by the computational comparison of the sequences of IRa and IRb, the indirect evaluation via the numerical and visual comparison of number, length, and location of all genes contained in the IRs. Specifically, the software conducts a two-step procedure in which equality in sequence, sequence length, and the number of genes is confirmed across the two IR regions, and the equality then visualized by connecting the matching genes of the two IRs via blue connector lines. To that end, PACVr computationally extracts the two IR regions from the input sequence record, stores their sequences as well as the names, start and end positions of all IR genes in separate data frames, and compares the exact number, length, and location of the genes across both regions. Any difference in sequence, sequence length, or gene complement between the two IRs results in a warning message to the user. PACVr then visualizes the equality between the IRs by connecting genes with identical names across the regions using blue connector lines. The lines hereby originate and end at the central nucleotide of each gene shared between the two IRs. The start and end width of these connector lines can be set to be uniform or

proportional to the length of the genes they connect. Any difference in name or location of the IR genes becomes visible through unequal or missing connector lines, thus enabling the visual assessment of equality among the IR regions regarding gene presence and synteny. This visualization of gene location and synteny contributes to the discovery of rules and patterns in genome orientation and rearrangements [73].

Visualization

PACVr employs RCircos [74] as the visualization engine. RCircos is an R implementation of the Circos environment [50] and is employed by PACVr to visualize the various aspects of plastome structure and coverage depth in four separate layers. In the first, outermost layer, PACVr displays length-labeled tick marks at each decile of total genome length to provide positional information across the genome. The layer also plots the names and relative positions of the individual regions of the quadripartite genome structure (i.e., LSC, SSC, IRa and IRb), with each region marked in a different color for easier delineation. If none or only one of the IRs are detected in the input genome, this layer displays a homogeneous color. In the second layer, PACVr plots the names and positions of all genes of the plastid genome, with gene positions indicated by their central nucleotide. In the third layer, PACVr plots the coverage depth of the plastid genome in the form of a circular histogram, with bars displaying one of two possible colors depending on their depth value relative to a user-defined threshold: bars with a coverage depth above the threshold are displayed in black, bars below the threshold in red. The threshold is by default specified relative to the average genome-wide coverage depth, but can optionally be set as an absolute value. Moreover, this layer indicates the average coverage depth of each of the four plastome regions via a horizontal, yellow line, which is missing in areas without coverage. In the fourth, innermost layer, PACVr plots blue connector lines that connect genes with identical names across the two IR regions, with lines originating and ending at the central nucleotide of each gene. At the lower left of the circular graph, PACVr prints a legend that displays the absolute and relative coverage depth threshold values below which histogram bars are highlighted, as well as the numeric values of average coverage depth of the four plastome regions. The name of the organism under study, which is parsed from the GenBank input file, is displayed as the figure title.

Accounting for quadripartite structure

The quadripartite structure of plastid genomes requires adjustments in the calculation of coverage depth and the visualization of IR equality compared to unpartitioned chromosomes. By default, PACVr calculates window-sized coverage depth values and, based on these, the region-wide average for each of the four plastome regions. However, PACVr would double-count the coverage of those windows that span across a region boundary, unless the coverage calculation included a special adjustment. Similarly, PACVr requires customization when visualizing the equality of gene position and synteny between the two IR regions. Natural expansions in IR size can cause genes located near the border of a single copy and an IR region to be displaced from one region into the other over time [75, 76]. Without a customized visualization, genes that are located primarily in the single copy region but span into the IR (or vice versa) would not be included in the IR equality visualization if the central nucleotide of genes used to connect the counterparts was located outside the IR. This can be particularly problematic with large plastome genes

such as *ycf1* and *ycf2*, which are located near the 5' end of the SSC and the IRa, respectively, in most angiosperms and represent nearly 10% of the unit-genome length [77]. A similar issue would arise with trans-spliced plastome genes whose exons were located in an SC and an IR region, respectively (e.g., *rps12* in many angiosperms [78]). Thus, the code of PACVr was customized to split genes that span more than one genome region into two separate parts along the region boundary and to treat both parts as separate units. PACVr tracks the position of these unequal units in relation to the region boundaries throughout software execution and corrects the location of the original genes and, by extension, their gene labels and histogram bars during the generation of the final plastome map using a size correction factor.

Installation, dependencies and usage

PACVr was written in R and can be installed via the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/) using the R command `install.packages('PACVr')`. It requires the presence of the R packages optparse [79], genbankr [63], and RCircos [74] as dependencies and employs several generic library functions developed for high-throughput genomic analysis [80]. Additionally, PACVr requires the software mosdepth [69] to be present on the system, which can be installed via the Unix shell command `conda install mosdepth`. The source code of PACVr is available via Github at https://github.com/michaelgruenstaeudl/PACVr. The technical documentation and a user tutorial (vignette) is distributed as part of the R package. The vignette provides example commands for the installation and execution of PACVr as well as for the generation of BAM files under different read filtering settings.

Two mandatory and nine optional input parameters can be specified when invoking PACVr. The mandatory input parameters are: the name of, and file path to, the input GenBank file, and the name of, and file path to, a sorted and indexed BAM file. The optional input parameters are: (i) the window size for calculating coverage depth, with a default value of 250; (ii) the shell command to execute mosdepth, with a default command of `mosdepth`; (iii) the coverage depth threshold above which histogram bars are plotted in red as opposed to the default black, with a default value of 0.5; (iv) the selection if the threshold value is specified relative to the average genome-wide coverage depth as opposed to representing an absolute value, with the default set to true; (v) the selection if the coverage depth values are to be log-transformed prior to visualization, with the default set to false; (vi) the selection if and what type of connector lines to draw between matching genes of the IRs, with the default line type displaying a start and end width proportional to the length of the genes that the lines connect; (vii) the size of all text elements of the resulting visualization relative to the maximum font size, with the default set to 0.5; (viii) the decision to remove all temporary files generated during the coverage depth calculation, with the default set to true; and (ix) the name of, and file path to, the output file, with the output saved as `./PACVr_output.pdf` by default. The software can be invoked either from within the R environment or directly from a Unix shell. A complete list of the short- and long-flag command-line (CLI) arguments available when invoking PACVr from the Unix shell is displayed via the shell command `Rscript ./inst/extdata/PACVr_Rscript.R -h/--help`. In the framework of an automated workflow (and upon setting the location of PACVr to a shell variable with the same

name), the following shell command can, for example, be used to execute PACVr on the empirical dataset co-supplied with the R package:

```
Rscript $PACVr/inst/extdata/PACVr_Rscript.R \
 -k $PACVr/inst/extdata/NC_045072/NC_045072.gb \
 -b $PACVr/inst/extdata/NC_045072/NC_045072_\
PlastomeReadsOnly.sorted.bam \
 -w 300 \
 -o NC_045072_PlastomeVisualization.pdf
```

Testing of software

To evaluate and demonstrate the functionality of PACVr, the software was tested under a variety of different settings. First, PACVr was tested on empirical data of four complete plastid genomes. Specifically, the software was employed for visualizing coverage depth and IR equality of the assemblies of two novel as well as two previously published plastid genomes. The novel plastid genomes represent the angiosperm species *Archidasyphyllum excelsum* (Asteraceae) and *Nuphar japonica* (Nymphaeaceae) and display a quadripartite genome structure as well as a genome size typical for the majority of angiosperms [2]. The previously published plastid genomes represent the angiosperm species *Pelargonium* x *hortorum* (Geraniaceae; [81]) and the non-photosynthetic green algae *Prototheca cutis* (Chlorellaceae; [82]) and display a genome size that substantially deviates from the typical size range of angiosperm [2] and green algae plastomes [83], respectively. Moreover, the plastid genome of *Prototheca cutis* naturally lacks the IR regions and, thus, a quadripartite genome structure [82, 84]. Details on the length and position of the different plastome regions present, the overall size of the genome, the GenBank accession number, and, in case of previous publication, the accession number of the original sequence reads are given in Table 1. The plastid genomes of *Archidasyphyllum excelsum* and *Nuphar japonica* were generated for this investigation via Illumina MiSeq sequencing following the sequencing protocol of [27] and the assembly workflow of [21]; the plastid genomes of *Pelargonium* x *hortorum* and *Prototheca cutis* were downloaded from GenBank. Information on coverage depth was generated for each plastid genome by mapping the original sequence reads to the complete genome sequence using BWA and samtools, which resulted in one sorted and indexed BAM file per genome. For each of the newly generated plastid genomes, spikes in the coverage depth were capped at a maximum of 20x to keep the size of the BAM files at a maximum of 2.5 megabytes per file and, thus, ensure a lightweight distribution of the R package once these BAM files were included in the package as example data. The cap was administered via script 'bbnorm.sh' of the software BBTools v.33.89 [85], which removes spikes in sequence coverage via a stochastic normalization procedure. Upon preparation of all input files, PACVr was employed on each of the four plastid genomes using default parameter values. Second, PACVr was tested on five different operating systems. Specifically, we tested the software on macOS 10.13.6 (High Sierra), macOS 10.14.6 (Mojave), Arch Linux 4.18, Debian 9.9, and Ubuntu 18.10. Under each system, PACVr was invoked both from within the R environment as well as directly from a Unix shell. Third, PACVr was compared to three software tools that are capable of visualizing plastome sequencing coverage. Specifically, we employed the tools Circleator v.1.0.2, FastPlast v.1.2.8, and IOGA v.20160910 to visualize the coverage depth of the two newly generated plastid genomes and compared their output to the default visualizations of PACVr.

Table 1 The GenBank and the sequence read archive (SRA) study accession numbers as well as details on genome size and structure of the four plastid genomes used to demonstrate the functionality of PACVr. All size and position values are given in bp. Abbreviations used: n.a. = not applicable; pos. = position

Species name	GenBank	SRA study	Total size	Pos. LSC	Pos. (size) IRb	Pos. (size) SSC	Pos. (size) IRa
Archidasyphyllum excelsum	MH899017	SRP253816	151,880	1..83,242	83,243..108,250 (25,007)	108,251..126,872 (18,621)	126,873..151,880 (25,007)
Nuphar japonica	NC_045072	SRP253784	160,761	1..90,653	90,654..116,283 (25,629)	116,284..135,131 (18,847)	135,132..160,761 (25,629)
Pelargonium x *hortorum*	NC_008454	SRP015898	217,942	1..59,710	59,711..135,457 (75,746)	135,458..142,209 (6,751)	142,210..217,942 (75,732)
Prototheca cutis	NC_037480	DRP002975	51,673	n.a.	n.a.	n.a.	n.a.

Results

Visualizations by PACVr

PACVr was successfully applied in the visualization of coverage depth and IR equality of the four complete plastid genomes used for evaluating the functionality of the software. Specifically, PACVr visualized the coverage depth of each plastid genome in relation to its circular, often quadripartite genome structure and illustrated the equality of its IR regions regarding gene position and synteny (if such regions existed in the genome). Based on the resulting visualizations, several important observations were made. First, the visualizations indicated differences in region-wide average coverage depth and the presence of genomic areas with markedly lower coverage depth compared to other areas of the same genome in each of the plastid genomes under study. In the plastid genome of *Nuphar japonica* (Fig. 2), for example, the average coverage depth of both IRs was detected to be 16x compared to 20x for the LSC and the SSC, respectively. Moreover, a window-sized coverage depth below 50% of the average genome-wide coverage depth was identified in several locations of the IRs, particularly at the 5' end of the IRb and, conversely, the 3' end of the IRa (assuming a single reading direction for the entire genome), which corresponds to the location of the ribosomal protein genes *rpl2* and *rpl23*. In the plastid genome of *Archidasyphyllum excelsum* (Additional file 1: Figure S2), the average coverage depth of the IRs was also detected to be lower than that of the LSC or the SSC. Moreover, a window-sized coverage depth below 50% of the average genome-wide coverage depth was identified in several locations of the IRs, particularly at or near the 5' end of the IRb (and, conversely, the 3' end of IRa), which corresponds to the location of gene *ycf2*; a suboptimal coverage depth was also detected in one calculation window of the LSC (near *trnD*-GUC, a gene encoding one of the transfer RNAs for aspartate). Successful visualizations of coverage depth were also conducted for the plastid genomes of *Pelargonium* x *hortorum* (Additional file 2: Figure S3) and *Prototheca cutis* (Additional file 2: Figure S4), despite their substantial deviations in genome size from the typical size range of angiosperm and green algae plastomes. The genome-wide average coverage depth of these genomes was calculated to be 4,569x and 615x, respectively, but could not be inferred by region, as the IR annotations of each genome were either unequal in length (*Pelargonium* x *hortorum*; Table 1) or missing to reflect the natural state (*Prototheca cutis*). The threshold value for highlighting histogram bars was set to 100% of the average genome-wide coverage depth for both genomes to contrast this option with the visualizations of the newly generated plastid genomes. Second, the visualizations by PACVr indicated strong gene synteny across the IRs of those plastid genomes under study that possess a quadripartite genome structure. Specifically, the symmetric display of equal gene position and length via blue, variable-width connector lines between the IRs of a genome indicated IR gene synteny in *Nuphar japonica*, *Archidasyphyllum excelsum*, and *Pelargonium* x *hortorum* (Figs. 2, S2, and S3). By contrast, PACVr automatically skipped the visualization of IR gene synteny for the plastid genome of *Prototheca cutis*, as this non-photosynthetic green alga does not possess IRs in its plastid genome or, by extension, the relevant IR feature annotations in its sequence record, rendering an evaluation and visualization of IR gene synteny obsolete. In summary, the IR regions of those plastid genomes with a quadripartite genome structure were found to display areas of reduced coverage depth, but equality in sequence length and gene position and, by extension, the presence of gene synteny between the IRs. Identical visualizations were retrieved when executing

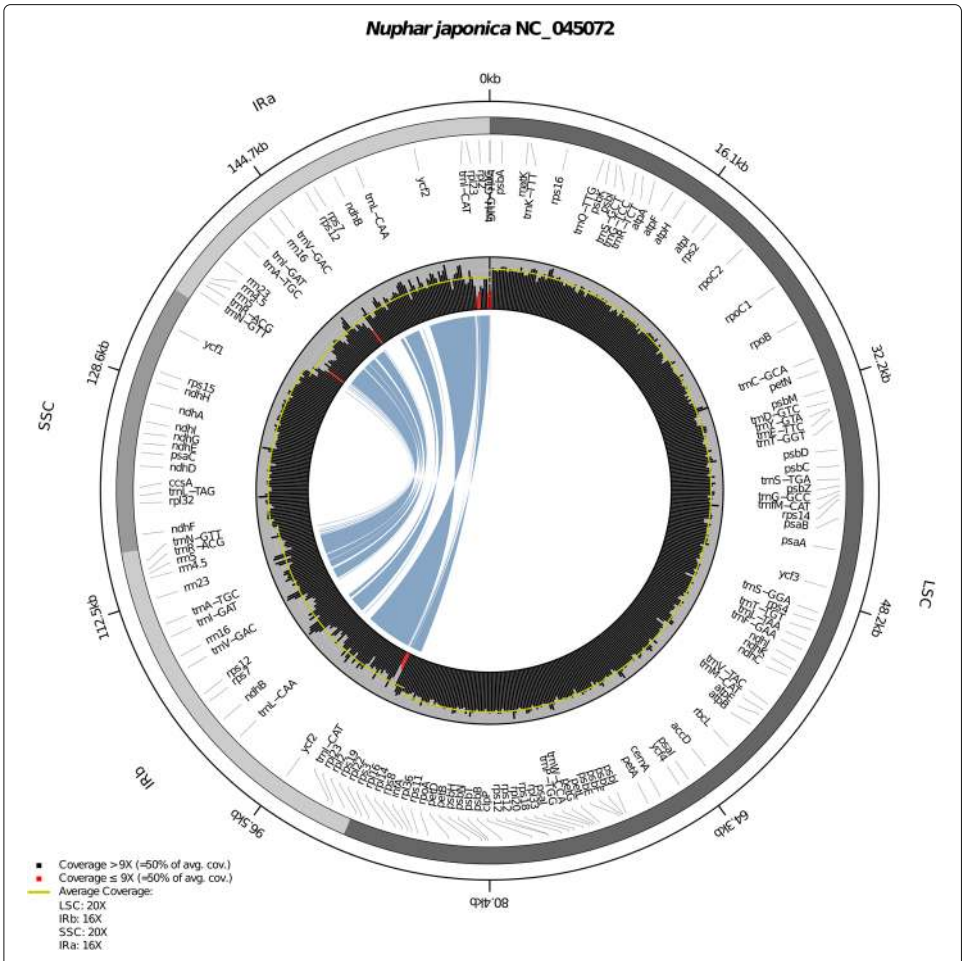

Fig. 2 Visualization of coverage depth and IR equality of the plastid genome of *Nuphar japonica* as generated via PACVr

PACVr on macOS or Linux, confirming the compatibility of PACVr to different operating systems.

Comparison to other software tools

The comparison of visualizations of coverage depth between PACVr and three other software tools recovered dissimilar coverage depth distributions. The graphs generated by the tools FastPlast, IOGA, and Circleator were dissimilar among each other and, in the case of FastPlast and IOGA, also dissimilar to PACVr. For the plastid genome of *Nuphar japonica*, FastPlast generated a linear plot of coverage depth that indicated a higher depth in the area corresponding to the IRs than in the large and small SC regions (Fig. 1a). The coverage depth of the IR regions was hereby often larger than 20x and, thus, larger than the manual cap instituted when generating the input files, indicating that the read mapping procedure of FastPlast allows a multiple counting of reads across the input genome. IOGA also generated a linear plot of coverage depth, which indicated a markedly higher coverage depth in an area that approximately corresponds to the SSC compared to other regions of the plastid genome (Fig. 1b); the precise location of this area in relation to the overall genome structure is uncertain, however, as IOGA generates coverage graphs

on the concatenation of individual contigs constructed during the assembly process, and these contigs may not be ordered according to their actual position in the genome. Similar to FastPlast, the coverage depth inferred by IOGA surpassed the maximum cap of 20x for certain areas of the genome assembly, indicating a multiple counting of reads. The graph generated by Circleator was the most similar representation of coverage depth compared to the visualization generated by PACVr. Visualized as a circular plot, it indicated areas of reduced coverage depth in the IRs compared to the SC regions and a largely homogeneous coverage depth across the SC regions (Fig. 1c). The precise locations of areas with reduced coverage depth were, however, difficult to determine due to missing references to the quadripartite genome structure and to gene positions. The visualizations of coverage depth for the plastid genome of *Archidasyphyllum excelsum* were also dissimilar among each other as well as to those of PACVr (Additional file 1: Figure S2). Moreover, the coverage graph generated by IOGA for this genome displayed coverage for only ca. 130 kb of the full genome length due to a missing contig (probably an IR) in the assembly product (Additional file 1: Figure S1b).

The dissimilarity of inferred coverage depth distributions between FastPlast, IOGA, Circleator, and PACVr may be the result of different visualization routines among these tools, but may also be impacted by the different plastome assembly procedures employed. The primary function of FastPlast and IOGA is the assembly of complete plastid genomes, with the ability to visualize coverage depth representing a peripheral function. Specifically, FastPlast and IOGA were designed to generate *de novo* plastid genome assemblies from sequence read data, map the sequence reads onto the inferred contigs, and then conduct visualizations of coverage depth to illustrate the assembly results. PACVr and Circleator, by contrast, were designed to visualize the coverage depth of plastid genomes that have been assembled independently of their own functionality. The observed differences in the coverage depth distributions may, therefore, reflect the idiosyncrasies of the genome assembly process as much as the differences in the respective coverage depth calculation and visualization routines. FastPlast, for example, generates plastome contigs via the assembler tools SPAdes [86] and afin [87] in an iterative assembly process, employs sequence coverage as an indicator for assembly accuracy, and calculates coverage depth via the software Jellyfish2 [88] using a fixed 25-mer sliding window. IOGA, by contrast, generates contigs iteratively via SOAPdenovo2 [89], selects the set of contigs with largest scaffold N50 as the new reference assembly, and then calculates coverage depth per base location on this assembly via the script 'bbmap.sh' of the software BBTools to illustrate the progressive improvement of the assembly output. The resulting visualizations of coverage depth of these software tools are, thus, different in both design and interpretation and can not be directly compared across tools.

Discussion

Importance of coverage information in plastid genomics

By visualizing coverage depth in relation to the quadripartite genome structure of plastid genomes and the location of individual genes, PACVr fills the need for a software tool that produces graphically intuitive visualizations for the identification of assembly regions with suboptimal coverage depth. Measuring the coverage depth is critical for the quality assessment of genome assemblies [38]. First, coverage depth is an essential metric for the identification of structural variation, as the depth of sequencing coverage drives the

power to detect sequence rearrangements and other structural variants [39]. Generally, greater coverage depth increases the chance that rearrangement endpoints are captured and confirmed by multiple independent reads [43]. This can be particularly relevant in the comparison of complete plastid genomes, which often differ structurally [73], not least in the precise start and end positions of the IR regions [76]. Second, coverage depth is an essential metric for the detection of sequence variation, as genomic regions with exceptionally high [90] or low [91] coverage depths become unreliable for variant calling. In plastid genomics, variant calling can be relevant to identify intra-individual polymorphisms, which are typically generated by the effects of heteroplasmy and common to organelle genomes [92]. Third, de novo assembly algorithms typically operate under the assumption of even coverage depth across the target genome [6, 44, 45], and errors in plastid genome sequences are often correlated with exceptionally high or low coverage depths [29]. The visualization of coverage depth of plastome assemblies, thus, represents an important tool in their quality assessment and should be conducted as early in their bioinformatic processing as possible in order to identify problematic assemblies before proceeding with subsequent analyses. Preferentially, such visualizations should be rapid, easily integrable into automated workflows, and suitable for the evaluation of a large cohort of genome assemblies [38].

Integration into automated pipelines

Given the demand for high throughput in bioinformatic workflows, individual software tools must be easy to integrate into automated analysis pipelines to be of lasting value for the research community. The integration of plastome assembly and annotation into automated or semi-automated workflows has been proposed and conducted by several investigations [21, 22, 93]. Such workflows are designed to deliver more consistent and repeatable results than the manual administration of individual software tools and provide an ideal platform for the integration of assembly quality tests. However, quality management has so far remained unimplemented in most plastid genome analysis pipelines (but see [21]). In fact, most quality control tools for plastid genome assembly in existence do not provide rapid visualizations of coverage depth. As a result, inaccurate or unsupported plastid genome assemblies may remain undetected and confound subsequent analyses, especially in large, composite investigations that compare hundreds if not thousands of plastid genomes (e.g., [14, 15]). Hence, it is critical to visualize the coverage profile of a plastid genome through an automated, yet user-friendly process that assists in highlighting genomic regions of interest to the researcher [59]. Strong emphasis was, thus, placed on the ability to integrate PACVr into automated bioinformatic pipelines easily. Given this objective, PACVr was customized for and submitted to CRAN, which tests incoming R packages to work on all major operating systems and, thus, ensures these packages to be platform-independent. Similarly, PACVr was designed to enable an operation directly from the Unix shell using CLI arguments, which allows easy integration into automated workflows.

Importance of open-source software in plastid genomics

Several previously available web-services for visualizing circular plastome maps have become inaccessible over recent years, highlighting the importance of open-source software development in plastid genomics. The development and release of PACVr as an

open-source software tool was one of the guiding principles in its development, as this allows other researchers to independently access its source code, customize the software, and extend its functionality. The aim of open-source development is particularly important in the field of plastid genomics, where several previously developed web-services have become inaccessible over recent years. In fact, several interactive web-based tools had been developed to visualize circular chromosomes and their associated metadata, including complete plastid genomes. However, many of these tools are no longer applied because their online interfaces have lost connectivity to the world wide web, and their source code has never been made publicly available. The online platform CARAS [57], for example, offered functionality to annotate and visualize complete plastid genomes and save the results in different output formats, but its web service has been inaccessible since at least February 2017. Similarly, the web platform CGAP [58] offered functionality to generate circular or linear genome maps, annotate assembled plastid genomes, and conduct comparative plastome analysis, but has been inaccessible since at least April 2017. Some of these online services provided installation-free alternatives to the limited number of visualization software tools for plastid genomics [94], and their inaccessibility should be considered a loss for the plant biological research community. Had these services been developed as open-source projects, other researchers would have had the opportunity to continue the maintenance and development of these resources [95]. Open-source development and public accessibility of software tools are, thus, considered critical aspects of bioinformatic software development [96–98]. Consequently, PACVr was developed as an open-source R package that is publicly available via both GitHub and CRAN.

Conclusions

Coverage depth is often used as an indicator of the quality of a plastid genome assembly. The R package PACVr was designed to visualize coverage depth of plastome assemblies in relation to the circular, quadripartite structure of plastid genomes, the location of individual plastome genes, different window calculation sizes, and user-defined threshold values for coverage depth. PACVr also enables the visual assessment of equality among the IR regions regarding gene presence and synteny. In tests on empirical data, the software successfully visualized the coverage depth and IR equality of complete plastid genomes of different plant lineages, which displayed total plastome sizes between 50 kb and 250 kb. Our evaluations also highlighted that alternative coverage visualization tools for plastid genomes generate incongruent depth visualizations on the same input data, which may be attributable to differences in the visualization process as well as the genome assembly routines. Given its design as an open-source R package with a Unix shell interface, PACVr allows easy integration into bioinformatic pipelines and, thus, provides an important tool for automated quality control in plastid genome sequencing.

Availability and requirements

Project name: PACVr
Project home page: https://cran.r-project.org/package=PACVr
Operating systems: Platform independent
Programming language: R (>= 3.3)

Other requirements: R packages BiocGenerics, Biostrings, GenomicAlignments, genbankr, optparse; mosdepth (>= 0.2.5)

License: BSD 3-Clause

Any restrictions to use by non-academics: none

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3475-0.

Additional file 1: Visualizations of coverage depth of the plastid genome of *Archidasyphyllum excelsum* (GenBank accession MH899017) as generated by the software tools FastPlast, IOGA, Circleator, and PACVr. (PDF 8211 kb)

Additional file 2: Visualizations of coverage depth and IR equality of the plastid genome of *Pelargonium* x *hortorum* (GenBank accession NC_008454) and *Prototheca cutis* (GenBank accession NC_037480). (PDF 5540 kb)

Abbreviations

bp: base pairs; CLI: command-line; IR: inverted repeat; kb: kilobases; LSC: large single copy; SSC: small single copy

Acknowledgements

The authors thank Yannick Hartmaring of the Freie Universität Berlin for assistance with testing the final software version. The authors acknowledge the high-performance computing service of the ZEDAT of the Freie Universität Berlin for providing allocations of computing time. The development of code for this R package constitutes part of a thesis by NJ toward a bachelor of science degree.

Authors' contributions

MG – Project idea and design, oversight of development, implementation, documentation, testing on empirical data, manuscript writing and manuscript revision. NJ – additional implementation, documentation and testing. Both authors read and approved the final manuscript.

Funding

This investigation was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 418670221 – and by a start-up grant of the Freie Universität Berlin (Initiativmittel der Forschungskommission), both to MG. The funding bodies did not play any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

PACVr is available under the BSD 3-Clause license from CRAN at https://cran.r-project.org/package=PACVr. All datasets analyzed during the present investigation are available from Zenodo at https://zenodo.org/record/3673838.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institut für Biologie, Systematische Botanik und Pflanzengeographie, Freie Universität Berlin, 14195 Berlin, Germany. [2]Institut für Bioinformatik, Freie Universität Berlin, 14195 Berlin, Germany.

Received: 6 August 2019 Accepted: 31 March 2020
Published online: 24 May 2020

References

1. Mower JP, Vickrey TL. Structural diversity among plastid genomes of land plants. Adv Bot Res. 2018;85:263–92. https://doi.org/10.1016/bs.abr.2017.11.013.
2. Ruhlman TA, Jansen RK. The plastid genomes of flowering plants. In: Maliga P, editor. Chloroplast Biotechnology, Methods in Molecular Biology (Methods and Protocols). Totowa: Humana Press; 2014. p. 3–38. https://doi.org/10.1007/978-1-62703-995-6.
3. Blazier JC, Jansen RK, Mower JP, Govindu M, Zhang J, Weng M, Ruhlman TA. Variable presence of the inverted repeat and plastome stability in*Erodium*. Ann Bot. 2016;117:1209–20. https://doi.org/10.1093/aob/mcw065.
4. Ruhlman TA, Zhang J, Blazier JC, Sabir JSM, Jansen RK. Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. Am J Bot. 2017;104:559–72. https://doi.org/10.3732/ajb.1600453.
5. Wicke S, Schneeweiss GM, de Pamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. Plant Mol Biol. 2011;76:273–97. https://doi.org/10.1007/s11103-011-9762-4.

6. Twyford AD, Ness RW. Strategies for complete plastid genome sequencing. Mol Ecol Resour. 2017;17:858–68. https://doi.org/10.1111/1755-0998.12626.
7. Gao L, Su YJ, Wang T. Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. J Syst Evol. 2010;48:77–93. https://doi.org/10.1111/j.1759-6831.2010.00071.x.
8. Gitzendanner MA, Soltis PS, Yi T, Li D-Z, Soltis DE. Plastome phylogenetics: 30 years of inferences into plant evolution. Adv Bot Res. 2018;85:293–313. https://doi.org/10.1016/bs.abr.2017.11.016.
9. Tonti-Filippini J, Nevill PG, Dixon K, Small I. What can we do with 1000 plastid genomes?,. Plant J. 2017;90:808–18. https://doi.org/10.1111/tpj.13491.
10. Bernhardt N, Brassac J, Kilian B, Blattner FR. Dated tribe-wide whole chloroplast genome phylogeny indicates recurrent hybridizations within Triticeae. BMC Evol Biol. 2017;17:141. https://doi.org/10.1186/s12862-017-0989-9.
11. Teisher JK, McKain MR, Schaal BA, Kellogg EA. Polyphyly of Arundinoideae (Poaceae) and evolution of the twisted geniculate lemma awn. Ann Botany. 2017;120:725–38. https://doi.org/10.1093/aob/mcx058.
12. Saarela JM, Burke SV, Wysocki WP, Barrett MD, Clark LG, Craine JM, Peterson PM, Soreng RJ, Vorontsova MS, Duvall MR. A 250 plastome phylogeny of the grass family (Poaceae): Topological support under different data partitions. PeerJ. 2018;6:4299. https://doi.org/10.7717/peerj.4299.
13. Huang B, Ruess H, Liang Q, Colleoni C, Spooner D. Analyses of 202 plastid genomes elucidate the phylogeny of *Solanum* section *Petota*. Sci Rep. 2019;9:7. https://doi.org/10.1038/s41598-019-40790-5.
14. Gitzendanner MA, Soltis PS, Wong GK-S, Ruhfel BR, Soltis DE. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. Am J Bot. 2018;105:291–301. https://doi.org/10.1002/ajb2.1048.
15. Li H-T, Yi T-S, Gao L-M, Ma P-F, Zhang T, Yang J-B, Gitzendanner MA, Fritsch PW, Cai J, Luo Y, Wang H, van der Bank M, Zhang S-D, Wang Q-F, Wang J, Zhang Z-R, Fu C-N, Yang J, Hollingsworth PM, Chase MW, Soltis DE, Soltis PS, Li D-Z. Origin of angiosperms and the puzzle of the Jurassic gap. Nat Plants. 2019;5:461–70. https://doi.org/10.1038/s41477-019-0421-0.
16. Ankenbrand MJ, Pfaff S, Terhoeven N, Gundel M, Weiss CL, Hackl T, Förster F. ChloroExtractor: Extraction and assembly of the chloroplast genome from whole genome shotgun data. J Open Source Softw. 2018;3:2016–8. https://doi.org/10.21105/joss.00464.
17. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017;45:18. https://doi.org/10.1093/nar/gkw955.
18. Izan S, Esselink D, Visser RGF, Smulders MJM, Borm T. De novo assembly of complete chloroplast genomes from non-model species based on a k-mer frequency-based selection of chloroplast reads from total DNA sequences. Front Plant Sci. 2017;8:1271. https://doi.org/10.3389/fpls.2017.01271.
19. McKain M, Wilson M. Fast-Plast v.1.2.6. 2017. https://github.com/mrmckain/Fast-Plast/. Accessed 04 Aug 2019.
20. Coissac E. Org.Asm: The genome ORGanelle ASseMbler v.1.0.3. 2019. https://pypi.org/project/ORG.asm. Accessed 04 Aug 2019.
21. Gruenstaeudl M, Gerschler N, Borsch T. Bioinformatic workflows for generating complete plastid genome sequences - An example from *Cabomba* (Cabombaceae) in the context of the phylogenomic analysis of the water-lily clade. Life. 2018;8:25. https://doi.org/10.3390/life8030025.
22. Jian J.-J., Yu W.-B., Yang J.-B., Song Y., Yi T.-S., Li D.-Z. GetOrganelle: a simple and fast pipeline for de novo assembly of a complete circular chloroplast genome using genome skimming data. bioRxiv. 2018;256479:. https://doi.org/10.1101/256479.
23. Wang X, Cheng F, Rohlsen D, Bi C, Wang C, Xu Y, Wei S, Ye Q. Organellar genome assembly methods and comparative analysis of horticultural plants. Hortic Res. 2018;5:3. https://doi.org/10.1038/s41438-017-0002-1.
24. Wu Z, Tembrock LR, Ge S. Are differences in genomic data sets due to true biological variants or errors in genome assembly: An example from two chloroplast genomes. PLOS ONE. 2015;10:0118019. https://doi.org/10.1371/journal.pone.0118019.
25. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: Automated circularization of genome assemblies using long sequencing reads. Genome Biol. 2015;16:294. https://doi.org/10.1186/s13059-015-0849-0.
26. Williams AV, Boykin LM, Howell KA, Nevill PG, Small I. Correction: The complete sequence of the *Acacia ligulata* chloroplast genome reveals a highly divergent *clpP1* gene. PLOS ONE. 2015;10:0138367. https://doi.org/10.1371/journal.pone.0138367.
27. Gruenstaeudl M, Nauheimer L, Borsch T. Plastid genome structure and phylogenomics of Nymphaeales: conserved gene order and new insights into relationships. Plant Syst Evol. 2017;303:1251–70. https://doi.org/10.1007/s00606-017-1436-5.
28. Amiryousefi A, Hyvoenen J, Poczai P. The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. PLOS ONE. 2018;13:0196069. https://doi.org/10.1371/journal.pone.0196069.
29. Kim K, Lee S-C, Lee J, Yu Y, Yang K, Choi S, Koh H-J, Waminal NE, Choi H-I, Kim N-H, Jang W, Park H-S, Lee J, Lee HO, Joh HJ, Ju H, Park JY, Perumal S, Jayakodi M, Lee YS, Kim B, Copetti D, Kim S, Kim S, Lim K-b, Kim Y-D, Lee J, Cho K-S, Park B-S, Wing RA, Yang T-J. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. Sci Rep. 2015;5:15655. https://doi.org/10.1038/srep15655.
30. Walker JF, Jansen RK, Zanis MJ, Emery NC. Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. Am J Bot. 2015;102:1751–2. https://doi.org/10.3732/ajb.1500299.
31. Wang W, Lanfear R. Long-reads reveal that the chloroplast genome exists in two distinct versions in most plants. Genome Biol Evol. 2019;11:3372–81. https://doi.org/10.1093/gbe/evz256.
32. Earl D, Bradnam K, John JS, Darling A, Lin D, Fass J, On H, Yu K, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung W-K, Ning Z, Haimel M, Simpson JT, Fonseca NA, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Koren S, Yang S-P, Wu W, Chou W-C, Srivastava A, Shaw TI, Ruby JG, Skewes-Cox P, Betegon M, Dimon MT, Solovyev V, Seledtsov I, Kosarev P, Vorobyev D, Ramirez-Gonzalez R, Leggett R, Maclean D, Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Yin S, Sharpe T, Hall G, Kersey PJ, Durbin R, Jackman SD, Chapman JA, Huang X, Derisi JL, Caccamo M, Li Y, Jaffe DB, Green RE, Haussler D, Korf I, Paten B.

Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Res. 2011;21: 2224–41. https://doi.org/10.1101/gr.126599.111.Freely.
33. Alhakami H, Mirebrahim H, Lonardi S. A comparative evaluation of genome assembly reconciliation tools. Genome Biol. 2017;18:93. https://doi.org/10.1186/s13059-017-1213-3.
34. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5. https://doi.org/10.1093/bioinformatics/btt086.
35. Udall JA, Dawe RK. Is it ordered correctly? Validating genome assemblies by optical mapping. Plant Cell. 2018;30: 7–14. https://doi.org/10.1105/tpc.17.00514.
36. Palmer JD. Chloroplast DNA exists in two orientations. Nature. 1983;301:92–3. https://doi.org/10.1038/301092a0.
37. Turmel M, Otis C, Lemieux C. Divergent copies of the large inverted repeat in the chloroplast genomes of ulvophycean green algae. Sci Rep. 2017;7:994. https://doi.org/10.1038/s41598-017-01144-1.
38. Pedersen BS, Collins RL, Talkowski ME, Quinlan AR. Indexcov: Fast coverage quality control for whole-genome sequencing. GigaScience. 2017;6:1–6. https://doi.org/10.1093/gigascience/gix090.
39. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: Key considerations in genomic analyses. Nature. 2014;15:121–32. https://doi.org/10.1038/nrg3642.
40. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14:157–67. https://doi.org/10.1038/nrg3367.
41. Ekblom R, Wolf JBW. A field guide to whole genome sequencing, assembly and annotation. Evol Appl. 2014;7: 1026–42. https://doi.org/10.1111/eva.12178.
42. McKain MR, Johnson MG, Uribe-Convers S, Eaton D, Yang Y. Practical considerations for plant phylogenomics. Appl Plant Sci. 2018;6:1038. https://doi.org/10.1002/aps3.1038.
43. Chen K, Wallis JW, Mclellan MD, Larson DE, Kalicki JM, Pohl CS, Mcgrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677–81. https://doi.org/10.1038/nmeth.1363.
44. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8. https://doi.org/10.1093/bioinformatics/bts174.
45. Olson ND, Treangen TT, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, Pop M. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinformatics. 2019;20:1140–50. https://doi.org/10.1093/bib/bbx098.
46. Bakker FT, Lei D, Yu J, Mohammadin S, Wei Z, van de Kerke S, Gravendeel B, Nieuwenhuis M, Staats M, Alquezar-Planas DE, Holmer R. Herbarium genomics: Plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. Biol J Linn Soc. 2016;117:33–43. https://doi.org/10.1111/bij.12642.
47. Lohse M, Drechsel O, Kahlau S, Bock R. OrganellarGenomeDRAW – A suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 2013;41:575–81. https://doi.org/10.1093/nar/gkt289.
48. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47:59–64. https://doi.org/10.1093/nar/gkz238.
49. Crabtree J, Agrawal S, Mahurkar A, Myers GS, Rasko DA, White O. Circleator: Flexible circular visualization of genome-associated data with BioPerl and SVG. Bioinformatics. 2014;30:3125–7. https://doi.org/10.1093/bioinformatics/btu505.
50. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, MA M. Circos: An information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45. https://doi.org/10.1101/gr.092759.109.19.
51. Shi C, Liu Y, Huang H, Xia E-H, Zhang H-B, Gao L-Z. Contradiction between plastid gene transcription and function due to complex posttranscriptional splicing: An exemplary study of *ycf15* function and evolution in angiosperms. PLOS ONE. 2013;8:59620. https://doi.org/10.1371/journal.pone.0059620.
52. Korotkova N, Nauheimer L, Ter-Voskanyan H, Allgaier M. Variability among the most rapidly evolving plastid genomic regions is lineage-specific: Implications of pairwise genome comparisons in Pyrus (Rosaceae) and other angiosperms for marker choice. PLOS ONE. 2014;9:112998. https://doi.org/10.1371/journal.pone.0112998.
53. Hu S, Sablok G, Wang B, Qu D, Barbaro E, Viola R, Li M, Varotto C. Plastome organization and evolution of chloroplast genes in *Cardamine* species adapted to contrasting habitats. BMC Genomics. 2015;16:306. https://doi.org/10.1186/s12864-015-1498-0.
54. Sato N, Ehira S. GenoMap, a circular genome data viewer. Bioinformatics. 2003;19:1583–4. https://doi.org/10.1093/bioinformatics/btg195.
55. Stothard P, Wishart DS. Circular genome visualization and exploration using CGView. Bioinformatics. 2005;21:537–9. https://doi.org/10.1093/bioinformatics/bti054.
56. Conant GC, Wolfe KH. GenomeVx: Simple web-based creation of editable circular chromosome maps. Bioinformatics. 2008;24:861–2. https://doi.org/10.1093/bioinformatics/btm598.
57. Li Y, Li H, Zhu Y, Li Z, Yin C, Lin X, Liu C. Development and implementation of CARAS algorithm for automatic annotation, visualization, and GenBank submission of chloroplast genome sequences. In: 2012 Computing, Communications and Applications Conference; 2012. p. 310–315. https://doi.org/10.1109/ComComAp.2012.6154863.
58. Cheng J, Zeng X, Ren G, Liu Z. CGAP: A new comprehensive platform for the comparative analysis of chloroplast genomes. BMC Bioinformatics. 2013;14:95. https://doi.org/10.1186/1471-2105-14-95.
59. Cruz A, Arrais J, Machado P. Interactive and coordinated visualization approaches for biological data analysis. Brief Bioinformatics. 2019;20:1513–23. https://doi.org/10.1093/bib/bby019.
60. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna: Computing, R Foundation for Statistical; 2019. https://www.r-project.org/. Accessed 13 Feb 2020.
61. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2006;35:21–5. https://doi.org/10.1093/nar/gkl986.

62. Lee T-H, Kim Y-K, Nahm BH. GBParsy: A GenBank flatfile parser library with high speed. BMC Bioinformatics. 2008;9:321. https://doi.org/10.1186/1471-2105-9-321.
63. Becker G, Lawrence M. Genbankr: Parsing GenBank files into semantically useful objects version 1.12.0. 2019. https://bioconductor.org/packages/release/bioc/html/genbankr.html. Accessed 04 Aug 2019.
64. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:. https://doi.org/10.1093/bioinformatics/btp352.
65. Schbath S, Martin V, Zytnicki M, Fayolle J, Loux V, Gibrat J-F. Mapping reads on a genomic sequence: An algorithmic overview and a practical comparative analysis. J Comput Biol. 2012;19:796–813. https://doi.org/10.1089/cmb.2012.0022.
66. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
67. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357. https://doi.org/10.1038/nmeth.1923.
68. Zhu X, Leung H, Wang R, Chin F, Yiu S, Quan G, Li Y, Zhang R, Jiang Q, Liu B, Dong Y, Zhou G, Wang Y. misfinder: Identify mis-assemblies in an unbiased manner using reference and paired-end reads. BMC Bioinformatics. 2015;16:16. https://doi.org/10.1186/s12859-015-0818-3.
69. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2017;34:867–8. https://doi.org/10.1093/bioinformatics/btx699.
70. Kearse M, Moir R, Wilson A, Stones-Havas S, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:1647–9. https://doi.org/10.1093/bioinformatics/bts199.
71. Karolchik D, Hinrichs AS, Kent WJ. The UCSC genome browser. Curr Protoc Bioinformatics. 2012;40:1–411433. https://doi.org/10.1002/0471250953.bi0104s40.
72. Phanstiel DH, Boyle AP, Araya CL, Snyder MP. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. Bioinformatics. 2014;30:2808–10. https://doi.org/10.1093/bioinformatics/btu379.
73. Wang D, Yu J. Plastid-LCGbase: a collection of evolutionarily conserved plastid-associated gene pairs. Nucleic Acids Res. 2014;43:990–5. https://doi.org/10.1093/nar/gku1070.
74. Zhang H, Meltzer P, Davis S. RCircos: an R package for Circos 2D track plots. BMC Bioinformatics. 2013;14:244. https://doi.org/10.1186/1471-2105-14-244.
75. Dugas DV, Hernandez D, Koenen EJM, Schwarz E, Straub S, Hughes CE, Jansen RK, Nageswara-Rao M, Staats M, Trujillo JT, Hajrah NH, Alharbi NS, Al-Malki AL, Sabir JSM, Bailey CD. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*. Sci Rep. 2015;5:16958. https://doi.org/10.1038/srep16958.
76. Zhu A, Guo W, Gupta S, Fan W, Mower JP. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytologist. 2016;209:1747–56. https://doi.org/10.1111/nph.13743.
77. Ruhlman T. A., Jansen R. K. Aberration or analogy? the atypical plastomes of geraniaceae. In: Chaw S.-M., Jansen R. K., editors. Plastid Genome Evolution, Advances in Botanical Research. Cambridge, MA: Academic Press; 2018. p. 223–262. https://doi.org/10.1016/bs.abr.2017.11.017.
78. Hildebrand M, Hallick RB, Passavant CW, Bourque DP. Trans-splicing in chloroplasts: the rps12 loci of *Nicotiana tabacum*. Proc Natl Acad Sci USA. 1988;85:372–6. https://doi.org/10.1073/pnas.85.2.372.
79. Davis TL. Optparse: Command Line Option Parser. v.1.6.2. 2019. https://CRAN.R-project.org/package=optparse. Accessed 04 Aug 2019.
80. Huber W, Carey V, Gentleman R, Anders S, Carvalho B, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen K, Irizarry R, Lawrence M, Love M, MacDonald J, Obenchain V, Ole? A, Morgan M. Orchestrating high-throughput genomic analysis with bioconductor. Nat Methods. 2015;12:115–21. https://doi.org/10.1038/nmeth.3252.
81. Chumley T, Palmer J, Mower J, Fourcade H, Calie P, Boore J, Jansen R. The complete chloroplast genome sequence of *Pelargonium x hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol. 2006;23:2175–90. https://doi.org/10.1093/molbev/msl089.
82. Suzuki S, Endoh R, Manabe R, Ohkuma M, Hirakawa Y. Multiple losses of photosynthesis and convergent reductive genome evolution in the colourless green algae prototheca. Sci Rep. 2018;8:11. https://doi.org/10.1038/s41598-017-18378-8.
83. Turmel M, Lemieux C, Vol. 85. Evolution of the plastid genome in green algae; 2018, pp. 157–93. https://doi.org/10.1016/bs.abr.2017.11.010.
84. Turmel M, Otis C, Lemieux C. Dynamic evolution of the chloroplast genome in the green algal classes pedinophyceae and trebouxiophyceae. Genome Biol Evol. 2015;7:2062–82. https://doi.org/10.1093/gbe/evv130.
85. Bushnell B. BBTools software package v.33.89. 2015. https://sourceforge.net/projects/bbmap. Accessed 04 Aug 2019.
86. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–477. https://doi.org/10.1089/cmb.2012.0021.
87. Wilson M. Afin: Assembly finishing v.2016.09.17. 2016. https://github.com/afinit/afin. Accessed 13 Feb 2020.
88. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70. https://doi.org/10.1093/bioinformatics/btr011.
89. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1:18. https://doi.org/10.1186/s13742-015-0069-2.

90. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30: 2843–51. https://doi.org/10.1093/bioinformatics/btu356.
91. Benjelloun B, Boyer F, Streeter I, Zamani W, Engelen S, Alberti A, Alberto F, BenBati M, Ibnelbachyr M, Chentouf M, Bechchari A, Rezaei H, Naderi S, Stella A, Chikhi A, Clarke L, Kijas J, Flicek P, Taberlet P, Pompanon F. An evaluation of sequencing coverage and genotyping strategies to assess neutral and adaptive diversity. Mol Ecol Res. 2019;19:48. https://doi.org/10.1111/1755-0998.13070.
92. Scarcelli N, Mariac C, Couvreur TLP, Faye A, Richard D, Sabot F, Berthouly-Salazar C, Vigouroux Y. Intra-individual polymorphism in chloroplasts from NGS data: Where does it come from and how to handle it?. Mol Ecol Res. 2016;16:434–45. https://doi.org/10.1111/1755-0998.12462.
93. McKain MR, Hartsock RH, Wohl MM, Kellogg EA. Verdant: Automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes. Bioinformatics. 2017;33:130–2. https://doi.org/10.1093/bioinformatics/btw583.
94. Sablok G, Mudunuri SB, Edwards D, Ralph PJ. Chloroplast genomics: Expanding resources for an evolutionary conserved miniature molecule with enigmatic applications. Curr Plant Biol. 2016;7-8:34–8. https://doi.org/10.1016/j.cpb.2016.12.004.
95. Huang X, Xie J, Otecko NO, Peng M. Accessibility and update status of published software: Benefits and missed opportunities. Front Res Metrics Anal. 2017;2:1. https://doi.org/10.3389/frma.2017.00001.
96. Ince DC, Hatton L, Graham-Cumming J. The case for open computer programs. Nature. 2012;482:485–8. https://doi.org/10.1038/nature10836.
97. Howison J, Bullard J. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. J Assoc Inf Sci Technol. 2016;67:2137–55. https://doi.org/10.1002/asi.23538.
98. Darriba D, Flouri T, Stamatakis A. The state of software for evolutionary biology. Mol Biol Evol. 2018;35:1037–46. https://doi.org/10.1093/molbev/msy014.

Publisher's Note