Received: 15 November 2021 | Revised: 23 October 2022 | Accepted: 1 November 2022

DOI: 10.1111/1755-0998.13729

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

CPGView: A package for visualizing detailed chloroplast genome structures

Shengyu Liu[1,2] | Yang Ni[1] | Jingling Li[1] | Xinyi Zhang[1] | Heyu Yang[1] | Haimei Chen[1] | Chang Liu[1]

[1]Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[2]Department of Medical Data Sharing, Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Correspondence
Chang Liu, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, China.
Emails: cliu6688@yahoo.com; cliu@implad.ac.cn

Funding information
National Science & Technology Fundamental Resources Investigation Program of China, Grant/Award Number: 2018FY100705; Innovation Funds for Medical Sciences from the Chinese Academy of Medical Sciences, (CIFMS), Grant/Award Number: 2021-I2M-1-022; National Mega-Project for Innovative Drugs of China, Grant/Award Number: 2019ZX09735-002; National Science Foundation of China Funds, Grant/Award Number: 81872966

Handling Editor: Suhua Shi

Abstract

Chloroplast genomes have been widely used in studying plant phylogeny and evolution. Several chloroplast genome visualization tools have been developed to display the distribution of genes on the genome. However, these tools do not draw features, such as exons, introns, repetitive elements, and variable sites, disallowing in-depth examination of the genome structures. Here, we developed and validated a software package called Chloroplast Genome Viewers (CPGView). CPGView can draw three maps showing (i) the distributions of genes, variable sites, and repetitive sequences, including microsatellites, tandem and dispersed repeats; (ii) the structure of the cis-splicing genes after adjusting the exon-intron boundary positions using a coordinate scaling algorithm, and (iii) the structure of the trans-splicing gene *rps12*. To test the accuracy of CPGView, we sequenced, assembled, and annotated 31 chloroplast genomes from 31 genera of 22 families. CPGView drew maps correctly for all the 31 chloroplast genomes. Lastly, we used CPGView to examine 5998 publicly released chloroplast genomes from 2513 genera of 553 families. CPGView succeeded in plotting maps for 5882 but failed to plot maps for 116 chloroplast genomes. Further examination showed that the annotations of these 116 genomes had various errors needing manual correction. The test on newly generated data and publicly available data demonstrated the ability of CPGView to identify errors in the annotations of chloroplast genomes. CPGView will become a widely used tool to study the detailed structure of chloroplast genomes. The web version of CPGView can be accessed from http://www.1kmpg.cn/cpgview.

KEYWORDS
cis-splicing genes, coordinate scaling algorithm, repeats, *rps12*, trans-splicing genes

1 | INTRODUCTION

Chloroplast genomes are widely used to study the phylogenetic classification and genome evolution of plants. Advancements in next-generation sequencing technology (NGS), third-generation DNA sequencing technologies, and bioinformatic tools have led to an influx of chloroplast genomes from various organisms (van Dijk et al., 2014). In December 2012, only 255 chloroplast sequences were available in GenBank. In July 2021, the database had more than 7000 chloroplast sequences, including those labelled "chloroplast

Shengyu Liu and Yang Ni contributed equally to the study.

genome" and "plastome." Ensuring the correct assembly and annotation of many chloroplast genome sequences has become increasingly challenging (Sandhya et al., 2020).

In the past few years, several annotations and visualization tools have been developed for organelle genomes; these tools include DOGMA (Wyman et al., 2004), CPGAVAS (Liu et al., 2012), Plann (Huang & Cronk, 2015), Verdant (McKain et al., 2017), GeSeq (Tillich et al., 2017), AGORA (Jung et al., 2018), OrganellarGenomeDRAW version 1.3.1 (OGDRAW) (Lohse et al., 2013), CPGAVAS2 (Shi et al., 2019), and Chloroplot (Zheng et al., 2020). CPGAVAS, GeSeq, and CPGAVAS2 can annotate the chloroplast genome and produce genome maps. By contrast, OrganellarGenomeDRAW and Chloroplot do not annotate the genome and only generate the genome map. In addition to predicting genes with simple structures, CPGAVAS2 can annotate genes with complex structures, such as those with small or trans-splicing exons. From the perspective of speed and efficiency for genome map generation, OGDRAW can batch process multiple GenBank files. Although these tools have been widely applied, they have several limitations.

First, repeats, including microsatellite, tandem repeats, and dispersed repeats, are widely used as genetic markers for species discrimination and understanding genome instabilities. Visualizing the repeat structures allows the users to identify the overall repeat patterns in the chloroplast genomes and generate hypotheses regarding their potential evolution. However, only CPGAVAS2 can generate a circular map with repeat distribution.

Second, chloroplast genomes are heteroplasmic due to the presence of multiple chloroplasts in a single cell (Lei et al., 2016). These heteroplasmic sites might have important functional implications. To our knowledge, no tools support the visualization of the heteroplasmic sites.

Third, the genome map produced by these tools can only show the general distribution of genes and does not display detailed gene structures, such as exon and intron boundaries for cis-splicing genes. The exon-intron boundaries remain the most error-prone regions for annotation (Shi et al., 2019). Previous studies on the numbers and distribution of chloroplast introns in different taxa suggest that splicing has evolved through different pathways in various chloroplast lineages (Schmitz-Linneweber & Barkan, 2007). As a result, visualization of exon-intron structures is critical for identifying potential annotation errors and determining possible intron evolution paths.

Lastly, the visualization of genes with complex structures needs further development, particularly for trans-splicing genes. Trans-splicing is a phenomenon that connects the exons across long distances and on different strands to form mature gene transcripts (Lasda & Blumenthal, 2011). Trans-splicing genes are one of the most error-prone features in chloroplast genome annotation. Visualizing trans-splicing genes can help control the annotation process and lay the basis for further understanding the mechanism and evolution of gene trans-splicing in chloroplast genomes.

To overcome these limitations, we developed a software package (Chloroplast Genome Viewer, CPGView) for the graphical representation of nongenic features and features below the gene level. CPGView contains three modules, which draw three gene maps showing (i) gene, repeat, and variable site distributions; (ii) the exon and intron structures for cis-splicing genes; and (iii) the detailed structure of the trans-splicing gene *rps12*. In particular, we developed a coordinate scaling algorithm (CSA) to solve the overlapping-label and one-page layout problems. We then tested CPGView with 31 newly sequenced and annotated chloroplast genomes. Lastly, CPGView was used to identify erroneous sequence annotations from 5998 chloroplast genome sequences released in GenBank. Overall, CPGView is the only chloroplast genome visualization tool that shows detailed genome structures and will become an indispensable tool for chloroplast genome research.

2 | MATERIALS AND METHODS

2.1 | Implementation of CPGView

CPGView has a command-line version and a web version. The command-line version was developed with PYTHON version 3.6.5 and R version 4.0.3 and has been packaged into a singularity container. The web version was developed with the Perl MVC framework. The third-party package CHLOROPLOT (Zheng et al., 2020) was modified to generate the general gene distribution map. We used several third-party packages including BIOPYTHON version 1.7.8 (Cock et al., 2009), GGPLOT2 version 3.3.3 (Wickham, 2011), and GGGENES version 0.4.1 (https://CRAN.R-project.org/package=gggenes) to draw cis- or trans-splicing gene maps. The repeat analysis pipeline was developed based on the following third-party packages: VMATCH version 2.3.0 (Kurtz, 2003), MISA version 1.0 (Beier et al., 2017), and TRF version 4.0.9 (Benson, 1999). The web version of CPGVIEW was successfully tested on commonly used browsers (e.g., Internet Explorer version 11.0, Firefox version 65.0, and Chrome version 72.0).

2.2 | Plant materials

We collected leaves from 31 plants freshly obtained from Central China Medicinal Botanical Garden (Enshi, HuBei, China, 109.750241°E, 30.175262°N) and Beijing Medicinal Plant Garden, Institute of Medicinal Plant Development (IMPLAD, Beijing, China, 116.27636°E, 40.035036°N). The plant samples were identified by Professors Zhao Zhang and Linfang Huang of IMPLAD. The samples were deposited to the Herbarium of IMPLAD. Detailed information of the plant materials is shown in Table S1.

2.3 | DNA extraction and sequencing

We extracted total genomic DNA by using a plant DNA extraction kit (Tiangen Biotech). In total, 1 μg of DNA from each plant sample was used for library construction with a template size of 500 bp.

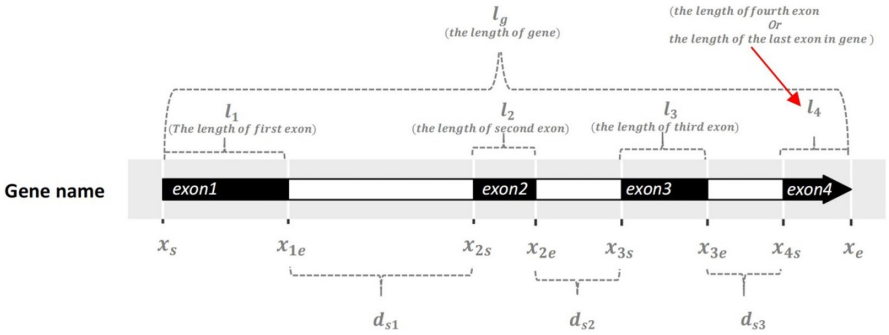

FIGURE 1 Definition of symbols used in CSA. l_n, the length of exon n; x_{is}, the start position of the i-th exon; x_{ie}, the end position of i-th exon; x_s, the start position of exon 1; x_e, the end position of the last exon; and d_{si}, the distance between i-th and (i + 1)-th exons.

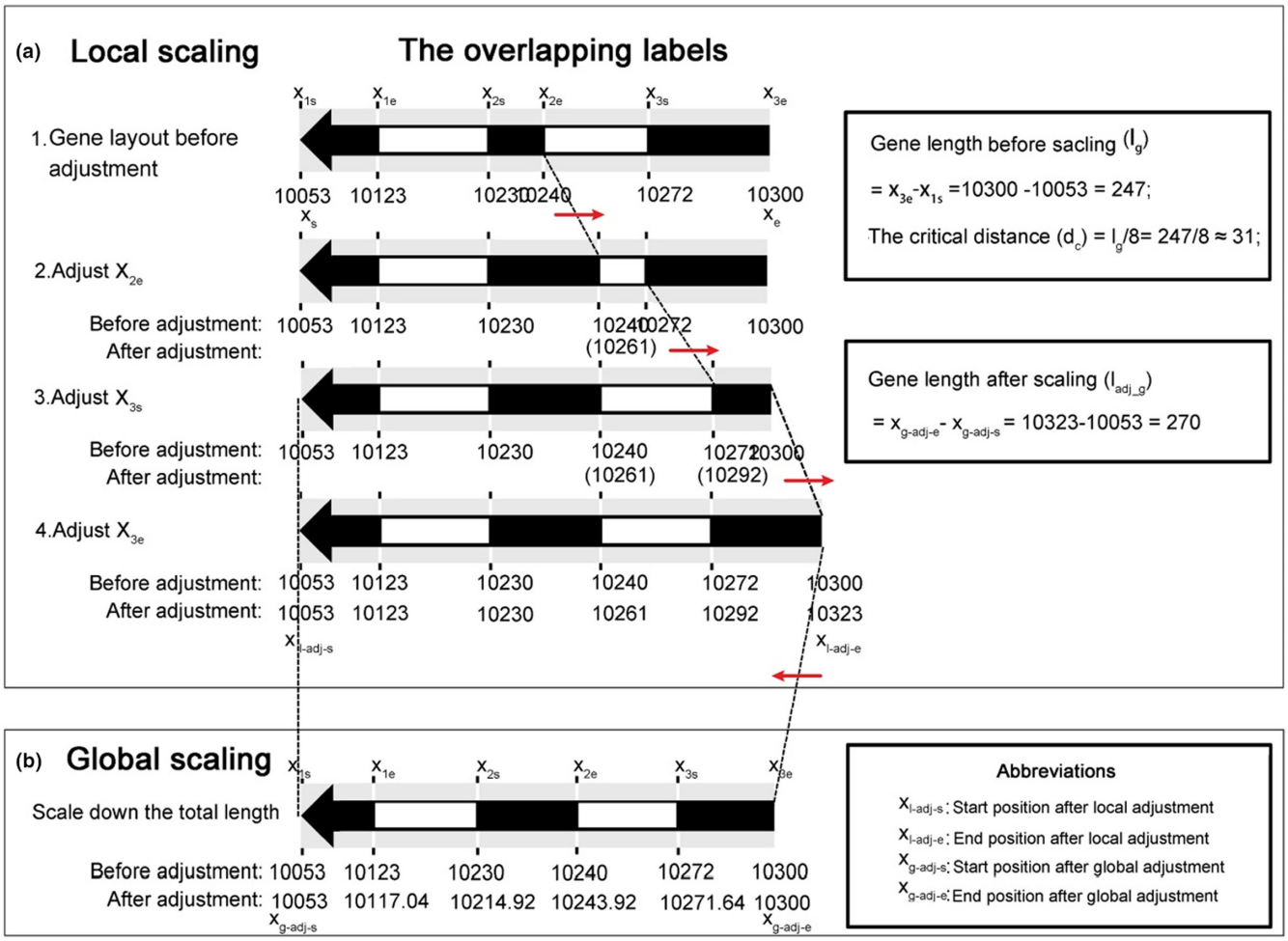

FIGURE 2 Graphic representation of the CSA. (a) Local scaling process. (b) Global scaling process. The numbers shown in black are the original and adjusted positions. The red arrows indicate the scaling direction of particular boundaries.

Each library was subjected to pair-end sequencing with the fragment size of 100 bp using a Hiseq 2500 platform (Illumina).

2.4 | Genome assembly and annotation

We generated a total of 5G data for each sample. We assembled the chloroplast genomes using GetOrganelle (version 1.6.4) (Jin et al., 2020) with the parameters "-R 15 -k 21,45,65,85,105 -F embplant_pt". GEPARD (version 1.40) software (Krumsiek et al., 2007) was used to select chloroplast genome conformations and draw dotplot pictures. The starting point of the chloroplast genome was adjusted using SEQKIT (version 0.12.1) (Shen et al., 2016), and the depth of coverage was detected using BWA (version 0.7.17) (Li & Durbin, 2010) and SAMTOOLS software (version 0.1.19) (Li et al., 2009). The reads mapped to the exon/intron junctions were identified using the TOPHAT software (Kim et al., 2013). We annotated the chloroplast genomes using CPGAVAS2 (Shi et al., 2019). We then manually

edited the annotation results by using APOLLO (version 1.11.8) software (Lewis et al., 2002) according to the multiple sequence alignment of homologous sequences.

2.5 | Detecting potential erroneous chloroplast genome records in GenBank

We downloaded the GenBank files for a total of 5998 chloroplast genomes in March 2021. These genomes were then analysed with CPGView. The Genbank accession numbers and taxonomic classifications of the 5998 chloroplast genomes are shown in Table S2.

3 | RESULTS

3.1 | Algorithms for drawing splicing genes

We encountered three problems in developing the package. The first problem was overlapping labels for the exon-intron boundary positions when drawing cis-splicing genes with small exons. The second problem was that the blank space for the graph must be minimized, so the entire graph occupied only one page for publication readiness. The third problem was selecting a set of complex structure models for the trans-splicing gene *rps12*. The algorithms to solve these problems are discussed below.

3.1.1 | Developing a coordinate scaling algorithm for drawing cis-splicing genes

In the cis-splicing gene map, we label the exon-intron boundaries with the start and end positions of the exons. However, the labels often overlap for genes with small introns (*rpl*16 gene, Figure S1A) and exons (*pet*D gene, Figure S1B). In addition, chloroplast genomes commonly have 12 cis-splicing genes: *atp*F, *clp*P, *ndh*A, *ndh*B, *pet*B, *pet*D, *rpl*16, *rpl*2, *rpo*C1, *rps*12, *rps*16, and *ycf*3. The structures of these genes need to be placed in one page. We developed CSA to solve the overlapping-label and one-page layout problems, which contain two scaling processes. The first process (local scaling) scales individual exons so the boundary labels do not overlap. The second process (global scaling) scales down all exons when the gene length is increased after local scaling.

We explain the meanings of all the symbols graphically in Figure 1. The local scaling and global scaling processes are described in steps 1–4 (Figure 2a) and in Figure 2b, respectively.

1. The length of a gene (l_g) is calculated, which equates the end position of the last exon (x_{ne}) minus the start position of exon 1 (x_{1s}).
2. After multiple trials, when the length of an exon or an intron is greater than one-eighth of the gene length, no position overlap occurs. Consequently, we set the critical distance (d_c) as $\frac{l_g}{8}$.
3. If the length of exon 1 ($l_1 = x_{1e} - x_{1s}$) is less than the critical distance (d_c), we scale the exon size by adding a difference of ρ_1 ($\rho_1 = d_c - l_1$) to obtain a length of d_c (Figure S2A). If l_1 is greater than or equal to d_c, then we do not change the end position of exon 1 (Figure S2B).
4. We examine the distance between exons 1 and 2 (d_{s1}). If d_{s1} is greater than or equal to d_c, then the start position (x_{2s}) of exon 2 does not need to be changed (Figure S3A). If $x_{1e} > x_{2s}$, or, d_{s1} is <0, we add a difference ρ_2 $(\rho_2 = d_c + |d_{s1}|)$ to x_{2s} (Figure S3B). If d_{s1} is greater than or equal to 0 and is less than d_c, then we add a difference of ρ_3 ($\rho_3 = d_c - d_{s1}$) to x_{2s}, bringing d_{s1} to at least d_c (Figure S3C). Subsequently, x_{2e} is adjusted as described in step 3. If the length of exon 2 (l_2) is less than d_c, we add a difference ρ_4 $(\rho_4 = d_c - l_2)$ to x_{2e}. After this step, x_{2s} and x_{2e} are adjusted to satisfaction.
5. This step is conducted when the exon number is greater than 2. It is an iteration of steps 3 and 4. This step is explained in detail using the *ndh*B gene as an example (Figure S4).
6. The locally adjusted length of the gene (l_{adj_g}) may be greater than l_g. Here, l_{adj_g} equates the adjusted end position of the last exon ($x_{l\_adj\_ne}$) minus the start position of exon 1 (x_{1s}). A global scaling process is developed to solve this problem (Figure 2b). If l_{adj_g} is greater than l_g, we use Formulas (van Dijk et al., 2014) and (Sandhya et al., 2020) to adjust the start and end positions of all exons:

$$\frac{q - x_{1s}}{l_{adj\_g}} = \frac{q_{adj} - x_{1s}}{l_g} \quad (1)$$

$$q_{adj} = \frac{l_g q - l_g x_{1s} + l_{adj\_g} x_{1s}}{l_{adj\_g}} \quad (2)$$

In Formula (1), q is the start or end position of an exon after local scaling. In particular, x_{1s} is the start position of exon 1, l_g is the initial gene length, l_{adj_g} is the locally adjusted gene length, and q_{adj} is the adjusted boundary position. Formula (2) is a transformation of Formula (van Dijk et al., 2014).

3.1.2 | Three-exon and two-exon models for drawing trans-splicing gene *rps12*

The *rps*12 gene has been reported as a trans-splicing gene in the chloroplast genomes. We analysed the *rps*12 genes from the publicly released chloroplast genomes and found that most *rps*12 genes contained either three or two exons.

To determine if the three-exon or two-exon model and experimental evidence support the two-exon models of the *rps*12 genes, we searched public databases for RNA-seq data. These data were then mapped to the *rps*12 gene sequences. We successfully found RNA-seq data for *Glycine max* (SRA accession number: SRR8447156) and *Cicer arietinum* (SRA accession number: SRR15808164). We then mapped the reads to the chloroplast genomes of *G. max*

(NC_007942.1) and *C. arietinum* (NC_011163.1). The mapping results supported that the rps12 gene had three exons transcribed into two transcripts in *G. max* (Figure S5). By contrast, the rps12 gene had two exons transcribed into two transcripts in *C. arietinum* (Figure S6).

Depending on whether or not *rps*12 exons are in the IR regions, there are four configurations for the *rps*12 genes. In configuration one, *rps*12 genes have three unique exons. Two are duplicated as they are located in the IR regions (Figure 3a). In configuration two, *rps*12 genes have two unique exons. One of them is duplicated as it is located in the IR regions (Figure 3b). In configuration three, *rps*12 genes have three unique exons. None of them are duplicated (Figure 3c). In configuration four, *rps*12 genes have two unique exons. None of them are duplicated (Figure 3d). The structures of *rps*12 genes in *G. max* and *C. arietinum* chloroplast genomes belong to configurations one and four, respectively.

3.2 | Overall architecture and workflow

The architecture and analysis workflow of CPGView are shown in Figure 4 and can be summarized as "Input", "CPGView analysis modules," and "output". CPGView takes two types of inputs and one type of optional input (VCF format file). The first type is the annotation results in the GenBank format. The second type is a GenBank accession number. When an accession number is provided, CPGView will download the GenBank file. Once the GenBank file is downloaded, it will be processed following the same workflow as the users submit the Genbank input file.

The analysis process of CPGView can be divided into six steps. In step 1, CPGView parses CDS information in the GenBank file, applies CSA, and draws the cis-splicing gene map. In step 2, CPGView parses the *rps*12 CDS information in the input file, compares the results with the four models, and selects the suitable one to draw the trans-splicing gene map. In step 3, CPGView identifies the repeat elements, removes the overlapping repeats, and plots the distributions of the repeat sequences. In step 4, CPGView extracts all the gene information and draws the gene distribution map. Step 5 is optional when the users provide an input file containing the site variation information in VCF (Danecek et al., 2011) format. CPGView extracts the alternative alleles and their frequencies and plots the frequencies of the alternative bases in a line chart. At the end of steps 3, 4, and 5, CPGView combines the three maps: the gene distribution map, the repeat distribution map, and the optional site variation maps into a single circular map called cpgenome_r. In step 6,

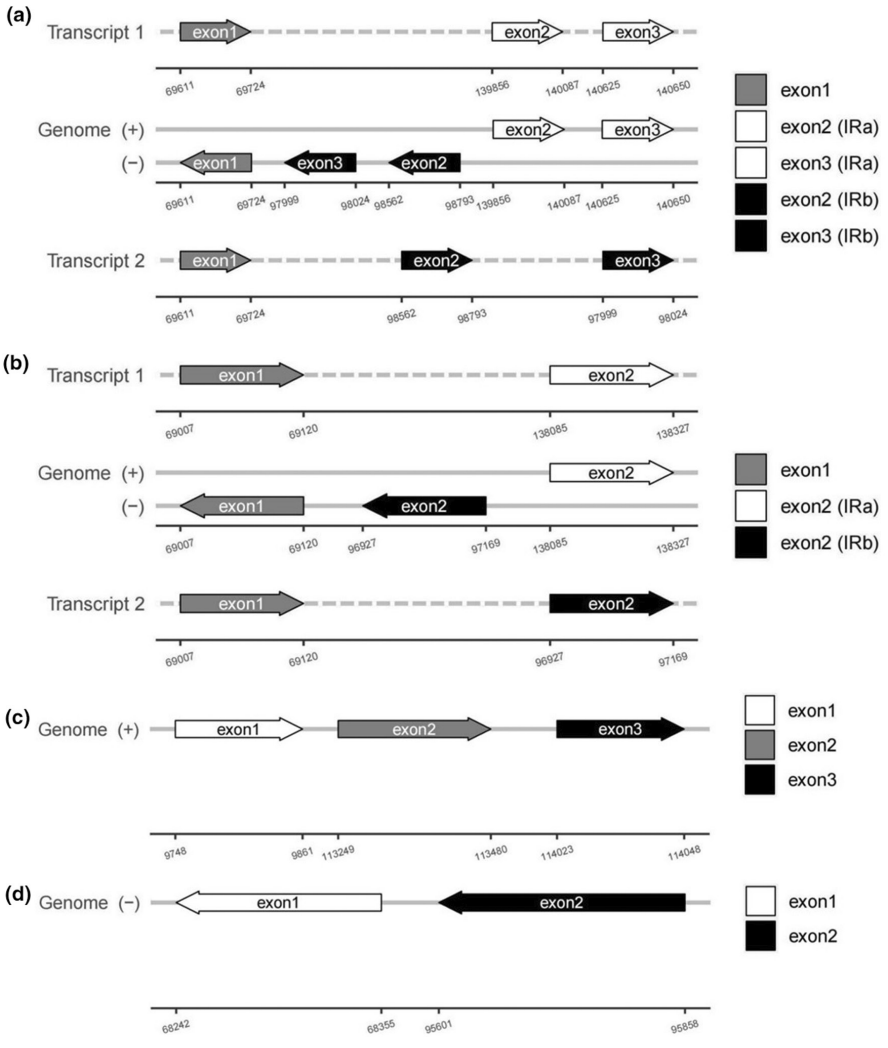

FIGURE 3 Schematic maps showing the trans-splicing gene *rps*12 in four different configurations. (a) Shows the *rps*12 gene structure in the *Arabidopsis thaliana* chloroplast genome. It has three unique exons. Two of them are duplicated as they are located in the IR regions. (b) Shows the *rps*12 gene structure in *Clerodendranthus spicatus*. It has two unique exons. One of them is duplicated as it is located in the IR regions. Each graph has three panels arranged from top to bottom, which display the structure of transcript 1, genome, and transcript 2, respectively. In the transcript panel, the line represents pre-mRNA. (c) Shows the *rps*12 gene structure in *Picea abies*. It has three unique exons. None of them are duplicated as they are not in the IR regions. (d) Shows the *rps*12 gene structure in *Cicer arietinum*. It has two unique exons. None of them are duplicated as they are not in the IR regions. The start and end positions on the pre-mRNA are shown below the line. The lines represent the genome plus (+) and minus (−) DNA strands. The arrowheads represent the corresponding exons of the *rps*12 genes.

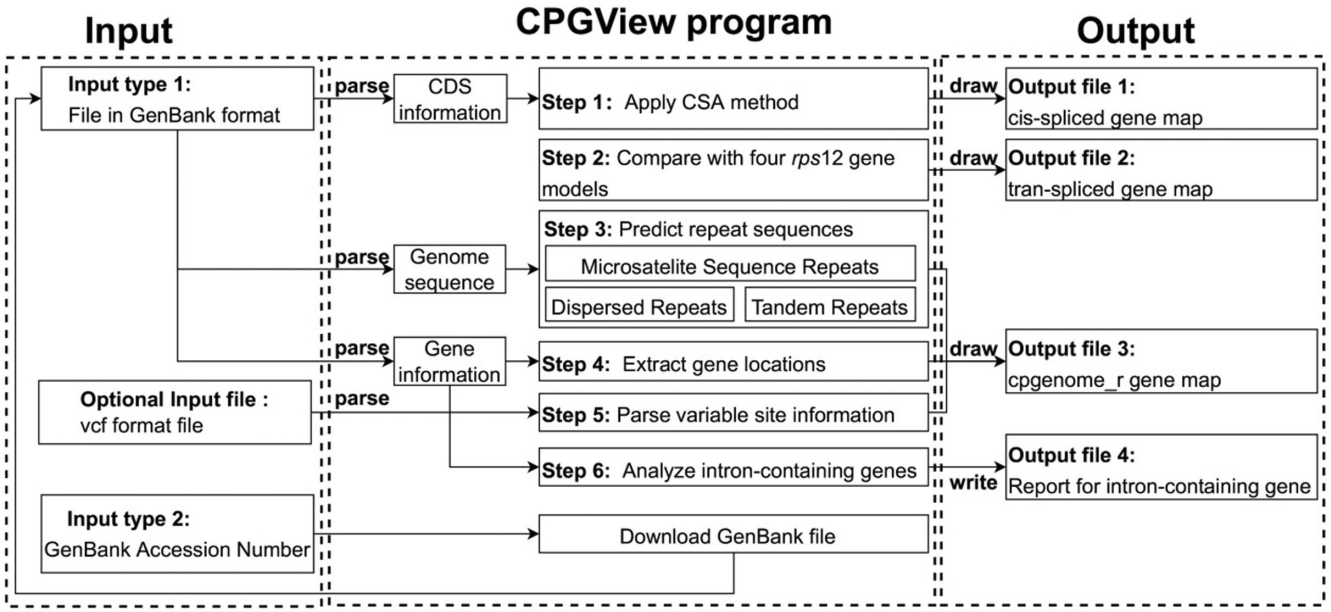

FIGURE 4 Overall of CPGView architecture and analysis workflow. The input data, drawing or analysis modules, and output files are shown on the left, in the middle, and on the right, respectively. The two types of input data are labelled with input types 1 and 2, respectively. Input type 1 is a file in GenBank format. Input type 2 is a GenBank accession number. The users can optionally provide a file in VCF format when the input type is 1. The six steps of the analysis process of CPGView are labelled with steps 1–6. The four output files are labelled as output files 1–4.

CPGView generates a report on intron-containing genes by using the same information used to draw the cis-splicing and trans-splicing gene maps.

3.2.1 | Drawing the cis-splicing gene map

CPGView first finds the coding sequence (CDS) in the GenBank file to create a cis-splicing gene map. The position information of the extrons and introns is extracted from the clauses with the "join" keyword. The keyword "join" indicates that the protein-coding sequence consists of multiple DNA fragments (exons). The clause might contain another keyword, "complement," which indicates that the protein-coding sequences are on the complementary strand (Rose, 2019). CSA is run to adjust the boundary positions so that the displayed coordinate labels will not overlap. The adjusted boundary information is then passed to gggenes for plotting.

3.2.2 | Drawing the trans-splicing gene map

CPGView will first determine how the exons of *rps12* are spliced to form the protein-coding sequence. If the exons in the protein-coding sequences are not arranged in the same order on the genome, then the *rps12* gene is considered trans-spliced. For example, in the GenBank file for the chloroplast genome of *Arabidopsis thaliana*, we found the following clause for the *rps12* gene: "join (97999...98024, 98562...98793, 69611...69724)." The numbers correspond to the positions on the chloroplast genome and indicate that the *rps12* gene has three exons (97999...98024), (98562...98793), and (69611...69724). They are connected in this order in the protein-coding sequence. Here, the end position of exon 1 (98024) is < the start position of exon 2 (98562), and the end position of exon 2 (98793) is > the start position of exon 3 (69611). This arrangement indicates that the three exons in the protein-coding sequence are not connected following their orders on the genome. Furthermore, parsing the GenBank file suggests that the two exons of *rps12* are both in the IRa and IRb regions. Thus, *rps12* is determined to be a trans-splicing gene. CPGView will transform this information into the corresponding three-exon or two-exon models and pass the information to gggenes for plotting.

3.2.3 | Drawing the cpgenome_r gene map

CPGView generated a circular map showing the distributions of all genes and repeats along the genome (cpgenome_r) with two modules. The first module was from the Chloroplot package with modifications. In particular, we adjusted the location of the track showing the LSC, SSC, and IR regions. In addition, the track showing the GC contents shrank to provide space for the track showing the variable sites (see below). The second module was modified from the repeat generation module from CPGAVAS2. In particular, the overlapped dispersed repeats were combined. The cpgenome_r module integrates all the structure information, including protein-coding genes, tRNA genes, rRNA genes, microsatellites, tandem repeats, and dispersed repeats, to draw the cpgenome_r map.

If the users provided the information for polymorphic sites or RNA editing site in VCF format, the frequencies of alternative bases at different sites were drawn in the cpgenome_r map. Given that the site variation information might not be available for all genomes under study, this function is only optional. The VCF file could be generated with software such as GATK (Nielsen et al., 2011). However, identification of polymorphic and RNA-editing sites was beyond the scope of this study and will not be discussed in detail here.

3.2.4 | Generating a genome report

CPGView finds the labels (CDS, rRNA, and tRNA) in the input file and extracts the location of those genes. All genes were summarized according to their labels and whether they contain introns. This report will help users quickly understand the contents of their submitted chloroplast genomes.

3.2.5 | Output

The output of CPGView includes three graphs: a cis-splicing gene map, a trans-splicing gene map, and a general gene distribution map cpgenome_r. An example of the cis-splicing gene map is shown in Figure 5. The structure follows the conventions depicting gene structures and should be self-explanatory. The graph's height is automatically changed depending on the number of genes displayed because different numbers of cis-splicing genes are present for various chloroplast genomes. In this particular example, 13 cis-splicing genes are shown.

An example of a trans-splicing gene map is shown in Figure 3. CPGView can draw the trans-splicing gene based on the three-exon (Figure 3a) and two-exon models (Figure 3b) with duplicated exons. CPGView can also draw the trans-splicing gene based on the three-exon (Figure 3c) and two-exon models (Figure 3d), with no duplicated exons.

An example of the circular map (cpgenome_r) is shown in Figure 6. The circular map contains seven circular tracks. From the centre going outward, the first three tracks show the distributions of the repeat sequences, which include distributed (Track A), tandem (Track B), and microsatellite (Track C) repeat sequences. The distribution of alternative bases is shown on Track D when the users provide the corresponding variable site information. The small single-copy (SSC), inverted repeat (IRa and IRb), and large single-copy (LSC) regions are shown on Track E. The GC contents are shown on Track F. Lastly, the distribution of genes is shown on Track G.

3.3 | Testing CPGView using 31 newly generated chloroplast genome sequences

To determine if CPGView can generate the graphic maps correctly, we sequenced, assembled, and annotated 31 chloroplast genomes. We did not use the chloroplast genome records from public databases for the validation because we could not be sure that chloroplast genome records were free of assembly and annotation errors. Instead, by generating new chloroplast genomes for testing, we adopted measures to ensure that these chloroplast genomes were correctly assembled and annotated. In particular, we mapped the reads to the assembled genomes. The assembly quality could then be determined based on the coverage depth (data not provided). In addition, we conducted multiple sequence alignment of all annotated genes with their homologous sequences. The results are shown in Appendix S1. Manual examination of the alignment supported that the annotations were free of errors.

The 31 chloroplast genomes belong to 31 genera from 22 families. The detailed information on the 31 chloroplast genomes, including the taxonomic classification of the source samples, the collection sites, and the sample identifiers, can be found in Table S1. The annotation results in the GenBank format were then subjected to visualization using CPGView.

CPGView generated 96 maps, including 31 general gene distribution maps, 31 cis-splicing gene maps, and 31 trans-splicing gene maps. We manually checked the correctness of the plotted maps. The exon and intron information for the 31 chloroplast genomes is shown in Table S3. The structure information of the *rps*12 genes is shown in Table S4. The output maps for the 31 chloroplast genomes are shown in Figures S7–S37. The circular cpgenome_r maps, cis-splicing gene maps, and trans-splicing gene maps are shown in panels A, B, and C of Figures S7–S37. Manual examination shows that the intron locations in the annotation files are consistent with those in the cis-splicing maps (panel B of Figures S7–S37). In addition, the structures of the rps12 gene shown in the annotation files are consistent with those in the corresponding trans-splicing maps (panel C in the Figures S7–S37). In summary, CPGView plotted the gene maps that were 100% consistent with those in the annotation files.

3.4 | Identifying problematic chloroplast genome records in public databases

At the time of this study, 5998 chloroplast genomes were available in GenBank. These genomes were from 2513 genera belonging to 553 families. We downloaded the 5998 chloroplast genomes and analysed them with CPGView. Considering the uncertainty of the error rate in these records, we did not use them to validate CPGView. Nevertheless, studying them with CPGView gave us some general ideas of the robustness of CPGView and allowed us to estimate the potential error rate in the GenBank record.

Out of these 5998 chloroplast genomes, CPGView generated maps for 5884 genomes and failed to generate any maps for the remaining 116 genomes (Table S5). Further analysis of the 116 chloroplast genomes showed that 46 had multiple "N," meaning that they had at least one gap. The 19 other chloroplast genomes had degenerate bases. We are currently determining why CPGView failed to generate maps for the remaining 49 chloroplast genomes. On the basis of these results,

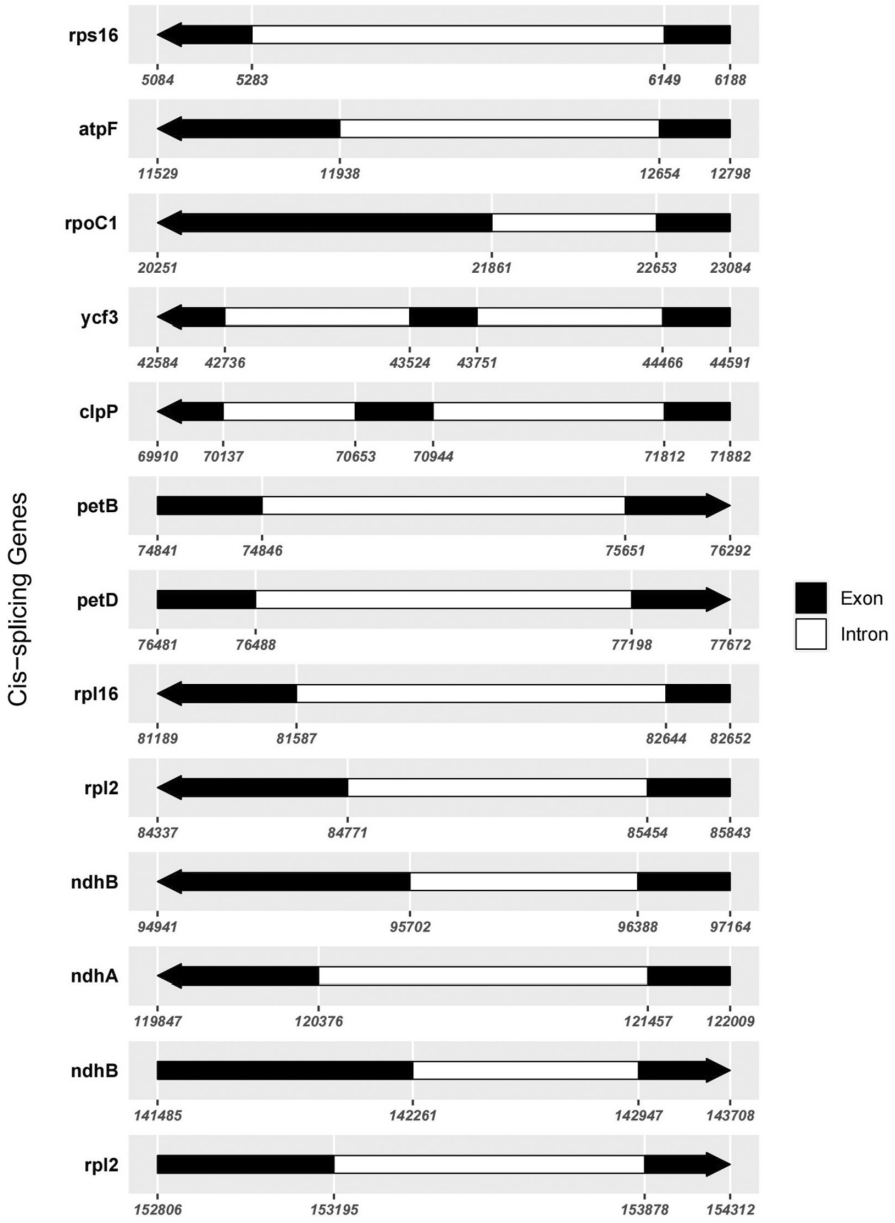

FIGURE 5 Schematic of the cis-splicing gene map generated for the chloroplast genome of *Arabidopsis thaliana*. The genes are arranged from top to bottom based on their order on the chloroplast genome. The gene names are shown on the left, and the gene structures are on the right. The exons are shown in black; the introns are shown in white. The arrow indicates the sense direction of the gene. Please note that lengths of exons and introns are not drawn to scale.

CPGView can plot 98% of all chloroplast genomes in GenBank. For those genomes that CPGView failed to create maps for, 1% of the 5998 chloroplast sequences appear to contain errors.

4 | DISCUSSION

With the rapid development of high-throughput DNA sequencing technology and bioinformatic software tools, the assembly and annotation of the chloroplast genome have become straightforward tasks. However, identifying the annotation errors remains challenging. The presence of sequences with errors in the database will propagate quickly, leading to the potential expansion of the errors. One effective method to counter this problem is to provide a visualization tool to help the users examine the detailed genome structures and identify any possible genome errors quickly.

Here, we developed CPGView, a visualization tool that can display the detailed chloroplast genome features, such as repeats, variable sites, cis-splicing genes, and trans-splicing genes. CPGView plotted the structures for 5884 out of these 5998 (98%) chloroplasts genomes, demonstrating its robustness. To our knowledge, CPGView is the only software package to draw the general gene distribution and the detailed structures of the cis-splicing and trans-splicing genes simultaneously. Furthermore, it can plot the repeat sequences and the frequencies of alternative bases at different sites.

In our validation experiment, CPGView failed to draw graphs for 116 out of 5998 (~2%). About 1% had sequence errors, such as the presence of "N" and degenerate nucleotides. One possible reason for the other 1% of chloroplast genomes that CPGView failed to plot maps is that these genomes contained additional errors. Additional optimization steps might be needed for CPGView. One of the reasons to develop CPGView is that the potential errors in the annotations can be

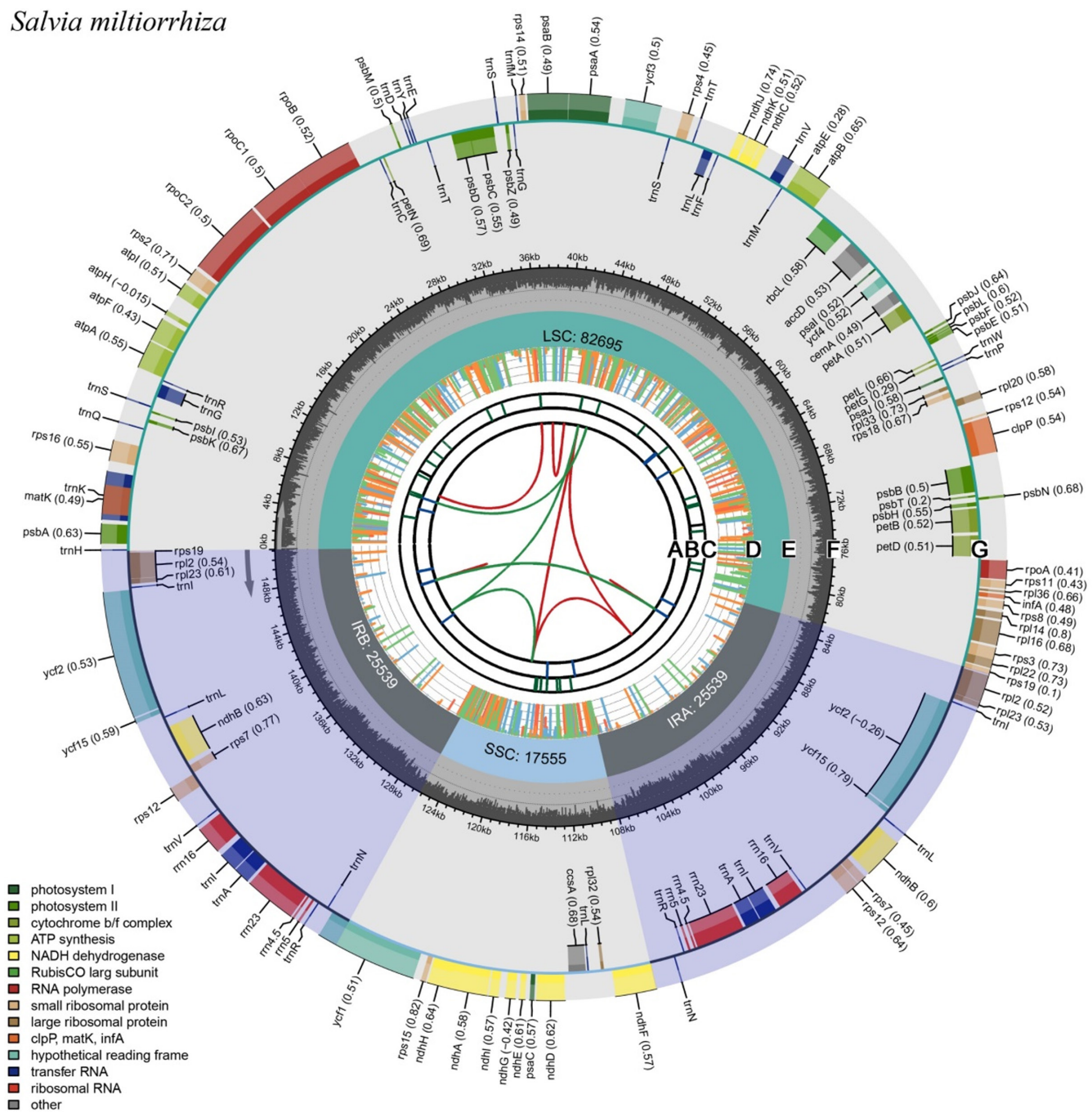

FIGURE 6 The cpgenome_r gene map of the *Salvia miltiorrhiza* chloroplast genome (NC_020431.1). The species name is shown in the left top corner. The map contains six tracks. From the centre outward, the first track (A) shows the forward and reverse repeats connected with red and green arcs. The second track (B) shows the tandem repeats as short blue bars. The third track (C) shows the microsatellite sequences also as green and yellow short bars. The line chart on the fourth track (D) shows the frequencies of alternative bases at particular sites. The frequencies of bases "A," "G," "C," and "T" are represented with red, blue, orange, and green lines, respectively. The length of the lines represents the substitute frequencies. The small single-copy (SSC), inverted repeat (IRa and IRb), and large single-copy (LSC) regions are shown on the fifth track (E). The GC content along the genome is plotted on the sixth track (F). The genes are shown on the seventh track (G). The optional codon usage bias is displayed in the parenthesis after the gene name. Genes are colour-coded by their functional classification. The transcription directions for the inner and outer genes are clockwise and anticlockwise, respectively. The functional classification of the genes is shown in the left bottom corner.

identified easily. We plan to examine these annotations further to determine possible errors or unique structures by either reassembling the genomes from the raw data or re-annotating the assembled genomes.

Several areas can be improved for CPGView in the future. First, additional modules might be needed to draw unique structures of complex genes, such as the *rps*12 genes. Our early study found that

most *rps*12 genes have either three or two unique exons, with or without duplication. However, some *rps*12 genes might have a structure that does not fit into these four models. When new structures of these genes are observed, we shall expand CPGView to draw the corresponding structures.

Second, the R modules appear less efficient than those implemented in other computational languages to draw the circular gene map. CPGView is significantly slower than several other comparative tools, such as GeSeq. Although this feature should not be a problem because speed is not critical for this type of work, future optimization of CPGView is needed.

Third, CSA can be further optimized. For example, global scaling might reduce the distance between the boundaries to be less than the critical values. Our preliminary analysis suggests that this problem might occur when a gene has more than seven exons. Although for chloroplast genes, very few genes, if any, have more than seven exons.

Fourth, alternative methods should be tested to solve the overlapping-label and one-page layout problems, while keeping the exon and intron lengths in proportion to their original lengths. CSA was developed for two reasons. Our original design for the cis-splicing and trans-splicing maps was to show the gene structures schematically for general examination, error detection, and publication. As a result, keeping the exon and intron lengths in proportion is not critical. In addition, the cis-splicing gene map and the trans-splicing map were drawn based on a third-party module gggenes. The module gggenes provides the functions we need, but it only supports drawing the labels right below the expected positions, causing overlapping labels. To continue using gggenes as our base module, we developed CSA. Alternative methods can be developed in the future to keep the exon and intron lengths in proportion.

Fifth, additional information can be added to the plot. For example, the substitution rates might vary among different genes or genomic regions. Plotting information such as substitution rates will help users identify correlation between them and particular genomic features. One problem to plot this information is the lack of well-accepted format describing the data. Thus, we have only implemented the function that can plot the site variation information.

Lastly, a quality check report is necessary to support the identification of unique features of a genome. This quality report should report the total numbers of protein-coding, rRNA, and tRNA genes; the list of lost genes and pseudogenes; and the presence of internal stop codons and alternative start codons. The combined use of the maps generated by CPGView and a quality check report will provide the users with specific directions to identify potential annotation errors.

AUTHOR CONTRIBUTIONS

Chang Liu conceived the study. Shengyu Liu developed the modules for drawing cis- and trans-gene maps. Chang Liu set up the webserver and modified the repeat identification module from the CPGAVAS2 package. Yang Ni performed the assembly, annotation, manual correction of the annotation results, and validation of the results generated by CPGView, for the 31 chloroplast genomes sequenced in this study. Jingling Li implemented the quality checking function. Xinyi Zhang implemented the drawing of the variable sites. Heyu Yang and Haimei Chen provided the raw data for implementing the variable site distribution. Yang Ni also tested CPGView with 5988 chloroplast genome sequences that were available in GenBank. All authors have read and agreed to the contents of the manuscript.

ACKNOWLEDGEMENTS

We thank Miss Lu Bai from Wake Forest University (NC, USA) for help in running the 5998 tests. We also thank Dr Shuyu Zheng from the University of Helsinki for help in debugging the Chloroplot programme.

CONFLICT OF INTEREST

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING INFORMATION

This study was funded by the Innovation Funds for Medical Sciences from the Chinese Academy of Medical Sciences, (CIFMS) (2021-I2M-1-022), the National Science Foundation of China Funds (81872966), and the National Science and Technology Fundamental Resources Investigation Programme of China (2018FY100705).

DATA AVAILABILITY STATEMENT

The genome sequence data that support the findings of this study have been made openly available in GenBank at https://www.ncbi.nlm.nih.gov under the accession no. (MZ636523–MZ636536, MZ636538–MZ636554). The associated BioProject, SRA, and Bio-Sample numbers are (PRJNA753690), (SRR15412841–SRR15412849, SRR15412851–SRR15412872), and (SAMN20691405, SAMN20691539–SAMN20691566, SAMN20691568–SAMN20691569), respectively. The source code for the command-line version can be obtained from https://github.com/Shengyu-Liu558/CPGView. The singularity container of CPGView can be obtained from the Figshare platform (https://figshare.com/articles/dataset/CPGView/20509272). The CPGView is publicly available under the GPL-2 license.

ORCID

Shengyu Liu https://orcid.org/0000-0002-5262-1744
Yang Ni https://orcid.org/0000-0002-3472-386X
Jingling Li https://orcid.org/0000-0002-1498-0260
Xinyi Zhang https://orcid.org/0000-0003-0060-5416
Heyu Yang https://orcid.org/0000-0003-3488-5487
Haimei Chen https://orcid.org/0000-0001-7100-5915
Chang Liu https://orcid.org/0000-0003-3879-7302

REFERENCES

Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISA-web: A web server for microsatellite prediction. *Bioinformatics*, *33*(16), 2583–2585.

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de MJL, H. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., De Pristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., & Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Huang, D. I., & Cronk, Q. C. B. (2015). Plann: A command-line application for annotating plastome sequences. *Applications in Plant Sciences*, *3*(8), 1500026.

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., de Pamphilis, C. W., Yi, T.-S., & Li, D.-Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1), 241.

Jung, J., Kim, J. I., Jeong, Y.-S., & Yi, G. (2018). AGORA: Organellar genome annotation from the amino acid and nucleotide references. *Bioinformatics*, *34*(15), 2661–2663.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), 1–13.

Krumsiek, J., Arnold, R., & Rattei, T. (2007). Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, *23*(8), 1026–1028.

Kurtz, S. (2003). The Vmatch large scale sequence analysis software. *Ref Type: Computer Program*, *412*, 297.

Lasda, E. L., & Blumenthal, T. (2011). Trans-splicing. *WIREs RNA*, *2*(3), 417–434.

Lei, W., Ni, D., Wang, Y., Shao, J., Wang, X., Yang, D., Wang, J., Chen, H., & Liu, C. (2016). Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of Astragalus membranaceus. *Scientific Reports*, *6*(1), 1–13.

Lewis, S. E., Searle, S. M. J., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglu, L., Birney, E., Crosby, M. A., et al. (2002). Apollo: a sequence annotation editor. *Genome Biology*, *3*(12), RESEARCH0082.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup GPDP. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., & Guan, X. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*, *13*(1), 715.

Lohse, M., Drechsel, O., Kahlau, S., & Bock, R. (2013). OrganellarGenomeDRAW—A suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research*, *41*(W1), W575–W581.

McKain, M. R., Hartsock, R. H., Wohl, M. M., & Kellogg, E. A. (2017). Verdant: Automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes. *Bioinformatics*, *33*(1), 130–132.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–451.

Rose, A. B. (2019). Introns as gene regulators: A brick on the accelerator. *Frontiers in Genetics*, *9*(672).

Sandhya, S., Srivastava, H., Kaila, T., Tyagi, A., & Gaikwad, K. (2020). Methods and tools for plant organelle genome sequencing, assembly, and downstream analysis. In M. Jain & R. Garg (Eds.), *Legume genomics: Methods and protocols* (pp. 49–98). Springer US.

Schmitz-Linneweber, C., & Barkan, A. (2007). RNA splicing and RNA editing in chloroplasts. In R. Bock (Ed.), *Cell and molecular biology of plastids* (pp. 213–248). Springer Berlin Heidelberg.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, *11*(10), e0163962.

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., & Liu, C. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Research*, *47*(W1), W65–W73.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq – Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*(W1), W6–W11.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*(9), 418–426.

Wickham, H. (2011). ggplot2. *WIREs Computational Statistics*, *3*(2), 180–185.

Wyman, S. K., Jansen, R. K., & Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, *20*(17), 3252–3255.

Zheng, S., Poczai, P., Hyvönen, J., Tang, J., & Amiryousefi, A. (2020). Chloroplot: An online program for the versatile plotting of organelle genomes. *Frontiers in Genetics*, *11*, 576124.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu, S., Ni, Y., Li, J., Zhang, X., Yang, H., Chen, H., & Liu, C. (2023). CPGVIEW: A package for visualizing detailed chloroplast genome structures. *Molecular Ecology Resources*, *00*, 1–11. https://doi.org/10.1111/1755-0998.13729