

Research paper

PlastidHub: An integrated analysis platform for plastid phylogenomics and comparative genomics



Na-Na Zhang (张娜娜)^{a,b,1}, Gregory W. Stull^{c,1}, Xue-Jie Zhang (张学杰)^{a,b},
Shou-Jin Fan (樊守金)^{a,b,*}, Ting-Shuang Yi (伊廷双)^{d,**}, Xiao-Jian Qu (曲小健)^{a,b,***}

^a College of Life Sciences, Shandong Normal University, Ji'nan 250358, China

^b Dongying Institute, Shandong Normal University, Dongying 257092, China

^c Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013, USA

^d Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China

ARTICLE INFO

Article history:

Received 13 March 2025

Received in revised form

6 May 2025

Accepted 18 May 2025

Available online 22 May 2025

Keywords:

Annotation

Comparative genomics

Plastid phylogenomics

Sequence processing

Visualization

ABSTRACT

The plastid genome (plastome) represents an indispensable molecular resource for studying plant phylogeny and evolution. Although plastome size is much smaller than that of nuclear genomes, accurately and efficiently annotating and utilizing plastome sequences remain challenging. Therefore, a streamlined phylogenomic pipeline spanning plastome annotation, phylogenetic reconstruction and comparative genomics would greatly facilitate research utilizing this important organellar genome. Here, we develop PlastidHub, a novel web application employing innovative tools to analyze plastome sequences. In comparison with existing tools, key novel functionalities in PlastidHub include: (1) standardization of quadripartite structure; (2) improvement of annotation flexibility and consistency; (3) quantitative assessment of annotation completeness; (4) diverse extraction modes for canonical and specialized sequences; (5) intelligent screening of molecular markers for biodiversity studies; (6) gene-level visual comparison of structural variations and annotation completeness. PlastidHub features cloud-based web applications that do not require users to install, update, or maintain tools; detailed help documents including user guides, test examples, a static pop-up prompt box, and dynamic pop-up warning prompts when entering unreasonable parameter values; batch processing capabilities for all tools; intermediate results for secondary use; and easy-to-operate task flows between file upload and download. A key feature of PlastidHub is its interrelated task-based user interface design. Give that PlastidHub is easy to use without specialized computational skills or resources, this new platform should be widely used among botanists and evolutionary biologists, improving and expediting research employing the plastome. PlastidHub is available at <https://www.plastidhub.cn>.

Copyright © 2025 Kunming Institute of Botany, Chinese Academy of Sciences. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Currently, analytical tools fall into two categories: web-based platforms and locally executed command-line scripts. The former is often user-friendly but typically limited in functionality,

requiring users to visit multiple platforms to fully process and analyze omics data. The latter rely on the operation of scripts/codes, which can enhance flexibility but demand computational expertise due to complexities in installation and operation. Therefore, integrating the strengths of both approaches through “one-stop” online platforms offers a solution to advance genomics research. This need is especially acute in plastome (plastid genomic) research, where few integrative analytical platforms are currently available.

Due to its relatively small genome size, conservative gene content, and moderate evolutionary rate, the plastome is an indispensable molecular data source that has been widely applied in phylogenomics, evolutionary biology, comparative genomics, population genetics, phylogeography, and genetic engineering

* Corresponding author. College of Life Sciences, Shandong Normal University, Ji'nan 250358, China.

** Corresponding author.

*** Corresponding author. College of Life Sciences, Shandong Normal University, Ji'nan 250358, China.

E-mail addresses: fansj@sdu.edu.cn (S.-J. Fan), tingshuangyi@mail.kib.ac.cn (T.-S. Yi), quxiaojian@sdu.edu.cn (X.-J. Qu).

Peer review under the responsibility of Editorial Office of Plant Diversity.

¹ These authors contributed equally to this work.

(Wicke et al., 2011; Jansen and Ruhlman, 2012; Straub et al., 2012; Jin and Daniell, 2015; Daniell et al., 2016; Li et al., 2019; Stull et al., 2021). Many specialized tools exist for plastome analyses. Assembly tools include SPAdes (Bankevich et al., 2012), IOGA (Bakker et al., 2016), ORG.Asm (Coissac et al., 2016), NOVOplasty (Dierckxsens et al., 2017), Organelle_PBA (Soorni et al., 2017), chloroExtractor (Ankenbrand et al., 2018), GetOrganelle (Jin et al., 2020), ptGAUL (Zhou et al., 2023), Oatk (Zhou et al., 2024), and TIPPO (Xian et al., 2025). Annotation tools include DOGMA (Wyman et al., 2004), CpGAVAS (Liu et al., 2012), Plann (Huang and Cronk, 2015), Verdant (McKain et al., 2017), GeSeq (Tillich et al., 2017), AGORA (Jung et al., 2018), CPGAVAS2 (Shi et al., 2019), and PGA (Qu et al., 2019). Visualization tools include OGDRAW (Greiner et al., 2019), CGView (Stothard et al., 2019), and CPGView (Liu et al., 2023). Phylogenetic tools include HomBlocks (Bi et al., 2018), ORPA (Bi et al., 2024), and OGU (Wu et al., 2024). Comparative genomic tools include VISTA (Frazer et al., 2004), progressiveMAUVE (Darling et al., 2010), IRscope (Amiryousefi et al., 2018), IRplus (Díez Menéndez et al., 2023), and CPJSdraw (Li et al., 2023). Despite these tools, there are still various technical issues regarding plastome sequence processing and utilization, including standardization of genome structure, evaluation of annotation quality, accurate extraction of target sequences, streamlined pre- and/or post-alignment sequence processing, automatic screening of molecular markers, and gene-level visual comparison of structural variations and annotation completeness. These issues are elaborated below.

First, accurate identification of the quadripartite structure of plastomes remains a challenge. The plastomes of most photosynthetic seed plants are highly conserved, with a quadripartite structure comprising large (LSC) and small (SSC) single-copy regions and two identical inverted repeat (IR) regions (Wicke et al., 2011; Jansen and Ruhlman, 2012). In some cases, two IR copies are not completely identical, due to the imperfections (mismatches) or heterogeneities introduced into genome sequences (Turudić et al., 2022; Díez Menéndez et al., 2023), suggesting there is still room for IR identification by current methods (Turudić et al., 2022). The standardization of quadripartite structure relies on precise IR boundary identification (Turudić et al., 2022). Current main plastome annotation tools, such as the web-based GeSeq (Tillich et al., 2017) and command-line PGA (Qu et al., 2019), incorporate the identification of IR copies, but lack quadripartite structure adjustment. Specialized tools like IRscope (Amiryousefi et al., 2018) and CPJSdraw (Li et al., 2023) focus solely on identifying IR boundary, with CPJSdraw resolving accurate LSC-IRa junctions; however, neither method produces standardized plastome sequences. Tools like NOVOWrap (Wu et al., 2021) and Plastaumatic (Chen et al., 2022) claim to have the functions of adjusting non-standardized form to standardized form. However, the “validate.py” script in NOVOWrap depends on reference sequences for adjustments, risking incorrect SSC orientation when using distantly related or unsuitable references. The “standardize_cpDNA.sh” bash script in Plastaumatic cannot be run without a Linux operating system.

Second, accurate annotation of whole plastomes remains a challenging task, despite the availability of multiple plastome annotation tools, including web tools such as GeSeq (Tillich et al., 2017), CPGAVAS2 (Shi et al., 2019), and Chloe (<https://chloe.plastid.org/annotate.html>), along the command-line tool PGA (Qu et al., 2019). Annotation principles of existing plastome annotation tools were reviewed and compared in Qu et al. (2023). Critical challenges include annotating pseudogenes, redundant genes, and the trans-splicing *rps12* gene. Pseudogenes are non-functional gene fragments derived from functional genes (Goodhead and Darby, 2015), arising from mutations like

premature stop codons or frameshifts (Xiao et al., 2016). Their presence complicates genome annotation (Qu et al., 2023), yet accurate identification of these evolutionary remnants is critical for investigating pseudogenization, novel gene origin, and structural rearrangements. However, no tool reliably annotates pseudogenes or evolutionary relics. Trans-splicing, unlike cis-splicing, joins exons derived from either distinct pre-mRNAs or duplicated regions within the same gene. The *rps12* gene—typically containing three exons (or occasionally two) distributed across two distinct pre-mRNAs—is prone to misannotation or incomplete annotation, as its fragmented exons make manual exon link particularly challenging (Tillich et al., 2017; Shi et al., 2019; Qu et al., 2023). Currently, only GeSeq and CPGAVAS2 have the ability to annotate *rps12*, but their accuracy and reliability require comprehensive validation. Although CPGview enables visual verification of *rps12* annotation accuracy, a comprehensive assessment of the hit rate and link accuracy of *rps12* in angiosperms and broader eukaryotes is lacking.

Third, tools for quantitatively evaluating plastome annotation quality remain underdeveloped. There are tools like BUSCO (Waterhouse et al., 2018) and CEGMA (Parra et al., 2007) that can assess nuclear genome assembly/annotation completeness, equivalent solutions for plastomes are absent. Our previous study proposed three methods for quantitatively evaluating plastome annotations, i.e., gene number comparison, gene length difference comparison, and gene sequence similarity comparison (Qu et al., 2023). However, no dedicated tools exist to implement these assessments. Additionally, gene-level visualization tools for direct comparison of annotation quality of plastomes are still lacking.

Fourth, accurate extraction of coding and noncoding sequences from plastomes is a complicated task. The genic composition of plastomes includes diverse elements: protein-coding (PCG) genes, RNA (tRNA and rRNA) genes, intron-containing vs. intron-lacking genes, cis-splicing vs. trans-splicing genes, non-overlapped/non-nested genes and overlapped/nested genes, single-copy vs. multi-copy genes, and genes spanning start-end position or the start-end position within the noncoding (intergenic/intronic) regions (Raubeson and Jansen, 2005; Qu et al., 2023). Despite the complexity, no dedicated tools offer customizable extraction modes to address these diverse demands.

Fifth, tools for pre- and post-alignment processing remain not integrated, and it is difficult to search customized tools to meet diverse or specialized needs. Although most land plants have conserved plastome structures, many lineages possess highly rearranged plastomes with disordered gene order (Knox, 2014; Mower and Vickrey, 2018; Qu et al., 2022), complicating the use of intergenic regions (Shaw et al., 2007). For plant taxonomists lacking programming expertise, tasks such as assessing codon status in PCGs, identifying missing taxa in sequence datasets, extracting codon sites or sub-alignments from alignments, generating alignments for PartitionFinder, and automatically screening markers based on sequence variation prove challenging. Therefore, the development of streamlined methods for pre- and/or post-alignment sequence processing are urgently needed.

Sixth, tools for micro-synteny comparison of plastomes are lacking. Given the limited number of genes in plastomes (Raubeson and Jansen, 2005), gene-level micro-synteny comparison is realistic and effective. This approach can reveal structural variations such as inversion, translocation, and duplication (Palmer, 1991; Mower and Vickrey, 2018), as well as annotation completeness issues such as gene loss/gain (Qu et al., 2023). Visualizing pairwise gene arrangements could further evaluate annotation quality at the gene level. However, specific tool remains underdeveloped for performing plastome-specific micro-synteny comparisons.

To address these critical challenges in plastome analysis, we developed PlastidHub, an online platform integrating an interconnected task-based user interface for end-to-end processing of plastome sequences. All tools in PlastidHub possess batch processing capability, critical for handling the large-scale datasets routinely analyzed by botanists and evolutionary biologists.

2. Materials and methods

2.1. Implementation

PlastidHub is a web application developed with Java, TypeScript, and Perl, designed to cater to users who favor browser-driven workflows or those lack familiarity with UNIX command-line tools. The front-end web interface was developed using vue3 (<https://cn.vuejs.org/>) and Element-Plus (<https://element-plus.org/>). PlastidHub is compatible with recently released modern browsers, including Google Chrome, Microsoft Edge, Mozilla Firefox, Opera, Apple Safari, 360 Speed Browser X, 360 Secure Browser, QQ Browser, and Huawei Browser. The back-end web framework uses SpringBoot v.3.3.2 (<https://docs.spring.io/spring-boot/index.html>) with RabbitMQ (<https://www.rabbitmq.com/>) as message queue and Redis (<https://redis.io/>) storing result status, which can manage data and parameters submitted by users, and then relay the analysis results (generated files) to the front-end interface for users to manage (batch selection, download, and deletion). The Perl-based (v.5.26.1) back-end architecture of PlastidHub supports execution across Windows, Linux, and Mac via command-line interfaces.

2.2. Analysis workflow

PlastidHub is an integrated plastome analysis platform including 10 tool kits: annotation, assessment, submission, visualization, extraction, pre-alignment, post-alignment, phylogeny, barcoding, genomics, each designed for specific tasks. Users can dynamically chain these toolkits to build customized workflows. For examples, “annotation → assessment”; “annotation → assessment → submission”; “annotation → assessment → visualization”; “annotation → assessment → extraction → pre-alignment → (post)-alignment → phylogeny”; “annotation → assessment → extraction → pre-alignment → (post)-alignment → barcoding”; and “annotation → assessment → plastomics”. Workflows are fully flexible, users can start or terminate analyses at any stage.

2.3. Supported file formats

PlastidHub supports GenBank (.gb/.gbf/.gbk), FASTA (.fasta/.fa/.fas/.fna), and TEXT (.txt) input files, as specified in the File Upload Panel of each tool. The format of output files is user-selectable and include GenBank, FASTA, BED (.bed), TABLE (.tbl), TAB-delimited (.txt), SVG (.svg), PNG (.png), Newick TREE (.tre), and CONFIG (.conf).

2.4. File number and size limitation

The file number and size allowed by PlastidHub are displayed in the static pop-up prompt box of each tool.

2.5. Special characters in filenames

Some functionalities in PlastidHub rely on third-party tools, so the use of non-English characters and certain special characters (~!@#%&*()+=[]{};'"<>/?|) is not recommended. In particular, we

recommend to use underscores (_) or dots (.) instead of spaces () or hyphens (–).

2.6. Help documents and notes

To facilitate use of this new platform, PlastidHub provides detailed user guides, test example files, and warning prompts when entering unreasonable parameter values. Help documents can be easily acquired from the document pages corresponding with each tool. Structured help documents can also be accessed in the supplemental information. Notes and warning prompts for each tool are shown in the tool interface. In general, use of unreasonable parameter values will prompt an error warning encouraging use of values within a specified range. Warnings will also appear in cases where files are incorrectly uploaded (e.g., due to wrong file format) or due to restrictions in file number or size. These resources streamline platform adaptation, enabling researchers to focus on biological insights.

2.7. Test data and tool evaluation and comparisons

To evaluate the performance of newly developed tools in PlastidHub, 15,206 eukaryotic plastome sequences were retrieved from the NCBI RefSeq database (<https://ftp.ncbi.nlm.nih.gov/refseq/release/plastid/>, accessed 2024.5.16). First, we count the proportion of non-standardized quadripartite structure across all 15,206 eukaryotic plastome sequences using the Quadripartition tool. Second, we evaluated the hit rate and link accuracy of *rps12* with the plastome annotation tool PGA v.2.0 across 15,206 eukaryotic and 13,731 angiosperm plastome sequences. To evaluate plastome annotation consistency of PGA v.2.0, the number of annotated genes in GenBank and PGA annotations were compared for 13,731 angiosperm plastid sequences. Third, we applied the tools of “Assess Gene Number” and “Assess Gene Length” to evaluate the annotation quality of the target species *Rosa roxburghii* using *Amborella trichopoda* as reference. Fourth, to evaluate the accuracy of the plastome Extraction tool, the number of extracted genes and PGA-annotated genes in the 13,731 angiosperm plastid sequences was compared. Fifth, to evaluate the effectiveness of the Barcoding tool in automatically screening molecular markers, plastomes from *Amaranthaceae* species including *Amaranthus caudatus*, *A. tricolor*, *Atriplex centralasiatica*, *Chenopodium album*, *C. quinoa*, *Salicornia bigelovii*, *S. europaea*, *Spinacia oleracea*, *Suaeda glauca*, and *S. salsa* were used. The choice of taxonomic level (e.g., order, family, genus, species, or mixed) does not affect barcoding results, as the final barcoding results are sorted by variable sites and parsimony informative sites—metrics that remain phylogenetically meaningful across all taxonomic scales. Thus, selection criteria remain consistent, regardless of whether the taxa are closely related or broadly sampled. Sixth, to evaluate the functionality (showing micro-syntenic structural variations and evaluating annotation completeness) of the Gene Homology tool, we conducted plastome comparisons of *Amborella trichopoda* against other angiosperms (*Rosa roxburghii* and *Salicornia bigelovii*) and a gymnosperm species (*Calocedrus decurrens*), as well as a comparison of the gymnosperm species *C. decurrens* and *C. rupestris*.

We conducted a series of comparisons with existing tools. For “Main Function 1”, we compared the consistency of quadripartite standardization between the Quadripartition tool in PlastidHub and the “validate.py” tool in NOVOWrap using 53 angiosperm plastomes from 48 orders. For “Main Function 2”, we have compared PGA v.1.0 with GeSeq using practical examples in our previous published paper (Qu et al., 2019). For further comparison of the newly developed function in PGA v.2.0, we use above

mentioned 53 angiosperm plastomes to compare the hit rate and link accuracy of *rps12* for PGA v.2.0 with GeSeq and CPGAVAS2. For “Main Function 4”, to demonstrate the accuracy and flexibility of the Extraction tool in PlastidHub, we use above mentioned 53 angiosperm plastomes to compare the accuracy of sequence extraction between PlastidHub and PhyloSuite. For “Main Function 3/5/6”, they are completely novel tools, without other existing tools to compare. Therefore, we use true data examples to demonstrate how these three tools combine simplicity and ease of use with accuracy and practicality in evaluating annotation quality, assisting primer design, and displaying micro-synteny structural variations.

For some PlastidHub tools that are similar to other existing popular tools, we have introduced them in the paragraph titled “Similar functions with other tools” within the Results and Discussion section, allowing users to choose their own preferences. For completing the streamlining analysis of plastid phylogenomics and comparative genomics, we have developed some personalized new functions that are not found in other published papers and introduced them in the paragraph titled “Streamlining analysis with personalized new functions in three toolkits: Pre-Alignment, (Post)-Alignment, and Phylogeny”.

2.8. Performance test

PlastidHub's scalability was benchmarked under simulated workloads, evaluating platform behavior across three escalating scenarios: light (20 users), moderate (60 users), and heavy (100 users) concurrent loads. In each scenario, virtual users randomly executed diverse tasks (e.g., annotation, extraction), submitting requests within a 10-s window to mimic real-world usage patterns.

3. Results and discussion

3.1. Interrelated user interface design of PlastidHub

To create an easy-to-use web-based tool kit for researchers engaged in phylogenomic and comparative genomic analyses using plastome data, we generated task-based interfaces where the users can easily choose a tool to accomplish their analysis goals (Fig. 1). PlastidHub contains three interfaces, i.e., “P-Utils” (tool interface), “Document” (help interface), and “My Files” (download interface).

For the “P-Utils” tool interface, the left panel provides links to task-running interfaces of 25 tools across 10 tool kits. The right panel displays a comprehensive list of all 25 tools in PlastidHub, enabling users to instantly navigate to any specific tool. Each tool has a brief introduction to help users to make a right choice. Within each tool's interface, there are interactive options for uploading input files (required), selecting parameter values (optional), submitting (required) or resetting (optional) the task, and linking to the help document for reference. To avoid submission errors, PlastidHub enforces file quantity limits, specified input formats (with permitted suffixes), and restricts parameters to predefined ranges. If the input files are omitted, the submit button remains disable, triggering a brief warning prompt (e.g., Please upload reference/target file first) for 3 s.

For the “Document” help interface, the left panel shows the links to toolkit-specific help documents for 10 tool kits. The right panel lists “Get Started”, “User Guide”, “Browser Compatibility”, “Tech Stack”, and “Content”. For the “Content” section, a hyperlink-based help document for each of 25 tools is listed. Within the help document of each tool, paragraph navigation links such as “General Introduction”, “Features”, “Command-line Tool”, and

“Example” are set up to help users to quickly locate the target paragraph.

An important customization feature of PlastidHub is the “My Files” download interface, allowing users to check the running status of the submitted tasks and download or delete the completed tasks at any time. Intermediate results are also provided to assist users to conduct custom post-processing. The validity period of the result files stored in the server database is one week. If the download button is clicked, a prompt such as “The result of output files will be stored for a week since task finished” will be displayed. Similarly, if the delete button is clicked, a warning prompt such as “Confirm that you are deleting the result (s)” will be displayed. Each tool kit has a corresponding download interface, and the tool interface is located below the corresponding download interface, making it convenient for users to access directly. It indicates that the tool interface can be freely switched not only with the help interface, but also with the download interface. This design enhances the flexibility of running tasks in PlastidHub.

The middle panel of the home page contains user access and usage statistics, including “Task Handled”, “Page Views”, and “Unique Visitors”, providing real-time feedback on user visits and completed tasks. The bottom panel of the home page contains a “Links” section, which includes “GitHub” and “Change Log”, and a “Community” section, which includes “Feedback”, “Get Involved”, and “Contact”, as well as a “Policy” section, which includes “Terms of Use” and “Privacy Policy”.

3.2. Functionalities of 25 tools within 10 designed tool kits

Full methodological details, use cases, and workflows for each tool are comprehensively documented in the supplementary file.

3.3. Main function 1: High proportion of non-standardized quadripartite structure of plastomes in NCBI RefSeq database identified by the Quadripartition tool

The quadripartition tool in PlastidHub (Fig. 2A) is newly developed based on an optimized IR-copy-identification method from PGA v.1.0 (Qu et al., 2019). Flexibility is our main consideration for adjusting quadripartition structure. Therefore, we allow the IRb and IRa to be set as completely (100%) or almost (99%) identical, the minimum allowed IR length to be changed manually (e.g., ≥ 1000 bp or ≥ 100 bp), the orientation of the whole plastome sequences or the SSC sequences to be reverse complemented or not, and the standardized plastome sequences as well as coordinates and sequences of LSC, IRb, SSC, and IRa to be outputted. This adaptability enables precise structural standardization, critical for resolving inconsistencies in RefSeq plastome annotations.

The consistency of quadripartite standardization was evaluated by comparing the Quadripartition tool in PlastidHub and the validate.py tool in NOVOWrap (Fig. 2B and Table S1). For 53 angiosperm plastomes, a total of 60 unique IRs were identified across four runs (two PlastidHub runs and two NOVOWrap runs), indicating inconsistent identification of IRs within and between tools. Specifically, there is 1 unique IR in each of PlastidHub run1/2, but 3 missing IRs and 3 unique IRs were identified in NOVOWrap run1/2, respectively. For PlastidHub run2, a total of 10 IRs with warnings “Nucleotide and/or length of two IR copies may not be exactly the same!” indicated different IRs that need to be checked (Table S1). Among these 10 different IRs, 4 are corrected to identical IRs by NOVOWrap run2, but 6 are failed to be corrected by NOVOWrap run2, indicating the inconsistency of NOVOWrap in IR identification. For 1 identical IR in PlastidHub run1/2 (*Pandanales_Pandanaceae_Pandanus_tectorius_NC_042747*), NOVOWrap run1/2

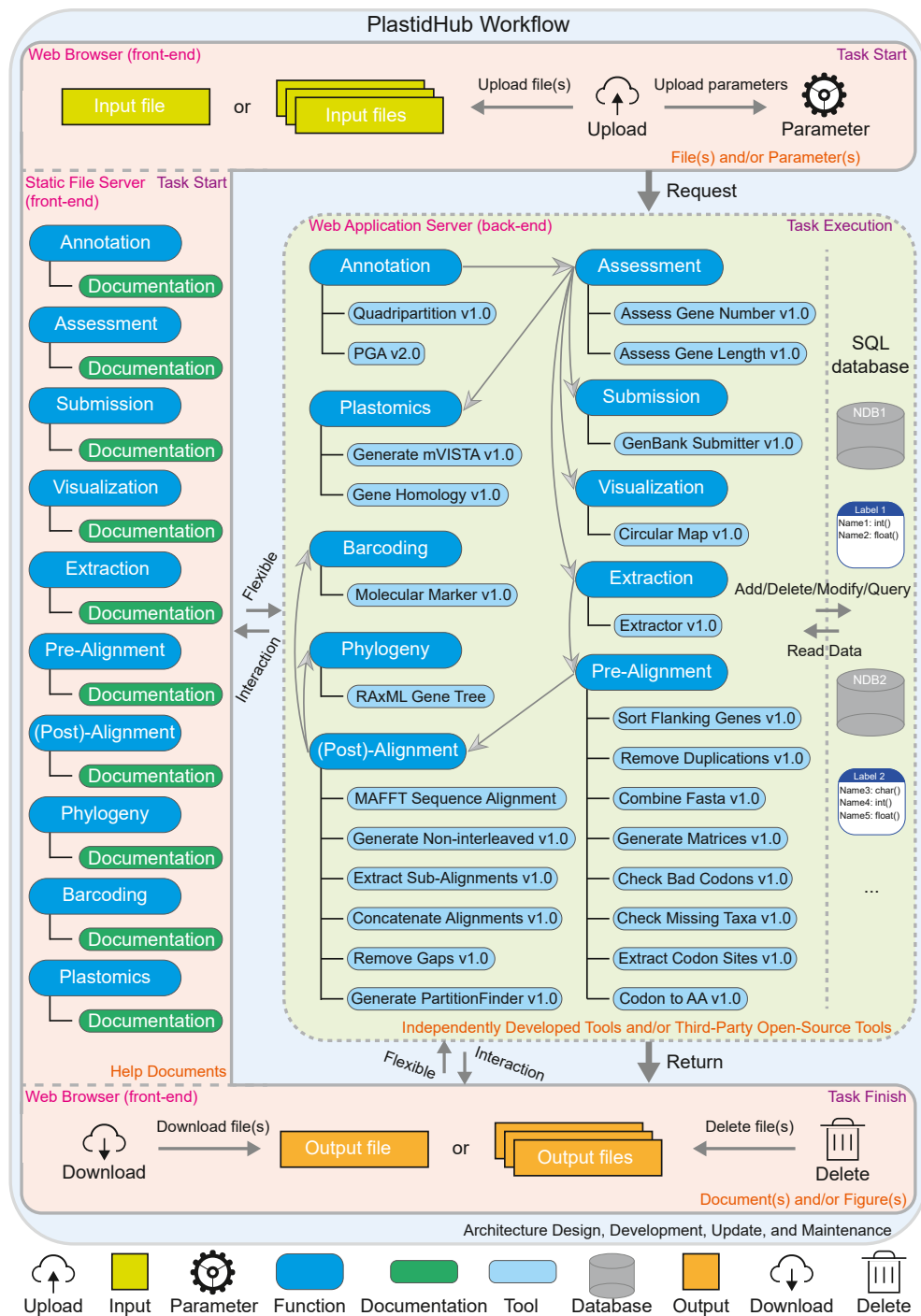


Fig. 1. Overview of interrelated task-based user interface design in PlastidHub. Task Start includes the front-end Web Browser for users to upload input Files and Parameters, and the front-end Static File Server for users to access Help Documents. Task Execution involves the back-end Web Application Server for users to execute analytical tools. Task Finish includes the front-end Web Browser for users to download or delete output Files (documents and figures). The C-shaped pink background indicates front-end interface, and the light-green background represents back-end interface. Request and Return between front-end and back-end interfaces are shown. The left Flexible Interaction allows switching between tool and help interfaces, and the bottom Flexible Interaction enables switching between download and tool interfaces. Within the tool interface, data can be stored and read interactively between the tools and the database.

identified it as a different IR. For the IR in a long plastid sequence (Vitales_Vitaceae_Vitis_romanetii_NC_056348), PlastidHub run1 identified a long and different IR containing four rRNA copies in two IR copies, while PlastidHub run2 identified a canonical and different IR containing two rRNA copies in two IR copies. NOVOWrap run1 did not identify IR in this long plastid sequence, but NOVOWrap run2 identified a long and identical IR. These results

demonstrate the Quadripartition tool in PlastidHub tends to consistently tolerate sequence differences between IRb and IRa, while the validate.py tool in NOVOWrap tends to become inconsistency in IR identification but its tolerance threshold is unknown. The proportion of non-standardized quadripartite structures were counted in 15,206 eukaryotic plastomes (Fig. 2C). If we set $IR \geq 1000$ bp, except 9 abnormal plastomes with run fault and null

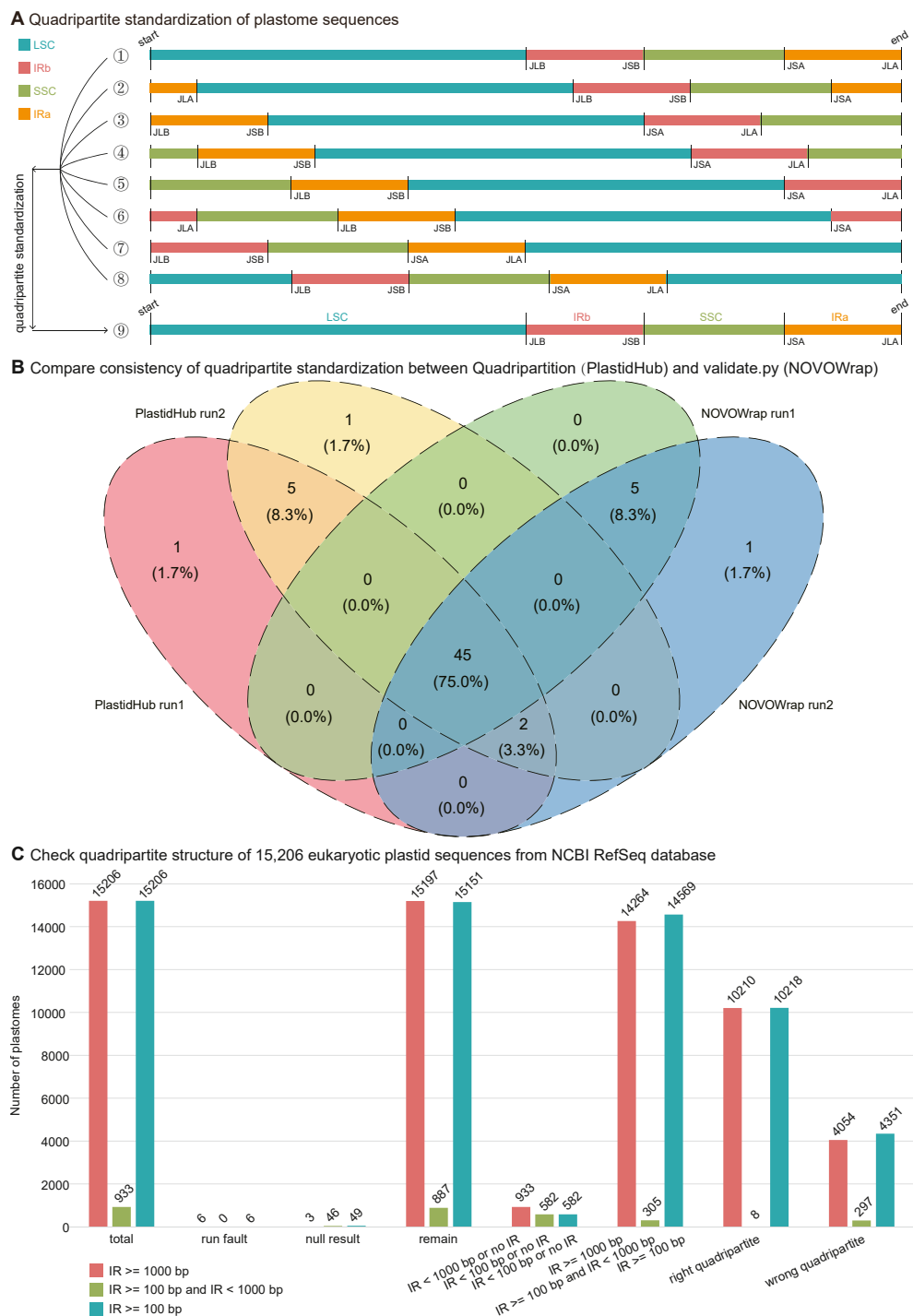


Fig. 2. Standardization of quadripartite structure for plastome sequences. (A) Standardization of plastome quadripartite structure. ①–⑧ represent unclassified structures, including non-canonical or partially standardized configurations. ⑨ depicts standardized quadripartite structure. (B) Comparison of the consistency of quadripartite standardization between the Quadripartition tool in PlastidHub and the “validate.py” tool in NOVOWrap. A total of 53 angiosperm plastomes were used to evaluate the IR identification. (C) Assessment of quadripartite structure in 15,206 eukaryotic plastomes from NCBI RefSeq database. Three scenarios are showed: IR ≥ 1000 bp, IR ≥ 100 bp and IR < 1000 bp, and IR ≥ 100 bp.

result, there are 933 plastomes with IR < 1000 bp or no IR and 14,264 plastomes (93.9% hit rate) with IR ≥ 1000 bp. For the 14,264 plastomes, there are 10,210 plastomes (71.6% accuracy rate) with correct quadripartite structure and 4054 plastomes with incorrect quadripartite structures. If we set IR ≥ 100 bp and IR < 1000 bp, except 46 abnormal plastomes with null result, there are 582 plastomes with IR < 100 bp or no IR and 305 plastomes (34.4% hit

rate) with IR ≥ 100 bp and IR < 1000 bp. For the 305 plastomes, there are 8 plastomes (2.6% accuracy rate) with the correct quadripartite structure and 297 plastomes with the incorrect quadripartite structure. In total, if we set IR ≥ 100 bp, except 54 abnormal plastomes with run fault and null result, there are 582 plastomes with IR < 100 bp or no IR and 14,569 plastomes (96.2% hit rate) with IR ≥ 100 bp. For the 14,569 plastomes, there are

10,218 plastomes (70.1% accuracy rate) with the correct quadripartite structure and 4351 plastomes with the incorrect quadripartite structure. These results confirm a high prevalence of structural non-standardizations in NCBI RefSeq, and the incorrect start-end position of circular plastomes is primarily caused by the offset of LSC-IRa junction (Turudić et al., 2022).

Quadripartite standardization hinges on the identification of inverted repeats (IR). From the perspective of bioinformatics, identifying IRs can be formulated as identifying the two longest reverse-complements in circular plastomes. While tools for repeat detection abound, Turudić et al. (2022) demonstrated that IR identification accuracy varied across six annotation tools due to pervasive imperfections in plastome assemblies (e.g., sequencing errors, assembly artifacts, or genomic heterogeneity). These imperfections necessitate tolerance for IR mismatches, as identical IRs may reflect artificial homogeneity, while divergent IRs may represent biological reality. Therefore, the newly developed Quadripartition tool in PlastidHub mainly considered consistency in IR identification. When discrepancies arise from nucleotide mismatches or length variations between IRb and IRa, customizable warnings (e.g., “Nucleotide and length of IRb and IRa may not be exactly the same!”, “Nucleotide of IRb and IRa may not be exactly the same!”, or “Length of IRb and IRa may not be exactly the same!”) are logged in warning.txt for user review. This transparency empowers researchers to manually validate IR boundaries and mitigate annotation errors stemming from assembly flaws.

3.4. Main function 2: Flexibility and consistency of the Annotation tool PGA v.2.0

We updated PGA v.1.0 to PGA v.2.0 by providing flexible options to annotate evolutionary remnants, namely pseudogenes and/or redundant genes (Fig. 3A). In some cases, it is difficult to distinguish pseudogenes from redundant genes, especially in highly divergent plastomes (Qu et al., 2023). PGA v.2.0 excels at detecting pseudogene remnants, which are critical markers for tracing the evolutionary history of these fast-evolving genes, especially highly degraded plastomes in heterotrophic plants (Goodhead and Darby, 2015; Hsu et al., 2016; Xiao et al., 2016). For regular phylogenetic analysis, we suggest to exclude annotations of redundant genes/pseudogenes to minimize alignment artifacts that compromise phylogenetic accuracy. For comparative genomics and RNA-editing studies, redundant genes/pseudogenes provide a chance to evaluate functional gene decay under relaxed selection.

PGA v.2.0 provided a flexible parameter to link or unlink *rps12*, addressing diverse annotation needs. The *rps12* gene is a trans-splicing gene, containing three exons across two pre-mRNAs, including exon 1 in LSC, and exon 2 and exon 3 in IR (Fig. 3B). Accurate *rps12* gene annotation remains challenging due to fragmented exons (Tillich et al., 2017; Shi et al., 2019; Qu et al., 2023). We compared PGA v.2.0 with GeSeq and CPGAVAS2 for *rps12* annotation accuracy using the aforementioned 53 angiosperm species (Fig. 3C and Table S2). Using plastome of *Amborella trichopoda* as reference, PGA v.2.0 and GeSeq successfully linked *rps12* in 51 of 53 species (Fig. 3C and Table S2). In contrast, CPGAVAS2 achieved 28 of 53 and 5 of 53 successes in *rps12* annotation using *Arabidopsis thaliana* and 2544 plastomes as reference, respectively. When no external reference sequence being specified, GeSeq successfully linked *rps12* in 22 of 53 species. Specifically, *rps12* in the long plastid sequence (Vitalis_Vitaceae_Vitis_romanetii_NC_056348) was not successfully linked by PGA v.2.0, GeSeq, and CPGAVAS2 (Table S2). The *rps12* with a 1-nucleotide intron (Zygophyllales_Zygophyllaceae_Zygophyllum_xanthoxylon_NC_052769) was successfully linked by PGA v.2.0 and CPGAVAS2, highlighting enhanced sensitivity to

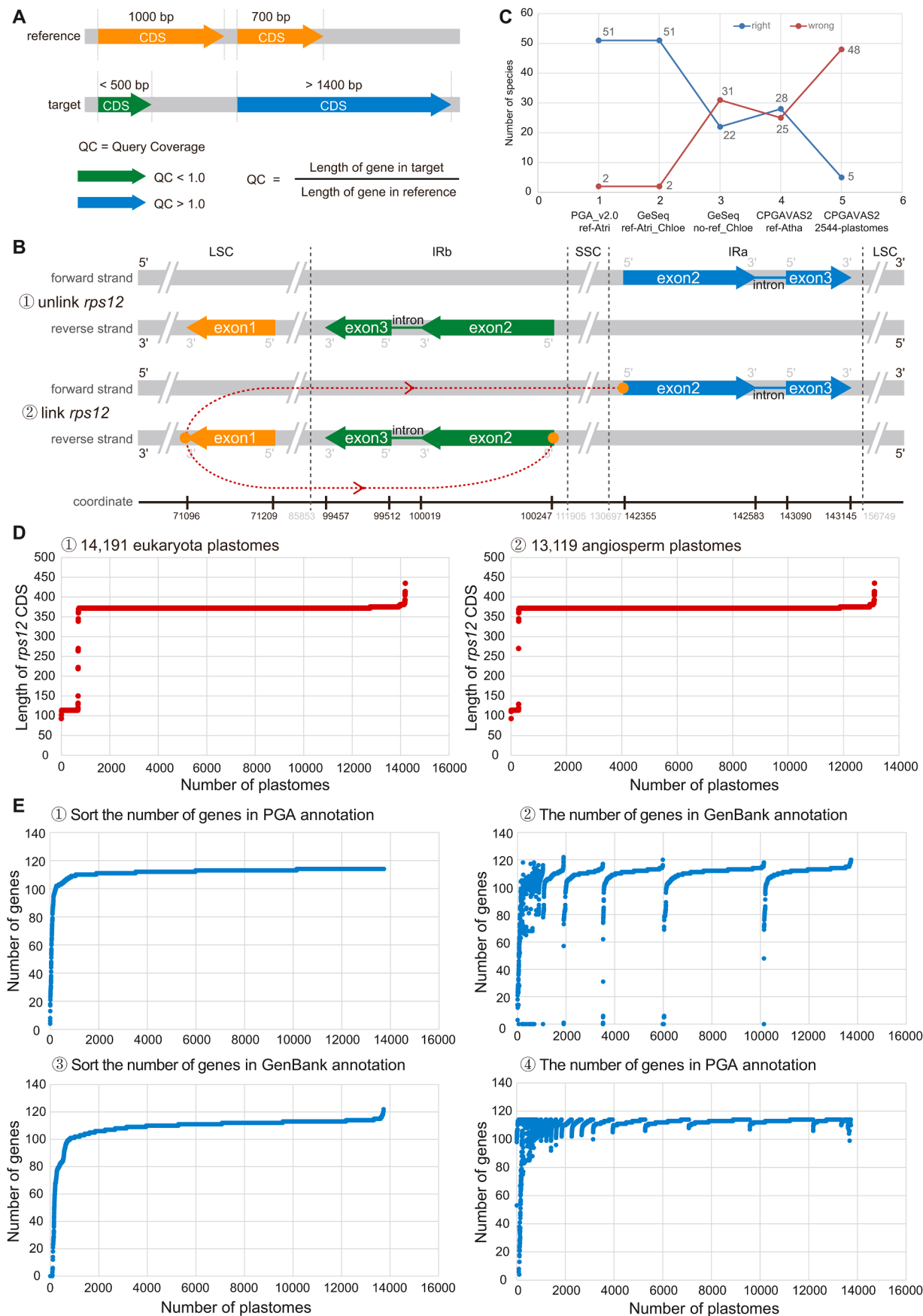
microstructural variations. CPGAVAS2 primarily failed due to missing exon3 annotations, and GeSeq (no external reference) simultaneously exhibited two annotations with incorrect boundaries for blatX annotations and correct boundaries for Chloe annotations (Table S2). Therefore, we recommend PGA v.2.0 and GeSeq for reliable *rps12* annotation, prioritizing reference-guided workflows to minimize errors.

To evaluate *rps12* annotation accuracy by at large scale, 15,206 eukaryotic plastid plastomes from NCBI RefSeq were re-annotated using PGA v.2.0 with default parameter values (Fig. 3D). Using *Amborella trichopoda* as reference, 15,200 plastomes were successfully annotated, while 6 from Chlorophyta failed. For 15,200 eukaryotic plastomes, 14,191 (93.4% hit rate) contained *rps12* annotations with 93–435 bp length variation, while 1009 lacked *rps12*. For 14,191 eukaryotic plastomes, 13,494 (95.1% link accuracy) had full-length *rps12* (339–435 bp), while 697 were only annotated with one exon (93–270 bp). For 13,731 angiosperm plastomes, 13,119 (95.5% hit rate) contained *rps12* (93–435 bp), while 612 lacked *rps12*. For 13,119 angiosperm plastomes, 12,845 (97.9% link accuracy) had full-length *rps12* (339–435 bp), while 274 sequences were only annotated with one exon (93–270 bp). These results confirm the robustness of PGA v.2.0, achieving high *rps12* detection and link accuracy even when annotating divergent eukaryotic plastomes with angiosperm-specific references.

To evaluate annotation flexibility and consistency, the number of annotated genes in GenBank and PGA v.2.0 were compared using the 13,731 angiosperm plastomes from NCBI RefSeq (Fig. 3E). GenBank annotations (original submissions) and PGA v.2.0 annotations (re-annotated with *Amborella trichopoda* as reference) were compared. The order of accession numbers for the angiosperm plastomes from PGA and GenBank was the same between ① and ② in Fig. 3E, as well as between ③ and ④ in Fig. 3E, ensuring direct comparability of annotation quality. Only 224 plastomes had gene numbers fewer than 100 in PGA annotations (Fig. 3E①), while 759 plastomes in GenBank annotations (Fig. 3E③). By comparing ② with ① in Fig. 3E, we found that the plastomes with gene numbers under 100 were scattered across taxa in GenBank annotations, reflecting inconsistent annotation quality. In contrast, by comparing ④ with ③ in Fig. 3E, we found that the plastid sequences with gene numbers under 100 in PGA annotations clustered in a fixed range, indicating systematic annotation rigor. Importantly, all PGA low-gene plastomes were subsets of GenBank low-gene plastomes, suggesting shared sequence features (e.g., fragmentation, pseudogenization) rather than annotation errors. This contrasts with GenBank's inconsistent annotation quality, likely reflecting heterogeneous standards across decades of plastome sequencing, assembly, and annotation practices. Despite RefSeq's curation, data quality variation persists due to evolving methodologies. PGA v.2.0's standardized workflow mitigates this, delivering highly consistent annotations for comparative studies.

3.5. Main function 3: Evaluating plastome annotation quality with the Assessment tool

We proposed two quantitative methods to evaluate the annotation quality of plastomes: gene number completeness and gene length accuracy (Qu et al., 2023). Using plastome of *Amborella trichopoda* as a reference, we evaluated the annotation quality of *Rosa roxburghii* (Fig. 4). For gene number assessment, the annotation completeness of genes can be judged by number of hitting (nHG), missing (nMG), and redundant (nRG) genes; percentage of hitting (pHG), missing (pMG), and redundant (pRG) genes; number of hitting (nHP, nHT, nHR), missing (nMP, nMT, nMR), and redundant (nRP, nRT, nRR) PCGs, tRNAs, and rRNAs; and names of hitting, missing, and redundant PCGs, tRNAs, and



A Gene number comparison

H: hitting, M: missing, R: redundant, n: number, p: percentage, P: PCGs, T: tRNAs, R: rRNAs

1. Number of hitting, missing, and redundant genes:

[114 total gene numbers in reference]
[115 total gene numbers in target]

[114 number of hitting genes (nH)]
[0 number of missing genes (nM)]
[1 number of redundant genes (nR)]

[100% percentage of hitting gene (pH)]
[0% percentage of missing genes (pM)]
[1% percentage of redundant genes (pR)]

2. Number of hitting, missing, and redundant PCGs, tRNAs and rRNAs:

[80 number of hitting PCGs (nHP)]
[0 number of missing PCGs (nMP)]
[1 number of redundant PCGs (nRP)]

[30 number of hitting tRNAs (nHT)]
[0 number of missing tRNAs (nMT)]
[0 number of redundant tRNAs (nRT)]

[4 number of hitting rRNAs (nHR)]
[0 number of missing rRNAs (nMR)]
[0 number of redundant rRNAs (nRR)]

reference (*Amborella trichopoda*)
vs
target (*Rosa roxburghii*)

3. Name of hitting, missing, and redundant PCGs, tRNAs and rRNAs:

[name of hitting PCGs]
accD atpA atpB atpE atpF atpH atpI ccsA cemA clpP infA matK ndhA
ndhB ndhC ndhD ndhE ndhF ndhG ndhH ndhI ndhJ ndhK petA petB
petD petG petL petN psaA psaB psaC psaI psaJ psbA psbB psbC
psbD psbE psbF psbH psbI psbJ psbK psbL psbM psbN psbT psbZ
rbcL rpl14 rpl16 rpl2 rpl20 rpl22 rpl23 rpl32 rpl33 rpl36 rpoA rpoB
rpoC1 rpoC2 rps11 rps12 rps14 rps15 rps16 rps18 rps19 rps2 rps3
rps4 rps7 rps8 ycf1 ycf15 ycf2 ycf3 ycf4
[name of missing PCGs]
None
[name of redundant PCGs]
ycf68

[name of hitting tRNAs]
trnA-UGC trnC-GCA trmD-GUC trnE-UUC trnF-GAA trnG-GCC
trnG-UCC trnH-GUG trnI-CAU trnI-GAU trnK-UUU trnL-CAA
trnL-UAA trnL-UAG trmM-CAU trnN-GUU trnP-UGG trnQ-UUG
trnR-ACG trnR-UCU trnS-GCU trnS-GGA trnS-UGA trnT-GGU
trnT-UGU trnV-GAC trnV-UAC trnW-CCA trnY-GUA trnM-CAU
[name of missing tRNAs]
None
[name of redundant tRNAs]
None

[name of hitting rRNAs]
rm16 rm23 rm4.5 rm5
[name of missing rRNAs]
None
[name of redundant rRNAs]
None

B Gene length difference comparison

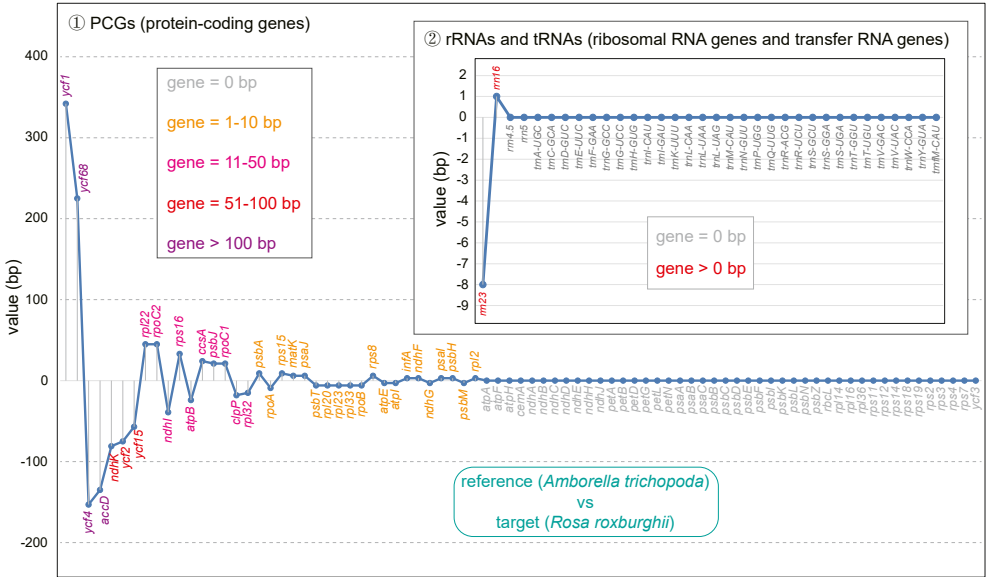


Fig. 4. Standards for quantitatively evaluating the annotation quality of plastomes. (A) Gene number comparison between target and reference plastomes. Annotation completeness of plastomes is assessed by comparing numbers and names of hitting, missing, and redundant genes. (B) Gene length difference comparison between target and reference plastomes. Annotation accuracy of plastomes is assessed by comparing length differences between genes in the target plastome and their corresponding genes in the reference plastome. The length differences of ① PCGs and ② rRNAs are separately compared. The gene length differences are sorted by magnitude, and those larger than a specified threshold (or those with the largest differences) are highlighted in different colors.

rRNAs (Fig. 4A). Since plastomes do not contain many genes, it would be reasonable for users to manually check annotations (hitting, missing, and redundant) of each gene to determine biological gene loss from annotation errors or pseudogene misclassification. For gene length assessment, length differences between target and reference genes are ranked by magnitude (Fig. 4B), and those with the largest differences (or those larger than a specified threshold) can be checked. Gene lengths of tRNAs and rRNAs are more conserved than those of PCGs across plant species, with rare exceptions, so the length differences of PCGs,

and tRNAs and rRNAs are separately compared. These tools empower users to refine gene annotations by identifying missing/poorly bounded genes and redundant false positives, ensuring high-quality plastome data.

3.6. Main function 4: Accuracy of the plastome Extraction tool

Extracting coding and noncoding sequences from plastomes is an important step for downstream analyses of selection tests or phylogenetic reconstructions. Given the circular structure and

Fig. 3. Improvement of annotation personalization for the plastome. (A) Annotation redundancy: annotate or unannotate evolutionary remains. The evolutionary remains (e.g., pseudogenes) can be unannotated by changing default parameter value (Y: annotate pseudogenes) in PGA v.2.0. (B) Annotation link: ① unlink *rps12* or ② link *rps12*. The trans-splicing gene *rps12* can be unlinked by modifying default parameter value (Y: link *rps12*) in PGA v.2.0. (C) Comparison of the effectiveness of PGA v.2.0 with GeSeq and CPGAVAS2 in annotating *rps12*. A total of 53 angiosperm plastomes were used to evaluate the link accuracy of *rps12*. (D) Evaluation of the annotation status of *rps12* by re-annotating 15,206 eukaryotic plastomes from NCBI RefSeq database with PGA v.2.0. (E) Comparison of the number of annotated genes in GenBank and PGA v.2.0 using 13,731 angiosperm plastomes from the NCBI RefSeq database.

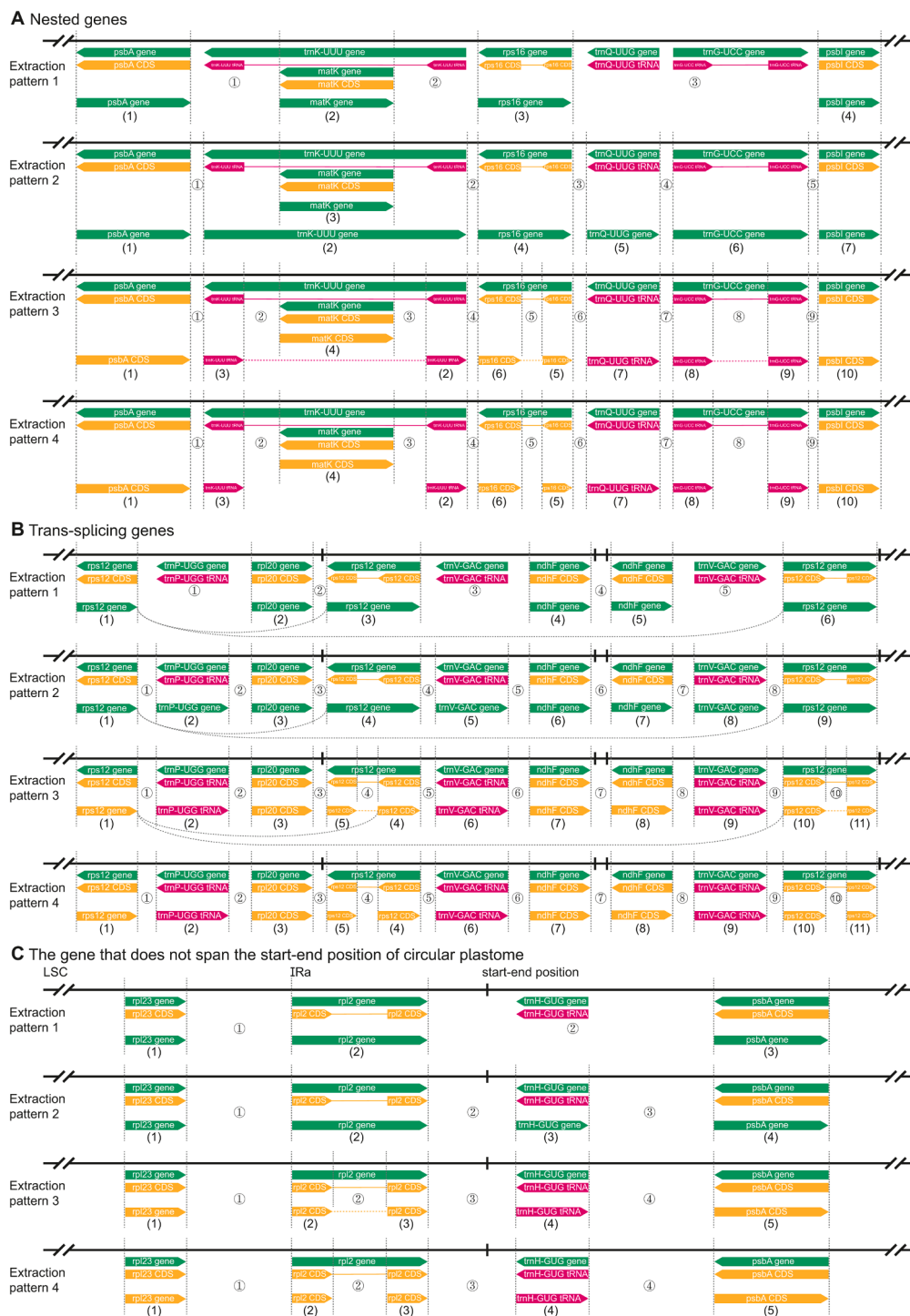


Fig. 5. Four extraction patterns of canonical and special coding and noncoding sequences in plastomes. (A) Nested genes. Pattern 1: (1)–(4), PCGs; ①–③, sequences among PCGs. Pattern 2: (1)–(7), genes; ①–③, intergenic sequences. Pattern 3: (1)–(10), coding sequences; ①–③, noncoding sequences. Exons of PCGs and tRNAs are connected with dash lines. Pattern 4: (1)–(10), coding sequences; ①–③, noncoding sequences. Exons of PCGs and tRNAs are connected with dash lines. (B) Trans-splicing genes. Pattern 1: (1)–(6), PCGs; ①–③, sequences among PCGs. Exon1 of *rps12* in LSC (1) and exon2, exon3, and intron of *rps12* in IRb (3) and IRa (6) are connected with curve line, respectively. Pattern 2: (1)–(9), genes; ①–③, intergenic sequences. Exon1 of *rps12* in LSC (1) and exon2, exon3, and intron of *rps12* in IRb (4) and IRa (9) are connected with curve lines, respectively. Pattern 3: (1)–(11), coding sequences; ①–③, noncoding sequences. Exon1 of *rps12* in LSC (1) and exon2 (4) and exon3 (5) of *rps12* in IRb and exon2 (10) and exon3 (11) of *rps12* in IRa are connected with curve line, respectively. Pattern 4: (1)–(11), coding sequences; ①–③, noncoding sequences. Exon1 of *rps12* in LSC (1) and exon2 (4) and exon3 (5) of *rps12* in IRb and exon2 (10) and exon3 (11) of *rps12* in IRa are separately presented. (C) Genes not spanning the start-end position of the circular plastome. Pattern 1: (1)–(3), PCGs; ① and ②, sequences among PCGs. Pattern 2: (1)–(4), genes; ①–③, intergenic sequences. Pattern 3: (1)–(5), coding sequences; ①–③, noncoding sequences. Exons of PCGs and tRNAs with introns are

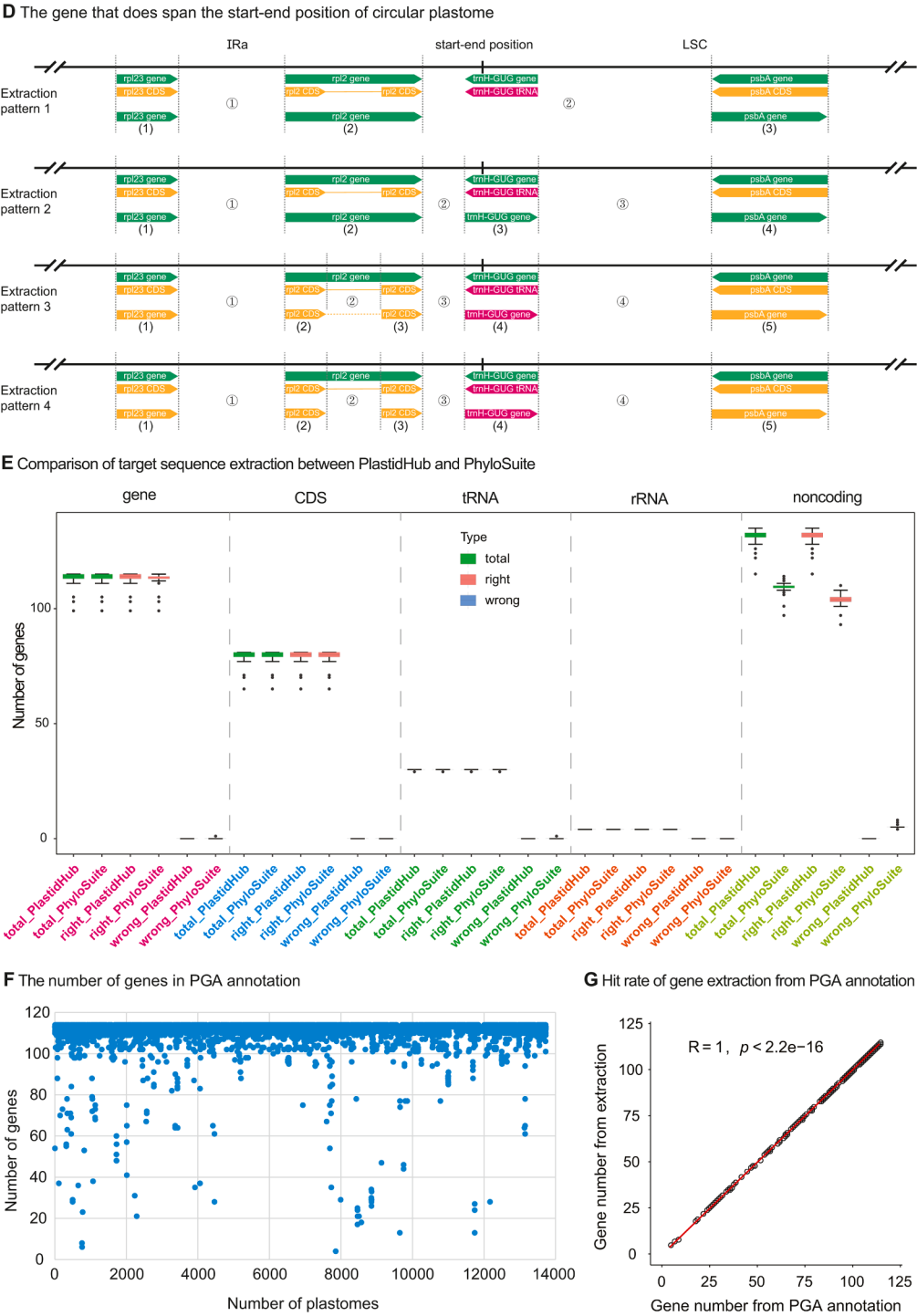


Fig. 5. (continued).

connected with dash lines. Pattern 4: (1)–(5), coding sequences; ①–④, noncoding sequences. Exons of PCGs and tRNAs with introns are separately presented. (D) Genes spanning the start-end position of the circular plastome. Pattern 1: (1)–(3), PCGs; ① and ②, sequences among PCGs. Pattern 2: (1)–(4), genes; ①–③, intergenic sequences. Pattern 3: (1)–(5), coding sequences; ①–④, noncoding sequences. Exons of PCGs and tRNAs with introns are connected with dash line. Pattern 4: (1)–(5), coding sequences; ①–④, noncoding sequences. Exons of PCGs and tRNAs with introns are separately presented. (E) Comparison of target sequence extraction between the Extraction tool in PlastidHub and the “Extract GenBank file” (EGBF) tool in PhyloSuite. A total of 53 angiosperm plastomes are used to evaluate the number and accuracy of different types of extracted sequences (gene, CDS, tRNA, rRNA, and noncoding regions). (F) The number of genes in PGA v.2.0 annotation. Accession numbers are sorted in ascending order. (G) Hit rate of gene extraction from PGA v.2.0 annotation. Extracted and PGA-annotated gene numbers is significantly positively correlated.

diverse gene types of plastomes, accurate extraction of all target sequences is not trivial. Our newly developed extraction tool in PlastidHub accommodates diverse extraction patterns (Fig. 5), including canonical genes with or without introns, and specific nested genes (Fig. 5A), trans-splicing genes (Fig. 5B), and genes avoiding (Fig. 5C) or spanning (Fig. 5D) the start-end position of circular plastome. In addition to PCGs with introns and sequences among PCGs, genes with introns and intergenic sequences are extracted. Furthermore, coding sequences with linked and unlinked CDSs and tRNAs are considered, as well as noncoding sequences. Customizable parameters enable tailored extraction, ensuring high accuracy for complex plastome structures.

To compare the Extraction tool in PlastidHub and the “Extract GenBank file” (EGBF) tool in PhyloSuite, we used aforementioned 53 angiosperm plastomes to evaluate the number and accuracy of different types of extracted sequences (gene, CDS, tRNA, rRNA, and noncoding regions) (Fig. 5E and Table S3). For gene, CDS, tRNA, and rRNA, PlastidHub and PhyloSuite extracted equal total number of target sequences (Fig. 5E and Table S3). Among the 53 species, the *trnH-GUG* gene in 8 species spans the start-end position, PhyloSuite extracted two partial sequences, while PlastidHub retrieved complete *trnH-GUG*. PhyloSuite failed to link exons of unlinked *rps12* exons, whereas PlastidHub output complete *rps12* regardless of link status. These results confirm PlastidHub’s superior handling of circular plastome complexities, ensuring accurate sequence retrieval for downstream analyses.

It is worth mentioning that the number of noncoding sequences extracted by PhyloSuite ($N_{\text{avg.}} = 109.0$) is much smaller than that of PlastidHub ($N_{\text{avg.}} = 132.2$). Notably, PlastidHub correctly extracted all noncoding sequences, while PhyloSuite generated an average of 4.1 erroneous noncoding sequences per species across the 53 analyzed plastomes (Fig. 5E and Table S3). This discrepancy arises because PhyloSuite fails to extract intron sequences unless they are explicitly labeled with the “intron” feature, resulting in the omission of 20 introns from 18 intron-containing genes. Additionally, PhyloSuite cannot handle nested genes well, such as *matK-trnK-UUU*, further contributing to incomplete noncoding sequence extraction. Similarly, Geneious, a widely used commercial platform, lacks functionality for extracting intergenic regions or noncoding sequences. These limitations highlight that PhyloSuite or Geneious are not optimized for comprehensive noncoding sequence extraction from plastomes, underscoring the need for tools like PlastidHub, which is specifically designed for this purpose.

GenBank file permits features (e.g., gene, CDS, tRNA, rRNA, etc.) to be listed in any order, as long as the provided coordinates are accurate. PhyloSuite requires features to follow physical gene order for correct intergenic region extraction. Disordered features in GenBank files lead to pervasive extraction errors in PhyloSuite. PlastidHub resolves this limitation, accurately processing both ordered and disordered GenBank features to ensure precise sequence extraction.

To validate the scalability of the Extraction tool in PlastidHub, the gene numbers from PGA v.2.0 annotations were analyzed across 13,731 angiosperm plastomes (Fig. 5F). Re-annotation with *Amborella trichopoda* as reference, gene numbers concentrated at ~110, aligning with typical gene numbers in angiosperm plastomes (Wicke et al., 2011; Jansen and Ruhlman, 2012). The hit rate of gene extraction from PGA annotation was used to evaluate the accuracy of the extraction tool in PlastidHub (Fig. 5G). A strong positive correlation (Pearson correlation: $R = 1$, $p < 2.2 \times 10^{-16}$) was observed between extracted and annotated gene counts, with ≤ 1 -gene discrepancies in most cases (Fig. 5G). These results affirm the extraction tool in PlastidHub as a state-of-the-art solution for

retrieving canonical and special coding and noncoding sequences from plastomes, adaptable to diverse research needs.

3.7. Main function 5: Automatic screening of molecular markers using the Barcoding tool

DNA barcodes is pivotal for species identification, phylogenetics, population genetics, phylogeography, conservation genetics, and biodiversity studies, but automated plastid marker screening remains challenging.

We developed a novel barcoding tool to address this issue by enabling batch calculation and extraction of variable, invariable, and gap sites from sequence alignments (Fig. 6). This tool streamlines the screening of molecular markers by integrating variable sites, parsimony informative sites, and total alignment length into a unified workflow, eliminating the need for third-party tools and accelerating downstream applications and analyses. Specifically, both the number and percentage of variable sites/parsimony informative sites within each alignment are calculated. In addition, both variable sites/parsimony informative sites and their positional data of nucleotides within each alignment are extracted. Most importantly, the alignments can be sorted by the percentage of variable or parsimony informative sites to prioritize informative markers (Fig. 6A–H). Notably, variable or parsimony informative sites across PCGs, RNAs, introns, and intergenic regions are separately sorted and categorized, with total alignment length incorporated as a criterion for marker selection (Fig. 6A–H), a critical factor in primer design (Ye et al., 2012). Boxplot analyses revealed that intergenic regions, introns, and PCGs exhibit significantly higher variability in both variable and parsimony informative sites compared to RNAs, tRNAs, and rRNAs (Fig. 6I), reflecting differences in evolutionary rates among genomic regions (Wolfe et al., 1987; Magee et al., 2010).

3.8. Main function 6: Micro-synteny structural comparison using the Gene Homology tool

Whole-plastome alignment typically highlights macro-scale structural variations. mVISTA (Frazer et al., 2004) is one of the most widely used tools for plastome comparison. However, its utility is hampered by the technical complexity of generating input files, which poses a significant barrier for researchers without bioinformatics expertise. There are other tools available for comparative analysis of plastomes, such as progressiveMAUVE (Darling et al., 2010), which are widely applied in studies focusing on plastomes with structural variations. However, tools capable of performing micro-synteny analysis in plastomes remain scarce.

We developed a novel tool “gene homology” to address this gap by generating pairwise circular maps of two plastomes that connect homologous genes and mark non-shared genes (Fig. 7). This tool enables pairwise visual comparison of each plastid gene, ultimately displaying micro-synteny structural variations and evaluating gene-level annotation completeness. It not only identifies structural variations such as inversions, translocations, and duplications, but also quantifies the extent of annotation completeness by accounting for gene losses/gains. Regardless of whether the reference and target plastomes have IR copies, including scenarios with two IRs or a single IR, this tool accommodates all structural possibilities (Fig. 7A). To demonstrate the flexibility of this tool, we compared plastome of *Amborella trichopoda* with other angiosperms (*Rosa roxburghii*, *Salicornia bigelovii*) and a gymnosperm species (*Calocedrus decurrens*), as well as gymnosperm species pair *Calocedrus decurrens* vs. *Calocedrus rupestris* (Fig. 7B). As far as we know, this is the first tool for pairwise

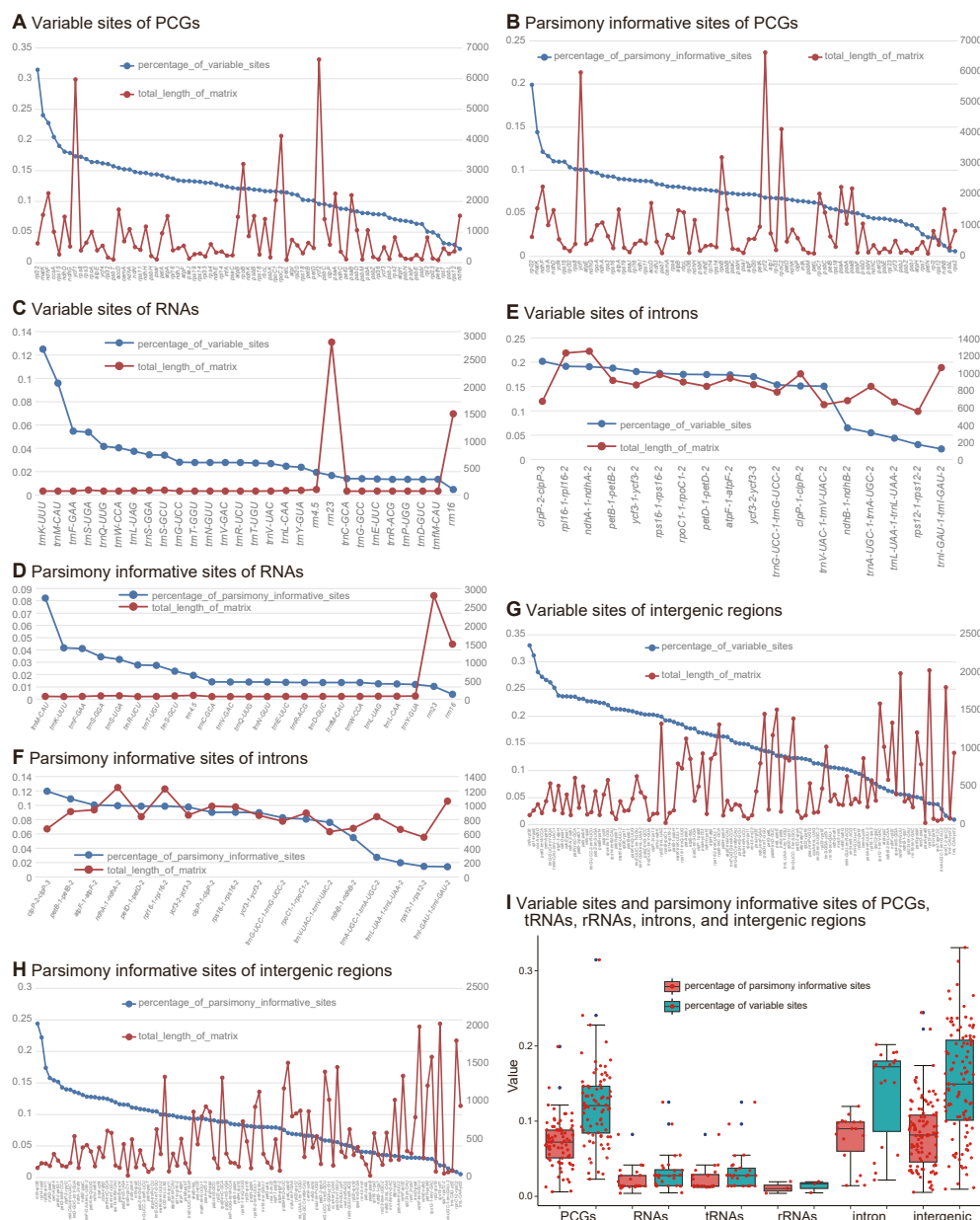


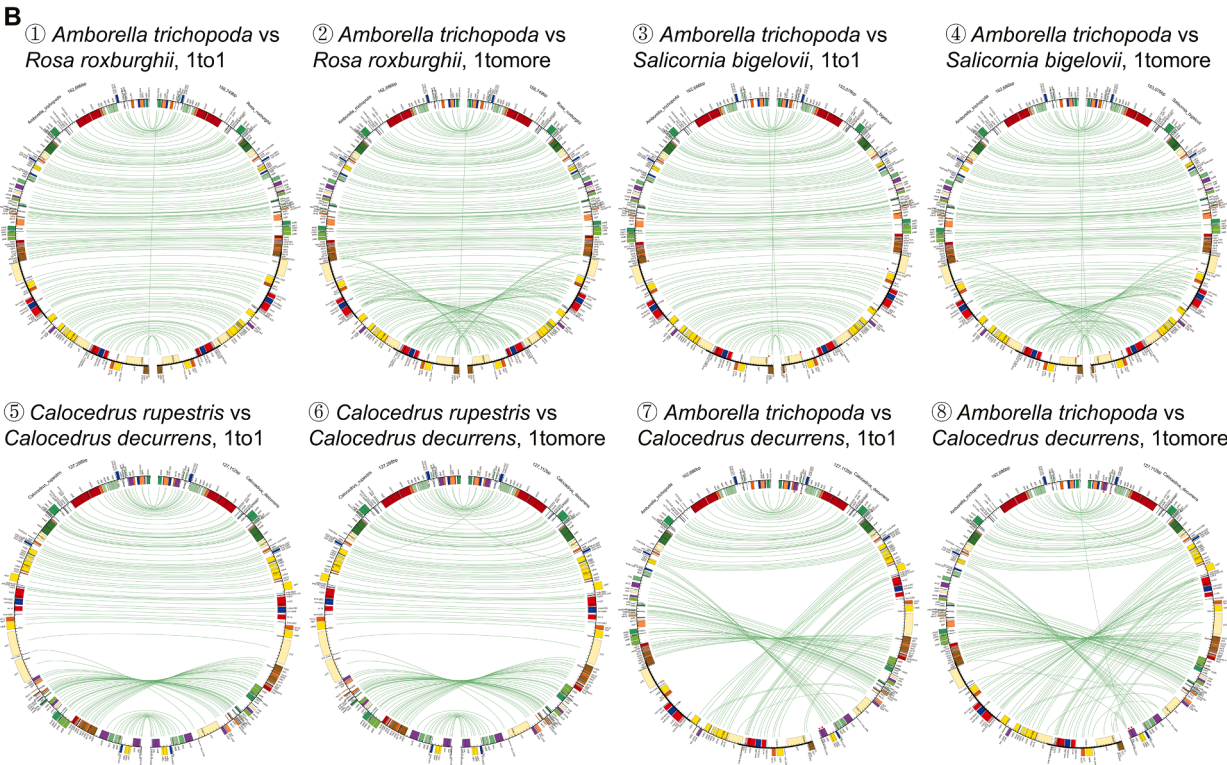
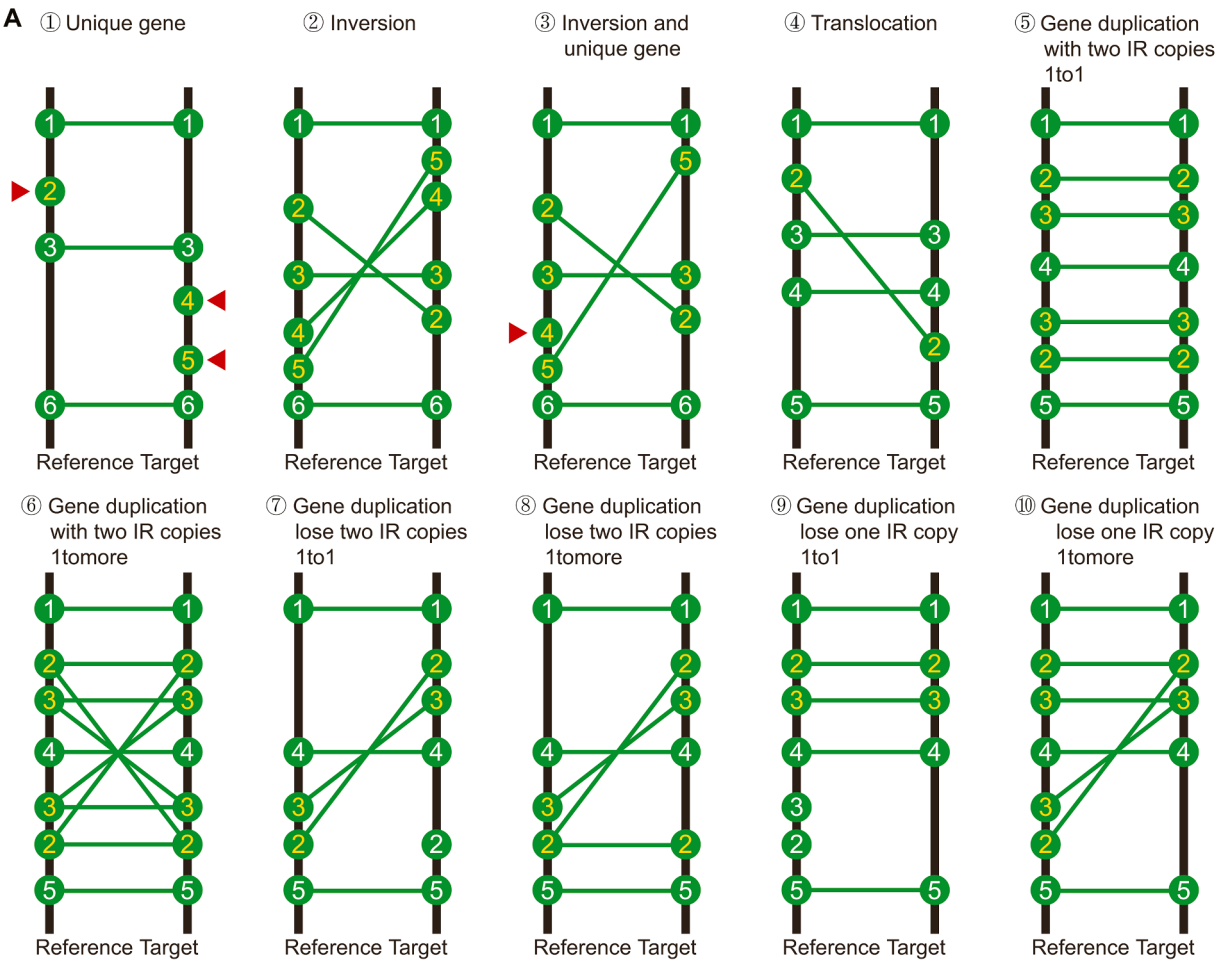
Fig. 6. Automatically screening molecular markers for plastomes. (A, C, E, G) Variable sites of PCGs, RNAs, introns, and intergenic regions, respectively. Alignment matrices are sorted in descending order based on percentage of parsimony informative sites. Total lengths of alignment matrices are displayed synchronously. (B, D, F, H) Parsimony informative sites of PCGs, RNAs, introns, and intergenic regions, respectively. (I) Boxplots of variable sites and parsimony informative sites for PCGs, RNAs, tRNAs, rRNAs, introns, and intergenic regions. Red points are shake scatter points of all sequences, and blue points are outlier values of sequences.

visualization of plastomes at the gene level, offering unprecedented resolution for micro-synteny structural variation analysis.

3.9. Similar functions with other tools: Submission and Visualization

As far as we know, there are two tools that can convert GenBank flatfiles to the tbl-formatted files required for NCBI/EMBL/DDJB database submission: gbf2tbl.pl (NCBI) and GeSeq (Tillich et al., 2017). There are six tools that can perform plastome visualization: OGDRAW v.1.3.1 (Greiner et al., 2019), CGView (Stothard et al., 2019), CPGAVAS2 (Shi et al., 2019), Chloroplot (Zheng et al., 2020), Chloe (<https://chloe.plastid.org/annotate.html>), and CPGView (Liu et al., 2023). GeSeq is the only tool that can conduct both format

conversion and gene visualization of annotated plastomes. PlastidHub complements these tools by providing user-friendly submission workflow and enhanced visualization capabilities. Its submission tool can batchly generate tbl-formatted files with 100% format compliance (validated on 200+ submissions), including mandatory fields like “gene”, “CDS”, “tRNA”, and “rRNA”, supporting multi-genome submissions. The submission tool is capable of recognizing *rps12* exon-linking algorithm, achieving 100% accuracy in resolving the *rps12* split issue (demonstrated across 200+ plastomes), thereby eliminating the need for manual corrections previously required for NCBI submissions. Its visualization tool uniquely shows linked dispersed repeats (forward and inverted), with different repeats including minimum allowed percent identity, minimum allowed length, and BLAST e-value to



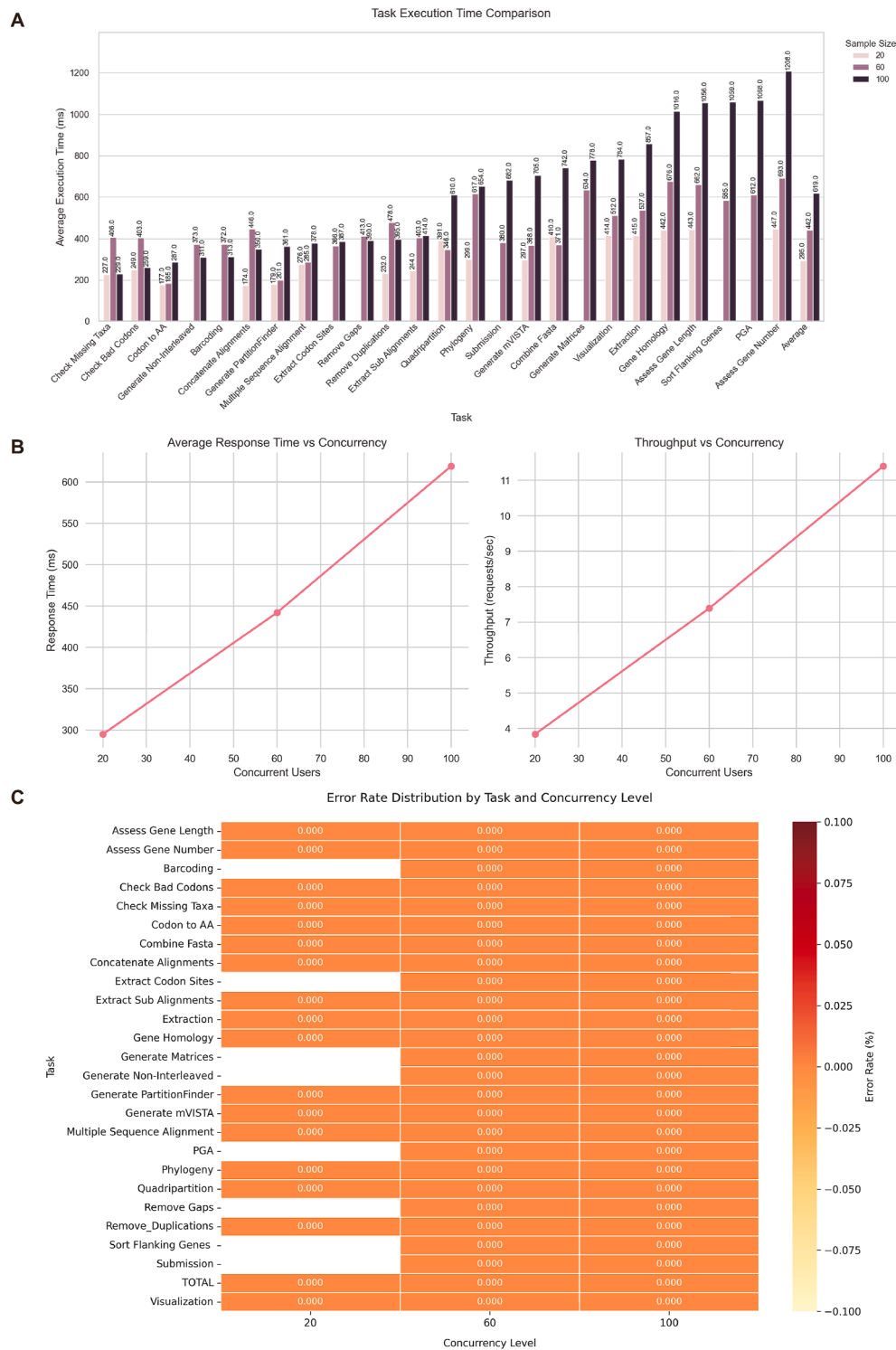


Fig. 8. Performance test to demonstrate PlastidHub's concurrent processing capability and scalability. (A) Response time analysis of PlastidHub. Performance test simulated three escalating scenarios: light (20 users), moderate (60 users), and heavy (100 users) concurrent loads. Virtual users randomly selecting different tasks and sending requests within 10 s. (B) Linear analysis of response time and throughput with concurrency for PlastidHub. (C) Error rate distribution by task and concurrency for PlastidHub.

Fig. 7. Micro-synteny comparison of each gene in target and reference plastomes. (A) Principle. ① Unique gene. Unique genes in target or reference plastomes are marked with red triangle symbols. ② Inversion. Inversions are shown by connecting genes in reciprocally inverted segments. ③ Inversion and unique gene. Not all genes in inverted segments are shared between target and reference, so unique genes in inverted segments are marked with red triangle symbol. ④ Translocation. ⑤ Gene duplication, with two IR copies, 1to1. The tool only connects shared genes with same order. ⑥ Gene duplication, with two IR copies, 1tomore. ⑦ Gene duplication, lose two IR copies, 1to1. ⑧ Gene duplication, lose two IR copies, 1tomore. ⑨ Gene duplication, lose one IR copy, 1to1. ⑩ Gene duplication, lose one IR copy, 1tomore. For ⑥, ⑧, and ⑩, the tool connects shared genes with both same and opposite order. For ⑦ and ⑨, the tool only connects shared genes with either same or opposite order. (B) Examples. ① *Amborella trichopoda* vs *Rosa roxburghii*, 1to1. ② *A. trichopoda* vs *R. roxburghii*, 1tomore. For ① and ②, no IR copies are lost from both species. ③ *A. trichopoda* vs *Salicornia bigelovii*, 1to1. ④ *A. trichopoda* vs *S. bigelovii*, 1tomore. For ③ and ④, no IR copies are lost from both species. ⑤ *Calocedrus rupestris* vs *C. decurrens*, 1to1. ⑥ *C. rupestris* vs *C. decurrens*, 1tomore. For ⑤ and ⑥, one IR copy is lost from both species. ⑦ *A. trichopoda* vs *C. decurrens*, 1to1. ⑧ *A. trichopoda* vs *C. decurrens*, 1tomore. For ⑦ and ⑧, one IR copy is lost from *C. decurrens*.

refine repeat annotations. This flexibility supports tailored analysis of structural variation, setting PlastidHub apart as a multifunctional platform for plastome curation.

3.10. Streamlining analysis with personalized new functions in three toolkits: Pre-alignment, (Post)-Alignment, and Phylogeny

PlastidHub introduces five customizable tools to optimize plastid phylogenomic workflows: “sort flanking genes”, “check bad codons”, “check missing taxa”, “extract sub-alignments”, and “RAxML gene tree”. The tool “sort flanking genes” was specifically developed for plastomes with significant structural rearrangements, such as inversions and translocations. This is because it is essential to adjust the orientations of intergenic regions (to a uniform direction) before alignment and phylogenetic reconstruction. This tool improves the alignment accuracy for intergenic regions in plastomes, and can also trace the evolutionary origins of gene remnants in intergenic regions, catering to users with specialized analytical requirements. The tool “check bad codons” detects problematic codons in batch-submitted of PCG matrices, allowing users to filter out low-quality sequences to improve alignment accuracy. Since reliable PCG matrices exclude internal stop codons, these codons serve as indicators of potential errors in PCG annotations. The tool “check missing taxa” identifies taxa absent in batch-submitted sequence matrices, generating a matrix that displays species and sequences coverage to help users to assess dataset integrity and detect potential false-positive issues linked to specific species or plastid regions. The tool “extract sub-alignments” accepts a user-defining species list to create sub-alignments from batch sequence submissions, while pruning gap-only sites in the sub-alignments. This tool is important for illuminating how species sampling influences phylogenetic results. The tool “RAxML gene tree” conducts ML phylogenetic inference using RAxML on batch sequence alignments, enabling users to infer gene and species trees for assessing phylogenetic conflicts.

3.11. Response time, throughput and scalability

3.11.1. Response time analysis

Under a load of 20 concurrent users, PlastidHub achieved an average response time of 295 ms (optimal performance) (Fig. 8A). At 60 users, response time rose to 442 ms (moderate increase), and at 100 users, reached 619 ms (sub-second response). All requests under peak load completed in < 1 s, demonstrating robust responsiveness and scalability under growing demand.

3.11.2. Throughput analysis

The throughput of PlastidHub increases linearly with the number of concurrent users, from 3.84 requests/second at 20 users to 11.40 requests/second at 100 users (Fig. 8B). This near-linear scalability indicates a well-designed architecture that efficiently utilizes server resources to handle increasing loads. Notably, PlastidHub maintained a 0% error rate across all loads, with no failed or timed-out requests even at peak load (Fig. 8C), confirming exceptional stability.

3.11.3. Scalability analysis

When response times or processing capacity approach bottlenecks, PlastidHub can quickly scale horizontally by adding more servers. Horizontal expansion ensures sustained performance. This elastic scaling reduces latency and multiplies task-processing capacity on demand.

Finally, the performance test results demonstrate that PlastidHub exhibits exceptional concurrent processing capability and robust scalability. Specifically, the system can: (1) stably handle high loads of up to 100 concurrent users; (2) maintain all request response times within acceptable thresholds; (3) achieve a 0% error rate across varying load conditions; (4) scale throughput linearly with added computational resources. These results confirm PlastidHub's reliability, operational efficiency, and readiness for deployment in production environments with demanding workloads.

4. Conclusions

While tools like GeSeq (Tillich et al., 2017) possess partial functionalities with PlastidHub, none match its breadth of customizable functions. PlastidHub currently offers a set of 25 tools spanning diverse plastome analyses, with 19/25 tools operating independently of third-party tools. The remaining six integrate BLAST + v.2.16.0 (Camacho et al., 2009), Circos v.0.66 (Krzywinski et al., 2009), MAFFT v.7.3 (Katoh and Standley, 2013), and RAxML v.8.2.10 (Stamatakis, 2014). PlastidHub will be continuously updated, adding new tools and incorporating emerging methods to serve the expanding needs of plastome research.

PlastidHub's cloud-based architecture eliminates the need for software installation/updates and high-performance computing investments. User experience drives platform development, exemplified by seamless switching between analysis, help, and download interfaces during workflow execution. User feedback and field advancements will guide daily updates and tool expansions, ensuring that PlastidHub remains at the forefront of plastome informatics.

CRedit authorship contribution statement

Na-Na Zhang: Writing – original draft, Visualization, Formal analysis, Data curation. **Gregory W. Stull:** Writing – review & editing, Supervision, Investigation. **Xue-Jie Zhang:** Writing – review & editing, Visualization. **Shou-Jin Fan:** Writing – review & editing, Supervision, Conceptualization. **Ting-Shuang Yi:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Xiao-Jian Qu:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Funding acquisition, Formal analysis, Conceptualization.

Data availability statement

All relevant data and code can be found within the article and on the PlastidHub website (<https://www.plastidhub.cn>) and GitHub (<https://github.com/quxiaojian/PlastidHub>).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Dan Zou, Yu-Liang Chen, Zhen-Xiang Feng, Yuan-Ting Gu, and Rui-Yu Zhang for testing the PlastidHub. We thank the editors and reviewers for constructive comments and suggestions on improving this study. This study was funded by the Natural Science Foundation of Shandong Province (ZR2020QC022), the

Science and Technology Basic Resources Investigation Program of China (No. 2019FY100900), the Major Program for Basic Research Project of Yunnan Province (202401BC070001), Yunnan Revitalization Talent Support Program: Yunling Scholar Project to Ting-shuang Yi, and the open research project of “Cross Cooperative Team” of the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.pld.2025.05.005>.

References

- Amiryousefi, A., Hyvönen, J., Pocai, P., 2018. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031.
- Ankenbrand, M.J., Pfaff, S., Terhoeven, N., et al., 2018. chloroExtractor: extraction and assembly of the chloroplast genome from whole genome shotgun data. *J. Open Source Softw.* 3, 464.
- Bakker, E.T., Lei, D., Yu, J., et al., 2016. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biol. J. Linn. Soc.* 117, 33–43.
- Bankevich, A., Nurk, S., Antipov, D., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477.
- Bi, G., Luan, X., Yan, J., 2024. ORPA: a fast and efficient phylogenetic analysis method for constructing genome-wide alignments of organelle genomes. *J. Genet. Genomics* 51, 352–358.
- Bi, G., Mao, Y., Xing, Q., et al., 2018. HomBlocks: a multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. *Genomics* (San Diego, Calif.) 110, 18–22.
- Camacho, C., Coulouris, G., Avagyan, V., et al., 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10, 1–9.
- Chen, W., Achakagari, S.R., Strömviik, M., 2022. Plastaumatic: automating plastome assembly and annotation. *Front. Plant Sci.* 13, 1011948.
- Coissac, E., Hollingsworth, P.M., Lavergne, S., et al., 2016. From barcodes to genomes: extending the concept of DNA barcoding. *Mol. Ecol.* 25, 1423–1428.
- Daniell, H., Lin, C.-S., Yu, M., et al., 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 17, 134.
- Darling, A.E., Mau, B., Perna, N.T., 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5, e11147.
- Dierckx, N., Mardulyn, P., Smits, G., 2017. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45, e18.
- Díez Menéndez, C., Pocai, P., Williams, B., et al., 2023. IRplus: an augmented tool to detect inverted repeats in plastid genomes. *Genome Biol. Evol.* 15, evad177.
- Frazier, K.A., Pachter, L., Poliakov, A., et al., 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279.
- Goodhead, I., Darby, A.C., 2015. Taking the pseudo out of pseudogenes. *Curr. Opin. Microbiol.* 23, 102–109.
- Greiner, S., Lehwark, P., Bock, R., 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64.
- Hsu, C.Y., Wu, C.S., Chaw, S.M., 2016. Birth of four chimeric plastid gene clusters in Japanese umbrella pine. *Genome Biol. Evol.* 8, 1776–1784.
- Huang, D.I., Cronk, Q.C., 2015. Plann: a command-line application for annotating plastome sequences. *Appl. Plant Sci.* 3, 1500026.
- Jansen, R.K., Ruhlman, T.A., 2012. Plastid genomes of seed plants. In: Bock, R., Knoop, V. (Eds.), *Genomics of Chloroplasts and Mitochondria*. Springer, Dordrecht (The Netherlands), pp. 103–126.
- Jin, J.-J., Yu, W.-B., Yang, J.-B., et al., 2020. GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 1–31.
- Jin, S., Daniell, H., 2015. The engineered chloroplast genome just got smarter. *Trends Plant Sci.* 20, 622–640.
- Jung, J., Kim, J.I., Jeong, Y.-S., et al., 2018. AGORA: organellar genome annotation from the amino acid and nucleotide references. *Bioinformatics* 34, 2661–2663.
- Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
- Knox, E.B., 2014. The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. *Proc. Natl. Acad. Sci. U.S.A.* 111, 11097–11102.
- Krzywinski, M., Schein, J., Birol, I., et al., 2009. Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645.
- Li, H., Guo, Q., Xu, L., et al., 2023. CPJSDraw: analysis and visualization of junction sites of chloroplast genomes. *PeerJ* 11, e15326.
- Li, H.T., Yi, T.S., Gao, L.M., et al., 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* 5, 461–470.
- Liu, C., Shi, L., Zhu, Y., et al., 2012. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics* 13, 1–7.
- Liu, S., Ni, Y., Li, J., et al., 2023. CPGView: a package for visualizing detailed chloroplast genome structures. *Mol. Ecol. Resour.* 23, 694–704.
- Magee, A.M., Aspinall, S., Rice, D.W., et al., 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* 20, 1700–1710.
- McKain, M.R., Hartsock, R.H., Wohl, M.M., et al., 2017. Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes. *Bioinformatics* 33, 130–132.
- Mower, J.P., Vickrey, T.L., 2018. Structural diversity among plastid genomes of land plants. *Adv. Bot. Res.* 85, 263–292.
- Palmer, J.D., 1991. Plastid chromosomes: structure and evolution. In: Hermann, R.G. (Ed.), *The Molecular Biology of Plastids: Cell Culture and Somatic Cell Genetics of Plants*. Springer, Vienna (Austria), pp. 5–53.
- Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067.
- Qu, X.-J., Moore, M.J., Li, D.-Z., et al., 2019. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* 15, 50.
- Qu, X.-J., Zhang, X.-J., Cao, D.-L., et al., 2022. Plastid and mitochondrial phylogenomics reveal correlated substitution rate variation in *Koenigia* (Polygonoideae, Polygonaceae) and a reduced plastome for *Koenigia delicatula* including loss of all *ndh* genes. *Mol. Phylogenet. Evol.* 174, 107544.
- Qu, X.-J., Zou, D., Zhang, R.-Y., et al., 2023. Progress, challenge and prospect of plant plastome annotation. *Front. Plant Sci.* 14, 1166140.
- Raubeson, L.A., Jansen, R.K., 2005. Chloroplast genomes of plants. In: Henry, R.J. (Ed.), *Plant Diversity and Evolution: Genotypic and Phenotypic Variation in Higher Plants*. CABI, Cambridge (UK), pp. 45–68.
- Shaw, J., Lickey, E.B., Schilling, E.E., et al., 2007. Comparison of whole chloroplast genomes to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am. J. Bot.* 94, 275–288.
- Shi, L., Chen, H., Jiang, M., et al., 2019. CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47, W65–W73.
- Soorani, A., Haak, D., Zaitlin, D., et al., 2017. Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC Genomics* 18, 1–8.
- Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313.
- Stothard, P., Grant, J.R., Van Domselaar, G., 2019. Visualizing and comparing circular genomes using the CGView family of tools. *Briefings Bioinf.* 20, 1576–1582.
- Straub, S.C.K., Parks, M., Weitemier, K., et al., 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *Am. J. Bot.* 99, 349–364.
- Stull, G.W., Qu, X.J., Parins-Fukuchi, C., et al., 2021. Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nat. Plants* 7, 1015–1025.
- Tillich, M., Lehwark, P., Pellizzer, T., et al., 2017. GeSeq-versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11.
- Turudić, A., Liber, Z., Grdiša, M., et al., 2022. Chloroplast genome annotation tools: prolegomena to the identification of inverted repeats. *Int. J. Mol. Sci.* 23, 10804.
- Waterhouse, R.M., Seppey, M., Simão, F.A., et al., 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* 35, 543–548.
- Wicke, S., Schneeweiss, G.M., dePamphilis, C.W., et al., 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297.
- Wolfe, K.H., Li, W.-H., Sharp, P.M., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U.S.A.* 84, 9054–9058.
- Wu, P., Xu, C., Chen, H., et al., 2021. NOVOWrap: an automated solution for plastid genome assembly and structure standardization. *Mol. Ecol. Resour.* 21, 2177–2186.
- Wu, P., Xue, N., Yang, J., et al., 2024. OGU: a toolbox for better utilising organelle genomic data. *Mol. Ecol. Resour.* 25, e14044.
- Wyman, S.K., Jansen, R.K., Moore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255.
- Xian, W., Bezrukov, I., Bao, Z., et al., 2025. TIPPO: A user-friendly tool for *de novo* assembly of organellar genomes with high-fidelity data. *Mol. Biol. Evol.* 42, msae247. <https://doi.org/10.1093/molbev/msae247>.
- Xiao, J., Sekhwal, M.K., Li, P., et al., 2016. Pseudogenes and their genome-wide prediction in plants. *Int. J. Mol. Sci.* 17, 1991.
- Ye, J., Coulouris, G., Zaretskaya, I., et al., 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13, 1–11.
- Zheng, S., Pocai, P., Hyvönen, J., et al., 2020. ChloroPlot: an online program for the versatile plotting of organelle genomes. *Front. Genet.* 11, 576124.
- Zhou, C., Brown, M., Blaxter, M., 2024. OatK: a *de novo* assembly tool for complex plant organelle genomes. *bioRxiv*. <https://doi.org/10.1101/2024.10.23.619857>.
- Zhou, W., Armijos, C.E., Lee, C., et al., 2023. Plastid genome assembly using long-read data. *Mol. Ecol. Resour.* 23, 1442–1457.