Received: 14 November 2021 | Accepted: 7 March 2022

DOI: 10.1002/aps3.11475

SOFTWARE NOTE

Applications in Plant Sciences

PhyloHerb: A high-throughput phylogenomic pipeline for processing genome skimming data

Liming Cai[1,2,3] | **Hongrui Zhang**[1] | **Charles C. Davis**[1]

[1]Harvard University Herbaria, 22 Divinity Avenue, Cambridge, Massachusetts 02138, USA

[2]Department of Integrative Biology, University of Texas at Austin, Austin, Texas 78712, USA

[3]Department of Botany and Plant Sciences, University of California, Riverside, California 92507, USA

Correspondence

Liming Cai, Department of Integrative Biology, 205 W. 24th Street, Room 201, University of Texas at Austin, Austin, Texas 78712, USA.
Email: lmcai@utexas.edu

Charles C. Davis, Harvard University Herbaria, 22 Divinity Avenue, Cambridge, Massachusetts 02138, USA.
Email: cdavis@oeb.harvard.edu

Abstract

Premise: The application of high-throughput sequencing, especially to herbarium specimens, is rapidly accelerating biodiversity research. Low-coverage sequencing of total genomic DNA (genome skimming) is particularly promising and can simultaneously recover the plastid, mitochondrial, and nuclear ribosomal regions across hundreds of species. Here, we introduce PhyloHerb, a bioinformatic pipeline to efficiently assemble phylogenomic data sets derived from genome skimming.

Methods and Results: PhyloHerb uses either a built-in database or user-specified references to extract orthologous sequences from all three genomes using a BLAST search. It outputs FASTA files and offers a suite of utility functions to assist with alignment, partitioning, concatenation, and phylogeny inference. The program is freely available at https://github.com/lmcai/PhyloHerb/.

Conclusions: We demonstrate that PhyloHerb can accurately identify genes using a published data set from Clusiaceae. We also show via simulations that our approach is effective for highly fragmented assemblies from herbarium specimens and is scalable to thousands of species.

KEYWORDS

herbariomics, high-throughput sequencing, mitochondria, plastome, ribosomal genes

Herbarium specimens provide the most reliable links between taxonomy, phenotypic traits, genetic information, and species distributions. Beyond their traditional uses, they are increasingly utilized to elucidate the impacts of global change (Meineke et al., 2018). The advent of high-throughput digitization and industrial-scale sequencing of herbarium specimens presents unparalleled opportunities to investigate species diversity in a phylo-spatio-temporal context. Recently, protocols allowing for massive DNA extraction and sequencing of herbarium specimens have been implemented in large-scale systematic (Nevill et al., 2020; Folk et al., 2021) and ecological investigations (Nitta et al., 2017). These studies often rely on cost-effective library reconstruction and sequencing strategies such as genome skimming, hybrid enrichment, or genotyping by sequencing. Genome skimming (Straub et al., 2012), in particular, is designed to target high-copy conserved regions including plastid, ribosomal, and mitochondrial loci. It can be applied to DNA from both fresh tissue and degraded herbarium materials (Bakker et al., 2016). The streamlined library preparation protocol also makes this technique easily automized in wet bench workflows (e.g., robot library preparation). Compared to hybrid enrichment techniques such as the Angiosperms353 kit (Johnson et al., 2019), genome skimming does not require upfront investment in primer design or prior knowledge from a reference genome, and only requires standard DNA isolation and library preparation (McKain et al., 2018). The resulting plastid and ribosomal regions have also been extensively used in plant systematics since the 1980s (Palmer and Zamir, 1982), especially for massive phylogenetic investigations (Ruhfel et al., 2014; Zanne et al., 2014; Li et al., 2021). However, these organellar loci are tightly linked within each cellular compartment (Doyle, 2022) and thus have limited power in addressing hybridization or polyploidization (McKain et al., 2018). Various assembly software have been

developed to harvest organelle genomes from short-read data, most notably GetOrganelle (Jin et al., 2020), FastPlast (McKain and Wilson, 2017), and NOVOPlasty (Dierckxsens et al., 2017). Annotation tools such as GeSeq (Tillich et al., 2017), Verdant (McKain et al., 2017), and PGA (Qu et al., 2019) have also been implemented in parallel to produce publication-quality annotations. However, there are several limitations associated with these annotation tools, hindering their ability to efficiently assemble phylogenetic matrices at a massive scale and for lower-quality data, and thus reducing their scalability and application for herbariomic investigations. First, assemblies often return as fragmented scaffolds owing to the generally degraded nature of herbarium DNA and low base coverage from genome skimming experiments (Forrest et al., 2019). These fragmented assemblies are prone to assembly errors, which may break the synteny of genes, posing additional challenges for accurate annotation. It is thus recommended to align the fragmented assembly against a reference genome to better establish homology, which is often not available (Qu et al., 2019). Second, many of these tools are web-based (e.g., GeSeq and Verdant) or do not allow multiple genomes to be analyzed simultaneously (Table 1). Consequently, they cannot support batched analyses including hundreds to thousands of species, and moreover, these tools present difficulties involving data transfer. Third, these existing tools output annotations in GenBank or general feature format (GFF) but often require a third-party tool to extract individual genes, further impeding the workflow. Furthermore, short tRNA genes and intergenic regions are often excluded from downstream phylogenetic analyses, which is problematic because loci that include such intergenic regions can frequently be more phylogenetically informative (e.g., the *trnL-trnF* spacer).

To bridge this impasse, we present PhyloHerb, a command line tool for the simultaneous annotation, alignment, and phylogenetic estimation of thousands of species using genome skimming data. The core function of PhyloHerb is to apply BLAST searches to identify locus boundaries using either its built-in database (plastid, nuclear ribosomal, and mitochondrial genes) or customized references specified by the user. This allows a user to extract gene and intergenic regions en masse from assemblies with marginal quality and directly output orthologous sequences into FASTA format for rapid downstream phylogenomic investigation. PhyloHerb also offers a wide array of functions to assist with genome assembly, evaluate assembly statistics, concatenate loci, generate gene partition files, and curate alignments for easier manual inspection. It is especially designed to work with lower-quality assemblies such as those derived from sequencing older herbarium specimens or mining "off-target" reads from hybrid enrichment data (Granados Mendoza et al., 2020). Our lab has been applying this tool to assemble phylogenetic data sets for published and ongoing systematic studies in flowering plants and algae (Marinho et al., 2019; Lyra et al., 2021; C. C. Davis, personal observation). These published and ongoing data sets include more than 1500 species with a median base coverage of 21.9× for the plastid genome. Within less than one hour of CPU time, PhyloHerb can compile orthologous FASTA sequences for 1000 species across 150 loci in the plastid, nuclear, and mitochondrial genomes. We also piloted this tool recently with a group of scientists, including many first-time users, at a day-long workshop hosted by our authorship team at the Botany 2021 meeting (Cai et al., 2021).

METHODS AND RESULTS

PhyloHerb is an open-source program (GNU General Public License) written in Python 3. The source code, user manual, as well as the example data set, are freely available at https://github.com/lmcai/PhyloHerb/. The software can be easily installed on Linux, OS X, and Windows systems by simply decompressing the source code package. Before implementing the software, users need to install the Python modules Biopython and ete3, and BLAST+ (Johnson et al., 2008). Specific installation instructions can be found on the Github repository.

Input preparation

The minimum input for PhyloHerb includes the raw assemblies of plastid, ribosomal, or mitochondrial genomes in FASTA format (Figure 1). We recommend GetOrganelle

TABLE 1 Comparison of existing plastome annotation tools. The execution time for PhyloHerb is estimated on the Lenovo SD650 NeXtScale server of the FASRC Cannon compute cluster at Harvard University. The execution time for all other software is cited from Qu et al. (2019).

Tools	User interface	Time	Output format	Accept multi-FASTA/fragmented assembly	References
Plann	Console	~30 s	tbl	No	Huang and Cronk (2015)
Verdant/annoBTD	Web/Console	10–30 min	GFF3	Currently not supported but can be incorporated (personal communication with author).	McKain et al. (2017)
GeSeq	Web	6 s–13 min	GenBank	Yes. One assembly per run for fragmented genomes.	Tillich et al. (2017)
PGA	Console	~20 s	GenBank	Yes, but not recommended. Batch processing.	Qu et al. (2019)
PhyloHerb	Console	2–30 s	FASTA	Yes. Batch processing.	This paper

(Jin et al., 2020) for de novo assembly of these three genomes, which has been demonstrated to be state-of-the-art for this initial step (Freudenthal et al., 2020). Once assemblies are obtained, users can implement the 'qc' function in PhyloHerb to evaluate their assembly quality. When providing assemblies alone, this function will generate a summary spreadsheet for the following information: assembly size in base pairs (bp), number of scaffolds, and GC content. If the assembly is generated from GetOrganelle, PhyloHerb will read the log files and output the following additional statistics: total input reads, number of reads in the target region, average base coverage, and whether the genome is circularized.

PhyloHerb relies on a reference database to identify orthologs using BLAST searches. A built-in reference plant database is included in the source code and is explained in detail below. This database is comprehensive for land plant organellar genes and the nuclear rRNA repeat region. Users can also specify their own customized references to define loci. This user-specified reference should be a single FASTA file containing sequences from all targeted loci. The sequence header should start with the locus name, followed by an underscore '_', and any additional characters to distinguish different copies (Figure 1). Users may consider applying customized references under the following circumstances: (1) to consolidate multiple short gene and intergenic regions to a longer locus for better BLAST results; (2) to specify only closely related species for more accurate ortholog identification; or (3) to harvest loci not included in the reference database. For example, PhyloHerb can be used to extract *LFY*—a single-copy nuclear phylogenetic marker (Frohlich and Meyerowitz, 1997)—from transcriptome assemblies when provided with a *LFY* reference sequence. The utility of this functionality in targeting non-organellar or nuclear rRNA will be dependent on the overall coverage of the genome skimming data and the complexity of the focal genome.

Locating locus boundaries

The 'ortho' function of PhyloHerb uses a reverse query-subject BLAST approach to locate loci within an assembly (Qu et al., 2019). Here, the BLAST database is constructed from the unannotated assemblies, while the reference nucleotide database is used for BLAST queries. The input files are genome assemblies in FASTA format, and the outputs are FASTA files of individual loci. Our built-in plastid database referenced above contains 98 genes (Appendix S1) from 355 land plants (Appendix S2). To prepare this plastid database, we downloaded all available plastid genome annotations in GenBank (accessed 17 June 2021) and selected one representative species per family. Therefore, the plastid genes in the database represent the union of all identified protein-coding genes and rRNAs across land plants (i.e., bryophytes, ferns, gymnosperms, and angiosperms). We also manually curated the database to correct synonymous gene and annotation errors. The mitochondrial and nuclear rRNA databases were similarly prepared. The mitochondrial database includes 71 genes (Appendix S3) from 68 species (Appendix S4), while the nuclear rRNA database of 18 S, 28 S, and 5.8 S includes 155 species (Appendix S5).

For each locus, reference sequences from all species in the database will be BLASTed to the unannotated assembly with an *E*-value threshold of 1e-20 and length threshold of 60 bp. The BLASTN hit with lowest *E*-value and longest alignment length will be used to establish gene boundaries.

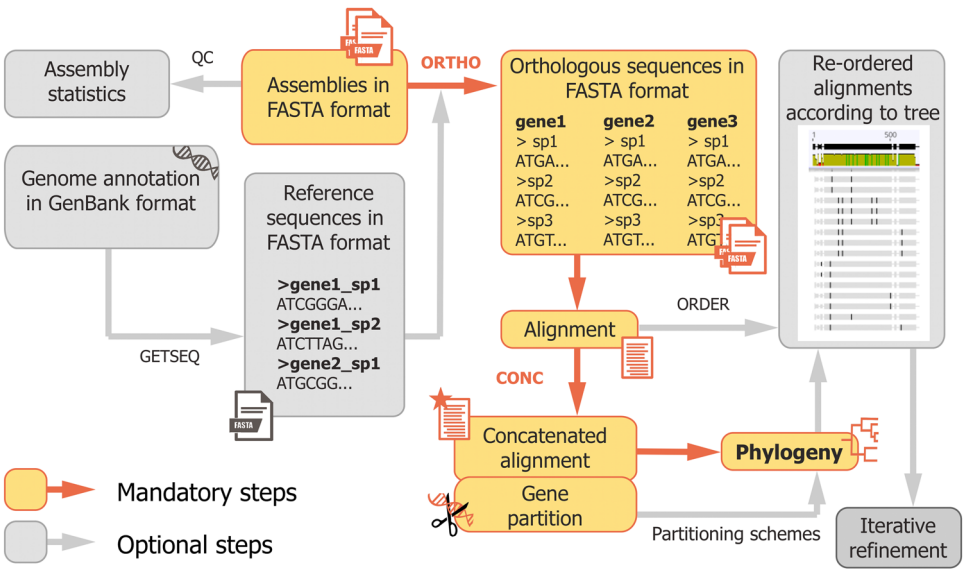

FIGURE 1 PhyloHerb workflow. The five main function modules of PhyloHerb, including qc, getseq, ortho, conc, and order, provide a versatile and efficient tool to curate and analyze genome skimming data.

A subset of genes and species can be included in the analysis by invoking the '-sp' and '-g' flags, respectively. The minimum length threshold can be adjusted using the '-l' flag, and the minimum *E*-value can be modified using the '-evalue' flag. For the two internal transcribed spacers (ITS), the gene locations of the three rRNAs are used to identify the start and end site of ITS. Here, we use gene synteny instead of sequence similarity to avoid spurious BLAST hits associated with high sequence divergence. The external transcribed spacer (ETS) and non-transcribed spacer (NTS) will not be automatically extracted by PhyloHerb.

To obtain intergenic regions, PhyloHerb uses a BLASTN search instead of gene synteny despite high sequence variation. This is because structural changes and fragmented assemblies can confound ortholog identification using empirical data. To more accurately determine locus boundaries, we recommend using closely related species as references and including conserved gene regions on both ends to define locus boundaries. For example, 12 short genes, including *psbJ*, *psbL*, and *rpl20*, are arranged linearly in a 5-kbp block in the plastid genome of *Arabidopsis thaliana* (L.) Heynh. (Figure 2A). To include intergenic regions, we can group these loci into two segments, each approximately 2.5 kbp in length and containing five to seven genes (LOC1 and LOC2 in Figure 2A). Here, the defined genetic block should not exceed 5 kbp in length for closely related species (e.g., species in closely related genera) and 3 kbp for more divergent lineages. This is because structural changes can break the synteny of the genome, leading to truncated BLAST hits. Once the boundaries of continuous genetic blocks are defined, PhyloHerb offers the function 'getseq' to extract corresponding regions from GenBank-formatted genome annotations. The input files include a genetic block definition file and GenBank annotations (Figure 2B). This function outputs FASTA files of the designated regions that can be directly used as reference in the 'ortho' function. The 'getseq' function offers two modes, 'genetic_block' and 'intergenic', which will include gene sequences on both ends or include intergenic regions only, respectively (Figure 2B). The 'genetic_block' mode is suitable for obtaining longer loci spanning multiple genes, and we anticipate that this functionality will be especially relevant for resolving clades at shallower phylogenetic depths. The 'intergenic' mode can be applied to obtain the more highly variable intergenic regions between two adjacent genes.

Utility functions for phylogenetic analysis

PhyloHerb offers several useful functions to assist alignment and phylogenetic reconstruction. We recommend MAFFT (Katoh and Standley, 2013) for aligning conserved genes, and PASTA (Mirarab et al., 2015) for aligning more variable intergenic regions. An example Bash file is provided in our source package (https://github.com/lmcai/PhyloHerb/blob/main/phyloherbLib/mafft_pasta.sh). Once individual alignments are generated, users can implement the 'conc' function in PhyloHerb to concatenate sequences. The list and order of loci to be included can be customized using the '-g' flag. The 'conc' function is especially suitable for creating large matrices with hundreds of species and genes, for which other GUI applications, such as MEGA (Tamura

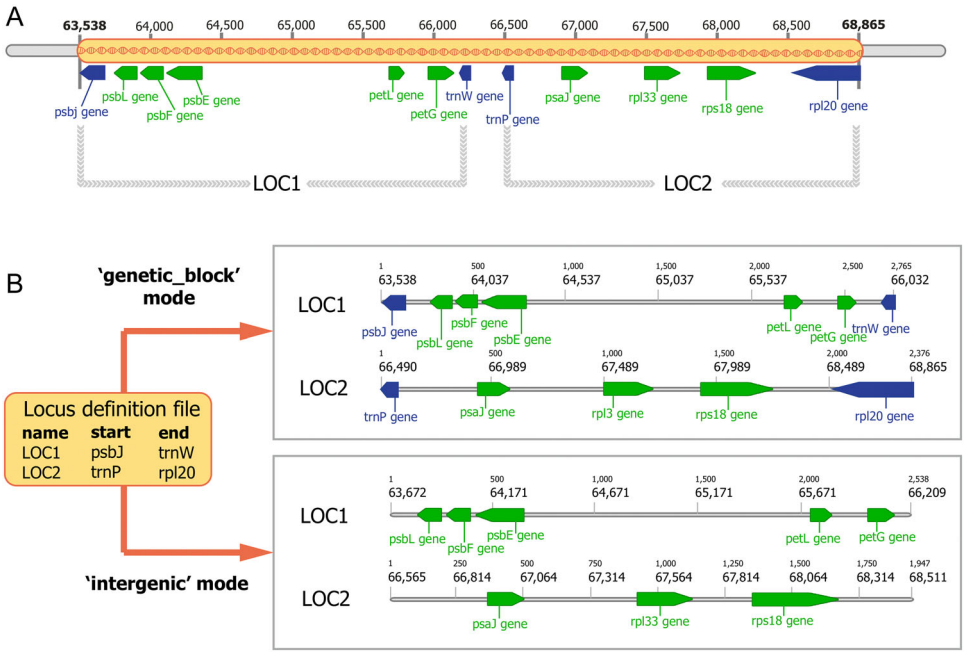

FIGURE 2 Defining and extracting genetic blocks with PhyloHerb. (A) A 5-kbp-long continuous genetic block on the plastid genome of *Arabidopsis thaliana* divided into two loci (LOC1 and LOC2). (B) The 'getseq' function of PhyloHerb can be used to extract sequences of predefined genetic blocks. The 'genetic_block' mode will include genes on both ends, while the 'intergenic' mode does not.

et al., 2007), often suffer from insufficient memory. PhyloHerb will also generate a gene partition file that can be directly input into PartitionFinder (Lanfear et al., 2017). The inferred partition scheme and the concatenated sequences can then be applied to phylogeny inference tools such as RAxML (Stamatakis, 2014), IQ-TREE (Minh et al., 2020), or ExaML (Kozlov et al., 2015).

Large-scale phylogenetic studies often require iterative alignment–phylogeny refinement practices to clean sequence data. To do so, researchers often visualize and edit the alignments in tools such as Geneious (https://www.geneious.com). Here, reordering sequences according to the species tree can help distinguish shared mutations between close relatives versus spuriously aligned regions arising from assembly or BLAST errors. Therefore, we developed the 'order' function of PhyloHerb, which takes a reference tree and reorders all input alignments based on the input phylogeny (Figure 1). It also offers the option to remove sequences with excessive missing data via the '-missing' flag. A float number from 0 to 1 can be used to indicate the maximum proportion of ambiguous sites allowed for each sequence. This function will generate an ordered alignment and a pruned species tree for each locus. The pruned species tree can be used to guide the PASTA alignment in the second round, which will significantly improve the alignment of especially intergenic regions (Mirarab et al., 2015).

Example workflow to harvest three cellular genomic compartments in Clusiaceae

We selected 10 species from three genera in the flowering plant family Clusiaceae using the published data set by Marinho et al. (2019; Appendix S6) to verify the utility of our pipeline. The input data were generated from a paired-end Illumina Hi-Seq 2 × 125 sequencing platform (Illumina, San Diego, California, USA), and their size ranged from 81.5 to 517.8 Mbp. These sequencing data included fresh tissue and degraded herbarium samples. The outputs of the PhyloHerb pipeline included alignments of plastid, mitochondrial, and nuclear ribosomal genes, as well as the species tree. All testing was conducted on the Lenovo SD650 NeXtScale server (Lenovo, Hong Kong, China) of the FASRC Cannon compute cluster at Harvard University. A detailed tutorial is provided in Appendix S7.

After generating genomes using GetOrganelle (Jin et al., 2020), we implemented the 'qc' function of PhyloHerb to summarize the assembly statistics, which took 0.53 s CPU time for 10 species. The plastid genome assemblies ranged from 128.2 to 164.3 kbp (Appendix S6). Three of 10 species have fully circularized plastid genomes. The coverage of rRNA is generally higher than the plastid organelle, and seven species have complete rRNA repeats assembled. The mitochondria assemblies are more fragmented, ranging from 6.5 to 359.3 kbp in size.

We then used the 'ortho' function of PhyloHerb to extract orthologous regions. For 10 species, this required 297.7 s CPU time and 159 MB peak memory using the built-in plastid database. When using a custom reference from *Garcinia gummi-gutta* (L.) Roxb. (Clusiaceae, GenBank accession number NC_047250.1), only 6.6 s CPU time and 153 MB peak memory were required for 10 species, demonstrating better scalability when more locally customized references are included. For nuclear rRNA, this step required 2.6 s CPU time and 47 MB peak memory, whereas for mitochondrial genes, it required 17.4 s CPU time and 108 MB peak memory. In the resulting sequence matrices, no missing data were identified in the plastid and nuclear ribosomal genes; however, the 53 mitochondrial genes contained from 0 to 90% missing data, reflecting the lower assembly quality of mitochondrial genomes. Due to these issues, we will only focus on the plastid data for phylogenetic reconstructions.

PhyloHerb recovered all 81 plastid genes when compared to the GenBank reference annotation of *G. gummi-gutta*, plus six additional genes. Five of these six genes (*psbG*, *ycf10*, *ycf15*, *ycf68*, and *ycf9*) were not annotated in the reference but do exist in the plastid genome, but the *rpl32* gene was misassigned by PhyloHerb. Here, all species had a false positive *rpl32* BLAST hit in their *rpl23* region, while *rpl32* was absent in Clusiaceae. Using gene tree phylogeny, we confirmed that no misleading paralogous copies were included in any of the extracted 87 plastid genes, including the *rpl32* gene. The resulting input and output files are available on Github (https://github.com/lmcai/PhyloHerb/blob/main/example/APPS_supplementary_data.zip). Therefore, in this case the incorrect gene assignment did not introduce spurious phylogenetic inference. We anticipate that shorter genes will have higher risks of such misassignment, but this is not unique to our pipeline (Qu et al., 2019). Based on our experience, loci shorter than 100 bp are especially prone to ortholog misassignment with low BLAST *E*-values. We therefore highly recommend users examine individual gene trees to identify such potential biases. One effective approach to mitigate false positive BLAST hits is to consolidate multiple adjacent short genes and intergenic regions into longer genetic blocks, a practice that PhyloHerb facilitates with its customizable reference database. Such practices have the additional benefit of including informative intergenic regions to improve phylogenetic resolution. For a detailed tutorial on this, please see the section "Combining short gene and intergenic regions to improve phylogeny" below.

To infer a species tree based on our assembled plastid genes, we used MAFFT v7.407 (Katoh and Standley, 2013) for alignment and then used the 'conc' function in PhyloHerb to concatenate individual alignments. This concatenation required 5.8 s CPU time with 64 MB peak memory for 87 plastid genes across 10 species. Finally, we used IQ-TREE v2.0.5 (Minh et al., 2020) to infer a species tree based on the concatenated alignment under the GTRGAMMA substitution model. We applied a partitioned analysis where each gene was assigned its own partition, and used 1000 ultrafast bootstrap replicates (UFBoot) to assess branch support and quantified genealogical concordance

using site concordance factor (sCF). All internal nodes were maximally supported except for *Tovomita acutiflora* M. S. Barros & G. Mariz and *Tovomita choisyana* Planch. & Triana, which received only 48 UFBoot and 27.81 sCF (Figure 3A). For other internal nodes, the sCF values ranged between 66.4 to 96.8 with an average value of 83.4.

Scalability

The two most resource-consuming steps in PhyloHerb are ortholog extraction and alignment concatenation. To demonstrate its scalability to massive data sets, we applied PhyloHerb to our unpublished data set for an ongoing project focusing on Malpighiales (C. C. Davis and L. Cai, personal observation). We randomly selected from this larger data set 1000 species sharing a most recent common ancestor of ~90 million years ago (Xi et al., 2012) and an average pairwise sequence divergence of 0.035 for plastid genes. Here, we used a single outgroup species *Vitis vinifera* L. (GenBank accession number: NC_007957.1) as reference to extract plastid genes. It required 357.1 s CPU time (~2.5 h wall-clock time for a single thread) with a peak memory of 53 MB to extract all 84 plastid genes from all 1000 assemblies. After aligning each of the gene sequences with MAFFT, we used the 'conc' function of PhyloHerb to create a concatenated matrix of 247-kbp aligned sites across 1000 species. This step required 1127.4 s CPU time with 4.43 GB peak memory. To assess its performance compared to other concatenation programs, we applied the same data set to Geneious, MEGA (Tamura et al., 2007), and SeaView (Gouy et al., 2010). Both MEGA and SeaView crashed shortly after we initiated the concatenation on our laptop (MacBook Pro 2.5 GHz Intel Core i7 [Apple Inc., Cupertino, California, USA] with 16 GB RAM). In contrast, Geneious was capable of concatenating this large matrix, which required 158.23 s CPU time and 2.14 GB peak memory on the MacBook. However, according to the user manual (https://www.geneious.com/), Geneious is not designed to work efficiently with more than 10,000 sequences (e.g., 100 loci × 100 species). Moreover, this concatenation function is also proprietary in Geneious and requires a paid subscription, which creates barriers for accessibility.

Combining short gene and intergenic regions to improve phylogeny

As we pointed out above, short loci have higher risks of ortholog misassignments. One effective way to reduce such risks is to combine multiple adjacent genes and intergenic regions into longer genetic blocks. This approach also

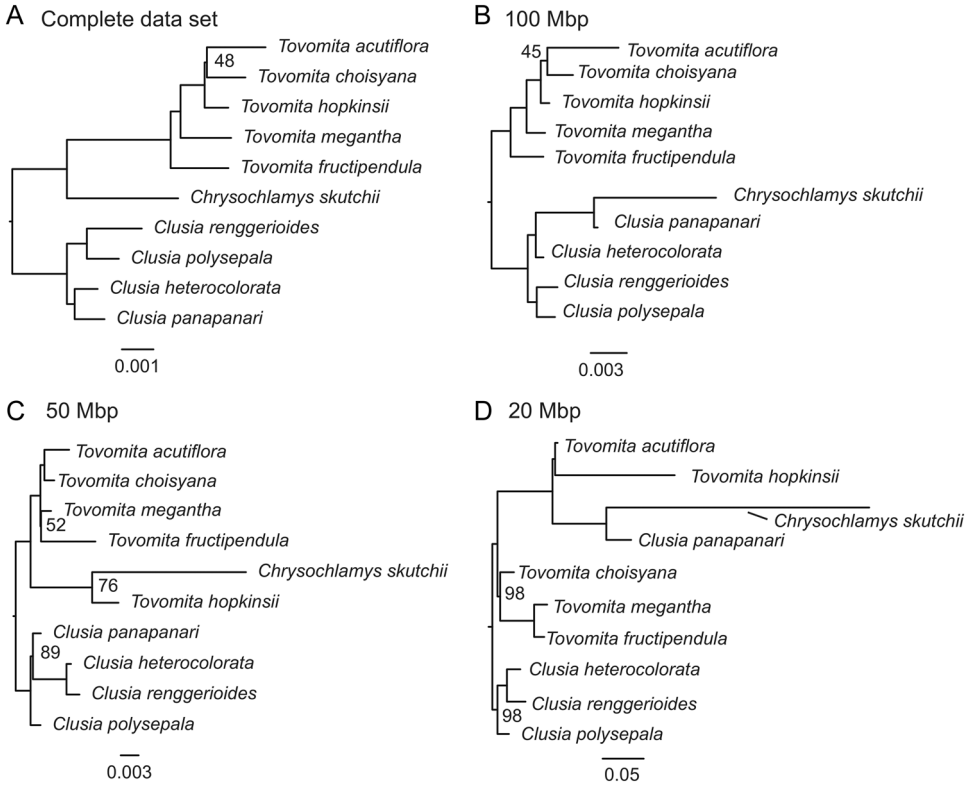

FIGURE 3 Phylogeny of 10 Clusiaceae species inferred from the complete (A) and subsampled plastid data sets (B–D). Raw reads were randomly subsampled to 100 Mbp (B), 50 Mbp (C), and 20 Mbp (D) to simulate decreasing base coverage in genome skimming. For all four analyses, a partitioned concatenated DNA alignment of 87 plastid genes was used to infer the species tree in IQ-TREE using the GTRGAMMA model. Nodal support was estimated from 1000 ultrafast bootstrap replicates (UFBoot). Unlabeled nodes indicate 100 UFBoot support. Note the unstable placement of *Chrysochlamys skutchii* in subsampled data sets.

includes variable intergenic regions to improve phylogenetic resolution, especially in rapidly diverging lineages (Gielly and Taberlet, 1994). PhyloHerb offers the 'getseq' functions to define and extract sequences of genetic blocks spanning multiple genes. The input is a locus definition file indicating the genes on both ends of each genetic block (Figure 2B). PhyloHerb will then extract sequences from GenBank files based on the definition and output into a FASTA file, which can be used as the reference for the 'ortho' function (Figure 1).

Using our Clusiaceae test data set as an example, we defined four plastid genetic blocks containing five to seven short genes (Appendix S8). The gene regions alone accounted for 569 to 2149 bp in each locus, but after adding intergenic regions, these loci increased to 2326 to 4044 bp. Importantly, the number of phylogenetically informative sites increased ninefold with the addition of intergenic regions. Consequently, the species tree inferred from both gene and intergenic regions had a higher average nodal support of 93 UFBoot (ranging between 50–100 UFBoot) compared to the species tree inferred from gene regions only (mean = 69 UFBoot, ranging between 47–100 UFBoot; Appendix S9). We also observed a slight increase in average sCFs from 71.4 to 72.9 after adding intergenic regions. This result demonstrated that the integration of intergenic regions effectively improves phylogenetic resolution by adding more rapidly evolving sites. One caveat here is that establishing site homology among highly variable intergenic regions is especially challenging when sampling deep and shallow phylogenetic depths simultaneously (e.g., both within genus and across families). In such cases, a profile alignment approach, which builds alignment subsets among closely related taxa and then merges them into a single aligned matrix, can be implemented (e.g., PASTA; Mirarab et al., 2015). For smaller data sets with less than 50 species or closely related genera such as in our Clusiaceae data set, a MAFFT-style alignment is sufficient.

Risks of low coverage: A simulation

To explore the limits of genome skimming techniques and artifacts attributed to low sequence coverage, we subsampled the Clusiaceae genome skimming data to include only 100 Mbp, 50 Mbp, and 20 Mbp reads. These data translate to an average base coverage of 8.4×, 4.5×, and 2.2× for the plastid genome as reported by the 'qc' function of PhyloHerb (Appendix S10). We applied the same genome assembly and gene extraction pipeline described above. The concatenated matrices from these three data sets were similar in length (approximately 96 kbp), but contained 14%, 29%, and 56% ambiguous characters, respectively. The quality of the alignments varied significantly across these subsampled data sets. When we manually inspected our complete alignment in Geneious, it required minimal adjustment with high sequence identity throughout (Appendix S11). When the input data size was reduced to 100 Mbp, we noticed more incidences of assembly or annotation errors requiring removal (green bars in Appendix S11). These biases, combined with the increasing amount of missing data, also reduced the sequence identity significantly. The same trend also applied to the 50 Mbp and 20 Mbp data sets. Without correction, species phylogenies built from these alignments have incorrect topologies and potentially spurious branch length distributions (Figure 3). In particular, species with excessive missing data often exhibit long branches (e.g., *Chrysochlamys skutchii* Hammel in Figure 3 and Appendix S10). Based on these results, we empirically conclude that 50 Mbp or 5× coverage is the lower limit for plastid genome assembly using the genome skimming technique. For such data sets, researchers need to apply more stringent filtering criteria, such as a smaller BLAST *E*-value threshold and using closely related custom references for accurate ortholog assignment.

CONCLUSIONS

As herbarium specimen sequencing and plastid genome–based phylogenomics become increasingly popular and greatly scalable for biodiversity research, sophisticated bioinformatic tools need to be developed in parallel to accommodate data sets of various sizes and quality. Working with fragmented organellar and ribosomal assemblies like those yielded from degraded herbarium materials is challenging owing to currently unsupported multi-FASTA file formats. PhyloHerb offers an easy-to-use tool that allows users to efficiently analyze assemblies of marginal quality at massive scale. This is likely to be especially useful in coming years as systematic biologists greatly expand their taxon and gene sampling using museum specimens. To facilitate this effort, our tool directly outputs orthologous sequences in FASTA format that can be used for downstream alignment and phylogenomic inference. Users can create custom references to extract intergenic regions, which will likely be crucial to resolve rapidly diverging lineages. However, data sets with less than 5× coverage should be processed very carefully because we demonstrated via data subsampling simulations that degraded data sets contain excessive assembly errors and require substantial manual cleaning. In addition, it should be noted that PhyloHerb is not designed to generate polished genome annotations but rather to generate alignments for phylogenomic purposes, which complement the functionality of existing tools such as GeSeq or PGA. In PhyloHerb, the accuracy of locus boundary determination is tied with the performance of BLAST searches. For conserved gene regions, PhyloHerb can confidently identify their locations given our comprehensive built-in reference database spanning land plants. For lineages or loci with high sequence divergence or in the presence of paralogs, however, we strongly recommend applying more closely related taxa as references. Moreover, genome structural modifications such as insertion,

deletion, and reversion will greatly impact the performance of PhyloHerb, especially when extracting genetic blocks spanning several genes. Where possible, we recommend checking gene synteny using completely circularized plastid genomes to avoid using regions prone to macrostructural changes. Even in cases where gene synteny is conserved, the custom genetic loci should not exceed 5 kbp or span more than 10 genes to avoid truncated BLAST hits. Finally, the plastid, mitochondrial, and ribosomal regions represent three tightly linked coalescent genes (Doyle, 2022), and thus have limited power in addressing more complex evolutionary scenarios involving introgression or polyploidization (McKain et al., 2018). Despite these caveats, compared to other annotation tools, PhyloHerb offers additional flexibility in input data quality while demonstrating high annotation accuracy. It represents a powerful and freely available software that allows researchers to rapidly assemble orthologous alignments from three cellular genomic components, providing recovered alignments that can generate robust phylogenies at various phylogenetic depths.

AUTHOR CONTRIBUTIONS

L.C. and C.C.D. planned and designed the research, L.C. and H.Z. analyzed the data, L.C. wrote the initial manuscript, and all authors contributed to revising and editing the text. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank the participants of the workshop we coordinated at the Botany 2021 meeting (Cai et al., 2021) who helped to prototype this tool and provided useful feedback to improve our software. We also thank Dr. Lucas C. Marinho (Universidade Federal do Maranhão) for supplying the Clusiaceae data set and revising the figures. We thank members of the Davis lab for insightful comments on the software and manuscript. Funding was provided from the National Science Foundation (grants DEB-0544039 and DEB-1355064) and startup funding from Harvard University to C.C.D. L.C. was supported by the Stengl Wyer Fellowship from the University of Texas at Austin.

OPEN RESEARCH BADGES

This article has been awarded an Open Materials badge. All materials are publicly accessible via the Open Science Framework at https://github.com/lmcai/PhyloHerb/. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

DATA AVAILABILITY STATEMENT

The PhyloHerb source code, as well as the example data set used to demonstrate its utility, are available at https://github.com/lmcai/PhyloHerb.

ORCID

Liming Cai http://orcid.org/0000-0002-8982-2435
Hongrui Zhang http://orcid.org/0000-0002-4845-260X
Charles C. Davis http://orcid.org/0000-0001-8747-1101

REFERENCES

Bakker, F. T., D. Lei, J. Yu, S. Mohammadin, Z. Wei, S. van de Kerke, B. Gravendeel, et al. 2016. Herbarium genomics: Plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society* 117: 33–43.

Cai, L., H. Zhang, and C. C. Davis. 2021. Herbariomics-based biodiversity research: from specimen to phylogeny. Botany 2021: Annual Meeting of the Botanical Society of America, held online [online abstract]. Website: https://2021.botanyconference.org/engine/search/index.php?func=detail%26aid=20 [accessed 19 April 2022].

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

Doyle, J. J. 2022. Defining coalescent genes: Theory meets practice in organelle phylogenomics. *Systematic Biology* 71: 476–489.

Folk, R. A., H. R. Kates, R. LaFrance, D. E. Soltis, P. S. Soltis, and R. P. Guralnick. 2021. High-throughput methods for efficiently building massive phylogenies from natural history collections. *Applications in Plant Sciences* 9: e11410.

Forrest, L. L., M. L. Hart, M. Hughes, H. P. Wilson, K.-F. Chung, Y.-H. Tseng, and C. A. Kidner. 2019. The limits of Hyb-Seq for herbarium specimens: Impact of preservation techniques. *Frontiers in Ecology and Evolution* 7: 439.

Freudenthal, J. A., S. Pfaff, N. Terhoeven, A. Korte, M. J. Ankenbrand, and F. Förster. 2020. A systematic comparison of chloroplast genome assembly tools. *Genome Biology* 21: 254.

Frohlich, M. W., and E. M. Meyerowitz. 1997. The search for flower homeotic gene homologs in basal angiosperms and Gnetales: A potential new source of data on the evolutionary origin of flowers. *International Journal of Plant Sciences* 158: S131–S142.

Gielly, L., and P. Taberlet. 1994. The use of chloroplast DNA to resolve plant phylogenies: Noncoding versus *rbcL* sequences. *Molecular Biology and Evolution* 11: 769–777.

Gouy, M., S. Guindon, and O. Gascuel. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* 27: 221–224.

Granados Mendoza, C., M. Jost, E. Hágsater, S. Magallón, C. van den Berg, E. M. Lemmon, A. R. Lemmon, et al. 2020. Target nuclear and off-target plastid hybrid enrichment data inform a range of evolutionary depths in the orchid genus *Epidendrum*. *Frontiers in Plant Science* 10: 1761.

Huang, D. I., and Q. C. Cronk. 2015. Plann: A command-line application for annotating plastome sequences. *Applications in Plant Sciences* 3: e1500026.

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

Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden. 2008. NCBI BLAST: A better web interface. *Nucleic Acids Research* 36: W5-W9.

Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigue, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68: 594–606.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Kozlov, A. M., A. J. Aberer, and A. Stamatakis. 2015. ExaML version 3: A tool for phylogenomic analyses on supercomputers. *Bioinformatics* 31: 2577–2579.

Lanfear, R., P. B. Frandsen, A. M. Wright, T. Senfeld, and B. Calcott. 2017. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* 34: 772–773.

Li, H.-T., Y. Luo, L. Gan, P.-F. Ma, L.-M. Gao, J.-B. Yang, J. Cai, et al. 2021. Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biology* 19: 232.

Lyra, G. d. M., C. Iha, C. J. Grassa, L. Cai, H. Zhang, C. Lane, N. Blouin, et al. 2021. Phylogenomics, divergence time estimation and trait evolution provide a new look into the Gracilariales (Rhodophyta). *Molecular Phylogenetics and Evolution* 165: 107294.

Marinho, L. C., L. Cai, X. Duan, B. R. Ruhfel, P. Fiaschi, A. M. Amorim, C. van den Berg, and C. C. Davis. 2019. Plastomes resolve generic limits within tribe Clusieae (Clusiaceae) and reveal the new genus *Arawakia*. *Molecular Phylogenetics and Evolution* 134: 142–151.

McKain, M., and M. Wilson. 2017. Fast-Plast: Rapid de novo assembly and finishing for whole chloroplast genomes. Available on GitHub: https://github.com/mrmckain/Fast-Plast [accessed 19 April 2022].

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

McKain, M. R., M. G. Johnson, S. Uribe-Convers, D. Eaton, and Y. Yang. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences* 6: e1038.

Meineke, E. K., C. C. Davis, and T. J. Davies. 2018. The unrealized potential of herbaria for global change biology. *Ecological Monographs* 88: 505–525.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37: 1530–1534.

Mirarab, S., N. Nguyen, S. Guo, L.-S. Wang, J. Kim, and T. Warnow. 2015. PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology* 22: 377–386.

Nevill, P. G., X. Zhong, J. Tonti-Filippini, M. Byrne, M. Hislop, K. Thiele, S. Van Leeuwen, et al. 2020. Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. *Plant Methods* 16: 1. https://doi.org/10.1186/s13007-019-0534-5

Nitta, J. H., J. Y. Meyer, R. Taputuarai, and C. C. Davis. 2017. Life cycle matters: DNA barcoding reveals contrasting community structure between fern sporophytes and gametophytes. *Ecological Monographs* 87: 278–296.

Palmer, J. D., and D. Zamir. 1982. Chloroplast DNA evolution and phylogenetic relationships in *Lycopersicon*. *Proceedings of the National Academy of Sciences, USA* 79: 5006–5010.

Qu, X.-J., M. J. Moore, D.-Z. Li, and T.-S. Yi. 2019. PGA: A software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* 15: 15.

Ruhfel, B. R., M. A. Gitzendanner, P. S. Soltis, D. E. Soltis, and J. G. Burleigh. 2014. From algae to angiosperms: Inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology* 14: 23.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

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

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24: 1596–1599.

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

Xi, Z., B. R. Ruhfel, H. Schaefer, A. M. Amorim, M. Sugumaran, K. J. Wurdack, P. K. Endress, et al. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences, USA* 109: 17519–17524.

Zanne, A. E., D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. FitzJohn, D. J. McGlinn, et al. 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506: 89–92.

SUPPORTING INFORMATION

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

Appendix S1. List of 98 plastid genes included in the PhyloHerb built-in reference database.

Appendix S2. Taxonomic information and GenBank accession numbers of 355 land plants included in the PhyloHerb built-in plastid reference database.

Appendix S3. List of 71 mitochondrial genes included in the PhyloHerb built-in reference database.

Appendix S4. Taxonomic information and GenBank accession numbers of 68 land plants included in the PhyloHerb built-in mitochondrial reference database.

Appendix S5. List of 155 land plants included in the PhyloHerb built-in ribosomal gene reference database.

Appendix S6. Summary statistics of the organellar and ribosomal assemblies from 10 Clusiaceae species,

Appendix S7. Example PhyloHerb workflow applied to Clusiaceae.

Appendix S8. Gene content of four pre-defined loci L1–L4.

Appendix S9. Intergenic regions in the plastid genome improve phylogenetic resolution.

Appendix S10. Plastid genome assembly statistics in the subsampled data sets.

Appendix S11. Increasing risks of systematic errors stemming from low-coverage data.

How to cite this article: Cai, L., H. Zhang, and C. C. Davis. 2022. PhyloHerb: A high-throughput phylogenomic pipeline for processing genome-skimming data. *Applications in Plant Sciences* 10(3): e11475. https://doi.org/10.1002/aps3.11475