Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

HomBlocks: A multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching

Guiqi Bi[a,b], Yunxiang Mao[a,b,c,*], Qikun Xing[a,b], Min Cao[a,b]

[a] Key Laboratory of Marine Genetics and Breeding (OUC), Ministry of Education, Qingdao, China
[b] College of Marine Life Sciences, Ocean University of China, Qingdao, China
[c] Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China

ARTICLE INFO

Keywords:
Organelle phylogenomics
Locally collinear blocks
Alignment construction
Efficient pipeline

ABSTRACT

Organelle phylogenomic analysis requires precisely constructed multi-gene alignment matrices concatenated by pre-aligned single gene datasets. For non-bioinformaticians, it can take days to weeks to manually create high-quality multi-gene alignments comprising tens or hundreds of homologous genes. Here, we describe a new and highly efficient pipeline, HomBlocks, which uses a homologous block searching method to construct multiple sequence alignment. This approach can automatically recognize locally collinear blocks among organelle genomes and excavate phylogenetically informative regions to construct multiple sequence alignment in a few hours. In addition, HomBlocks supports organelle genomes without annotation and makes adjustment to different taxon datasets, thereby enabling the inclusion of as many common genes as possible. Topology comparison of trees built by conventional multi-gene and HomBlocks alignments implemented in different taxon categories shows that the same efficiency can be achieved by HomBlocks as when using the traditional method. The availability of Homblocks makes organelle phylogenetic analyses more accessible to non-bioinformaticians, thereby promising to lead to a better understanding of phylogenic relationships at an organelle genome level.
Availability and implementation: HomBlocks is implemented in Perl and is supported by Unix-like operative systems, including Linux and macOS. The Perl source code is freely available for download from https://github.com/fenghen360/HomBlocks.git, and documentation and tutorials are available at https://github.com/fenghen360/HomBlocks.
Contact: yxmao@ouc.edu.cn or fenghen360@126.com

1. Introduction

The discord between gene trees and species trees is a common phenomenon when utilizing one or a few genes to infer species relationships [2,11]. Nevertheless, systemic error and the probability of false tree exploration will decline to satisfactory levels when a sufficiently long sequence length is used in alignment [5,12]. Therefore, the combination of multiple gene sequences has become the mainstream approach in phylogenetic studies. Owing to the characteristics of a high mutation rate and the near-absence of genetic recombination, along with the development of genome sequencing technology, organelle genomes are widely used in phylogenic and phylogeographic studies. Because genome rearrangements are frequent in some categories of taxa, it is impossible in most cases to carry out genome alignment directly. Manually constructing multi-gene alignments based on concatenation of pre-aligned single organelle gene datasets is a time-consuming and error-prone procedure, particularly when handling several dozens of genomes or genomes of large size, such as those of chloroplasts (typically greater than 100 kb with more than 70 common genes).

With the aim of improving the efficiency of sequence matrix construction derived from multitudes of organelle genomes, we developed a time-saving and accurate method that can be utilized in phylogenomics studies. In this pipeline, the core conserved fragment (protein coding genes, conserved non-coding regions, and rRNA genes) will be extracted and integrated into a long sequence from the same genome. This method avoids the time-consuming sequence alignment of every single gene and can generate a phylogenetically informative and high-quality data matrix. In contrast to days-long manual work, it typically takes less only than an hour to construct the HomBlocks matrix with approximately two dozen organelle genomes. In addition, HomBlocks produces sequence optimal partition schemes and models of sequence evolution for RAxML, which are important in downstream phylogenic analysis.

* Corresponding author.
E-mail address: yxmao@ouc.edu.cn (Y. Mao).

http://dx.doi.org/10.1016/j.ygeno.2017.08.001
Received 23 May 2017; Received in revised form 21 July 2017; Accepted 2 August 2017

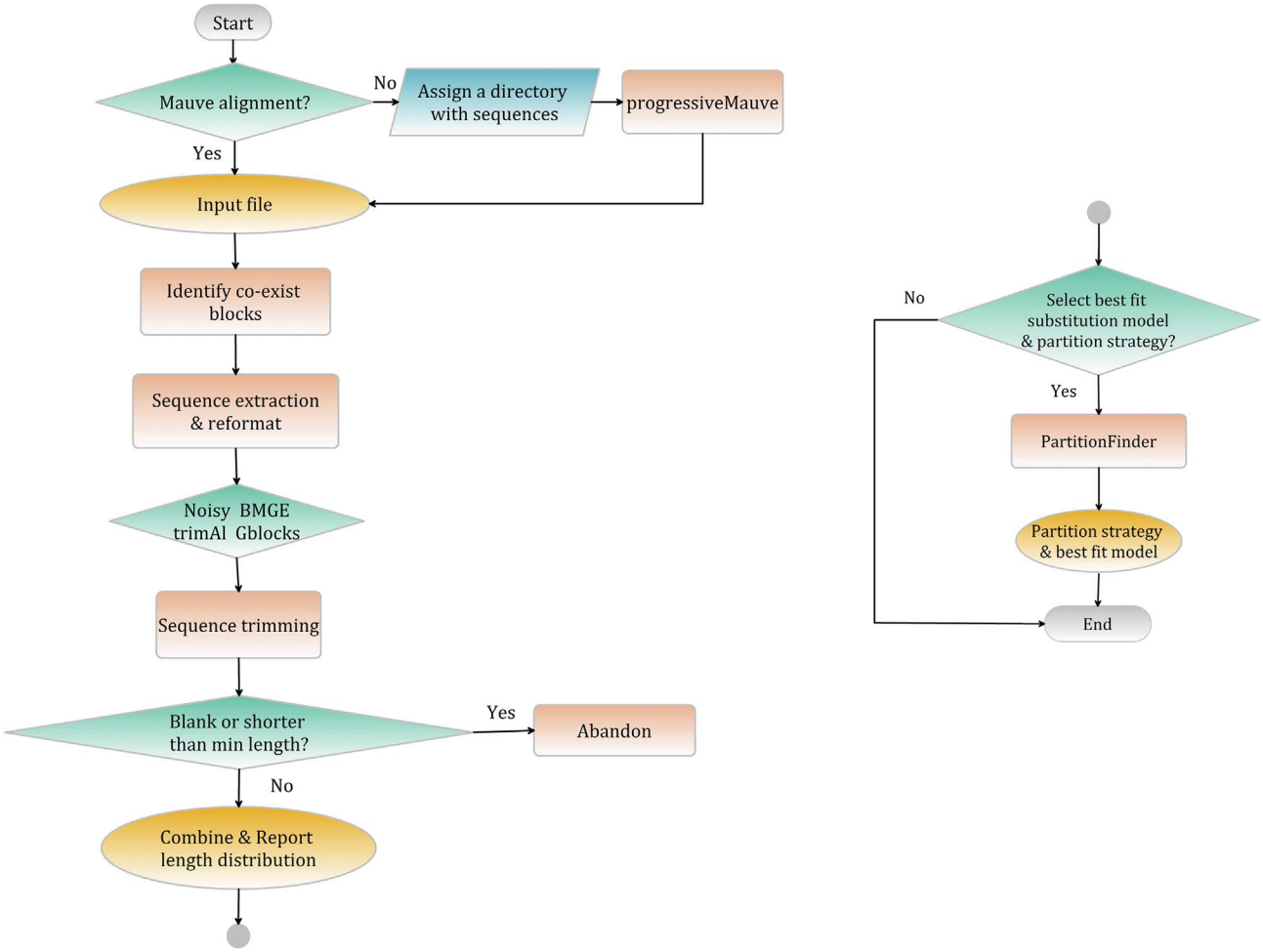

Fig. 1. Workflow diagram of HomBlocks. HomBlocks utilizes progressive Mauve to identify locally collinear blocks (LCBs) shared by organelle genomes (chloroplast and mitochondrial genomes) with default parameters. The co-exist LCBs among all organelle genomes will be extracted and trimmed to find the phylogeny informative regions. The final alignment composed of trimmed LCBs could be used in downstream analysis.

2. Methods

The framework of HomBlocks is implemented by Perl. It utilizes progressive Mauve [7], which applies an anchored alignment algorithm, to identify locally collinear blocks (LCBs) shared by organelle genomes (chloroplast and mitochondrial genomes). The LCBs co-existing among all organelle genomes will be extracted and trimmed to screen out phylogenetically informative regions. HomBlocks offers four different methods for LCB trimming: Gblocks [3], trimAl [4], noisy [8], and BMGE [6]. Without settings, the default trimming method is Gblocks. The final alignment composed of trimmed LCBs can be used in downstream analysis. Additional parameters are provided for users to select the best-fit DNA substitution model and optimal partition schemes and models of sequence evolution for RAxML with the final alignment by PartitionFinder [10].

A working flow diagram is shown in Fig. 1.

3. Example applications

We demonstrated the accuracy and efficiency of HomBlocks by comparing phylogenic trees inferred from traditionally concatenated gene alignments and HomBlocks alignments, respectively. Comparisons were composed of datasets derived from the mitochondrial genomes of 41 red algae, chloroplast genomes of 52 higher plants [15] and mitochondrial genomes of 36 xenarthrans [9]. A concise overview of the species utilized and their data sources is provided in Supplementary Tables S1–3.

For algal mitochondrial genomes, phylogenetic analyses were implemented through alignments composed of 13 pre-aligned protein coding genes and HomBlocks alignments by RAxML [14] and MrBayes 3.2.5 [13]. The topology comparison is shown in Fig. 2. Trees from two other datasets were built directly by ML and Bayes methods using HomBlocks alignments and were compared to topologies derived from corresponding references. These results are provided in Figs. 3, 4 and Supplementary Fig. S1–2, respectively. The tree topologies built using traditional and HomBlocks alignments are consistent, with the exception of certain minor differences in the bootstrap values on nodes. Moreover, all constructions of HomBlocks alignments were completed in less than half a day.

To elucidate the adaptability of HomBlocks when applied to different taxa, using an NJ tree inferred from HomBlocks matrices of the chloroplast genomes of 37 red algae (Supplementary Table S4) as a reference and the chloroplast genome of *Wildemani schizophylla* as an initial genome, we ran Homblocks iteratively by addition of a closely related species each time. These 36 alignments constructed by HomBlocks were blasted against the initial genome using BRIG [1] to determine the origins of the alignments. The results presented in Fig. 5 show that HomBlocks is adaptive and can retain different common genes when handling various taxa.

In conclusion, HomBlocks facilitates organelle phylogenomics analysis by application based on the locally collinear block searching method.

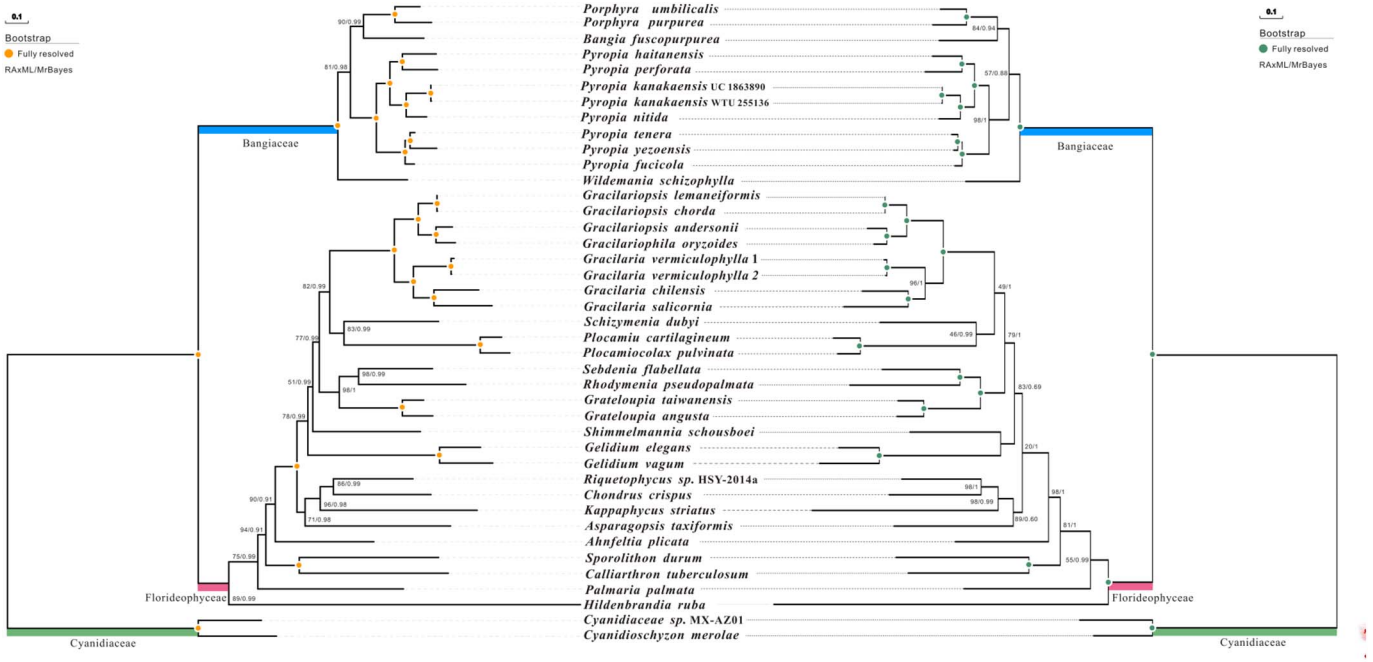

Fig. 2. Topology comparison between conventional tree (left) and HomBlocks tree (right) of 41 red algae mitochondrial genomes. Phylogenetic trees were inferred by maximum likelihood (ML) and Bayesian inference (BI) methods with both conventional alignment (13 concatenated nucleotide acid sequences of protein coding genes) and HomBlocks alignment. Numbers on the nodes represent support values inferred from RAxML (left) and Bayesian posterior probability (right), respectively. Fully resolved nodes were labeled by orange and green points, respectively.

4. Implementation and features

HomBlocks is a command-line tool and should be functional under any version of Unix or Linux, including macOS. There is no requirement for external installation, with the exception of directory uncompression. The pipeline's documentation provides a typical LCB alignment construction tutorial with test datasets.

The main features of HomBlocks compared with the traditional manual construction method are described briefly below:

- Fast and efficient. The runtimes and memory requirements of HomBlocks are highly dependent on the use cases. The pipeline took only 20 min to construct whole LCB alignments with mitochondrial genomes from 36 xenarthran species (Fig. 3 & Supplementary Table S3) using a desktop computer (IntelV R CoreTM i7-3770 CPU 3.40 GHz, 8G of RAM).
- Comprehensive. Conserved sequence fragments, including non-coding regions, unannotated coding genes, and rRNA genes, are also taken into account in the final alignments (Fig. 4).
- Adaptive. Genes shared by organelle genomes of different taxa are

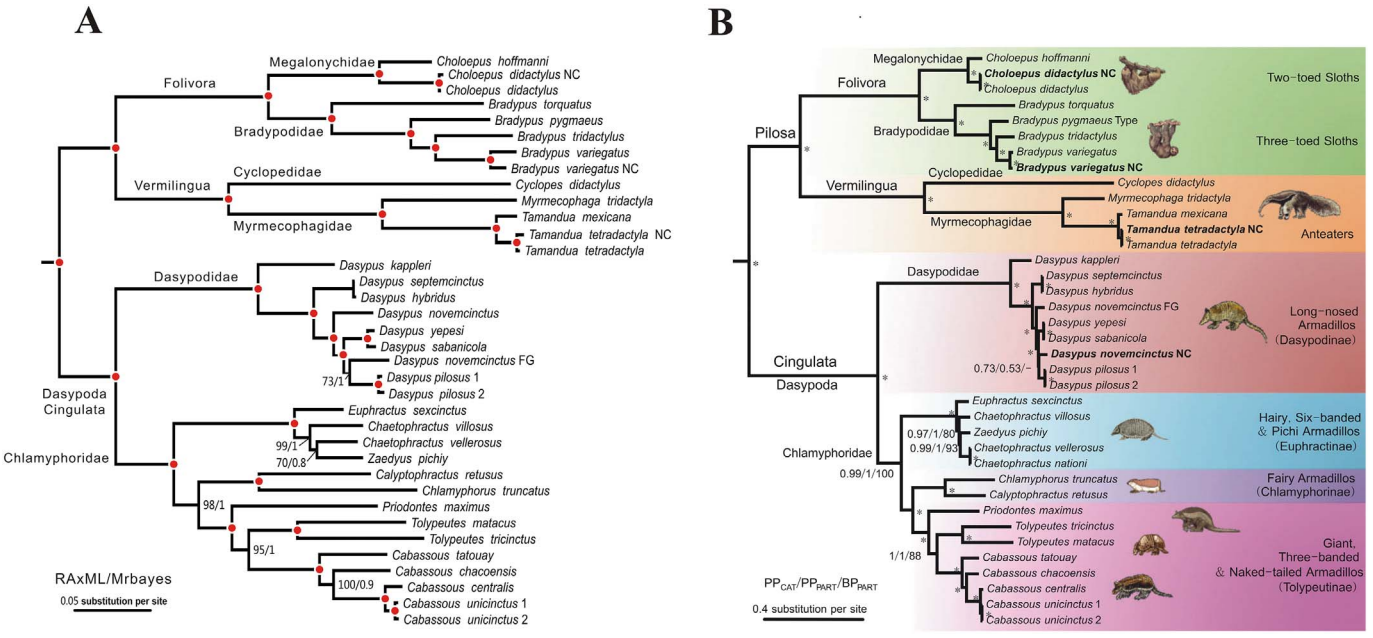

Fig. 3. Topology comparison between conventional tree (right) and HomBlocks tree (left) of 36 xenarthran mitochondrial genomes. (A) Phylogenetic tree was inferred by maximum likelihood (ML) and Bayesian inference (BI) methods with HomBlocks alignment (15,170 characters). Numbers on the nodes represent support values inferred from RAxML (left) and Bayesian posterior probability (right), respectively. Nodes with red points are fully resolved. (B) Original tree cited from the reference [9].

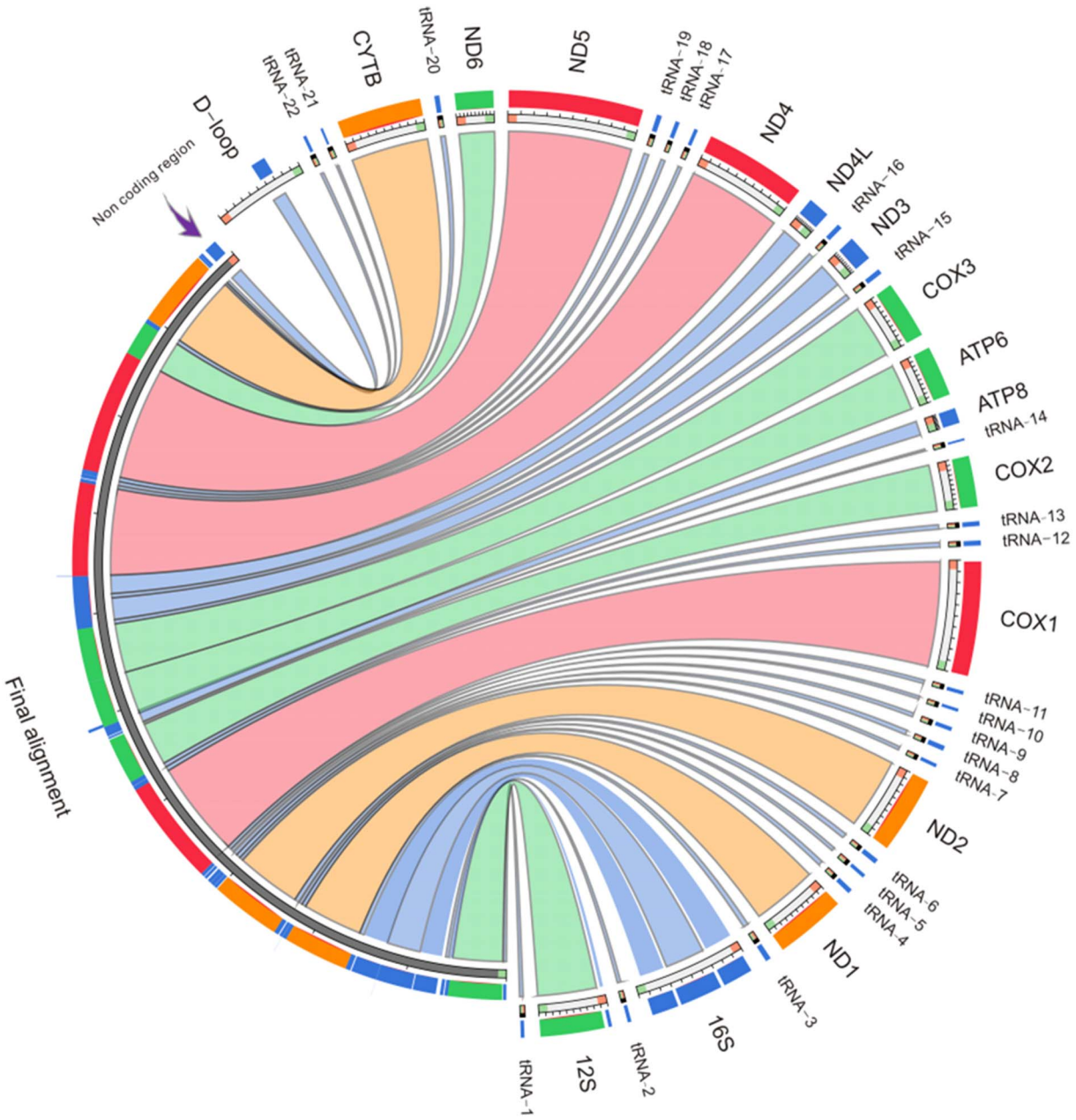

Fig. 4. Visualization of circos plot indicating genes included in concatenated sequence in example data of 36 xenarthran mitochondrial genomes. All protein coding genes and tRNAs and partial D-loop regions were integrated into the HomBlocks alignment. Links were colored by scaled matched length.

diverse and non-stable. HomBlocks consistently attempts to restore as many common genes as possible when dealing with variant datasets (Fig. 5).

- Accurate. Abundant sites guarantee high accuracy when building phylogenic trees.
- Convenient. Genome annotation is not required before implementation of HomBlocks. HomBlocks supports sequences in fasta or genbank format.

Funding

This work was supported by funding from National Natural Science Foundation of China (Grant No. 31372517), Scientific and Technological Innovation Project Financially Supported by the Qingdao National Laboratory for Marine Science and Technology (No. 2015ASKJ02), Project of National Infrastructure of Fishery Germplasm Resources (2016DKA30470), Fundamental Research Funds for the Central Universities (201762016) and Program for Chinese Outstanding Talents in Agriculture Scientific Research..

Acknowledgements

We would like to thank Chengjie Chen (College of Horticulture, South China Agricultural University), Penghao Yu (Institute of Genetics and Developmental Biology, Chinese Academy of Sciences), and Xiwen Xu (College of Informatics, HuaZhong Agricultural University) for their aid in technical support and valuable suggestions.

Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2017.08.001.

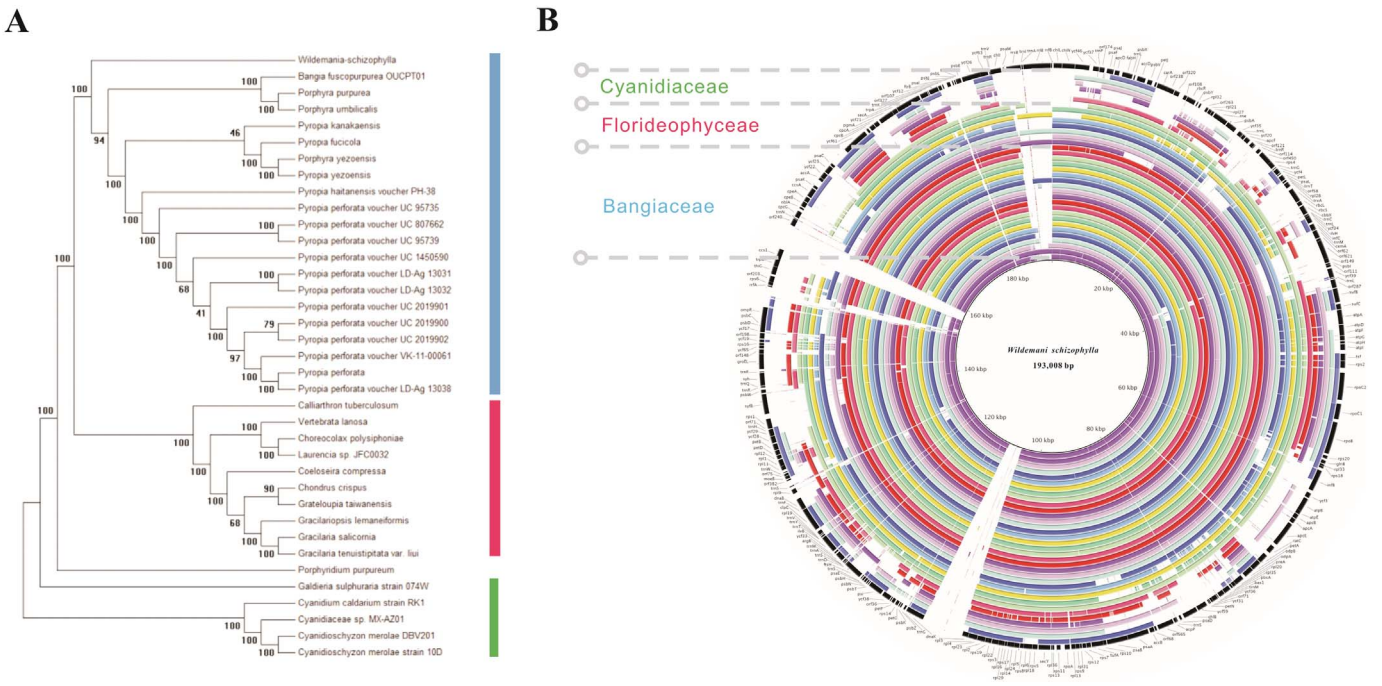

Fig. 5. Adaptability of HomBlocks when constructing alignment among different taxa. (A) The reference tree inferred by NJ method using MEGA 6 from HomBlocks alignment of 37 red algae chloroplast genomes, including three main subclasses labeled with different color strips. (B) Results of BLAST Ring Image Generator [1] showing locations of these alignments on the chloroplast genome.

References

[1] N.F. Alikhan, et al., BLAST ring image generator (BRIG): simple prokaryote genome comparisons, BMC Genomics 12 (1) (2011) 402.
[2] E. Bapteste, et al., Do orthologous gene phylogenies really support tree-thinking? BMC Evol. Biol. 5 (1) (2005) 33.
[3] J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, Mol. Biol. Evol. 17 (4) (2000) 540–552.
[4] S. Capella-Gutiérrez, J.M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, Bioinformatics 25 (15) (2009) 1972–1973.
[5] T.M. Collins, et al., Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics, Syst. Biol. 54 (3) (2005) 493–500.
[6] A. Criscuolo, S. Gribaldo, BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments, BMC Evol. Biol. 10 (1) (2010) 210.
[7] A.C. Darling, et al., Mauve: multiple alignment of conserved genomic sequence with rearrangements, Genome Res. 14 (7) (2004) 1394–1403.
[8] A.W. Dress, et al., Noisy: identification of problematic columns in multiple sequence alignments, Algorithms for Molecular Biology 3 (1) (2008) 7.
[9] G. Gibb, et al., Shotgun mitogenomics provides a reference phylogenetic framework and timescale for living xenarthrans, Mol. Biol. Evol. 33 (3) (2016) 621–642.
[10] R. Lanfear, et al., PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses, Mol. Biol. Evol. 29 (6) (2012) 1695–1701.
[11] W.P. Maddison, Gene trees in species trees, Syst. Biol. 46 (3) (1997) 523–536.
[12] D.D. Pollock, Genomic biodiversity, phylogenetics, and coevolution in proteins, Appl. Bioinforma. 1 (2) (2002) 81.
[13] F. Ronquist, et al., MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space, Syst. Biol. 61 (3) (2012) 539–542.
[14] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (9) (2014) 1312–1313.
[15] D. Zhang, et al., The complete plastid genome sequence of the wild rice *Zizania latifolia* and comparative chloroplast genomics of the rice Tribe Oryzeae, Poaceae, Front. Ecol. Evol. 4 (2016) 88.