realDB: a genome and transcriptome resource for the red algae (phylum Rhodophyta)

Fei Chen1,*,†, Jiawei Zhang1,†, Junhao Chen1,2,†, Xiaojiang Li1,†, Wei Dong1, Jian Hu1, Meigui Lin1, Yanhui Liu1, Guowei Li3,4, Zhengjia Wang2 and Liangsheng Zhang1,*

1State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Key Laboratory of Genetics, Breeding and Multiple Utilization of Corps, Ministry of Education, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou 350002, China, 2State Key Laboratory of Subtropical Silviculture, School of Forestry and Biotechnology, Zhejiang Agriculture and Forestry University, Hangzhou 311300, China, 3Biotechnology Research Center, Shandong Academy of Agricultural Sciences, Jinan 250100, China and 4College of Life Science, Shandong Normal University, Jinan 250014, China

*Corresponding author: Tel.: +86 17706919003; Fax: 0591 88202367; Email: frank@fafu.edu.cn
Correspondence may also be addressed to Liangsheng Zhang. Tel.: +86 591 88202367; Fax: +86 17720805596; Email: fafuzhang@163.com
†These authors contributed equally to this work.

Citation details: Chen,F., Zhang,J., Chen,J. et al. realDB: a genome and transcriptome resource for the red algae (phylum Rhodophyta). Database (2018) Vol. 2018: article ID bay072; doi:10.1093/database/bay072

Received 29 December 2017; Revised 11 June 2018; Accepted 15 June 2018

Abstract

With over 6000 species in seven classes, red algae (Rhodophyta) have diverse economic, ecological, experimental and evolutionary values. However, red algae are usually absent or rare in comparative analyses because genomic information of this phylum is often under-represented in various comprehensive genome databases. To improve the accessibility to the ome data and omics tools for red algae, we provided 10 genomes and 27 transcriptomes representing all seven classes of Rhodophyta. Three genomes and 18 transcriptomes were de novo assembled and annotated in this project. User-friendly BLAST suit, Jbrowse tools and search system were developed for online analyses. Detailed introductions to red algal taxonomy and the sequencing status are also provided. In conclusion, realDB (realDB.algaegenome.org) provides a platform covering the most genome and transcriptome data for red algae and a suite of tools for online analyses, and will attract both red algal biologists and those working on plant ecology, evolution and development.

Database URL: http://realdb.algaegenome.org/
**Introduction**

Red algae (phylum Rhodophyta) have various values in our daily life. They are important sources of food, such as nori used in sushi and pudding made of Irish moss. The high content of vitamins and proteins of red algae-derived foods has made them attractive and popular in east Asia for >1000 years (1). Red algae have valuable ecological roles, such as producing oxygen in the seawater while some species are important in the formation of tropical reefs. In many Pacific atolls, red algae have contributed far more to reef structure than other organisms including corals (2). In the oceans, various species of red algae are primary producers eaten by fish, crustaceans, worms and gastropods.

Red algae occupy the second basal branch in the green lineage following the Glaucophyta algae (3). Some red algal species have important evolutionary value for studying basic biological questions such as the origin of multicellularity (4), symbiosis (5) and evolution of photosynthesis. There are about 6000 species of red algae [Source: AlgaeBase (6), www.algaebase.org], ranging from single-celled species to complex, multi-cellular, ‘plant-like’ organisms. They are also excellent material to study symbiosis, since many are inexorably associated with other organisms. Some species are used to produce agars, which are gelatinous food additives and in science labs as a support substance in culture media (7).

The current available red algae related data, such as those included in AlgaeBase (www.algaebase.org) and Porphyra website (http://www.porphyra.org/), are limited to morphological descriptions. The integration of genome data and morphological data is in its beginning stage. For instance, the comprehensive database phytozome V12 (phytozome.jgi.doe.gov/pz/portal.html), plant genome duplication database (PGDD, chibba.agtec.uga.edu/duplication), plant genome database (PlantGDB, plantgdb.org) (release V187) and plant genome duplication database (PGDD, chibba.agtec.uga.edu/duplication), plant genome database (PlantGDB, plantgdb.org) (release V187) and plant genome and systems biology (PGSB, pgsb.helmholtz-muenchen.de/plant) database do not include any red algae genome. The pico-Plaza 2.0 (bioinformatics.psb.ugent.be/plaza/versions/pico-plaza/) database include one red algal genome, while CoGe database has two and Ensembl Plant database (plants.ensembl.org) has three (Figure 1). This dearth of information leads to the underestimation of the biological importance of red algae. Comparative analysis of red algae species, such as the evolutionary studies of genes families (8, 9), non-coding genes and small RNAs (10) lag far behind in plant science, partly because of the difficulty in obtaining red algae genomic information. In breeding, an open platform integrating various omics data and species information is the demand for scientists and breeders (11). To meet the ever-rapid increasing amounts of genomic and transcriptome data and their tremendous potential in understanding developmental process (12) and to assist molecular breeding, we build an online, searchable platform for integrating the omic data with the use of multiple omics tools. The Information gained will be a valuable to boost the understanding red algae genomes and the evolution of plant genomes.

**Data description**

**Dataset**

Sequences from six genomes (including partial genomes) and seven transcriptomes, together with annotation data were downloaded directly from public available websites. These data were shared freely on these websites, which provided no or little online analysis tools. Raw reads of four genomes and 20 transcriptomes were downloaded from NCBI-SRA (www.ncbi.nlm.nih.gov/sra) database without annotations, thus were de novo assembled and annotated in this study (Supplementary Table S1). The red algal transcription factors were predicted in this study, relying on the HMMsearch tool from the HMMER software (hmmer.org) with default parameters and homology seeds from Pfam database (pfam.xfam.org).

**Assembly and annotation of genomes and transcriptomes**

All the original reads from the downloaded raw data were filtered using Trimmomatic (13) (https://github.com/tim/flutre/trimmomatic) to remove the adapters and low-quality reads. These clean reads were then de novo assembled using the software Trinity (14) (https://github.com/trinityrnaseq/trinityrnaseq). Trinity produced the transcriptome files in FASTA format and the assembled sequences were then used for gene identification. TransDecoder was integrated in Trinity software and was employed for detecting gene regions (https://github.com/TransDecoder/TransDecoder/). Kyoto Encyclopedia of Genes Genomes (KEGG) and Enzyme Commission data were both obtained by BLAST genes with the KEGG database (https://www.kegg.jp/kegg/).

**Database construction**

The realDB database employs Aliyun, one of the largest cloud server providers in the world, thus facilitates realDB outstanding advantages such as (i) scalability in easily expanding its storage size and computing ability, (ii) more stability and (iii) simple to maintain. The realDB relies on the Linux Ubuntu Server 14.04.4, Apache2.4.18, Java
Results and discussions

An updating timeline for the sequenced red algae

To attract more visits to our online platform, we created an updating timeline system on the homepage of realDB that updates the recently sequenced genome or transcriptome of red algal species (Figure 2). This timeline system consists of >1000 lines of code adapted from vis.js (http://visjs.org/), dedicated to providing multiple forms of information, including the release time, genome size, reference and authors. User can click the hyperlink to browse the reference or related linked websites for additional information. The defining feature of this timeline tool is its dynamics and interactive features with species information. Users can move the timeline space and zoom in or zoom out of the timeline by dragging and scrolling in the species timeline zone. The time-scale on the axis is adjusted automatically (http://almende.github.io/chap-links-library/graph.html), supporting scales ranging from milliseconds to years. We will create new items when genome or transcriptome from other red algal species become available.

Bootstrap boosted framework for various display facilities

Users are able to check our genome updates and news via mobile phone using the Bootstrap framework, which is the world’s leading framework for building responsive, mobile-first sites (15). Users are able to check updates of genome releases or website news of realDB via mobile, ipad, laptop and desktop using all popular web browsers including Google Chrome, Safari, Firefox, Internet Explorer, etc. without any display difficulty (Figure 2).

Various introductions to red algae for wide readership

The lack of red algal genome sequences in various databases is partly due to the limited knowledge of red algae. The molecular biological studies of red algae provide many useful results, such as information on systematics, physiology, ecology and evolution. Concise introductions to each species assist visitors with different backgrounds to quickly decide which species to analyze. Most of the information for each species was provided and cited from the book ‘Red Algae in the Genomic Age’ (16), including descriptions of life histories, forms and styles, genomic information and data sources. We also provided the description and classification of red algae on the website because general researchers and comparative genomic biologists usually do not have extensive knowledge of red algae classification or morphology.

realDB covers the largest number of red algae with ome data

The current realDB V1.0 gathered 10 available genomes and 27 transcriptomes, representing all the 7 classes in the Rhodophyta. Among this dataset, we de novo assembled the genomes of Galdieria phlegrea, Gracilariaopsis lemaneiformis and Porphyridium cruentum, and 18 transcriptomes (Table 1). realDB has provided 37 ome
datasets, including genome and transcriptome sequences (Figure 2). In comparison, the green lineage oriented genome database Phytozome (phytozome.jgi.doe.gov) does not contain any genome/transcriptome data of red algae. Furthermore, the algae-oriented database Pico-PLAZA (bioinformatics.psb.ugent.be/plaza_versions/pico-plaza) harbors only one red algal genome, and the plant-specific database Ensemble Plant (plants.ensembl.org) has included only three red algal genomes (Figure 2). In realDB, we selected Chondrus crispus, Cyanidioschyzon merolae, Galdieria sulphuraria, G. phlegrea as flagship red algae with the best genome sequencing and assembly.

A suite of toolbox for online analysis

Besides the downloadable dataset, online tools would facilitate data retrieval and comparative analyses. Currently, realDB provides a complete suite of BLAST tools (Figure 3) consisting of BLASTn, BLASTx, BLASTp, tBLASTn and tBLASTx. This BLAST suit was constructed using the sequenceserver tool (www.seqenceserver.com/). A list of 21 advanced parameters such as -evalue 1.0e-10 -max_target_seqs 10 are optional for searches. For the nucleotides, coding sequences (CDS) and genomes were separated and could be individually selected. Users will find the GenBank style formatted BLAST results easy to use and download hits in FASTA format, and align data in tab-delimited or XML formats.

JBrowse was incorporated into realDB (Figure 3), allowing users to instantly browse, visualize, and retrieve sequence data. Currently, we provided the Jbrowse tool for C. crispus, C. merolae, G. sulphuraria, which are well assembled and annotated genomes. Using Jbrowse tool, users can easily browse and analyze these genomes at various scales with a graphic interface. Detailed gene information could be conveniently viewed and fetched by zooming in and out the interested genomic region, to view the information such as location, annotation and sequences by clicking on the corresponding tracks.
The search tools in realDB provide a series of search service for CDS, protein, gene annotation, gene family, transcription factors and miRNA information (Figure 4). These information will be useful for both wet lab and dry lab biologists. Gene families, especially transcription factor families, control various physiological processes and are breeding targets (17–27). miRNAs have been extensively studied in land plants and green algae (12, 28, 29). However, little is known about its function and evolutionary trajectory in red algae. We incorporated four miRNA datasets that have been experimentally validated from *Porphyridium purpureum* (30) (Porphyridiophyceae), *C. crispus* (31) (Florideophyceae), *Eucheuma denticulatum* (32) (Florideophyceae), *P. yezoensis* (33) (Bangiophyceae) into realDB. Users can easily discover a miRNA and related information through our search system.

### Conclusion and future perspectives

Red algae (Rhodophyta) have a critical place in plant evolution as the second branch after Glaucophyta, attracting...
thousands of scientists in areas of ecology, evolution and genomics. They are also attractive to people working on bioengineering, medicine and food science. This study of red algae is facing the rapid development of genomics. Facilitated by low-cost and fast sequencing technologies, more and more red algae have their genomes and transcriptomes sequenced. realDB is dedicated to being the leading platform for analyzing red algae genomes by providing the latest omics data and online analysis tools. Currently, we provide the most genome and transcriptome data for 37 red algae that are freely available to all researchers. The realDB Version 1.0 database is the first release and will be updated when new datasets are available. Furthermore, we will incorporate additional bioinformatics tools for easier
data access and online analyses. Since its release in September 2017, realDB has attracted the attention of scientists from around the world, and the website has been visited by researchers from 27 countries (April 2018). All people interested in realDB are encouraged to contact us for data sharing and collaboration. We are dedicated to collaborating with international teams to collect more data and develop more tools, hoping to make realDB the most influential database for red algae studies.

Supplementary data
Supplementary data are available at Database Online.

Funding
E.C. is supported by a grant from natural science foundation of Fujian Province (2018J01603) and a grant from State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops (SKB2017004). L.Z. is supported by the National Natural Science Foundation of China (81502437), and a start-up fund from Fujian Agriculture and Forestry University. G. L. is supported by a grant from Shandong Province Natural Science foundation (ZR2014YL043). Funding to pay the Open Access publication charges for this article was provided by a grant from natural science foundation of Fujian Province (2018J01603).

Conflict of interest. None declared.

References
1. Nisizawa, K., Noda, H., Kikuchi, R. et al. (1987) The main seaweed foods in Japan. Hydrobiologia, 151–152, 5–29.
2. Abbott, I.A. (1999) Marine Red Algae of the Hawaiian Islands. Bishop Museum Press, Honolulu.
3. Burki, F., Alegado, R.A., King, N. et al. (2014) The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb. Perspect. Biol., 6, a016147.
4. Grosberg, R.K. and Strathmann, R.R. (1998) One cell, two cell, red cell, blue cell: the persistence of a unicellular stage in multicellular life histories. Trends Ecol. Evol., 13, 112–116.
5. Lee, J.J., Cervasco, M.H., Morales, J. et al. (2010) Symbiosis drove cellular evolution: symbiosis fueled evolution of lineages
of Foraminifera (eukaryotic cells) into exceptionally complex giant protists. *Symbiosis*, 51, 13–25.

6. Guiry,M.D., Guiry,G.M., Morrison,L. et al. (2014) AlgaeBase: an on-line resource for algae. *Cryptogam. Algol.*, 35, 105–115.

7. Dahl,A.L. and Dixon,P.S. (1974) Biology of the Rhodophyta. *Taxon*, 23, 391–392.

8. Zhang,L. and Ma,H. (2012) Complex evolutionary history and diverse domain organization of SET proteins suggest divergent regulatory interactions. *New Phytol.*, 195, 248–263.

9. Chen,F., Zhang,L. and Cheng,Z.-M. (2017) The calmodulin fused kinase novel gene family is the major system in plants converting Ca<sup>2+</sup> signals to protein phosphorylation responses. *Sci. Rep.*, 7, 4127.

10. Taylor,R.S., Tarver,J.E., Hiscock,S.J. et al. (2014) Genome-wide analysis and transcriptome analysis of the grapevine (*Vitis vinifera*) WRKY gene family. *Hort. Res.*, 3, 16056.

11. Adam-Blondon,A.-F., Alaux,M., Pommier,C. et al. (2017) The WRKY transcription factor family in model plants and crops. *Crit. Rev. Plant Sci.*, 36, 311–335.

12. Liu,J., Chen,N., Chen,F. et al. (2014) Genome-wide analysis and expression profile of the bZIP transcription factor gene family in grapevine (*Vitis vinifera*). *BMC Genomics*, 15, 281.