scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level assembly and gene annotation of *Kappaphycus striatus* genome

Zhiyin Zhou[1,2,3], Yu Ma[1,2,3], Jie Zhang[1,2], Muhammad Firdaus[4], Michael Y. Roleda[5] & Delin Duan[1,2] ✉

Kappaphycus striatus **is one of the carrageenan-producing red algae, and found primarily in tropical and subtropical coastal regions. Its global distribution is mainly in the Philippines, Indonesia, and Malaysia, among other locations. Here, through the high-quality chromosome-level genome sequences and assembly with PacBio HiFi and Hi-C sequencing data, we assembled one genome with a total of 211.46 Mb in size, containing a contig N50 length of 5.04 Mb and a scaffold N50 length of 5.39 Mb. After Hi-C assembly and manual adjustment to the heatmap, we deduced that 199.42 Mb of genomic sequences were anchored to 33 presumed chromosomes, which accounting for 94.31% of the entire genome. One total of 14,596 protein-coding genes and 1,673 non-coding RNAs were identified, and the 100.96 Mb of repetitive sequences accounting for 47.73% of the assembled genome. Our chromosome-level genome assembly data provide valuable references for *K. striatus* future nursery and breeding, and will be useful for the functional genomics interpretations and evolutionary studies of eukaryotes.**

Background & Summary

As one of important red seaweeds, the eucheumatoids are members of the family Solieriaceae found in the tropical and subtropical regions worldwide and contributed significantly to the global carrageenan production. Taxonomically, it includes the genera *Kappaphycus* (kappa-carrageenan) and *Eucheuma* (iota-carrageenan), and were used in the food and cosmetic industries due to their different chemical properties[1].

In Southeast Asian countries, the commercial cultivation and processing of eucheumatoids provided pivotal livelihood sources for coastal communities[2–4], and their annual cultivation output has become the highest in the seaweed aquaculture[5]. Besides, the red alga was served as an ideal material for eukaryotic phylogenesis, particularly in illustrating endosymbiotic evolution, and morphological diversity and ecological functions[6]. *K. striatus* (known as "Green Sacol" variety), was a primary source for κ-carrageenan extraction[1]. Morphologically, *K. striatus* is characterized by densely packed thick cylindrical branches and blunt, bifurcated tips, with a diameter not exceeding 5 mm. Generally it exists in two forms: one is erect with irregular branching and acute branch axils, and the other forms a dense, decumbent clump with dichotomous branching[7]. The molecular analysis of eucheumatoids is unclear due to the absence of records on chromosome-level genomic studies. Only one genome draft of *K. alvarezii* assembled with PacBio and HiSeq sequencing data were reported[8], which is incomplete and lacks comprehensive genomic annotations[9]. Therefore, there is a need to decipher a high-quality reference genome for its genomic structure and subsequent genetic and evolutionary study.

Using Illumina short-reads sequencing, PacBio long-reads sequencing, and high-throughput chromosomal conformation capture (Hi-C) analysis, we constructed one high-quality *K. striatus* reference genome on the chromosomal level. Our yielded data will be positive to the understanding of micro-revolution and will be useful to the tropical seaweed nursery and breeding in the future.

[1]Key Lab of Breeding Biotechnology & Sustainable Aquaculture, Shandong Province Key Lab of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, 266071, P. R. China. [2]Lab for Marine Biology and Biotechnology, Qingdao Marine Science and Technology Center, Qingdao, 266237, P. R. China. [3]University of Chinese Academy of Sciences, Beijing, 100049, P. R. China. [4]Research Center for Marine and Land Bioindustry, National Research and Innovation Agency (BRIN), Lombok Utara, Nusa Tenggara Barat, 83352, Indonesia. [5]Marine Science Institute, University of the Philippine, Diliman, 1101, Philippines. ✉e-mail: dlduan@qdio.ac.cn

Fig. 1 *K. striatus* sample used for sequencing.

Platform	Insert size (bp)	Clean data (Gb)	Average read length (bp)	N50 read length (bp)	Coverage (X)
Illumina	350	48.45	150	150	239
PacBio	15 K	18.48	14,625	18.27 K	91
Hi-C	350	99.50	150	150	470
RNA-sequence	350	7.58	150	150	36

Table 1. Statistics of the genome sequencing of *K. striatus*.

Methods and Result

Sample preparation and nucleic acid preparation. *K. striatus* was collected from a seaweed farm in Seriwe Bay, Seriwe village, East Lombok, West Nusa Tenggara, Indonesia (Fig. 1). The algal thallus was cut for nucleic acid extraction and followed by library construction. Total genomic DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method[10]. The quality of the extracted DNA was assessed on the 0.5% agarose gel electrophoresis and was quantified on the Qubit 4.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), totally 97.2 ng/μL of DNA was collected for sequenced, assembly and annotation. RNA extraction was conducted with Polysaccharide Polyphenol Plant Total RNA Extraction Kit (DP441, TianGen, China). The quality of the extracted RNA was determined on the a Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA) and a Qubit 4.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), with the RNA detected as 149.6 ng/μL.

Library construction and sequencing. For Illumina sequence data obtaining, a 350 bp paired-end library was constructed followed the manufacturer's protocol (Illumina, San Diego, CA, USA) and was sequenced on the Illumina NovaSeq 6000 platform. This yielded 48.45 Gb of clean data, with approximately 239 × coverage of the estimated *K. striatus* genome size (Table 1).

For the PacBio sequencing, one 15Kb library was constructed using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA). The concentration and size of the library were assessed using the Qubit and Agilent 2100. The prepared library was then bound to primers and polymerase (Pacbio, USA) with the PacBio Binding Kit (Pacbio, USA) for the subsequent sequencing. The reaction products were purified using AMpure PB Beads (Pacbio, USA) for sequencing analysis on the PacBio Revio platform. Finally, a total of 18.48 Gb of clean data were generated, covering approximate 91× of the *K. striatus* genome, with an N50 read length of 18.27 Kb and an average read length of 14.63 Kb (Table 1).

For the Hi-C library construction, the DNA fixed using formaldehyde was digested with the restriction enzyme (*Dpn*II), which creates sticky ends. The digested fragments were biotin-labeled for the end repair and subsequently circularized. The constructed libraries were go through the sequenced on the Illumina NovaSeq 6000 platform. After removing the adapters, primer sequences and filtering low-quality data, we obtained 99.50 Gb of Hi-C clean reads, it covered 470 × estimated *K. striatus* genome (Table 1).

The RNA-seq libraries were also constructed with the Illumina standard protocol (San Diego, CA, USA), and were sequenced on the Illumina NovaSeq 6000 platform. We totally obtained 7.58 Gb of clean reads for subsequent gene prediction and annotation (Table 1).

***De novo* genome assembly.** Prior to genome assembly, we estimated genome size and heterozygosity with the k-mer analysis. The short-reads from Illumina NovaSeq X plus platform were subjected to quality filtration using fastp. We counted 21-mers using Jellyfish software (https://github.com/gmarcais/Jellyfish), and analyzed

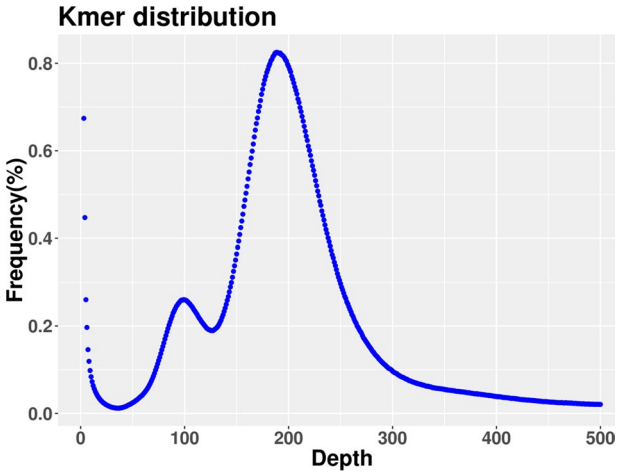

Fig. 2 The 21-mer analysis of the genome.

The 21-mer analysis		
Estimated genome size (Mb)	202.87	
Heterozygosity	0.48%	
Repeat rate	40.08%	
GC content	43.67%	
Genome assembly		
	PacBio assembly	Hi-C assembly
Total length (bp)	211,461,192	211,462,692
Contig/Scaffold number	210	198
Contig/Scaffold N50 (bp)	5,039,840	5,386,916
Contig/Scaffold N90 (bp)	1,601,280	3,550,198
Max length (bp)	7,918,904	17,093,109
GC content	45.48%	45.48%

Table 2. Statistics of the genome assembly of *K. striatus*.

genome characteristics with Genomescope software (https://github.com/tbenavi1/genomescope 2.0). The estimated *K. striatus* genome size was about 202.87 Mb, with a repeats of 40.08%, and 0.48% heterozygosities (Fig. 2 & Table 2).

We applied the 18.48 Gb of HiFi long-read data and employed Hifiasm (v0.19.8; l = 2, n = 4)[11] software for *de novo* assembly of *K. striatus* genome. By aligning the *K. striatus* genome to the NCBI NT database, and with mitochondrial (https://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/) and plastid (https://ftp.ncbi.nlm.nih.gov/refseq/release/plastid/) databases, we filtered out any NT contamination and organelles sequences. The final assembled genome was 211.46 Mb with 213 contigs and an N50 of 5.04 Mb (Table 2).

For anchored contigs screening, 332,238,328 clean reads pairs were generated from the Hi-C libraries, and were mapped to the polished genome using BWA (bwa-0.7.17)[12] with the default parameters. The paired reads with mates were mapped with different contigs were associated with the Hi-C scaffolding. Self-ligation, non-ligation and other invalid reads were filtered, such as Start Near Rsite, PCR amplification, random break, Large *Sma*I Fragments and Extreme Fragments. Finally, we clustered 83 scaffolds on the 33 groups with the agglomerative hierarchical clustering method in Lachesis[13]. Subsequent manual adjustment and inspection were conducted to refine the chromosome-level genome to *K. striatus*. One total of 199.42 Mb of genomic sequences were mapped on the 33 level chromosomes (94.31% coverage). The scaffold N50 is 5.39 Mb, chromosome sizes range from 3.38 Mb to 17.95 Mb (Fig. 3a,b & Table 3).

Repetitive and non-coding gene prediction. To screen the repetitive sequences on the *K. striatus* genome, we adopted the *de novo* and homology-based approaches. The *de novo* prediction was applied with Repeat Modeler (v2.0.1)[14] (http://www.repeatmasker.org/RepeatModeler/), with RECON (v1.0.8)[15] and Repeat Scout (v1.0.6)[16] software. The yielded sequences data were classified into repeat families with Repeat Classifier with the Dfam (v3.5) database. Long terminal repeats (LTRs) were predicted with LTR harvest (v1.5.10)[17] and LRT_FINDER (v1.07)[18], and the predictions were integrated using LTR_retriever (v2.9.0)[19]. When merging data, the co-estimations and plotting, the resolving redundancies in both *de novo* predictions and known databases, one species-specific repeat sequence database was finally concluded. The Repeat Masker (v4.1.2)[20] was employed to predict the transposable elements (TEs) with the constructed repeat sequence database.

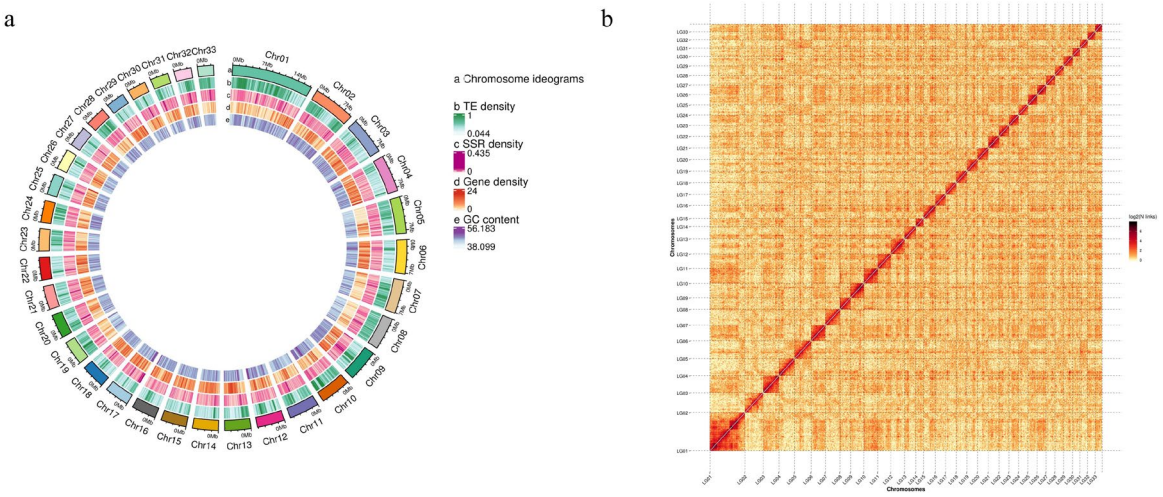

Fig. 3 Characteristics of *K. striatus* genome assembly. (**a**) A circos plot of 33 chromosomes in *K. striatus* genome. a, chromosome ideograms. (**b**) TE density. c, SSR density. d, GC content. (**b**) Hi-C assembly of chromosome interactive heatmap. Distinct groups of 33 chromosomes are clearly visible. Within each group, interactions along the diagonal are more intense than those off-diagonal.

Chromosome ID	Length (bp)	Number of Scaffolds
1	17,947,393	7
2	9,564,713	7
3	7,951,944	2
4	8,012,697	3
5	7,760,998	2
6	7,188,659	1
7	7,135,519	1
8	6,816,312	1
9	6,667,489	2
10	6,827,162	3
11	6,515,965	2
12	5,811,166	2
13	5,738,876	3
14	5,495,554	4
15	5,386,188	1
16	5,265,096	2
17	6,599,392	12
18	5,142,211	1
19	5,108,739	1
20	5,071,504	2
21	5,773,119	3
22	4,765,071	1
23	4,803,715	3
24	4,644,658	2
25	4,463,540	1
26	4,373,571	1
27	4,313,296	3
28	4,209,468	1
29	3,886,276	1
30	3,966,074	2
31	3,800,127	1
32	5,040,784	3
33	3,375,065	2
Total	199,422,341	83
Mean	6,043,101	3

Table 3. Statistics of 33 chromosomes of *K. striatus* genome.

Type	Number	Length (bp)	Percent (%)
DNA	29,104	12,074,977	5.71
DIRS	1,535	1,941,557	0.92
LINE	34,770	16,419,794	7.76
SINE	638	74,245	0.04
LTR	71,774	67,858,401	32.08
Tandem Repeat	26,367	2,587,113	1.22
Unknown	4	169	0
Total	164,192	100,956,256	47.73

Table 4. Statistics of repeat elements in *K. striatus* genome.

Type	Number	Length (bp)	Percent
rRNA	593	732,506	0.00346%
tRNA	1080	81,645	0.00038%
miRNA	0	0	0
snRNA	0	0	0
snoRNA	0	0	0

Table 5. Statistics of ncRNAs in *K. striatus* genome.

Tandem repeat sequences were predicted using the MicroSatellite identification tool (MISA v2.1)[21] and the Tandem Repeat Finder (TRF v4.09)[22]. Finally, we identified one total of 100.96 Mb of repetitive sequences, which account for 47.73% of the assembled genome. Among all the repetitive sequences, the most abundant elements were LTRs, comprising 67.86 Mb and representing 32.08% of the genome, followed by the long interspersed nuclear elements (LINEs), which accounted for 16.42 Mb (7.76%), and DNA transposons 12.07 Mb (5.71%) (Table 4).

We either employed various strategies to predict the noncoding RNA (ncRNA) in the genomes. The RNAscan-SE (v1.3.1)[23] algorithms with default parameters were used for identifying the tRNA; and rRNA were detected with barrnap (v0.9) (http://lup.lub.lu.se/student-papers/record/8914064); miRNAs, snoRNAs, and snRNAs were applied for prediction on the Rfam (v14.5)[24] database using Infernal (v1.1)[25]. Totally, 1080 tRNAs, 593 rRNAs were identified (Table 5).

Gene prediction and annotation. We integrated the *de novo* prediction, homologous searching and transcriptome-assisted approaches for annotating the protein-coding sequences. The *de novo* gene models were predicted using two ab initio gene-prediction software tools, Augustus (v3.1.0)[26] and SNAP(2006-07-28)[27]. For homology-based prediction, the protein sequences of *Arabidopsis thaliana* (https://www.arabidopsis.org/download/list?dir=Genes%2FTAIR10_genome_release), *Chondrus crispus* (GCA_000350225.2), *Gracilariopsis chorda* (GCA_003194525.1), *Gracilaria domingensis* (GCA_022539475.1), *Neopyropia yezoensis* (GCA_009829735.1) were collected from the NCBI database and was aligned with *K. striatus* genome with GeMoMa (v1.7)[28]. For transcriptional analysis, RNA-sequencing data were mapped to the reference genome using Hisat (v2.1.0)[29] and assembled by Stringtie (v2.1.4)[30]. GeneMarkS-T (v5.1)[31] was used to predict those genes based on the transcription data. The PASA (v2.4.1)[32] software was used to predict genes based on the unigenes and full-length transcripts from the PacBio (ONT) sequencing, which was assembled by Trinity (v2.11)[33]. Gene models from these different approaches were combined using the EVM software (v1.1.1)[32] and updated with PASA (v2.4.1). Among the 15,341 predicted protein-coding genes, the average gene length was 1,804.72 bp, with an average coding length of 1,476.69 bp and an average of 1.43 coding exons per gene (Table 6).

Gene functions were inferred according to the best match of the alignments to the NR (ftp://ftp.ncbi.nih.gov/blast/db/), EggNOG (http://eggnog5.embl.de/#/app/home)[34], KOG (http://www.ncbi.nlm.nih.gov/KOG/), TrEMBL (http://www.uniprot.org/), InterPro(https://www.ebi.ac.uk/interpro/) and Swiss-Prot (http://www.uniprot.org/) protein databases using diamond blastp (diamond v0.9.29.130) and the KEGG (http://www.genome.jp/kegg/)[35] database with an E-value threshold of 1E-3. The protein domains were annotated with InterProScan (v5.34-73.0) based on InterPro protein databases. GO IDs for each gene were obtained from TrEMBL, InterPro and EggNOG. Totally, 14,596 (about 96.14%) of predicted protein-coding genes were annotated with known genes (Table 6).

Usually, pseudo genes share similar sequences with functional genes, but may lose their biological function due to the mutations, insertions and deletions during the genetic exchange process. The GenBlastA (v1.0.4)[36] was applied for scanning the whole genomes after functional genes predictions. Those putative candidates were analyzed for stop condons and frame-shift mutations using GeneWise (v2.4.1)[37]. As a results, there were 151 pseudo genes predicted (Table 6).

Gene structure annotation	
Protein-coding gene number	15,341
Gene length (bp)	27,686,232
Average gene length (bp)	1,804.72
Exon number	21,966
Exon length (bp)	24,829,735
Average exon number	1.43
Average exon length (bp)	1,618.52
CDS number	21,883
CDS length (bp)	22,653,972
Average CDS length (bp)	1,476.69
Pseudogene number	151
Pseudogene length (bp)	549,323
Average pseudogene length (bp)	3,637.90
Gene function annotation	
Annotation database	Number (Percent)
GO	11,111 (72.43%)
KEGG	8,867 (57.80%)
KOG	7,913 (51.83%)
Pfam	12,054 (78.57%)
Swissprot	7,735 (50.42%)
TrEMBL	14,038 (91.51%)
eggNOG	8,796 (57.34%)
NR	12,390 (80.76%)
Annotated	14,596 (96.14%)

Table 6. Statistics of gene structure and functional annotation of the *K. striatus* genome.

	Number	Percent
Complete BUSCO (c)	200	78.43%
Complete and single-copy BUSCOs (S)	192	75.29%
Complete and duplicated BUSCOs (D)	8	3.14%
Fragmented BUSCOs (F)	14	5.49%
Missing BUSCOs (M)	41	16.08%
Total BUSCOs	255	100%

Table 7. Universal single copy ortholog (BUSCO) assessment of *K. striatus*.

Data Records

All sequencing data have been uploaded to the NCBI SRA database under BioProject accession number PRJNA1150769. The Illumina sequencing data for genomic survey has been deposited in SRR30806644[38]. The PacBio, and Hi-C sequencing data has been deposited in the NCBI SRA database under the accession numbers SRR31859881[39], SRR30806642[40], respectively. The RNA sequencing read for gene annotation has been deposited in the NCBI Sequence Read Archive with accession number SRR30806641[41].

The genome assembly has been uploaded to the GenBank database under the accession JBHZSV000000000[42]. Moreover, the genome annotations are available from the Figshare[43] repository.

Technical Validation

To evaluate the quality of the assembled genome, its completeness was assessed using BUSCO (v5.2.2) with eukaryota_odb10 database[44]. Approximately, 78.43% of the 255 single-copy orthologs were identified in the *K. striatus* genome (Table 7). To assess the completeness and sequencing coverage uniformity, Illumina short reads were aligned to the assembled genome using bwa, and it yielded 307,960,284 reads (account for 97.00%) mapping to the reference genome. HiFi reads were aligned using Minimap2 (v2.14-r883)[45], with 1,234,100 reads (account for 97.64%) successfully mapping (Fig. 4). The distributed read depth coverage for PacBio sequencing reads showed a Poisson distribution patterns. Additionally, we analyzed the read depth and GC content across 10 kb windows to check for significant GC bias or sample contamination (Fig. 5). The results indicated that the assembled genome was free of contamination, and one high-quality genome assembly for *K. striatus* was obtained.

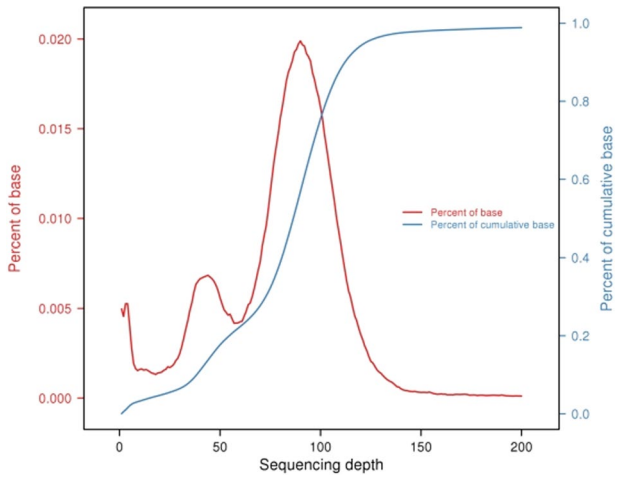

Fig. 4 Distribution of read depth coverage for PacBio reads. The x-axis represents sequencing depth (units X), and the y-axis shows the proportion of bases at that depth relative to the total number of bases. The distribution approximates a Poisson distribution, indicating high assembly quality.

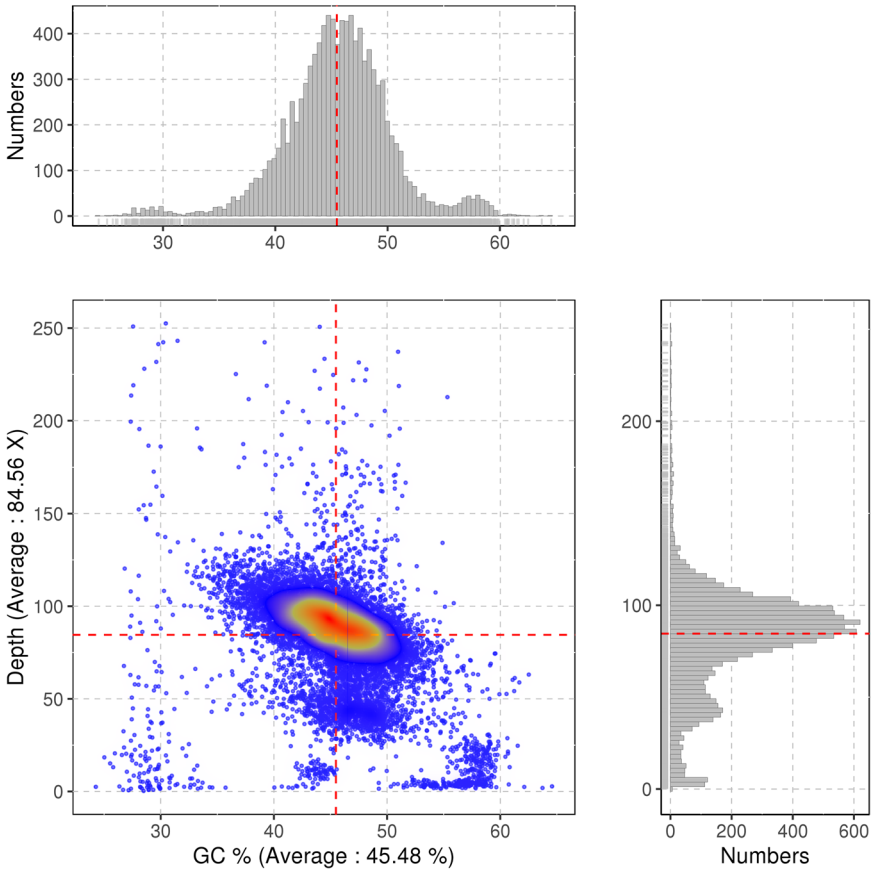

Fig. 5 GC content and depth distribution. The x-axis represents GC content and the y-axis represents coverage depth. The right panel depicts contig coverage depth distribution, the top panel shows GC content distribution, and the central scatter plot illustrates the relationship between GC content and coverage depth, with color intensity indicating point density.

Code availability

Genome assembly:

(1) hifiasm: parameters: hifiasm -t 36 --primary.
(2) bwa: parameters: bwa index & & bwa mem -t 16.
(3) minimap2: parameters: minimap2 -I 20 G --MD -ax asm20 -t 4. Genome annotation:
(1) RepeatModeler: parameters: BuildDatabase -name & & RepeatModeler -pa 12.
(2) LRT_FINDER: parameters: ltr_finder -w 2 -C –D.
(3) TRF: parameters: trf 2 7 7 80 10 50 500 -d –h.
(4) RepeatMasker: parameters: repeatmasker -nolow -no_is -norna -engine wublast-parallel 8 –qq.
(5) GeMoMa: parameters: mmseqs.
(6) Hisat: parameters: hisat2 --dta -p 10.
(7) Stringtie: parameters: stringtie -p 2.
(8) Trinity: parameters: Trinity --genome_guided_bam.
(9) gmap: parameters: gmap --cross-species --nthreads = 4 -f 2.
(10) barrnap: parameters: barrnap --kingdom euk–threads 1.
(11) Infenal: parameters: cmscan --cpu 3 --rfam.
(12) GenBlastA: parameters: genblasta -P wublast -pg tblastn.
(13) GeneWise: parameters: genblasta -P wublast -pg tblastn.
(14) InterProScan: parameters: interproscan.sh -iprlookup -pa -f xml -dp -t p -cpu 10.
(15) hmmscan: parameters: hmmscan -E 0.001 --domE 0.001 --cpu 6.
(16) diamond: parameters: diamond blastp --masking 0 -e 0.001.
(17) eggnog-mapper: parameters: emapper.py -m diamond.

For analysis modules where specific parameters were not mentioned, default settings were applied. The custom scripts used in this analysis are detailed in the methods sections.

Received: 4 November 2024; Accepted: 3 February 2025;
Published online: 12 February 2025

References

1. Farah Nurshahida, M. S., Nazikussabah, Z., Subramaniam, S., Wan Faizal, W. I. & Nurul Aini, M. A. Physicochemical, physical characteristics and antioxidant activities of three edible red seaweeds (*Kappaphycus alvarezii*, *Eucheuma spinosum* and *Eucheuma striatum*) from Sabah, Malaysia. *IOP Conf. Series: Materials Science and Engineering.* **991**, 012048 (2020).
2. Bindu, M. S. & Levine, I. A. The commercial red seaweed *Kappaphycus alvarezii*-an overview on farming and environment. *J Appl Phycol.* **23**, 789–796 (2011).
3. Hurtado, A. Q., Neish, I. C. & Critchley, A. T. Phyconomy: the extensive cultivation of seaweeds, their sustainability and economic value, with particular reference to important lessons to be learned and transferred from the practice of eucheumatoid farming. *Phycologia.* **58**, 472–483 (2019).
4. Doty, M. S. & Alvarez, V. B. Status, problem, advances and economics of *Eucheuma* farms. *Marine Technology Society Journal.* **9**, 30–35 (1975).
5. FAO. 2024. The State of World Fisheries and Aquaculture 2024 - Blue transformation in action. Rome. https://doi.org/10.4060/cd0683en
6. Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology & Evolution.* **21**, 809–818 (2004).
7. Trono, G. C. Jr *Eucheuma* and *Kappaphycus*: taxonomy and cultivation. *Bulletin of Marine Sciences & Fisheries, Kochi University.* **12**, 51–65 (1992).
8. Jia, S. *et al*. High-quality *de novo* genome assembly of *Kappaphycus alvarezii* based on both PacBio and HiSeq sequencing. Preprint at https://doi.org/10.1101/2020.02.15.950402 (2020).
9. Borg, M. *et al*. Red macroalgae in the genomic era. *New Phytologist.* **240**, 471–488 (2023).
10. Sahu, S. K., Thangaraj, M. & Kathiresan, K. DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. *Isrn Mol Biol.* **2012**, 205049 (2012).
11. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods.* **18**, 170–175 (2021).
12. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* **25**, 1754–1760 (2009).
13. Burton, J. N. *et al*. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol.* **31**, 1119–1125 (2013).
14. Bao, Z. & Eddy, S. R. Repeat Modeler2 for automated genomic discovery of transposable element families. *Genome Res.* **12**, 1269–1276 (2002).
15. Bao, Z. & Eddy, S. R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
16. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics.* **21**, i351–i358 (2005).
17. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTR harvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics.* **9**, 18 (2008).
18. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
19. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
20. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics.* **Chapter 4**, Unit 4.10 (2004).
21. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics.* **33**, 2583–2585 (2017).
22. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

23. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
24. Griffiths-Jones, S. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2004).
25. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *BMC Bioinformatics.* **29**, 2933–2935 (2013).
26. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *BMCBioinformatics.* **24**, 637–644 (2008).
27. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics.* **5**, 59 (2004).
28. Keilwagen, J. *et al.* Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89–e89 (2016).
29. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* **12**, 357–360 (2015).
30. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* **33**, 290–295 (2015).
31. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78–e78 (2015).
32. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidence Modeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
33. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* **29**, 644–652 (2011).
34. Huerta-Cepa, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2018).
35. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2015).
36. She, R., Chu, J. S. C., Wang, K., Pei, J. & Chen, N. genBlastA: Enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
37. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
38. *NCBI sequence read archive* https://identifiers.org/ncbi/insdc.sra:SRR30806644 (2024).
39. *NCBI sequence read archive* https://identifiers.org/ncbi/insdc.sra:SRR31859881 (2024).
40. *NCBI sequence read archive* https://identifiers.org/ncbi/insdc.sra:SRR30806642 (2024).
41. *NCBI sequence read archive* https://identifiers.org/ncbi/insdc.sra:SRR30806641 (2024).
42. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBHZSV000000000 (2024).
43. Zhou, Z. Annotations of *Kappaphycus striatus* genome. *Figshare* https://doi.org/10.6084/m9.figshare.28004447 (2024).
44. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copyorthologs. *Bioinformatics.* **31**, 3210–3212 (2015).
45. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**, 3094–3100 (2018).

Acknowledgements

This research is supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB42030203) and the Asia Collaboration Project on Development of Ecological Marine Ranching.

Author contributions

D.L.D. conceived, designed and supervised the study. Z.Y.Z., Y.M. and M.F. prepared the sample. Z.Y.Z., Y.M., J.Z. performed bioinformatics analysis and prepared the results. Z.Y.Z. drafted the manuscript. D.L.D. and M.Y.R. edited and improved the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.