



## ORIGINAL RESEARCH

# Multi-omics Analyses Provide Insight into the Biosynthesis Pathways of Fucoxanthin in *Isochrysis galbana*



Duo Chen<sup>1,#</sup>, Xue Yuan<sup>1,#</sup>, Xuehai Zheng<sup>1,#</sup>, Jingping Fang<sup>1,#</sup>, Gang Lin<sup>1</sup>, Rongmao Li<sup>2</sup>, Jiannan Chen<sup>1</sup>, Wenjin He<sup>1</sup>, Zhen Huang<sup>1</sup>, Wenfang Fan<sup>1</sup>, Limin Liang<sup>1</sup>, Chentao Lin<sup>1</sup>, Jinmao Zhu<sup>1</sup>, Youqiang Chen<sup>1</sup>, Ting Xue<sup>1,\*</sup>

<sup>1</sup> The Public Service Platform for Industrialization Development Technology of Marine Biological Medicine and Products of the State Oceanic Administration, Center of Engineering Technology Research for Microalga Germplasm Improvement of Fujian, Fujian Key Laboratory of Special Marine Bioresource Sustainable Utilization, Fujian Key Laboratory of Developmental and Neural Biology, Southern Institute of Oceanography, College of Life Sciences, Fujian Normal University, Fuzhou 350117, China

<sup>2</sup> Fujian Fishery Resources Monitoring Center, Fuzhou 350003, China

Received 17 July 2021; revised 27 April 2022; accepted 8 May 2022  
 Available online 13 August 2022

Handled by Kai Ye

## KEYWORDS

*Isochrysis galbana*;  
 Fucoxanthin;  
 Whole-genome duplication;  
 Metabolome;  
 Transcriptome

**Abstract** *Isochrysis galbana* is considered an ideal bait for functional foods and nutraceuticals of humans because of its high fucoxanthin (Fx) content. However, multi-omics analysis of the regulatory networks for Fx biosynthesis in *I. galbana* has not been reported. In this study, we report a high-quality genome assembly of *I. galbana* LG007, which has a genome size of 92.73 Mb, with a contig N50 of 6.99 Mb and 14,900 protein-coding genes. Phylogenetic analysis confirmed the monophyly of Haptophyta, with *I. galbana* sister to *Emiliania huxleyi* and *Chrysochromulina tobinii*. Evolutionary analysis revealed an estimated divergence time between *I. galbana* and *E. huxleyi* of ~ 133 million years ago. Gene family analysis indicated that lipid metabolism-related genes exhibited significant expansion, including *IgPLMT*, *IgOAR1*, and *IgDEGS1*. **Metabolome** analysis showed that the content of carotenoids in *I. galbana* cultured under green light for 7 days was higher than that under white light, and  $\beta$ -carotene was the main carotenoid, accounting for 79.09% of the total carotenoids. Comprehensive multi-omics analysis revealed that the content of  $\beta$ -carotene, antheraxanthin, zeaxanthin, and Fx was increased by green light induction, which was significantly

\* Corresponding author.

E-mail: [xueting@fjnu.edu.cn](mailto:xueting@fjnu.edu.cn) (Xue T).

# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

<https://doi.org/10.1016/j.gpb.2022.05.010>

1672-0229 © 2022 The Authors. Published by Elsevier B.V. and Science Press on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

correlated with the expression of *IgMYB98*, *IgZDS*, *IgPDS*, *IgLHCX2*, *IgZEP*, *IgLCYb*, and *IgNSY*. These findings contribute to the understanding of Fx biosynthesis and its regulation, providing a valuable reference for food and pharmaceutical applications.

## Introduction

Fucoxanthin (Fx) is widely distributed in algae and some invertebrate cells, including *Phaeodactylum tricorutum*, *Ectocarpus siliculosus*, *Thalassiosira pseudonana*, *Isochrysis galbana*, and *Nannochloropsis gaditana* [1,2]. Fx can be assembled with chlorophyll into Fx-chlorophyll protein (FCP) with certain proteins, and FCP exists in thylakoid membranes of algae and acts as a light-capturing antenna [3]. Fx has excellent capabilities of blue-green light harvesting and photoprotection, which could help algae make full use of solar energy in different bands at different depths of seawater [3,4]. Fx, an oxygenated carotenoid, exhibits potential advantages with various pharmacological activities, including anti-inflammatory, anti-tumor, anti-obesity, anti-oxidant, anti-diabetic, anti-malarial, and anti-lipid effects [5]. However, the utilization of Fx as a nutraceutical in food and nutrient supplements is limited because of its low production level and poor stability. Microalgae are regarded as the most promising alternative Fx production algae with multiple biotechnological advantages, such as short growth cycle, easy handling, and large-scale artificial cultivation [6]. Compared with other microalgae, such as *P. tricorutum* and *N. gaditana*, *I. galbana* lacks cell walls and is much easier to be digested and handled, making it a good initial food source for the larvae of aquatic animals [7,8]. Moreover, *I. galbana* is a marine single-cell microalga rich in Fx (more than 10% of dry weight biomass) and lipid (7.0%–20% of dry weight biomass), which is considered an ideal material for the development of functional foods for humans [9]. Additionally, we found that the Fx content of *I. galbana* LG007 was the highest in different strains or species, which can be used as an ideal material for follow-up research (Figure S1).

Although the process of Fx synthesis has not been fully elucidated, several studies have attempted to reveal genes or proteins involved in Fx biosynthesis [10,11]. Genes involved in Fx biosynthesis have been identified, including  $\beta$ -carotene, phytoene synthase (*PYS*), phytoene desaturase (*PDS*), 15-*cis*- $\zeta$ -carotene isomerase (*ZISO*),  $\zeta$ -carotene desaturase (*ZDS*), carotenoid isomerase (*CRTISO*), and lycopene  $\beta$ -cyclase (*LCYb*) [10,11]. However, some enzymes participating in the final step of Fx biosynthesis have not been discovered [11]. There are two generally accepted hypotheses for the final step of Fx biosynthesis from violaxanthin to Fx in Fx-producing algae: 1) violaxanthin is a precursor of Fx, which is converted by phycoxanthin; and 2) neoxanthin is the precursor of Fx, which is formed by the ketonization of the neoxanthin and the acetylation of an intermediate [12,13]. Studies have focused on the key genes related to the Fx biosynthesis pathway, mainly involved in the expression of some genes in the Fx biosynthesis pathway by external inducing factors (light intensity, methyl jasmonate, and arachidonic acid). Zhang et al. revealed the change in Fx content and gene expression pattern of the Fx biosynthesis pathway under conditions of high irradiance stress in the diatom *P. tricorutum*, showing an evident linear relationship between Fx content and the expression levels of

*PYS* and zeaxanthin epoxidase (*ZEP*) [14]. Yu et al. reported that the expression level of *LCYb* could be significantly increased by treatment with methyl jasmonate and arachidonic acid, and the content of Fx in *P. tricorutum* was significantly higher than that in the control group, which showed that *LCYb* played an important role in the Fx biosynthesis [15]. However, the aforementioned studies are mainly reflected in *P. tricorutum*, and few studies are on enzyme genes related to the Fx biosynthesis pathway of *I. galbana*. Because of limited genome information, how *I. galbana* regulates Fx biosynthesis at the DNA and RNA levels remains unclear. Draft genome sequences of *I. galbana* were generated in 2014 based on next generation sequencing. However, incomplete genome assemblies produced short contigs and scaffolds, causing problems for the follow-up research of *I. galbana* [16]. Additionally, high-quality genomic resources can enable the breeding of novel *I. galbana* strains with higher Fx content in industrial practice for commercial use. But up to now, a systematic analysis of the regulatory networks for Fx biosynthesis in *I. galbana* has not been performed using genome, transcriptome, and metabolome data according to our review of the literature.

In this study, we generated a high-quality genome assembly and annotation of *I. galbana* LG007 by using the third-generation sequencing (PacBio Sequel platform). The high-quality *I. galbana* LG007 genome provides a valuable resource for studying the evolutionary events and genomic characteristics of aquatic algae. A previous study revealed the influence of spectral intensity and quality of blue-green light on Fx content in Chlorophyceae, but it did not involve the role of green light as a single source for Fx biosynthesis [17]. Our previous results suggested that Fx content could be increased under green light conditions, which is a special simulating factor that occurs during the cultivation of *I. galbana* (Figure S2). Transcriptomic and metabolomic analyses were performed on algae cells at different stages of cultivation (3, 5, 7, and 9 days) under different light-quality conditions (white and green) to reveal key genes or metabolic products that are potentially related to the accumulation and regulation of Fx biosynthesis.

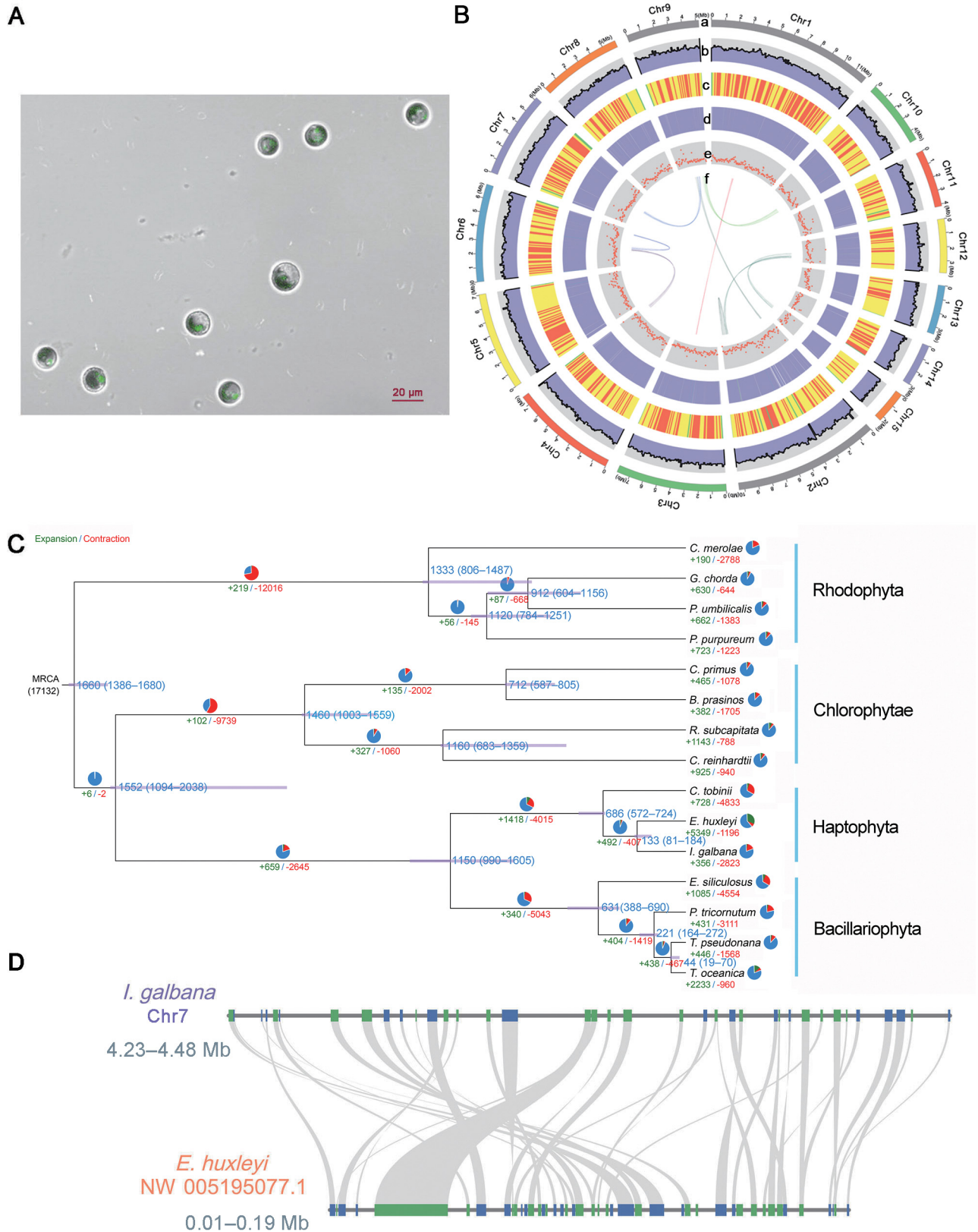
## Results

### Assembly of a high-quality genome of *I. galbana* LG007

About 15.5 Gb of PacBio long reads and 8.92 Gb of Illumina clean reads were generated (Table S1). The total length of all reads assembled from the *I. galbana* LG007 genome contained 353 contigs was 92.59 Mb, with a contig N50 of 666.7 kb, GC content of 58.44%, and the longest contig length of 2.93 Mb (Figure 1A and B, Figure S3; Table S2). The size of the assembled genome was close to that estimated by flow cytometry and 17-mer (Figures S4 and S5). Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis of our present assembly showed that  $\sim 83.8\%$  of the plant orthologs were included in the assembled sequences (Table S3). Likewise,  $\sim 98.4\%$  of Illumina clean reads and  $\sim 99.78\%$  of PacBio long reads could

be mapped to the genome, respectively (Tables S4 and S5). These metrics implied that the assembled genome is credible and can be used for subsequent analysis. Using the 3D-DNA

and LACHESIS workflow, 98.22% (90.95 Mb) of the genome was successfully anchored onto 15 superscaffolds (Figures S6 and S7; Table S6). The scaffold N50 of the *I. galbana* LG007



genome after high-throughput chromatin conformation capture (Hi-C) assisted assembly reached 6.99 Mb (Table S7), generating a high-quality genome assembly for *I. galbana*.

### Gene prediction and annotation

A total of 14,900 protein-coding genes were predicted in *I. galbana* LG007 genome by combining the homology-based search, *de novo* prediction, and transcriptome evidence. The protein-coding genes had an average gene length of 1789 bp, and an average coding sequence (CDS) length of 1428 bp (Table S8). We functionally annotated 9161, 12,469, 3773, 4977, and 9161 genes to eggNOG, Non-Redundant (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and Clusters of Orthologous Groups (COG), respectively, leading to ~ 83.89% (12,500 genes) of the total genes with at least one match to the known public databases (Table S9). A total of 439 transcription factors (TFs) were distributed in 20 families, including 198 protein kinase family proteins, 55 heat shock factor (HSF) family proteins, 49 zinc finger WD40 (ZFWD) family proteins, and 44 v-myb avian myeloblastosis viral oncogene homolog (MYB) family proteins (Table S10). In addition, we also identified 95 transfer RNAs (tRNAs), 58 ribosomal RNAs (rRNAs), and 4 small nuclear RNAs (snRNAs) in an *I. galbana* LG007 genome (Table S11). Moreover, ~ 46.82% of the assembled *I. galbana* LG007 genome comprised repetitive sequences. Long terminal repeat (LTR) retrotransposons spanned 15.36% of the assembled genome with 1.08% Ty1/Copia and 4.53% Ty3/Gypsy. Non-LTR elements accounted for 12.54% of the genome, including 11.62% long interspersed nuclear elements (LINEs) and 0.91% short interspersed nuclear elements (SINEs). Tandem Repeats Finder identified over 43,633 tandem repeats, spanning 4.3% of the *I. galbana* LG007 genome (Table S12).

### Gene family expansion associated with lipid metabolism

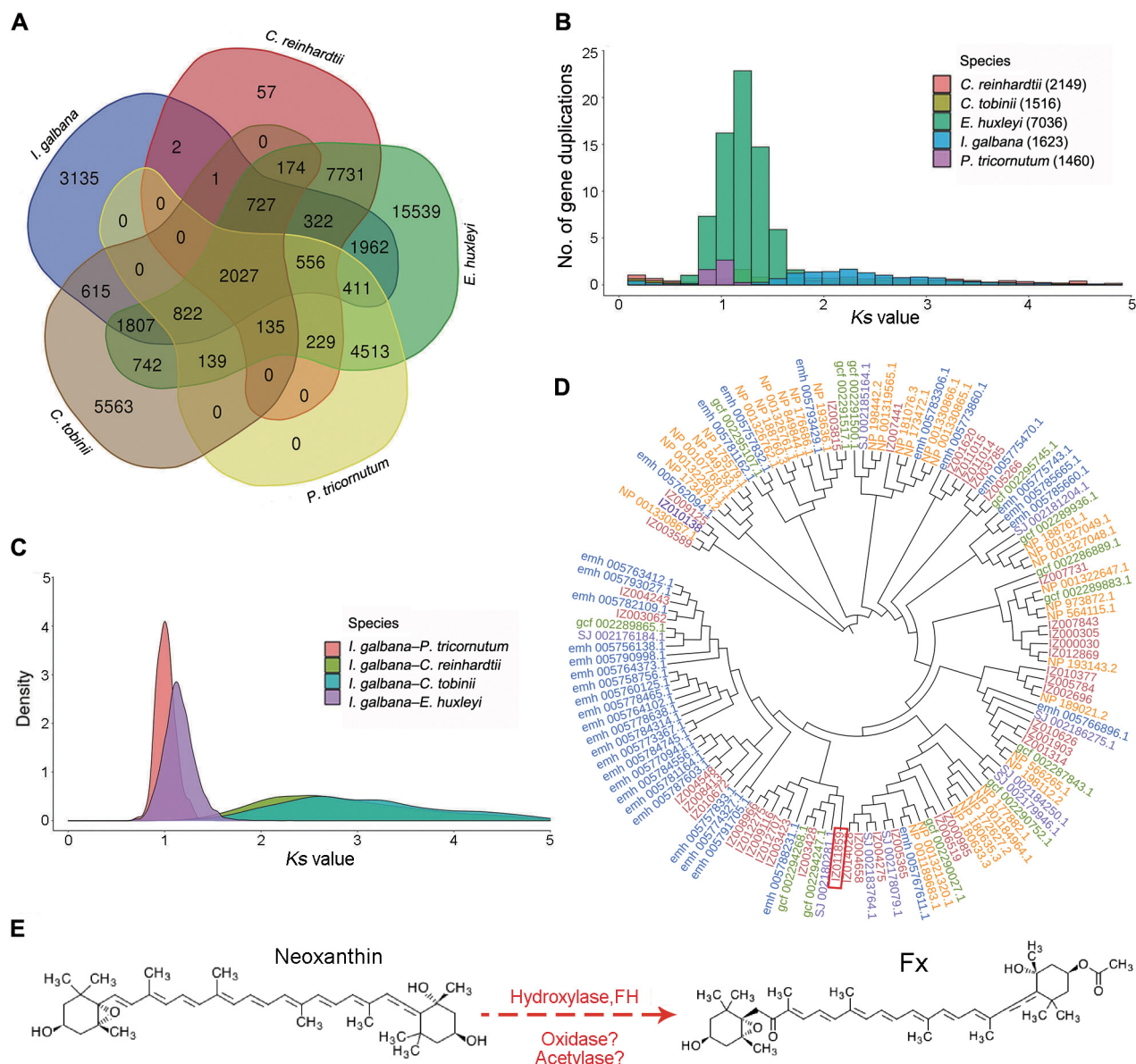
A total of 179 single-copy homologous genes were identified among 15 genomes by using OrthoFinder (version 2.3.12) and were used to reconstruct a phylogenetic tree (Figure 1C). Phylogenetic analysis showed that *I. galbana* diverged into the Haptophyta branch ~ 133 million years ago (MYA) after the divergence of the Rhodophyta (1333 MYA), Chlorophytae (1460 MYA), and Bacillariophyta (1150 MYA). These results support the view that *Emiliania huxleyi*, *I. galbana*, and *Chrysochromulina tobinii* as monophyletic groups share a common Haptophyta ancestor [18–20]. Although *E. huxleyi* and *I. galbana* genomes appear to share only limited collinearity, this

observation may stem from the lower quality of the *E. huxleyi* genome assembly (Figure 1D). The results showed that *I. galbana* and *E. huxleyi* may have a close relationship and are sisters in coccolithophores, which is consistent with the findings of phylogenetic analysis. Comparative genomic analysis showed that gene family expansions outnumbered contractions in *Raphidocelis subcapitata*, *Thalassiosira oceanica*, and *E. huxleyi*. We also discovered 356 expanded and 2823 contracted gene families in *I. galbana* LG007. KEGG pathway analysis showed that the expanded gene families were specifically enriched in signal transduction, purine metabolism, lipid metabolism, and ABC transporters (Table S13). GO analysis showed that these expanded genes were related to signaling, metabolic processes, stimulus response, and catalytic activity (Table S14). Interestingly, lipid metabolism-related genes (*IgPLMT*, *IgOAR1*, and *IgDEGSI*) exhibited significant expansion, indicating that the expansion of these genes in *I. galbana* could enhance the regulation and biosynthesis of carotenoid, resulting in a high content of Fx in *I. galbana*. A total of 2823 contracted gene families highlighted the functions pertaining to signal transduction, starch metabolism, and biosynthesis of other secondary metabolites (Table S15). GO terms of the contracted genes were associated with binding, catalytic activity, transporter activity, transcription regulator activity, stimulus response, signaling, and biosynthesis of secondary metabolites (Table S16). Thus, it is likely that the contractions of secondary metabolite biosynthesis- and stimulus-related genes could affect the accumulation of other secondary metabolites and resistance in *I. galbana*, resulting in a relatively good photoprotection capabilities only in the blue-green light of seawater. A comparison of *E. huxleyi*, *C. tobinii*, *P. tricornutum*, *Chlamydomonas reinhardtii*, and *I. galbana* LG007 revealed that 2027 (31.37%) of the 12,387 *I. galbana* LG007 gene families were common to other four species, whereas 3135 gene families were specific to *I. galbana* LG007 (Figure 2A). GO enrichment analysis showed that the functions of these specific genes mainly included metabolic process, catalytic activity, biological regulation, stimulus response, developmental process, binding, pigmentation, molecular transducer activity, and transcription regulator activity (Table S17). KEGG enrichment analysis showed enrichment of the calcium signaling pathway, the cGMP-PKG signaling pathway, fatty acid biosynthesis, signal transduction, energy metabolism, and terpenoid metabolism (Table S18).

To infer the whole-genome duplication (WGD) event in *I. galbana*, we calculated the synonymous substitution rate ( $K_s$ ) by a mixture model implemented in the R package. 2149, 1516, 7038, 1460, and 1623 genes were used to calculate the  $K_s$  value for *C. reinhardtii*, *C. tobinii*, *E. huxleyi*,

### Figure 1 Genomic characteristics of *I. galbana*

A. Confocal laser scanning microscopic images of *I. galbana* cells. Scale bar, 20  $\mu\text{m}$ . B. Images of *I. galbana* assembly. a, assembled superscaffolds; b, distribution of GC content; c, density of genes; d, expression values; e, percent coverage of transposable elements in nonoverlapping windows; f, syntenic blocks within the genome. C. Phylogenetic tree of *I. galbana* and other 14 species. Numbers in green and red represent the numbers of expanded and contracted gene families, respectively. The divergence time (MYA) is denoted at each node in blue. D. Microsynteny analysis of *I. galbana* superscaffolds and *E. huxleyi* scaffolds. MYA, million years ago; MRCA, most recent common ancestor; *I. galbana*, *Isochrysis galbana*; *C. merolae*, *Cyanidioschyzon merolae*; *G. chorda*, *Gracilariopsis chorda*; *P. umbilicalis*, *Porphyra umbilicalis*; *C. primus*, *Chloropicon primus*; *B. prasinos*, *Bathycoccus prasinos*; *R. subcapitata*, *Raphidocelis subcapitata*; *C. reinhardtii*, *Chlamydomonas reinhardtii*; *C. tobinii*, *Chrysochromulina tobinii*; *E. huxleyi*, *Emiliania huxleyi*; *E. siliculosus*, *Ectocarpus siliculosus*; *P. tricornutum*, *Phaeodactylum tricornutum*; *T. pseudonana*, *Thalassiosira pseudonana*; *T. oceanica*, *Thalassiosira oceanica*.



**Figure 2** Evolution of the *I. galbana* LG007 genome

**A.** Shared and unique gene families among five species. **B.** *Ks* distributions for duplicated gene pairs in *I. galbana*, *P. tricornutum*, *C. reinhardtii*, *E. huxleyi*, and *C. tobinii*. **C.** *Ks* distributions between *I. galbana*, *P. tricornutum*, *C. reinhardtii*, *E. huxleyi*, and *C. tobinii*. **D.** Evolutionary tree of hydroxylase genes in *I. galbana* (IZ, red), *P. tricornutum* (SJ, purple), *E. huxleyi* (emh, blue), *C. tobinii* (gcf, green), and *A. thaliana* (NP, yellow). *IgFH* gene is marked with the red box. **E.** Chemical reactions possibility in the process of synthesis according to the structure of neoxanthin and Fx. *Ks*, synonymous substitution rate; Fx, fucoxanthin; FH, Fx hydroxylase; *A. thaliana*, *Arabidopsis thaliana*.

*P. tricornutum*, and *I. galbana*, respectively. The sharp peak of distribution in *I. galbana* has a median *Ks* of  $\sim 2.53$ , which is higher than the ortholog divergences of *I. galbana* and *P. tricornutum* (*Ks*,  $\sim 0.95$ ), and the divergences of *I. galbana* and *E. huxleyi* (*Ks*,  $\sim 1.38$ ) (Figure 2B and C). Comparative analysis of *I. galbana* and the other four species genomes provided evidence of the two WGD events according to the sharp peaks in *Ks* values ( $\sim 0.95$  and  $\sim 2.53$ ) (Figure 2C). The distribution of *Ks* values suggested that divergence between *I. galbana* and *E. huxleyi* occurred at  $\sim 133$  MYA (*Ks*,  $\sim 1.38$ ), later than the ancient WGD in *I. galbana* (*Ks*,  $\sim 2.53$ ), which was dated at  $\sim 245$  MYA. This finding indicated that *I. galbana*, *P. tri-*

*cornutum*, and *E. huxleyi* experienced long-term divergence and speciation.

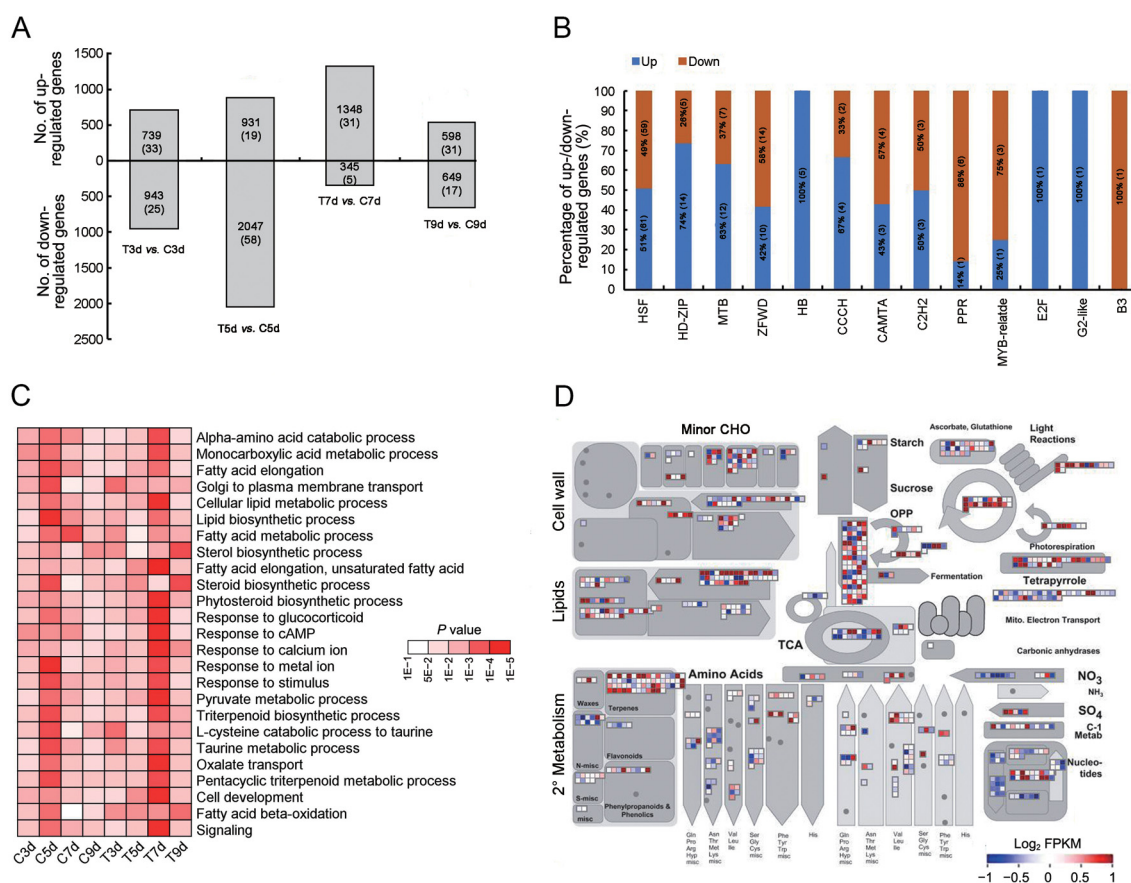
#### *IgFH* as a candidate gene could play a key role in Fx biosynthesis of *I. galbana*

The metabolic processes of violaxanthin, neoxanthin, diadinoxanthin, and Fx may be very complicated, which involved oxidation, isomerization, acetylation, deepoxidation, hydrogenation, and hydroxylation chemical reactions according to the structure of substrates or products. To efficiently identify candidate genes involved in Fx biosynthesis, we identified 39

hydroxylase genes by a combination of direct screening of the genome assembly annotations and conserved domain BLAST searches. One of the 39 hydroxylase genes catalyzed the hydroxylation of hydrophobic substrates, which has a similar chemical reaction according to the structure of diadinoxanthin, neoxanthin, and Fx, suggesting that the Fx hydroxylase gene (*IgFH*, IZ011859) might have a function similar to that of the characterized neoxanthin–Fx or diadinoxanthin–Fx as candidate genes (Figure 2D and E, Figure S8). We could not detect Fx or new products after incubating recombinant IgFH proteins with diadinoxanthin by enzyme activity assay. However, overexpression and enzyme activity assays confirmed that neoxanthin could be further catalyzed by the IgFH proteins, implying that *IgFH* could play a key role in Fx biosynthesis (Figure S9).

Two distinct interconversion cycles of zeaxanthin to violaxanthin [violaxanthin de-epoxidase (*VDE*) and *ZEP*] and

diatoxanthin to diadinoxanthin [diadinoxanthin de-epoxidase (*DDE*) and diatoxanthin epoxidase (*DEP*)] containing epoxidase and de-epoxidase were involved in the same type of catalytic reaction. We identified 6, 16, 4, and 1 epoxidase family proteins in *I. galbana*, *Arabidopsis thaliana*, *P. tricornutum*, and *C. reinhardtii*, respectively. The number of epoxidase gene proteins from *I. galbana* (6) was lower than that of *A. thaliana* (16), but was close to that of *P. tricornutum* (4) (Figures S10–S12). Additionally, we constructed a phylogenetic tree based on the identified de-epoxidase proteins from 15 amino acid sequences of *I. galbana* (5), *A. thaliana* (4), *P. tricornutum* (5), and *C. reinhardtii* (1). The number of de-epoxidase gene families in *I. galbana* (5) was significantly higher than that in *C. reinhardtii* (1). We speculated that the presence of most epoxidase and de-epoxidase gene proteins could play a role in the Fx metabolic synthesis of *I. galbana* (Figures S13–S15).



**Figure 3** Differential gene expression in the treated group as compared with control group at different stages

**A.** Number of up- and down-regulated genes at different cultivation time points in the treated group (green light) and control group (white light). The number of up- or down-regulated TF genes at different cultivation time points in the treated group is also given in the parenthesis. **B.** Percentage of up- and down-regulated genes in different TF gene families in the treated group on 7d. **C.** GO analysis of DEGs (biological process) at different cultivation time points in the control and treated groups. **D.** Metabolic pathways with differential expression profiles in treated group as compared with control group on 7d. DEGs between the treated group and control group on 7d were loaded into the MapMan software. Red and blue colors indicate high and low expression, respectively. TF, transcription factor; GO, Gene Ontology; DEG, differentially expressed gene; C3d, control group on day 3; T3d, treated group on day 3; C5d, control group on day 5; T5d, treated group on day 5; C7d, control group on day 7; T7d, treated group on day 7; C9d, control group on day 9; T9d, treated group on day 9; CHO, carbohydrate; TCA, tricarboxylic acid cycle; OPP, pentose phosphate pathway.

### Fx-related genes of *I. galbana* showed high expression levels under green light condition

Among the annotated genes, 12,093 (96.74%) genes were expressed in 24 samples. Some genes were highly expressed on day 7 (7d) with green light (Figure S16), which has a similar trend to the phenotype of Fx content (Figure S17). To explore the differentially expressed genes (DEGs) involved in the Fx biosynthesis, DEGs of pairwise comparisons between control and treatment were analyzed (e.g., control 3d vs. treated 3d, control 5d vs. treated 5d). The number of stage-specific genes [fragments per kilobase of exon per million fragments mapped (FPKM)  $\geq 8000$ ] varied from 646 to 807 for the control group and 448 to 802 for the treated group (Figure S18). The number of stage-specific genes in the treatment group changed little on 7d, but decreased significantly on 9d. The number of stage-specific genes in the control group fluctuated on 7d and increased sharply on 9d (Figure S18). These results indicate that 7 days would be an important period for Fx biosynthesis. GO enrichment analyses of these stage-specific genes between the control and treated groups showed a representation of genes associated with various biological regulations, carboxylic acid biosynthetic process, fatty acid metabolic process, stimulus response, and catalytic process (Figure S19).

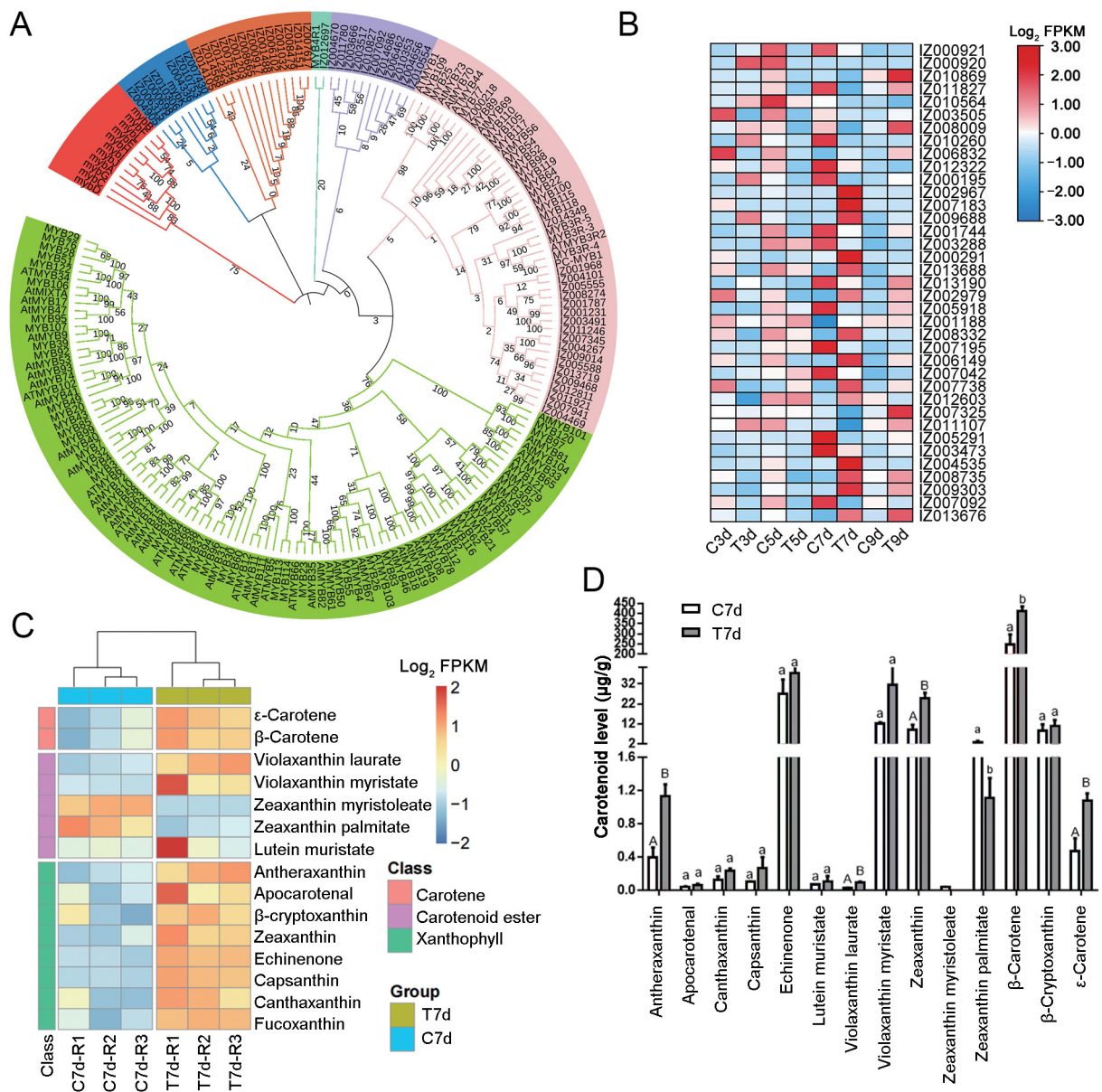
In total, 3730 genes exhibited significantly higher expression, and 4089 genes exhibited significantly lower expression at different stages in the treated group than in the control group (Figure 3A). Among these comparisons, the down-regulated and up-regulated expression between the control and treated groups was greatest on 5d and 7d, indicating a difference in the transcription levels at the 5d and 7d stages with green light irradiation. Some TFs also exhibited a significant difference between the control and treated groups. For example, the members of the MYB, homeobox (HB), and homeobox-leucine zipper (HD-ZIP) families involved in pigment accumulation and resistance stress showed significantly higher expression (Figure 3B). Fatty acid elongation, steroid biosynthesis, signaling, stimulus response, and cell development were enriched in the DEGs, particularly at the 7d stage with green light irradiation (Figure 3C). To explore the metabolic pathways responsible for the differences between the control and treated groups, we analyzed the expression profiles of DEGs using the MapMan tool. We found that the genes involved in lipid metabolism, light reactions, and pyruvate oxidation were more active in the treated group at the 7d stage, suggesting higher energy and more synthetic substrates for the metabolism of terpenoids and the  $\beta$ -carotene pathway (Figure 3D). With regard to MYB proteins, 52 and 146 MYB proteins were identified in *I. galbana* and *A. thaliana*, respectively (Figure 4A, Figure S20). R2R3-MYB TFs are related to the biosynthesis of pigment, suggesting a close relationship with the accumulation of Fx in *I. galbana* under light-induced conditions. Among the 114 TF genes with a relatively high expression, we found that *IgMYB98* (IZ007092) encodes an R2R3-MYB TF, which is significantly down-regulated ( $P = 1.99E-9$ ) in the synthesis of Fx and may be a key gene for negative regulation of Fx biosynthesis in *I. galbana* under light-induced conditions (Figure 4B; Table S19).

### Green light had a significant effect on the metabolism of carotenoids in *I. galbana*

In order to understand the biosynthesis pathways of the Fx accumulation under different light qualities (white and green), we compared the carotenoids between the control and treated groups on 7d (C7d vs. T7d). Fifteen carotenoids were identified in the comparison of C7d vs. T7d groups, including three types of carotenes, carotenoid esters, and xanthophylls (Figure 4C; Table S20). Phytoene,  $\zeta$ -carotene, neurosporene, lycopene,  $\gamma$ -carotene, violaxanthin, and neoxanthin involved in carotenoid biosynthesis were not detected between the T7d and C7d groups, indicating that these types of carotenoids may be prone to degradation or rapid conversion in *I. galbana*. The main carotenoids that accumulated in the T7d group were  $\beta$ -carotene, echinenone, violaxanthin myristate, zeaxanthin, and  $\beta$ -cryptoxanthin, among which  $\beta$ -carotene was the main carotenoid, accounting for 79.09% of the total carotenoid. Echinenone had the second-highest content in the T7d and C7d groups, accounting for 8.65% and 7.17% of the total carotenoid content, respectively. Heatmap analysis revealed that the content of carotenoids in the T7d group was significantly higher than that in the C7d group, including  $\beta$ -carotene (1.64-fold increase), lutein-myristate (1.56-fold increase),  $\beta$ -cryptoxanthin (1.29-fold increase), capsanthin (2.43-fold increase), and zeaxanthin (2.67-fold increase) (Figure 4D, Figure S17B). We identified eight differentially accumulated carotenoids (DACs) in C7d vs. T7d, including seven up-regulated DACs and one down-regulated DACs. The number of up-regulated DACs was much higher than that of down-regulated DACs in the comparison of T7d vs. C7d, suggesting the abundant diversity of carotenoids present under green light. Notable increases in carotenoids from the C7d to T7d samples included those in  $\epsilon$ -carotene (2.28-fold increase,  $P = 0.008$ ), violaxanthin-laurate (3.50-fold increase,  $P = 0.003$ ), violaxanthin-myristate (2.58-fold increase,  $P = 0.008$ ), antheraxanthin (2.86-fold increase,  $P = 0.02$ ), capsanthin (2.56-fold increase,  $P = 0.001$ ), zeaxanthin (2.67-fold increase,  $P = 0.002$ ), and Fx (2.14-fold increase,  $P = 0.009$ ) (Table S21). These results suggest that green light has a significant effect on the metabolism of carotenoids in *I. galbana*.

### Gene co-expression network involved in Fx accumulation

To identify the hub genes, we performed weighted gene co-expression network analysis (WGCNA) for the control and treated groups on 7d, separately. A total of 25 modules (comprising 31–2830 genes) were identified in the control group, and 24 modules (comprising 30–3333 genes) were recognized in the treated group (Figure 5A and B, Figure S21). Notably, the red co-expression module of the control group and turquoise co-expression module of the treated group showed a relatively high correlation ( $r \geq 0.60$ ) with Fx content (Figure 5C and D). GO and KEGG pathway enrichment analyses of relatively higher correlation modules highlighted key DEGs and biological processes with Fx content (Figure 5E and F). For example, the GO and KEGG analyses showed that the red module of the control group included most of the genes involved in metabolic processes, stimulus response, biological regulation, biosynthetic process, biosynthesis of secondary metabolites, fatty acid biosynthesis, and carotenoid



**Figure 4** Transcriptomic and metabolomic analyses of Fx accumulation in *I. galbana*

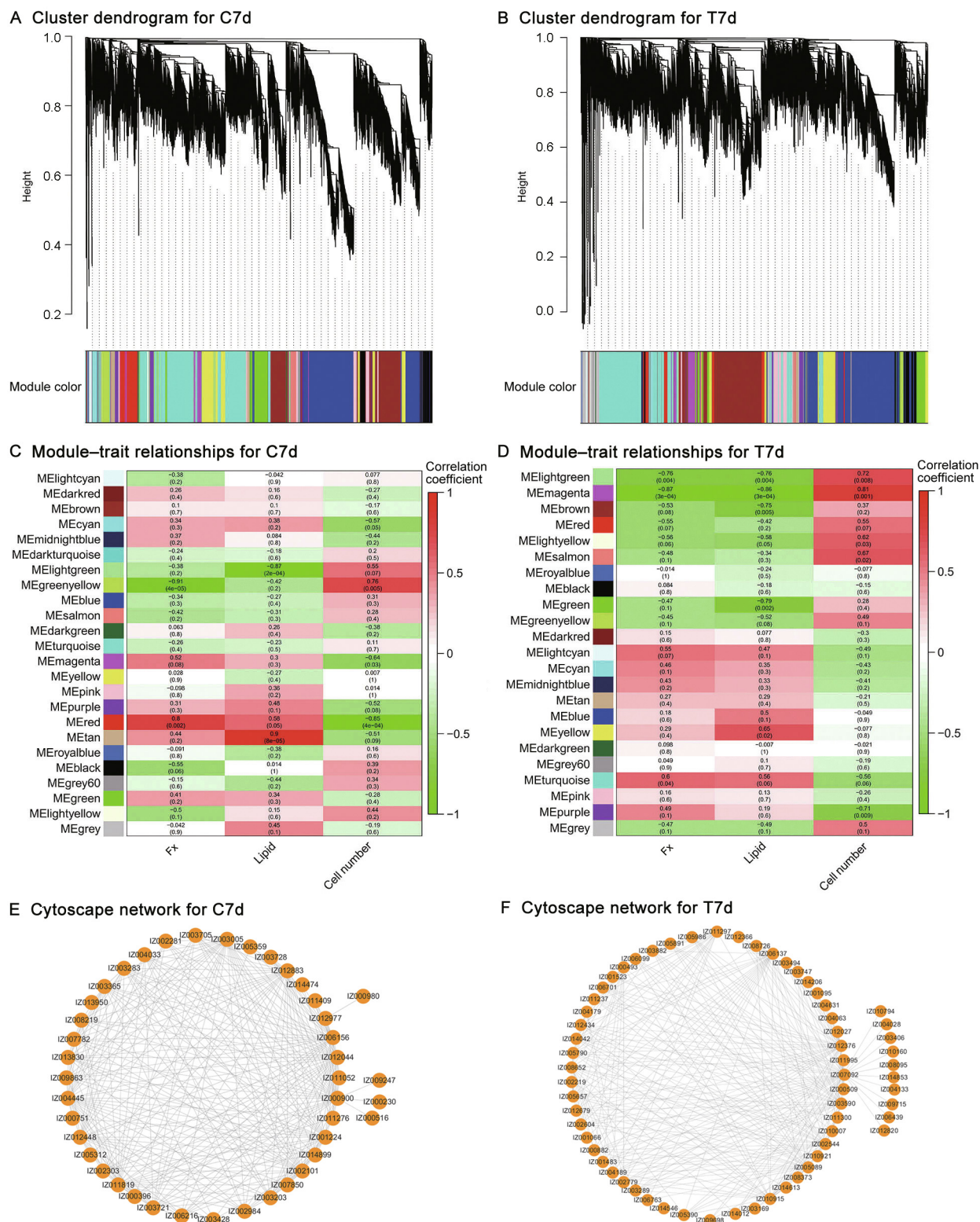
**A.** Evolutionary tree of MYB proteins in *I. galbana* (IZ) and *A. thaliana* (At). **B.** Heatmap of the DEGs in the treated and control groups at different stages. Each box represents an individual gene, and the red and blue colors indicate high and low expression of genes, respectively. **C.** Heatmap showing the changes of carotenoids in the treated and control groups on 7d. **D.** Content of major carotenoids in *I. galbana* under different light qualities. Error bar indicates standard deviation from three replicates. Capital letters and small letters indicate the significances at the 0.01 and 0.05 levels, respectively. MYB, v-myb avian myeloblastosis viral oncogene homolog.

biosynthesis (Figures S22 and S23). The turquoise module associated with green light irradiation for Fx content showed enrichment of GO terms and KEGG pathways related to biological processes, metabolic processes, biosynthetic processes, catabolism processes, metabolic pathways, biosynthesis of secondary metabolites, and carbon metabolism (Figures S24 and S25). Next, we studied the preservation of co-expression modules between the control and treated groups (Figure S26). We identified a midnight-blue module (35 genes) between the control and treated groups, and the genes of this module were enriched in metabolic processes, negative regulation of biological process, and stimulus response (Figure S26C and D). Taken together, hub gene analysis identified  $\zeta$ -carotene desat-

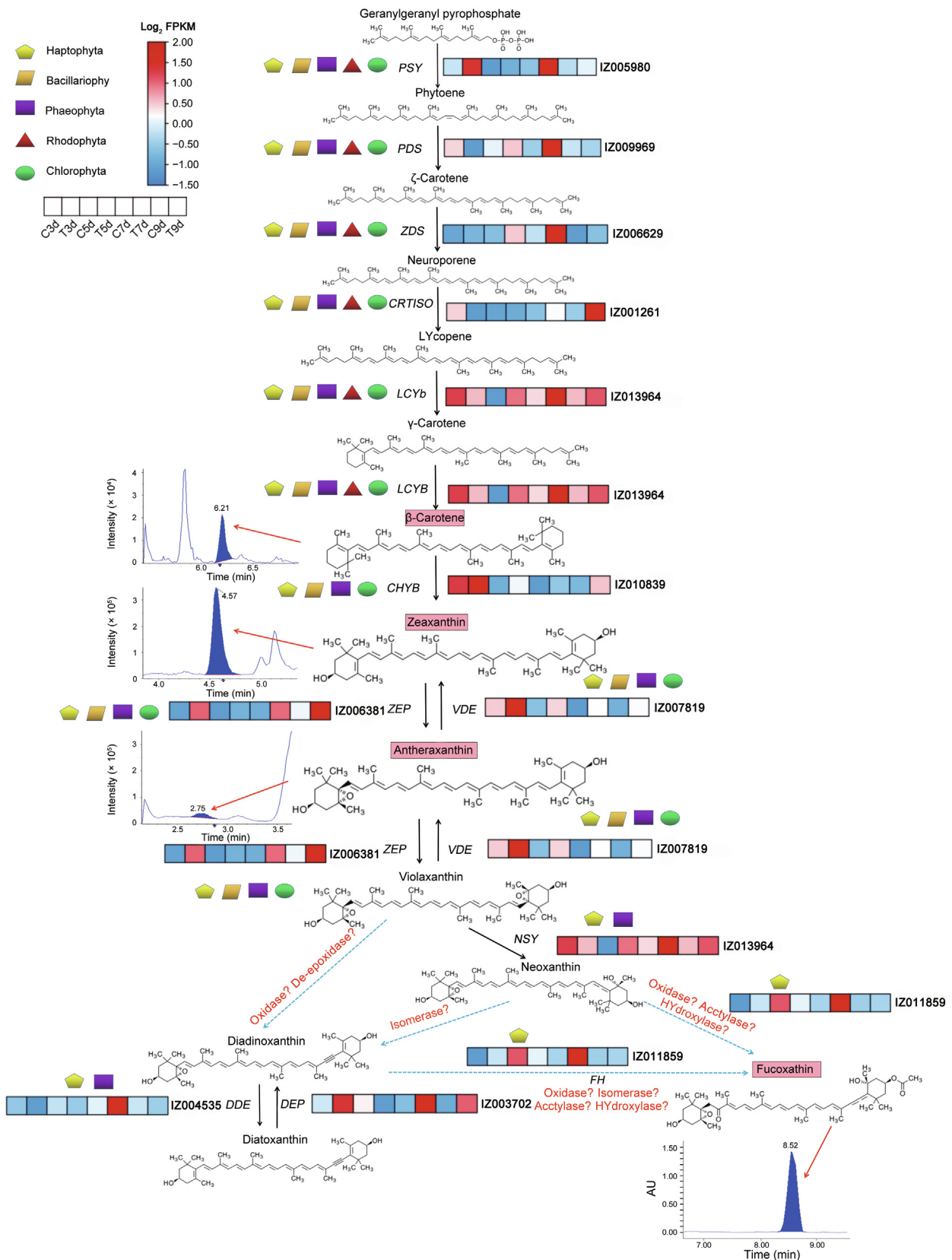
urase (*IgZDS*, IZ006629), phytoene desaturase (*IgPDS*, IZ009969), and Fx-chlorophyll a (*IgLHCX2*, IZ013244) in the red, turquoise, and midnight-blue modules, which are involved in the biosynthetic pathway of  $\beta$ -carotene (Fx synthesis).

#### Accumulation of $\beta$ -carotene, antheraxanthin, zeaxanthin, and Fx showed a strong correlation with the expression of genes related to carotenoid biosynthesis

To explore the relationship between genes and metabolites involved in Fx synthesis under different light qualities (white



**Figure 5** Co-expression network involved in Fx accumulation under different light qualities  
**A.** Hierarchical clustering from WGCNA in control group on 7d. **B.** Hierarchical clustering from WGCNA in treated group on 7d. **C.** Heatmap showing the correlation of modules in control group on 7d. **D.** Heatmap showing the correlation of modules in treated group on 7d. **E.** Transcriptional regulatory network between genes and module membership in control group on 7d. **F.** Transcriptional regulatory network between genes and module membership in treated group on 7d. In (C) and (D), red represents a positive correlation, and green represents a negative correlation. WGCNA, weighted gene co-expression network analysis.



**Figure 6** Diagram of the Fx metabolic pathway based on the results of transcriptional regulation and carotenoid changes in *I. galbana*. DEGs are shown in red (up-regulated) and blue (down-regulated). Heatmap showing log<sub>2</sub> values of transcripts. Chemical reaction possibilities in the Fx pathway are shown in red font. Blue dotted arrows indicate predicted or unknown reactions. Black boxes refer to the content of carotenoid was detected and increased by 7-day green light induction in the Fx pathway. Red arrows indicate chromatogram of the corresponding carotenoid. *PSY*, β-carotene, phytoene synthase; *PDS*, phytoene desaturase; *ZDS*, ζ-carotene desaturase; *CRTISO*, carotenoid isomerase; *LCYb*, lycopene β-cyclase; *CHYB*, lycopene β-cyclase; *ZEP*, zeaxanthin epoxidase; *VDE*, violaxanthin de-epoxidase; *NSY*, neoxanthin synthase; *DDE*, diadinoxanthin de-epoxidase; *DEP*, diatoxanthin epoxidase.

and green), the pathway of DEGs and DACs related to Fx was constructed (Figure 6). Genes involved in the Fx biosynthesis pathway exhibited very high expression levels in the treated group on 7d according to the transcriptome data. Of these, nine DEGs were up-regulated in the T7d group compared to the C7d group, including *IgPSY* (IZ005980), *IgPDS* (IZ009969), *IgZDS* (IZ006629), *IgLCYb* (IZ013964), *IgZEP* (IZ006381), *IgNSY* (IZ013964), *IgDDE* (IZ004535), *IgDEP* (IZ003702), and *IgFH* (IZ011859), while *IgCRTISO* (IZ001261), *IgCHYB* (IZ010839), and *IgVDE* (IZ007819) were down-regulated. There were both  $\beta$ - and  $\epsilon$ -branches of carotenoid biosynthesis in the comparison of C7d vs. T7d, and an abundance of  $\beta$ -carotene as well as small amounts of antheraxanthin, zeaxanthin, and Fx. The content of one carotene ( $\epsilon$ -carotene), two carotenoid esters (violaxanthin-laurate and violaxanthin-myristate), and four xanthophylls (antheraxanthin, capsanthin, zeaxanthin, and Fx) increased, and the content of zeaxanthin-palmitate decreased with green light induction. Taken together,  $\beta$ -carotene, antheraxanthin, zeaxanthin, and Fx involved in Fx biosynthesis were found to be accumulated and up-regulated by green light induction, which showed a trend similar to that of *IgPSY*, *IgLCYb*, *IgNSY*, *IgDDE*, *IgDEP*, *IgFH*, *IgMYB98*, *IgZDS*, *IgPDS*, and *IgLHCX2* (Figure 6). The results showed that the up-regulation of these genes in the T7d group led to the enhancement of the Fx biosynthesis pathway. Therefore, we hypothesized that green light can enhance Fx and  $\beta$ -carotene in the carotenoid pathway. Four unigenes (*IgMYB98*, *IgZDS*, *IgPDS*, and *IgLHCX2*) were selected for expression analysis (Figure S27).

## Discussion

We reported a 92.59-Mb high-quality genome assembly of *I. galbana* with the contig N50 and scaffold N50 both of 6.99 Mb. Moreover, the contig N50 of the assembled genome was 15.68 folds higher than that of prior short-read assemblies (6.99 Mb vs. 419 kb) [16]. These results provide the foundation for studying the regulation of Fx accumulation in Haptophyta and potential applications for other Fx-producing algae. Phylogenetic and collinearity analyses showed that *E. huxleyi*, *I. galbana* and *C. tobinii* as monophyletic groups share a common Haptophyta ancestor. Comparative analysis provided evidence of the WGD event, which was dated at  $\sim 245$  MYA and earlier than the divergence time of *E. huxleyi* and *I. galbana* ( $\sim 133$  MYA). In *I. galbana*, most genes involved in metabolic regulation have a relatively conserved structure or function, and a few TF gene families have even expanded a subset of duplicates. For example, overexpression of *MYB7* could activate the promoter of the lycopene- $\beta$ -cyclase (*AdLCY- $\beta$* ) gene in the carotenoid biosynthesis pathway of kiwifruit, altering the content of carotenoid and chlorophyll [21]. The MYB gene family (*IgMYB98*) maintained a higher number or expansion in Haptophyta and Bacillariophyta, and exhibited higher transcriptional activity, indicating that the latest algae rich in carotenoid linked to responses to water stress environmental stimuli exhibit lineage-specific gene expansions in environmental adaptation and metabolic regulation. Notably, lipid metabolism-related genes exhibit significant expansion, including *IgPLMT*, *IgOARI*, and *IgDEGS1*. These results indicate

that the expansion of lipid metabolism-related genes in *I. galbana* could enhance the regulation and biosynthesis of Fx.

Although Fx plays an important role as a photosensor of blue-green light and an effector of carotenoid-dependent, the metabolic pathway of Fx remains unclear, and there are many unknown steps in the process of violaxanthin to Fx [3,22]. We identified one domain from *I. galbana* genome by comparing with the hydroxylase function domain and predicted the *IgFH* gene, which is closely related to the chemical reaction according to the structure of diadinoxanthin and Fx, suggesting that it might function similarly to the characterized diadinoxanthin and Fx as a candidate gene. The increase in Fx content between the control and treated groups indicated that 7d with green light would be an important period for Fx biosynthesis. GO enrichment showed that these stage-specific genes in the control and treated groups were related to various biological regulation and fatty acid metabolic processes. Several TFs have been implicated in carotenoid accumulation; however, the members of MYB, bHLH, and HB families involved in pigment accumulation and resistance stress showed differential regulation response to the different light qualities [23–25]. We performed WGCNA analysis to identify gene modules. DEGs and TFs were significantly correlated with Fx synthesis. These results suggested that the identified TFs may be related to the accumulation and regulation of Fx production in *I. galbana* by the induction of green light. Transcriptomic data suggest that transcriptional profiling and phenotypic data methods can be beneficial to identify the most promising candidate genes involved in the Fx biosynthesis.

Studies have shown that multiple key genes are related to the Fx biosynthesis pathway, including *PYS*, *PDS*, *ZISO*, *ZDS*, *CRTISO*, and *LCYb* [10,11,14]. Comprehensive analysis of multi-omic data helps reveal the underlying accumulation of carotenoids [26–29]. For example, Jia et al. revealed the molecular mechanism of white petal color in *Brassica napus* by metabolomic and transcriptomic analyses, mining several candidate genes involved in carotenoid biosynthesis (*BnWRKY22* and *BnNCED4b*) [27]. Xia et al. found that DEGs and DACs involved in carotenoid biosynthesis were significantly up-regulated and accumulated more in yellow flower petals than in the green bud petals and white flower petals, indicating a predominantly promotion function for color transition in *Lonicera japonica* [29]. Thirteen genes (*PSY1*, *PSY2*, *PDS1*, *PDS2*, *ZDS*, *CYCB*, *LCYB1*, *LCYB2*, *LCYE*, *CHYB*, *LUT1*, *VDE*, and *ZEP*) were related to the carotenoid biosynthesis, which were strongly correlated with the changes in lycopene,  $\beta$ -carotene, and  $\beta$ -cryptoxanthin, providing an insight into controlling fruit color in papaya fruit [28]. Up- or down-regulated DEGs involved in the carotenoid biosynthetic pathway greatly affect the content of *trans*- $\beta$ -carotene, *trans*- $\beta$ -cryptoxanthin, and 5,8-epoxy- $\beta$ -carotene, resulting in a striking difference between peel and flesh tissue during on-tree loquat development [26]. Although most studies on the carotenoid biosynthesis have focused on the color transition of fruits and flower petals, combined metabolomic and transcriptomic analyses of carotenoid biosynthesis in *I. galbana* have not been reported yet. In this study, we identified 12 DEGs and 4 DACs involved in Fx biosynthesis by metabolomic and transcriptomic analyses. Notable increases in carotenoids involved in Fx biosynthesis in the T7d group compared to the C7d group included  $\beta$ -carotene, antheraxanthin, zeaxanthin and Fx,

suggesting the abundant diversity of carotenoids present under green light. *LCYb* catalyzes the formation of  $\beta$ -carotene and its oxides from lycopene, which is a key step in the synthesis of  $\beta$ -carotene [28]. *ZEP* plays a key function in the xanthophyll cycle of plants, catalyzing the conversion of zeaxanthin to antheraxanthin and violaxanthin [30]. *NXS* catalyzes the conversion of the double-epoxidation precursor violaxanthin into lutein with equilibrated double bonds, representing the classic end of the formation of plant xanthophyll [31]. Therefore, we hypothesized that green light can accumulate  $\beta$ -carotene, zeaxanthin, and Fx by activating the xanthophyll cycle process in the Fx pathway. The results of genome and transcriptome indicate how the genome of *I. galbana* provides a useful model for studying the evolution of Fx-producing algae and the mechanism of Fx biosynthesis.

## Conclusion

In summary, we report a high-quality genome of *I. galbana* LG007 by using the PacBio Sequel platform and Hi-C technology. Domain identification of a novel gene that encodes neoxanthin-Fx hydroxylase was analyzed. Fx content could be increased under green light condition, which is a special simulating factor that occurs during the cultivation of *I. galbana*. Metabolomic analysis indicated that the T7d group accumulated a higher content of carotenoids than that of the C7d group, and  $\beta$ -carotene was the main carotenoid, accounting for 79.09% of the total carotenoids. Multi-omic analysis revealed several DEGs (including TF genes) significantly correlated with the accumulation and regulation of Fx synthesis, including *IgMYB98*, *IgZDS*, *IgPDS*, and *IgLHCX2*. Therefore, our findings advance the understanding of Fx biosynthesis and its regulation, providing an important resource for food and pharmaceutical applications.

## Materials and methods

### Sample materials and genome sequencing

*I. galbana* LG007 was separated from the near sea area of Chuanshi Island in Fujian and deposited with the Southern Institute of Oceanography, Fujian Normal University, China. The seawater used for culture was collected from the near sea area of Chuanshi Island in Fujian, with a salinity of 28‰. The algae were cultured in 100 ml f/2 medium and incubated at  $23\text{ }^{\circ}\text{C} \pm 1\text{ }^{\circ}\text{C}$  by shaking the bottle manually 6 times per day under continuous light of  $100\text{ }\mu\text{mol photons}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$  with fluorescent lamps [32]. Genomic DNA was prepared using the cetyltrimethylammonium bromide (CTAB) method to construct the PacBio and Illumina libraries. Concentrated DNA was applied to select size with BluePippin; they were repaired, tailed, adaptor-ligated, and used for library construction in accordance with the protocol released by PacBio. Next,  $\sim 15.5$  Gb of clean data were obtained from the PacBio sequencing, and used to estimate genome size.

### Genomic size estimation

The BD FACSCalibur cytometer (Becton Dickinson, San Jose, CA) was used to estimate the genome size of *I. galbana* LG007,

which was calculated as a ratio of the average fluorescence. We further used Illumina short reads and a *K*-mer-based method to estimate the genome size, heterozygosity, and repeat content of *I. galbana* LG007. Approximately 8.92 Gb of Illumina data were generated and used to calculate the abundance of 17-mers by GenomeScope software (version 2.0) (Table S1).

### Genome assembly and completeness assessment

After removing the low-quality short reads and sequencing adaptors, the clean data were corrected, trimmed, and assembled by using Canu software with default parameters [33]. For improving the accuracy of base-pair correction, preliminary assembled contigs were polished by the BWA and Pilon software using  $\sim 8.92$  Gb Illumina data [34]. Summary statistics of the assembled genome are presented in Table S1. Assessment of the completeness of *I. galbana* LG007 genome was evaluated through BUSCO using eukaryotic models [35]. Illumina short reads and PacBio long reads were properly mapped to the genome via Bowtie2 [36] and Minimap2 [37], respectively.

### Superscaffold construction using Hi-C technology

The nuclear integrity of samples was examined by 6-diamidino-2-phenylindole (DAPI) staining to guarantee the quality of the Hi-C procedure [38,39]. After filtering adapter sequences and low-quality pair-end reads,  $\sim 12.35$  Gb of clean data were generated (Table S1). The Hi-C clean data were properly mapped to the *I. galbana* LG007 by BWA (version 0.7), and then erroneous mappings and duplicates were filtered by the Juicer pipeline [40]. The output of the Juicer pipeline was used for 3D-DNA analysis with default parameters, including misjoin correction, ordering, and orientation [41]. To ensure the accuracy of assembly, the assembled contigs combined with Hi-C data were ordered and clustered into the superscaffolds by using LACHESIS based on the relationships among valid reads [39], and then the invalid read pairs were filtered by HiC-Pro (version 2.7.8) [42].

### Gene and repetitive sequence annotation

LTR\_FINDER, Tandem Repeats Finder, and RepeatMasker were used to identify the repeat sequences in the *I. galbana* LG007 genome, as previously described [43,44]. We then performed annotation of the *I. galbana* LG007 genome assembly by combining the homology-based search, *de novo* prediction, and transcriptome evidence. *E. huxleyi*, *C. tobinii*, *P. tricoratum*, *E. siliculosus*, and *C. reinhardtii* were selected to perform the homology annotation. We predicted the coding genes with MAKER pipeline (version 2.31.9) by using transcript sequences from RNA sequencing (RNA-seq) [45]. The protein-coding genes were compared to the content of eggNOG, GO, COG, and KEGG by using BLASTP with an E-value cutoff of  $1\text{E}-5$  [46–48]. Non-coding RNAs (ncRNAs) and small RNAs were identified by searching from the Rfam and microRNA (miRNA) databases, respectively [49]. In addition, other types of ncRNAs, including miRNA and snRNA, were predicted by alignment to the Pfam database using Infernal software.

## Genome evolution analysis

Single-copy genes were identified among 15 genomes by using OrthoFinder and downloaded from the National Center for Biotechnology Information (NCBI) database, including *E. huxleyi*, *C. tobinii*, *P. tricornutum*, *E. siliculosus*, *C. reinhardtii*, *Pennisetum purpureum*, *Chloropicon primus*, *Bathycoccus prasinus*, *Porphyra umbilicalis*, *Gracilariopsis chorda*, *Cyanidioschyzon merolae*, *R. subcapitata*, *T. pseudonana*, and *T. oceanica* [50]. Based on the identified single-copy protein sequences, a phylogenetic tree was constructed by using RAxML software with *P. purpureum*, *P. umbilicalis*, *G. chorda* and *C. merolae* as the outgroup [50]. The divergence time of each tree node was calculated using the TimeTree database and the MCMCtree software. We used CAFÉ software (version 3.1) to identify the expansion and contraction of gene families with the criterion of  $P < 0.05$  [51]. GO terms for genes were obtained from the corresponding InterPro or Pfam entries. KEGG terms were assigned to the KEGG pathway database (<https://www.genome.jp/kegg>). Enrichment analyses of KEGG pathways and GO terms were performed using the OmicShare tools (<https://www.omicshare.com/tools>).

### Identification of candidate genes related to Fx pathway in *I. galbana*

The identification of orthologs of previously known functional genes in the Fx pathway was performed by combining the results of the genome assembly annotations, transcriptional expression levels, and conserved domain BLAST searches. The orthologs and previously known functional genes in the Fx pathway exhibited identity scores  $> 85\%$ , suggesting that these genes are functionally similar and can be used as candidate genes involved in Fx biosynthesis. For gene family analysis related to Fx pathway in *I. galbana*, BLASTP and HMMER were used to search for homologous proteins of related gene families in *I. galbana* LG007 (E-value  $< 1E-10$ ), and then homologous proteins were further confirmed using the NCBI conserved domain database tool [52]. The final deduced homologous protein sequences were aligned by using the ClustalW software [53]. RAxML software was used to construct a phylogenetic tree via the maximum-likelihood method with 1000 bootstrap iterations [54].

### Analysis of WGD and gene synteny

For detecting the polyploidization events in the *I. galbana* LG007 genome, the protein sequences from *I. galbana* LG007 were intercompared to identify conserved paralogs by using BLASTP with an E-value  $\leq 1E-5$ . *E. huxleyi*, *C. tobinii*, *P. tricornutum*, and *C. reinhardtii* were also analyzed and used for comparison. We identified the collinear blocks by using MCScanX and calculated the nonsynonymous substitution rate ( $Ka$ ),  $Ks$ , and  $Ka/Ks$  values for syntenic gene pairs by using KaKs\_Calculator software (version 2.0) [55,56]. Syntenic blocks between *I. galbana* LG007, *E. huxleyi* and *C. tobinii* were identified by using MCScanX [55].

### RNA-seq

Our previous results suggested that the green light could promote Fx synthesis at the 7d stage ( $P < 0.05$ , 14.06% higher

(Figure S2). To investigate the transcriptome dynamics and response of Fx accumulation under different light qualities in *I. galbana* LG007, we performed transcriptomic analysis of the simulated cells under white and green light conditions at different stages of cultivation. *I. galbana* LG007 was cultured in 100 ml f/2 medium and incubated at  $23\text{ }^{\circ}\text{C} \pm 1\text{ }^{\circ}\text{C}$  by shaking the bottle manually 4–6 times per day under continuous light of  $100\text{ }\mu\text{mol photons}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$  with a 12 h:12 h light:dark cycle [33]. The culture ( $1 \times 10^6$  cells/ml) was evenly divided into eight groups and cultured in a spectrum-adjustable plant growth box (Catalog No. AKF-KYG04-600DZ, Anhui Ancorgreen Optoelectronics Technology, Hefei, China) for 3 days, 5 days, 7 days, and 9 days, respectively (Figure S16A). Four treated groups (T3d, T5d, T7d, and T9d) were treated with green light irradiation  $100\text{ }\mu\text{mol photons}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$  [green light source: LED circular lamp beads (Catalog No. SZG05A0A, Seoul Semiconductor, Siheung-si, Korea); spectrum: 525 nm], and four control groups (C3d, C5d, C7d, and C9d) with a white light of  $100\text{ }\mu\text{mol photons}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$  [white light source: LED circular lamp beads (Catalog No. LH351H-D, Samsung LED, Tianjin, China); spectral range: 400–700 nm] were used as controls. Total RNA from each sample was extracted using a TransZol Up Plus RNA Kit (Catalog No. ER501-01, TransGen Biotech, Beijing, China), and the corresponding cDNA library was constructed for RNA-seq.

### Gene expression analysis

Approximately 307.77 Gb of high-quality transcript data were produced and processed by Trimmomatic (version 0.36). The high-quality filtered reads were mapped onto the genome by using HISAT2 with the default parameters [57,58]. FPKM values were calculated using Stringtie and Ballgown [59,60]. DEGs between the control and treated groups were analyzed using DESeq2 based on criteria of fold change  $\geq 1$  and false discovery rate  $\leq 0.05$  [61], followed by KEGG and GO enrichment analyses. The ratio of each sample to the genome is more than 90%, and the number of reads per sample was estimated to range from 32,792,958 to 59,111,858 (Tables S22 and S23).

### Metabolite profiling and statistical analysis

To explore the metabolites of *I. galbana* LG007 under white and green light conditions, we collected C7d and T7d samples with three biological replicates. Stock solution of Fx was prepared by dissolving 0.5 mg Fx in 50 ml methanol solution. Stock solutions of Fx standard were gradient diluted as follows: 5  $\mu\text{g/ml}$ , 10  $\mu\text{g/ml}$ , 20  $\mu\text{g/ml}$ , 50  $\mu\text{g/ml}$ , and 100  $\mu\text{g/ml}$ . Fx production was detected according to the linear relationship between the peak areas of the samples and a standard curve ( $R^2 = 0.999$ ). According to the aforementioned method of *I. galbana* LG007 fermentation, 10 ml of the fermentation mixture was centrifuged at 8000 r/min for 20 min, followed by removal of the supernatant and washing with distilled water for three times. After vacuum freeze-drying, 1 ml of acetone was added to the freeze-dried algae to extract the total carotenoids. The supernatant was harvested by centrifugation (8000 r/min, 15 min) and filtered (0.25- $\mu\text{m}$  filter membrane). The supernatant was analyzed by high-performance liquid chromatography (HPLC) using a Waters e2695 Liquid

Chromatograph (2695 type, Waters, Guangzhou, China) equipped with a Waters 2998 Photodiode Array (2998 type, Waters) detector and separated on a SunFire C18 HPLC column (250 × 4.6 mm; 5 μm) (SunFire C18 column, Waters). The mobile phase consisted of ternary solvents of water (A)/methanol (B)/acetonitrile (C) (15:30:55, v/v/v) and the flow rate of the mobile phase was 1 ml/min. The Fx content was detected using a PDA detector at 447 nm. Except for Fx, other carotenoids are obtained using Metware (<https://www.metware.cn/>) according to the following method. 1) The vacuum freeze-dried algae were crushed using a grinding mill (Catalog No. MM400, Retsch, Arzberg, Germany) at 30 Hz for 1.5 min. Powder (100 mg) was dissolved in 1.2 ml of 70% methanol, vortexed for 30 s, and stored at 4 °C overnight. 2) The mixture was centrifuged at 12,000 r/min for 10 min, and then filtered through a 0.22-μm membrane to obtain the supernatant for subsequent analysis [62]. 3) Carotenoid content was detected using an Ultra Performance Liquid Chromatography (UPLC) system (ExionLC AD, Sciex, Framingham, MA) and a Tandem Mass Spectrometry (MS/MS) system (QTRAP 6500+, Sciex), which was equipped with an atmospheric pressure chemical ionization (APCI)+ and controlled by Analyst (version 1.6.3) software. DACs were determined by fold change  $\geq 1$  and  $P < 0.05$ . Identified metabolites were annotated and mapped by using the KEGG compound and pathway databases, respectively.

### Co-expression network analysis

Based on  $\log_2(1 + \text{FPKM})$  values, WGCNA was performed by using a minimum module size of 30, with a soft power of 11 (control group), 8 (treated group), and 14 (control-treated group), as well as a merge cut height of 0.25 (Figure S28). Eigengene values of WGCNA module were calculated and associated with lipid and Fx content at different culture stages [63]. Each module gene was analyzed by GO enrichment and visualized using Cytoscape [64].

### qRT-PCR validation and *in vitro* experiments

qRT-PCR experiment was performed using SYBR Green PCR Master Mix (Catalog No. P2091, ThermoFisher Scientific, Waltham, MA) in an Applied Biosystems 7300 Real-Time PCR System (Catalog No. 7300, ThermoFisher Scientific) [65]. The *IgHF* gene was amplified from *I. galbana* LG007, and ligated into the pTrc99a vector (Figure S29). After transformation into *Escherichia coli* K-12 MG1655 cells (ThermoFisher Scientific), recombinant protein expression was induced by 0.2 mM isopropyl-thio-β-D-galactopyranoside with vigorous shaking at 220 r/min for 24 h at 37 °C. Thirty optical density (OD) cells were harvested by centrifugation at 8000 r/min for 10 min, and then induced with 10 ml Tris-HCl lysis buffer (50 mM, pH 7.5), 10% (v/v) glycerol, and 1.67 μM neoxanthin (Sigma-Aldrich, Louis, MO) with vigorous shaking at 220 r/min at 37 °C for 12 h. Then, 2 ml cultures were centrifuged at 8000 r/min for 5 min, and suspended in 2 ml methanol (chromatographic grade) by ultrasonic crushing at 60 Hz for 20 min. The supernatant was obtained by centrifugation (8000 r/min, 5 min) and filtration (0.22-μm filter membrane) and further used for HPLC analysis.

### Data availability

The assembled genome sequences have been deposited in BioProject at the NCBI (BioProject: [PRJNA669236](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA669236)), which are accessible at <https://www.ncbi.nlm.nih.gov/bioproject>. Raw data of RNA-seq have been deposited in the Genome Sequence Archive [66] at the National Genomics Data Center (NGDC), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS) / China National Center for Bioinformatics (CNCB) (GSA: [CRA003291](https://ngdc.cncb.ac.cn/gsa)), which are accessible at <https://ngdc.cncb.ac.cn/gsa>. The whole-genome sequence data reported in this study have been deposited in the Genome Warehouse [67] at the NGDC, BIG, CAS / CNCB (GWH: [GWH: GWHZHV000000000](https://ngdc.cncb.ac.cn/gwh)), which are accessible at <https://ngdc.cncb.ac.cn/gwh>.

### CRedit author statement

**Duo Chen:** Methodology, Formal analysis, Writing - original draft. **Xue Yuan:** Formal analysis, Data curation. **Xuehai Zheng:** Methodology, Validation, Data curation. **Jingping Fang:** Formal analysis, Resources, Data curation. **Gang Lin:** Resources. **Rongmao Li:** Resources. **Jiannan Chen:** Resources. **Wenjin He:** Investigation. **Zhen Huang:** Investigation. **Wenfang Fan:** Formal analysis. **Limin Liang:** Validation, Data curation. **Chentao Lin:** Resources. **Jinmao Zhu:** Writing - review & editing. **Youqiang Chen:** Writing - review & editing. **Ting Xue:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition. All authors have read and approved the final manuscript.

### Competing interests

The authors have declared no competing interests.

### Acknowledgments

We thank Feng Qi (Fujian Normal University) for sharing the plasmid and method that were used for the catalytic reaction of the *IgFH* gene in *E. coli*. This work was supported by the National Natural Science Foundation of China (Grant No. 42006087) and the Sugar Crop Research System, China (Grant No. CARS-170501).

### Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.gpb.2022.05.010>.

### ORCID

ORCID 0000-0002-2491-8532 (Duo Chen)  
 ORCID 0000-0003-0441-1985 (Xue Yuan)  
 ORCID 0000-0002-2243-6150 (Xuehai Zheng)  
 ORCID 0000-0003-4037-5862 (Jingping Fang)  
 ORCID 0000-0002-8989-3647 (Gang Lin)  
 ORCID 0000-0003-3927-4845 (Rongmao Li)

ORCID 0000-0003-4846-4245 (Jiannan Chen)  
 ORCID 0000-0002-7782-5276 (Wenjin He)  
 ORCID 0000-0002-8908-0355 (Zhen Huang)  
 ORCID 0000-0001-5628-0986 (Wenfang Fan)  
 ORCID 0000-0001-9079-8471 (Limin Liang)  
 ORCID 0000-0001-5582-6641 (Chentao Lin)  
 ORCID 0000-0002-7328-8821 (Jinmao Zhu)  
 ORCID 0000-0002-3836-1608 (Youqiang Chen)  
 ORCID 0000-0003-0727-9158 (Ting Xue)

## References

- [1] Xia S, Wang K, Wan LL, Li AF, Hu Q, Zhang C. Production, characterization, and antioxidant activity of fucoxanthin from the marine diatom *Odontella aurita*. *Mar Drugs* 2013;11:2667–81.
- [2] Peng J, Yuan JP, Wu CF, Wang JH. Fucoxanthin, a marine carotenoid present in brown seaweeds and diatoms: metabolism and bioactivities relevant to human health. *Mar Drugs* 2011;9:1806–28.
- [3] Wang W, Yu LJ, Xu C, Tomizaki T, Zhao S, Umena Y, et al. Structural basis for blue-green light harvesting and energy dissipation in diatoms. *Science* 2019;363:eaav0365.
- [4] Wang W, Zhao S, Pi X, Kuang T, Sui SF, Shen JR. Structural features of the diatom photosystem II-light-harvesting antenna complex. *FEBS J* 2020;287:2191–200.
- [5] Maeda H, Kanno S, Kodate M, Hosokawa M, Miyashita K. Fucoxanthinol, metabolite of fucoxanthin, improves obesity-induced inflammation in adipocyte cells. *Mar Drugs* 2015;13:4799–813.
- [6] Zhu Y, van Eyk PJ, Bomanc C, Broström M, Kirtaniad K, Piotrowska P, et al. Preliminary understanding on the ash behavior of algae during co-gasification in an entrained flow reactor. *Fuel Process Technol* 2018;175:26–34.
- [7] Zhao DH, Zhao LD, You H, Qin S, Wang YC, Jiao XD. Breeding and feeding characteristics of high-temperature-resistant strains of *Isochrysis galbana*. *Oceanol Limnol Sinica* 2021;1:206–12.
- [8] Kim SM, Kang SW, Kwon ON, Chung D, Pan CH. Fucoxanthin as a major carotenoid in *Isochrysis* aff. *galbana*: characterization of extraction for commercial application. *J Korean Soc Appl Biol Chem* 2012;55:477–83.
- [9] Koster M, Sietmann R, Meuche A, Paffenhofer GA. The ultrastructure of a doliolid and a copepod fecal pellet. *J Plankton Res* 2011;33:1538–49.
- [10] Liang MH, Zhu J, Jiang JG. Carotenoids biosynthesis and cleavage related genes from bacteria to plants. *Crit Rev Food Sci* 2018;58:2314–33.
- [11] Mikami K, Hosokawa M. Biosynthetic pathway and health benefits of fucoxanthin, an algae specific xanthophyll in brown seaweeds. *Int J Mol Sci* 2013;14:13763–81.
- [12] Lohr M, Wilhelm C. Xanthophyll synthesis in diatoms: quantification of putative intermediates and comparison of pigment conversion kinetics with rate constants derived from a model. *Planta* 2001;212:382–91.
- [13] Dambek M, Eilers U, Breitenbach J, Steiger S, Büchel C, Sandmann G, et al. Biosynthesis of fucoxanthin and diadinoxanthin and function of initial pathway genes in *Phaeodactylum tricoratum*. *J Exp Bot* 2012;63:5607–12.
- [14] Zhang NN, Luo L, Chen Z, Yang ZF, Huang FH, Wan X, et al. Biosynthesis pathway of fucoxanthin and expression levels of key genes for fucoxanthin synthesis in response to high irradiance in *Phaeodactylum tricoratum*. *Chinese J Oil Crop Sci* 2017;39:128–36.
- [15] Yu K, Gong YF, Zhu SQ, Liu H, Wang HY. Effects of different exogenous elicitors on *lyb* gene transcription and fucoxanthin content in *Phaeodactylum tricoratum*. *J Agric Biotechnol* 2017;25:2009–17.
- [16] Wang Y. Genome assembly and annotation of *Isochrysis galbana*. California: San Marcos; 2014.
- [17] McGee D, Archer L, Fleming GTA, Gillespie E, Touzet N. Influence of spectral intensity and quality of LED lighting on photoacclimation, carbon allocation and high-value pigments in microalgae. *Photosynth Res* 2020;143:67–80.
- [18] Fixen KR, Starkenburg SR, Hovde BT, Johnson SL, Deodato CR, Daligault HE, et al. Genome sequences of eight bacterial species found in coculture with the haptophyte *Chrysochromulina tobin*. *Genome Announc* 2016;4:e01162-16.
- [19] Fulton JM, Fredricks HF, Bidle KD, Vardi A, Kendrick BJ, DiTullio GR, et al. Novel molecular determinants of viral susceptibility and resistance in the lipidome of *Emiliania huxleyi*. *Environ Microbiol* 2014;16:1137–49.
- [20] Zheng M, Yue Y, Wan W, Liang K, Zheng L. Cloning and stress-responsive transcription of 3-ketoacyl-CoA thiolase gene of *Isochrysis galbana* (Haptophyta). *J Appl Phycol* 2013;25:695–702.
- [21] Ampomah-Dwamena C, Thrimawithana AH, Dejnopratt S, Lewis D, Espley RV, Allan AC. A kiwifruit (*Actinidia deliciosa*) R2R3-MYB transcription factor modulates chlorophyll and carotenoid accumulation. *New Phytol* 2018;221:309–25.
- [22] Kwon DY, Vuong TT, Choi J, Lee TS, Um JI, Koo SY, et al. Fucoxanthin biosynthesis has a positive correlation with the specific growth rate in the culture of microalga *Phaeodactylum tricoratum*. *J Appl Phycol* 2021;33:1473–85.
- [23] Chen K, Du L, Liu H, Liu Y. A novel R2R3-MYB from grape hyacinth, MaMybA, which is different from MaAN2, confers intense and magenta anthocyanin pigmentation in tobacco. *BMC Plant Biol* 2019;19:390.
- [24] Lotkowska ME, Tohge T, Fernie AR, Xue GP, Balazadeh S, Mueller-Roeber B. The *Arabidopsis* transcription factor MYB112 promotes anthocyanin formation during salinity and under high light stress. *Plant Physiol* 2015;169:1862–80.
- [25] Zhu Z, Chen G, Guo X, Yin W, Yu X, Hu J, et al. Overexpression of SIPRE2, an atypical bHLH transcription factor, affects plant morphology and fruit pigment accumulation in tomato. *Sci Rep* 2017;7:5786.
- [26] Hadjipiropi M, Georgiadou EC, Marin A, Diaz-Mula HM, Goulas V, Fotopoulos V, et al. Metabolic and transcriptional elucidation of the carotenoid biosynthesis pathway in peel and flesh tissue of loquat fruit during on-tree development. *BMC Plant Biol* 2017;17:102.
- [27] Jia L, Wang J, Wang R, Duan M, Qiao C, Chen X, et al. Comparative transcriptomic and metabolomic analyses of carotenoid biosynthesis reveal the basis of white petal color in *Brassica napus*. *Planta* 2021;253:8.
- [28] Shen YH, Yang FY, Lu BG, Zhao WW, Jiang T, Feng L, et al. Exploring the differential mechanisms of carotenoid biosynthesis in the yellow peel and red flesh of papaya. *BMC Genomics* 2019;20:49–60.
- [29] Xia Y, Chen W, Xiang W, Wang D, Xue B, Liu X, et al. Integrated metabolic profiling and transcriptome analysis of pigment accumulation in *Lonicera japonica* flower petals during colour-transition. *BMC Plant Biol* 2021;21:98.
- [30] Hoang MH, Kim HS, Zulfugarov IS, Lee CH. Down-regulation of zeaxanthin epoxidation in vascular plant leaves under normal and photooxidative stress conditions. *J Plant Biol* 2020;63:331–6.
- [31] Bouvier F, D'harlingue A, Backhaus RA, Kumagai MH, Camara B. Identification of neoxanthin synthase as a carotenoid cyclase paralog. *Eur J Biochem* 2000;267:6346–52.
- [32] Smith WL, Chanley MH. Culture of marine invertebrate animals. 1st ed. Boston: Springer; 1975.
- [33] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* 2017;27:722–36.

- [34] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;9:e112963.
- [35] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31:3210–2.
- [36] Langdon WB. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Minn* 2015;8:1.
- [37] Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–100.
- [38] Yang X, Yue H, Li H, Ding W, Chen G, Shi T, et al. The chromosome-level quality genome provides insights into the evolution of the biosynthesis genes for aroma compounds of *Osmanthus fragrans*. *Hortic Res* 2018;5:72.
- [39] Burton JN, Adey A, Patwardhan RP, Qiu RL, Kitzman JO, Shendure J. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol* 2013;31:1119–25.
- [40] Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* 2016;3:95–8.
- [41] Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 2017;356:92–5.
- [42] Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 2015;16:259.
- [43] Xu Z, Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 2007;35:W265–8.
- [44] Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;27:573–80.
- [45] Protein maker and gene regulator. *Nature* 2011;473:127.
- [46] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;25:25–9.
- [47] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;42:D199–205.
- [48] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000;28:33–6.
- [49] Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* 2016;44:W54–7.
- [50] Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;16:157.
- [51] Xu Y, Hou H, Liu Q, Liu J, Dou L, Qian G. Removal behavior research of orthophosphate by CaFe-layered double hydroxides. *Desalin Water Treat* 2016;57:7918–25.
- [52] Marchler-Bauer A, Lu SN, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 2011;39:D225–9.
- [53] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and clustal X version 2.0. *Bioinformatics* 2007;23:2947–8.
- [54] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–3.
- [55] Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucl Acids Res* 2012;40:e49.
- [56] Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs\_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* 2010;8:77–80.
- [57] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–20.
- [58] Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* 2016;11:1650–67.
- [59] Keel BN, Snelling WM. Comparison of Burrows-Wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: application to illumina data for livestock genomes. *Front Genet* 2018;9:35.
- [60] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;33:290–5.
- [61] Nikolayeva O, Robinson MD. edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. *Methods Mol Biol* 2014;1150:45–79.
- [62] Cao X, Liu M, Hu Y, Xue Q, Yao F, Sun J, et al. Systemic characteristics of biomarkers and differential metabolites of raw and ripened pu-erh teas by chemical methods combined with a UPLC-QQQ-MS-based metabolomic approach. *LWT* 2021;136:110316.
- [63] Peter L, Steve H. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559.
- [64] Michael ES, Keiichiro O, Johannes R, Wang PL, Trey L. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011;27:431–2.
- [65] Borecká-Melkusová S, Moran GP, Sullivan DJ, Kucharíková S, Chorvát Jr D, Bujdákova H. The expression of genes involved in the ergosterol biosynthesis pathway in *Candida albicans* and *Candida dublimiensis* biofilms exposed to fluconazole. *Mycoses* 2009;52:118–28.
- [66] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinformatics* 2021;19:578–83.
- [67] Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome Warehouse: a public repository housing genome-scale data. *Genomics Proteomics Bioinformatics* 2021;19:584–9.