# scientific **data**

Check for updates

# Transcriptome profiling and co-expression network analysis of 96 *Haematococcus pluvialis* samples

Bin Hu[1], Kai Liu[1], Feichao Du[1], Hailiang Xing[1], Xiaoyuan Su[1], Tao Hu[1], Feng Chen[1], Yujie Qin[1], Zhaozhen Dong[1], Xue Sun[1], Liuquan Zhang[1], Jagroop Pandhal[2], Nianjun Xu[1 ✉] & Chaoyang Hu[1 ✉]

*Haematococcus pluvialis* is a microalga known for producing the red carotenoid, astaxanthin. Key research areas to improve productivity include optimizing vegetative biomass, enhancing astaxanthin content, and controlling secondary cell wall formation. In this study, RNA-seq was conducted on 96 *H. pluvialis* samples under various treatments and time points, generating 96.7 GB of high-quality data (2,080,511,173 clean reads, quality score > 30). Gene expression was quantified as transcripts per million (TPM), and genes with similar expression patterns were clustered. A co-expression network, constructed with a soft threshold of $\beta = 7$ ($R^2 > 0.9$), identified 39 co-expression modules, each averaging 732 genes. Additionally, the re-annotation process provided functional descriptions for 8,598 previously uncharacterized genes. Gene functional classification analysis revealed their involvement in metabolic pathways related to genetic information processing and primary/secondary metabolite metabolism. This dataset provides a valuable resource for exploring the molecular mechanisms involved in critical processes such as growth regulation, astaxanthin accumulation, and secondary cell wall synthesis in *H. pluvialis*.

## Background & Summary

*Haematococcus pluvialis* is a freshwater unicellular green microalga belonging to the class Chlorophyceae, order Volvocales, and family Haematococcaceae. It has been cultivated on an industrial scale[1]. This species is wildly distributed across the world, often dominating temporary freshwater habitats such as shallow pools, rocky depressions, and birdbaths[2]. *H. pluvialis* garners significant interest due to its status as the richest natural source of astaxanthin, a potent antioxidant with nutraceutical and pharmacological applications[3]. Astaxanthin can be obtained from both natural sources and chemical synthesis. While chemical synthesis method allows for large-scale production at a lower cost, the resulting product often contains a mixture of isomers and by-products. These impurities can reduce the antioxidant capacity and bioavailability of synthetic astaxanthin compared to its natural counterpart. Additionally, the use of petrochemical sources in its production raises concerns about food safety, potential toxicity in the final product. Hence, synthetic astaxanthin has not been proved for direct human consumption, and is primarily used in industrial applications such as aquaculture feed for coloration purposes[4]. In contrast, natural astaxanthin is free from harmful by-products and contaminants, and has been approved for direct human consumption[5]. In the aquaculture industry, *H. pluvialis* powder, rich in natural astaxanthin, can also been used as feed to enhance coloration and improve the growth and immune capacity of crabs, shrimp, and fish[6–9]. In humans, natural astaxanthin-based dietary supplements have been shown to boost immune response and help prevent various human diseases, including cardiovascular disease, ocular diseases, cancer, diabetes, and UV-induced skin damage[10]. As a result, natural astaxanthin is indispensable and is expected to experience substantial demand and growth in the coming years[11]. Additionally, the significantly higher price of natural astaxanthin compared to its synthetic counterpart has drawn increased attention to its production[5].

As a unicellular microalga, *H. pluvialis* is highly sensitive to environmental changes. Under optimal conditions, *H. pluvialis* cells are green, motile, and rapidly increase in biomass through cell division. Sodium acetate is one of the most effective organic carbon sources for enhancing the growth rate and biomass of *H. pluvialis*[12].

[1]Key Laboratory of Aquacultural Biotechnology of Ministry of Education, School of Marine Sciences, Ningbo University, Ningbo, 315832, China. [2]Department of Chemical and Biological Engineering, University of Sheffield, Sheffield, United Kingdom. ✉e-mail: xunianjun@nbu.edu.cn; huchaoyang@nbu.edu.cn

Under stress conditions, the cells undergo significant physiological and morphological transformations, including cessation of cell division, loss of flagella, accumulation of astaxanthin, and formation of a thick secondary cell wall (SCW). This process ultimately results in the development of astaxanthin-rich aplanospores[13,14]. The production of astaxanthin from *H. pluvialis* typically follows a two-stage process to optimize yield, as the conditions favourable for vegetative growth differ from those required for astaxanthin accumulation[15]. In the first stage, *i.e.*, the vegetative growth phase, biomass production is maximized by cultivating algal cells in nutrient-rich media under low-light conditions. Optimal biomass productivity has been achieved using mixotrophic culture systems, which include sodium acetate as an organic carbon source and carbon dioxide as an inorganic carbon source[16]. The second stage, *i.e.*, the astaxanthin accumulation phase, is triggered by subjecting the cells to stress conditions, such as high light intensity, nutrient deprivation, UV-B irradiation, drought, or salinity stress[17,18]. Among these, high light exposure is one of the most effective and wildly used methods for inducing astaxanthin accumulation in *H. pluvialis*[19]. Combining high light stress with chemical modulators such as sucrose[20], potassium iodide[21], trisodium citrate[22], and sodium acetate[11], or with other stress conditions such as nitrogen deprivation[23], and salinity stress[24], can further enhance astaxanthin accumulation in *H. pluvialis*. Elucidating the mechanisms by which inducers promote the growth and astaxanthin accumulation in *H. pluvialis* holds significant biological importance.

The SCW of *H. pluvialis* exhibits a tri-layered structure consisting of the trilaminar sheath, the secondary wall, and the tertiary wall[25,26]. The trilaminar sheath is primarily composed of algaenan, a biopolymer resistant to acetylation. The secondary wall is mainly constituted of mannan, while the tertiary wall comprises heterogenous polysaccharides. This complex and durable SCW provides *H. pluvialis* cells with considerable resistance to chemical and physical damage, allowing them to endure various abiotic and biotic stresses. However, this structural robustness significantly hinders the digestion and absorption of astaxanthin within aplanospores for both humans and animals[27]. Furthermore, it complicates the extraction of astaxanthin through physical and chemical methods. Therefore, understanding the mechanisms by which stress conditions influence SCW biosynthesis, as well as the relationship between SCW formation and astaxanthin accumulation, is crucial.

Transcriptome analysis using rapid and cost-effective next-generation sequencing (NGS) enables precise, high-throughput profiling of gene expression. It has become a standard tool in life sciences research, enhancing our understanding of molecular mechanisms, particularly in plant responses to environmental factors, by providing an accurate depiction of cellular processes[28,29]. In this study, the transcriptome dataset of *H. pluvialis* using 96 RNA samples collected from two stages—vegetative growth and astaxanthin accumulation—under multiple treatments and at various time points were analysed. This dataset offers a comprehensive overview of gene expression in *H. pluvialis* under different conditions. Furthermore, it serves as a valuable tool for exploring the regulatory mechanisms behind cell growth, astaxanthin accumulation, and SCW biosynthesis. This can be achieved by examining the differential growth rates, variations in astaxanthin and SCW accumulation across multiple samples, and by gaining insights from weighted gene co-expression network analysis (WGCNA).

## Methods

**Design and sample processing.**    *H. pluvialis* (strain number: NBU489) was obtained from the Algae Collection Lab at Ningbo University, which was originally collected from Dongqian Lake in Ningbo, Zhejiang, China. The algal cultivation was carried out in two-stage method[30]. Algal cells were initially cultivated under low light conditions to promote biomass growth, and subsequently transferred to high light conditions to induce astaxanthin accumulation. In the low light stage, *i.e.*, the vegetative growth stage, cells were grown in 250 mL Erlenmeyer flasks containing 150 mL of basal medium, under a light intensity of 25–30 $\mu$mol photon m$^{-2}$ s$^{-1}$ with a 12-hour light/dark cycle at 24 °C. The initial cell density was about $2.5 \times 10^4$ cells/mL. The basal medium used was NBU3$^\#$, containing the following components: NaNO$_3$ (100 mg/L); Na$_2$EDTA (20 mg/L); K$_2$HPO$_4$ (10 mg/L); FeSO$_4$·7H$_2$O (2.5 mg/L); MnSO$_4$ (0.25 mg/L); vitamin B$_{12}$ ($5 \times 10^{-4}$ mg/L); and vitamin B$_1$ ($1.2 \times 10^{-3}$ mg/L). In the high light stage, *i.e.*, astaxanthin accumulation stage, flasks containing *H. pluvialis* cells at the plateau stage from first stage, with a cell density of approximately $2.5 \times 10^5$ cells/mL, were exposed to continuous illumination at a higher light intensity of 140 $\mu$mol photon m$^{-2}$ s$^{-1}$ [31].

Based on this two-stage cultivation strategy, samples for the sequencing project were collected at different time points from both the low and high light stages under various treatment conditions. At the low-light stage, algal cells were collected from both the logarithmic phase (sample IDs: 1–3) and plateau phase (sample IDs: 4–6) of the basal medium group, as well as from the logarithmic phase of the 1.8 g/L sodium acetate treatment group (sample IDs: 7–9). As described in the cultivation method above, the three samples collected during the low-light plateau phase (sample IDs: 4–6) can be regarded as the day 0 samples for high light exposure. During the high-light stage, samples were collected from different groups at various time points. On day 1 of high light exposure, samples were collected from the control group (sample IDs: 10–15), the 1 mM potassium iodide treatment group (sample IDs: 43–45), the nitrogen deprivation group (1/4 NBU3# without NaNO$_3$, sample IDs: 52–54), the nitrogen repletion group (1/4 NBU3# plus 25 mg/L NaNO$_3$, sample IDs: 58–60), the 1 mM fructose diphosphate treatment group (sample IDs: 64–66), and the 5 mM fructose diphosphate treatment group (sample IDs: 70–72). On day 2 of high light exposure, samples were collected from the control group (sample IDs: 16–18) and the 0.5 g/L NaCl plus 0.5 mM cytidine 5′-monophosphate (CMP) treatment group (sample IDs: 76–78). At the 60-hour of high light exposure, samples were collected from the control group (sample IDs: 19–21) and a mutant strain (RMS 10, sample IDs: 79–81). On day 3 of high light exposure, samples were collected from the control group (sample IDs: 22–30), the 1 mM xanthosine treatment group (sample IDs: 82–87), and the 1 mM taurine treatment group (sample IDs: 88–90). On day 5 of high light exposure, samples were collected from the control group (sample IDs: 31–39), the 1 mM potassium iodide treatment group (sample IDs: 46–48), the nitrogen deprivation group (1/4 NBU3$^\#$ without NaNO$_3$, sample IDs: 55–57), the nitrogen repletion group (1/4 NBU3$^\#$ plus 25 mg/L NaNO$_3$, sample IDs: 61–63), the 1 mM fructose diphosphate treatment group (sample IDs: 67–69),
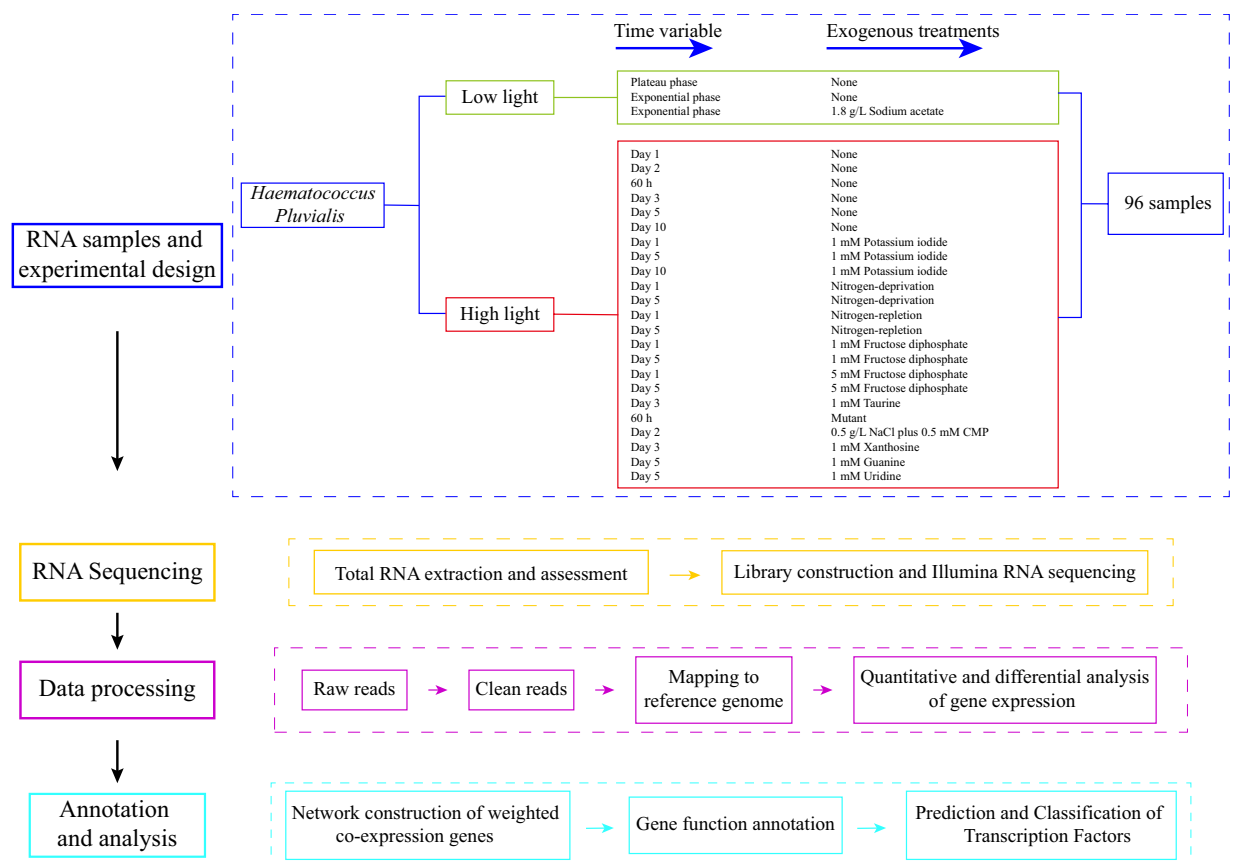
**Fig. 1** Overview of the experimental design and data analysis pipeline. This study comprised several key stages: Experimental design, RNA sequencing, Data processing, annotation, and analysis, each marked with distinct colours. Briefly, RNA from *H. pluvialis* under 26 different treatments was collected for Illumina RNA-seq. Then the sequencing data were subjected to quality control, comparison, quantification, and differential analysis. Finally, all genes underwent comprehensive re-annotation, co-expression networks were constructed, and transcription factors were predicted.

the 5 mM fructose diphosphate treatment group (sample IDs: 73–75), the 1 mM uridine treatment group (sample IDs: 91–93), and the 1 mM guanine treatment group (sample IDs: 94–96). Finally, on day 10 of high light exposure, samples were collected from the control group (sample IDs: 40–42) and the 1 mM potassium iodide treatment group (sample IDs: 49–51). The control group had the same culture conditions (basal medium, light intensity, photoperiod, etc.) as the experimental group in the same batch, except for the absence of exogenous inducer addition prior to high light stress. These samples were collected from ten batches of experiment. These experiments were designed to achieve three primary objectives: (1) enhancing biomass production (batches II, VII, IX and X), (2) increasing astaxanthin accumulation (batches I, III, IV, V, VI and IX), and (3) suppressing SCW synthesis (batches I, III, IV, V, VI, VII, VIII, IX and X). The sample information arranged by experimental batches was listed in Supplementary Table S1. The morphological and physiological traits of these samples were also described in Supplementary Table S1. Algal cells were harvested by centrifugation, immediately frozen in liquid nitrogen, and stored at $-80\,°C$ for subsequent transcriptomic profiling. An overview of the experimental workflows is provided in Fig. 1.

**RNA extraction, library preparation, and sequencing.** Total RNA of each sample was extracted using Trizol (Sangon Biotech, China) in accordance with the manufacturer's instructions. RNA concentration and quality were assessed using a NanoDrop ND1000 spectrophotometer (Thermo Scientific, Waltham, MA, USA) and an Agilent Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). RNA samples meeting the following criteria were used for library preparation and sequencing: concentration $> 200\,ng/\mu L$, $OD_{260}/OD_{280} > 1.8$, $OD_{260}/OD_{230} > 1.8$, $RIN > 6.5$, and 28S/18S ratio $> 1.0$. After passing quality control, RNA sequencing libraries were constructed using the MGIEasy™ mRNA Library Preparation Kit (MGI, Shenzhen, China). mRNA was enriched using oligo(dT) magnetic beads, which bind to the polyA tail of mRNA through A-T base pairing. The enriched mRNA was fragmented and reverse transcribed into second-strand cDNA using N6 random primers. The cDNA was then purified using DNA magnetic beads and subjected to PCR amplification. After amplification, the library was purified again using DNA magnetic beads. The quality and size of the libraries were evaluated using a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and an Agilent 2100 system (Agilent Technologies, Santa

Clara, CA, USA). Sequencing was performed by Wuhan BGI (Wuhan, China) on the MGISEQ-2000 platform, generating approximately 20 million clean reads per sample using single-end sequencing.

**Quality control and processing of RNA-seq data.** The quality of RNA-seq data was assessed using Fastp software[32], which filtered out low-quality reads based on several criteria: bases with a quality score below Q20, reads shorter than 15 bp, reads containing more than 5% N bases, and other default settings. Following filtering, FastQC was employed to generate quality control reports, which were subsequently consolidated using MultiQC v1.22.3 with default settings[33,34]. Clean reads from each sample were then aligned to the *H. pluvialis* genome (GenBank accession: GCA_011766145.1)[35] using HISAT2 v2.2.1[36] in end-to-end and sensitive mode. The alignment results were sorted and converted to BAM format using SAMtools[37]. Raw read counts were obtained and normalized to quantify transcript expression levels using transcripts per million (TPM) values, with featureCounts v2.0.1 of the Subread package used for quantification[38]. TPM normalization is based on raw counts and gene length information by using Trimmed Means of M-values method (TMM). For each sample, gene raw counts are divided by the gene length, and the resulting values are divided by the sum of gene-length normalized values for all genes and multiplied by one million[39]. Differential expression analysis was carried out with DESeq2 v1.34.0[40], and the p-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure within DESeq2 to control the false discovery rate (FDR). Functional enrichment analysis for the genes, including both unknown genes and module genes, was conducted using the R package clusterProfiler[41]. Differentially expressed genes (DEGs) were selected based on two criteria, namely, an absolute value of $\log_2$(fold change) $>1$ and a corresponding FDR value $< 0.05$. Pearson correlation analysis was conducted using the built-in cor function in R v4.21 software, with the method parameter set to "pearson".

**Construction of weighted gene co-expression network.** Weighted Gene Co-expression Network Analysis (WGCNA) was performed to identify modules of highly correlated genes and to summarize their interactions, with the aim of uncovering regulatory mechanisms involved in cell growth, astaxanthin accumulation, and secondary cell wall biosynthesis[42]. The "WGCNA" package (version 1.72–5) in R software was used to construct and visualize the network. All samples were included, and the TPM values of all genes were utilized to generate gene co-expression networks. Firstly, the hclust function from the WGCNA package was employed for sample clustering to detect any significant outliers. Next, the pickSoftThreshold function was applied to select candidate power values ranging from 1 to 20. The appropriate soft-thresholding power (β) was chosen based on the scatter plot of the fit index versus the power value. A regulatory network of all genes was constructed to conform to a scale-free distribution based on connectivity, and an adjacency matrix was generated. This matrix was then transformed into a topological overlap matrix (TOM). Subsequently, the TOM-based dissimilarity metric was used to group genes with similar expression patterns into gene modules via average linkage hierarchical clustering with the following parameters: deepSplit = 2, minModuleSize = 30, and mergeCutHeight = 0.25. The resulting gene co-expression network was visualized as a heatmap based on the TOM dissimilarity, accompanied by a hierarchical clustering dendrogram using the TOMplot function. Module correlations were analyzed by extracting module eigengenes (MEs) from the network and calculating inter-module correlations using the cor function in R software. Additionally, the number of genes in each module was counted and visualized using a bar plot, generated with ggplot2 (version 4.3.3), and the module correlation heatmap was generated with corrplot (version 0.95).

**Gene function annotation and Transcription factors identification.** The reference genome of *H. pluvialis* (GenBank accession: GCA_011766145.1) contains many genes annotated as hypothetical or uncharacterized. To address this, a more comprehensive annotation of gene functions was conducted. Gene functions were assigned based on the best hits to proteins annotated in evolutionary genealogy of genes: Non-supervised Orthologous Groups database (the eggNOG[43] v5.0, updated on 2022-07), Non-Redundant Protein database (NR[44], updated on 2024-02), Swiss Protein database (Swiss-Prot[45], updated on 2024-04), and Translated European Molecular Biology Laboratory database (TrEMBL[46], updated on 2024-04), using Diamond Blast[47] to search these databases with the target proteins as query sequences (cut-off E-value $\leq$ 1e-5), identifying homologous matches between the hypothetical proteins and known proteins, with the best match selected as the final annotation. Motifs and domains were annotated using InterProScan[48] (version 5.67) by searching against publicly available InterPro databases[49], including Pfam[50], PRINTS[51], PROSITE[52], ProDom[53], and SMART[54]. Gene Ontology (GO) information was retrieved from InterPro. GhostKOALA[55] was used to map the genes of *H. pluvialis* to KEGG pathways by querying the Kyoto Encyclopedia of Genes and Genomes database (KEGG, updated on 2024-04) and identify the best hit for each node. Transcription factors (TFs) were annotated using the iTAK[56] and PlantTFDB[57] websites.

## Data Records

The RNA-seq clean data and gene expression levels for the 96 samples analysed in this study have been deposited in the NCBI Sequence Read Archive (SRA) under the project accession number PRJNA1138290[58] and the Gene Expression Omnibus (GEO) repository under the accession number of GSE289633[59], respectively. Furthermore, supplementary materials, including the sample information table (**Table S1 Sample information**), gene expression level, the gene co-expression network, and gene functional annotations (**Table S2 Gene expression and function**), differential gene expression analysis (**Table S3 DEGs**), KEGG pathway enrichment analyses of genes in each modules (**Table S4 module enrichment**), correlation coefficient and p-value in module correlation analysis (**Table S5 module correlation**), KEGG pathway classification of the 8,600 newly annotated genes (**Table S6 New KEGG classification**), GO annotation classification of the 8,600 newly annotated genes (**Table S7 New**
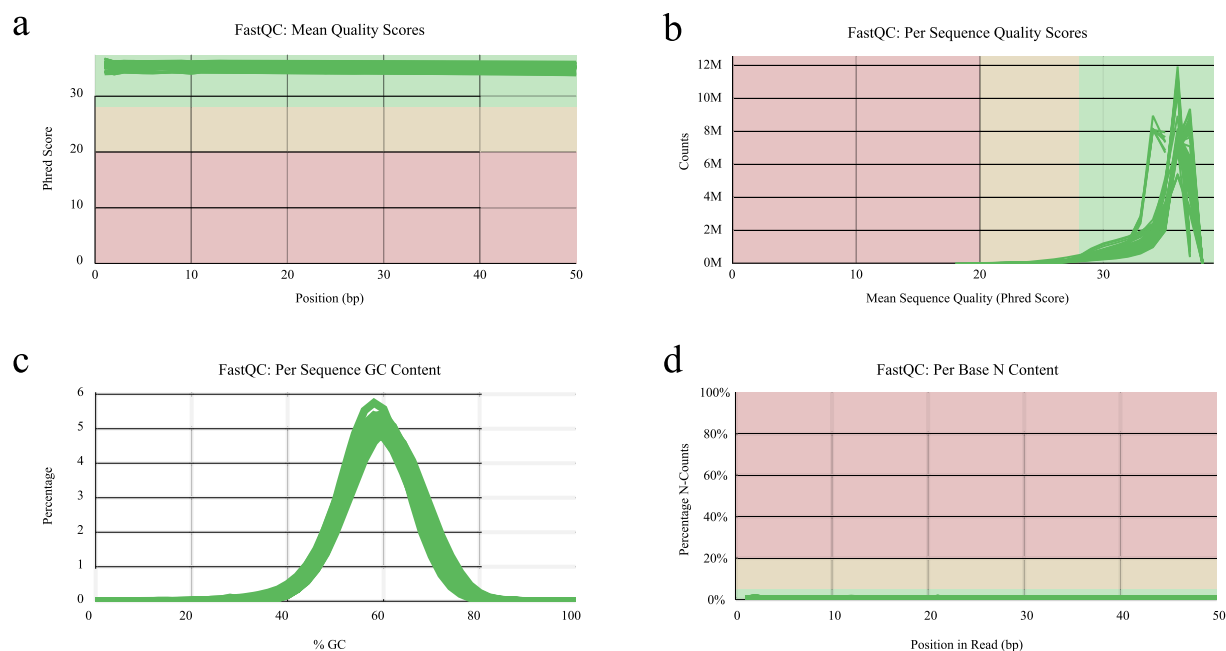
**Fig. 2** Sequence quality assessment metrics of filtered and trimmed RNA-Seq data using MultiQC. (**a**) Mean quality scores. (**b**) Per sequence quality scores. (**c**) Per sequence GC content. (**d**) Per base N content.

**GO classification**), predicted transcription factors (**Table S8 TFs**), and R code for constructing the WGCNA co-expression network (**R for WGCNA**), are accessible via the Figshare data management platform[60].

## Technical Validation

**Illumina RNA-seq quality validation.** Quality assessment of the raw reads from Illumina RNA-seq data for each sample was conducted using FastQC, which encompassed evaluation of mean quality scores (Fig. 2a), per sequence quality scores (Fig. 2b), per sequence GC content (Fig. 2c), and per base N content (Fig. 2d), followed by integration the results of 96 samples with MultiQC. Clean reads meet the following criteria to be retained for subsequent analysis: both the base quality score and the sequence quality scores exceed 30, and there should be no 'N' base in the sequence. Post the quality assessment procedures, a total of 96.7 gigabases (GB) of clean reads were yielded from these samples. The GC content of the samples exhibited a normal distribution ranging from 57% to 60% (Fig. 2c). The read quality of the samples indicated that the RNA-seq reads in this study were of high quality and devoid of contamination. Clean reads from the 96 samples were aligned to the reference genome using HISAT2, with an average mapping rate of 63.31% (Table 1). Gene expression levels were quantified based on the reads mapped to the reference genome. To mitigate the effects of gene length and data size on the calculated gene expression levels, the Transcripts Per Million (TPM) method was employed as a standardized approach for estimating gene expression levels. The distribution of normalized TPM values for all genes in each sample is depicted in a boxplot (Fig. 3a). The rectangles in the boxplot represent the interquartile range (IQR), capturing the central 50% of gene expression values, which primarily lie between 1.78 to 21.4 TPM. The median, marked by a bold line within each box, is positioned towards the lower end of the box, indicating a general trend of low gene expression. All samples exhibit a similar distribution, with the except of HL-Day3-Xan-2, which appears as an outlier. This consistency in low expression range, along with the alignment of IQRs across samples, suggests that gene expression distributions are comparable across different samples and that sequencing depth is consistent. The reliability of the RNA-seq data across the 96 samples was also assessed using Pearson correlation analysis (Fig. 3b). The correlation analysis shown that the correlation coefficients between different biological replicates of the same group approximately 0.950, indicating excellent sample reproducibility. The average correlation coefficient between samples under high light conditions was 0.758 and that under low light conditions was 0.841, while the average correlation coefficients between high-light and low-light samples was only 0.558. This result indicated that although the addition of different chemical reagents had a significant impact on the transcriptome of *H. pluvialis*, the effect of light intensity on the transcriptome of *H. pluvialis* was the most pronounced. The results from the above mentioned analytical methods demonstrated that the biological replicates in this study are highly consistent, indicating that the data obtained in this study could be used for subsequent research.

**Transcriptome data analysis.** Pearson's correlation coefficient was employed to cluster the samples and generate a sample clustering tree. By setting the height parameter to 1,500,000, one outlier (HL-Day3-Xan-2) was identified and subsequently removed, while the remaining samples were retained for further analysis (Fig. 4a). Using the pickSoftThreshold function, an optimal soft threshold of 7 was determined to ensure that the scale-free topology index ($R^2$) exceeded 0.9 and the mean connectivity was less than 100 (Fig. 4b), indicating that the network exhibits a high degree of randomness. To verify whether the network constructed at this soft threshold met

| Group | Clean reads | Total mapped reads | Mapping rate (%) | Uniquely mapped ratio (%) | Multi mapped ratio (%) |
|---|---|---|---|---|---|
| LL-Plateau | 21,712,009 | 14,107,131 | 64.97% | 50.07% | 14.90% |
| LL-Exponential | 21,274,334 | 14,582,793 | 68.55% | 51.41% | 17.14% |
| LL-Exponential-NaAc | 21,199,530 | 14,295,510 | 67.43% | 50.95% | 16.48% |
| HL-Day1 | 21,741,5 46 | 14,219,016 | 65.41% | 50.62% | 14.80% |
| HL-Day2 | 21,571,323 | 13,677,234 | 63.40% | 48.77% | 14.64% |
| HL-60h | 22,816,936 | 13,412,442 | 58.67% | 45.30% | 13.37% |
| HL-Day3 | 21,722,016 | 13,642,382 | 62.86% | 48.29% | 14.52% |
| HL-Day5 | 21,695,676 | 13,842,286 | 63.80% | 49.48% | 14.33% |
| HL-Day10 | 22,220,335 | 14,228,627 | 64.04% | 49.84% | 14.20% |
| HL-Day1-1mM-KI | 21,797,321 | 14,058,327 | 64.50% | 50.85% | 13.65% |
| HL-Day5-1mM-KI | 21,790,772 | 14,101,501 | 64.71% | 50.76% | 13.95% |
| HL-Day10-1mM-KI | 21,798,272 | 14,179,204 | 65.05% | 50.76% | 14.29% |
| HL-Day1-ND | 21,553,344 | 13,808,073 | 64.07% | 49.68% | 14.39% |
| HL-Day5-ND | 21,555,210 | 13,885,533 | 64.42% | 49.80% | 14.62% |
| HL-Day1-NR | 21,552,977 | 14,114,746 | 65.49% | 50.51% | 14.97% |
| HL-Day5-NR | 21,066,896 | 12,989,057 | 61.67% | 47.99% | 13.67% |
| HL-Day1-1mM-FDP | 21,622,648 | 14,804,140 | 68.47% | 51.74% | 16.73% |
| HL-Day5-1mM-FDP | 21,619,961 | 14,689,178 | 67.94% | 52.05% | 15.89% |
| HL-Day1-5mM-FDP | 21,594,611 | 14,747,192 | 68.29% | 51.20% | 17.10% |
| HL-Day5-5mM-FDP | 21,618,925 | 14,368,473 | 66.46% | 51.19% | 15.27% |
| HL-Day2-NaCl-CMP | 21,574,148 | 14,080,241 | 65.27% | 50.07 | 15.20% |
| HL-60h-Mutant | 22,304,121 | 13,065,487 | 58.36% | 45.22% | 13.15% |
| HL-Day3-Xan | 21,370,474 | 14,101,627 | 66.60% | 50.24% | 16.36% |
| HL-Day3-Tau | 21,630,082 | 13,529,582 | 62.55% | 47.9 | 14.65% |
| HL-Day5-Uri | 21,600,247 | 13,470,299 | 62.36% | 48.26% | 14.10% |
| HL-Day5-Gua | 21,552,594 | 13,562,964 | 62.93% | 48.73% | 14.20% |

**Table 1.** Summary of RNA-sequencing.

the scale-free topology, the softConnectivity function was used for validation. The $\log_{10}(k)$ of the number of connections for genes and the $\log_{10}(p(k))$ of the occurrence probability of such nodes showed a negative correlation, with $R^2 = 0.88$ (Fig. 4c). The larger the $R^2$ value, the closer the network is to the characteristics of a scale-free network[61]. These results indicated that the selected soft threshold of 7 can be used to construct a scale-free network. Next, the dynamic tree-cutting algorithm was applied for module identification, grouping genes with similar expression patterns into distinct clusters (Fig. 4d). This analysis identified 39 modules, each represented by a different colour (Fig. 4e). The module with the fewest genes, firebrick3, contained 31 genes, while the largest module, bisque4, comprised 5,312 genes, and the average number of genes per module was 732. To gain insights into the main functions of genes in each module, KEGG pathway enrichment analysis on these genes were performed, and the results were presented in Supplementary Table S4. The interactive relationships among all modules and genes were visualized using a heatmap based on the Topological Overlap Matrix (TOM), which confirmed the robustness of the module delineation. The analysis revealed strong correlations within modules, demonstrating that the modules are interrelated rather than independent (Fig. 4f). Furthermore, the gene expression patterns within each module were visualized (Fig. 5a). Subsequently, the correlations between modules were analysed to reveal the connections among them. The dynamic expression patterns varied significantly across different modules. For instance, the β-carotene ketolase gene (*BKT3*), a rate limiting gene for astaxanthin biosynthesis[62], was assigned to the salmon4 module, and its expression pattern (red line in Fig. 5a) closely mirrored the overall expression trend of the salmon4 module (salmon4 line in Fig. 5a). The genes in salmon4 module were significantly enriched in 24 KEGG pathways (Supplementary Table S4, salmon4 sheet), including "pyruvate metabolism", "fatty acid biosynthesis", "carotenoid biosynthesis", "terpenoid backbone biosynthesis", which were involved in the biosynthesis of astaxanthin and astaxanthin fatty acid esters in *H. pluvialis*[20,63]. The genes in modules that are significantly and positively correlated with salmon4 module, such as coral module ($r = 0.61$, $p$-value $< 1.0e-10$), and ivory module ($r = 0.44$, $p$-value $< 8.9e-6$), may also be associated with astaxanthin biosynthesis in *H. pluvialis*.

Mannose-1-phosphate guanylyltransferase gene (*GMPP*), mannosyltransferase genes (*OCH1* and *MNNs*), and mannan synthase genes were involved in SCW biosynthesis in *H. pluvialis*, most of which were found in bisque4 module. The genes in bisque4 module were significantly enriched in nine pathways including "galactose metabolism", "mannose type O-glycan biosynthesis", "fructose and mannose metabolism", "fatty acid elongation", "other glycan degradation", and "glycosylphosphatidylinositol (GPI)-anchor biosynthesis", which were involved in SCW biosynthesis in *H. pluvialis*[13,23]. The genes in modules that significantly and positively correlated with bisque4 module, such as cyan module ($r = 0.85$, $p$-value $< 1.0e-10$), darkturquoise module ($r = 0.81$, $p$-value $< 1.0e-10$), and lightsteelblue module ($r = 0.71$, $p$-value $< 1.0e-10$), which may also be associated with SCW biosynthesis in *H. pluvialis*.

a



b
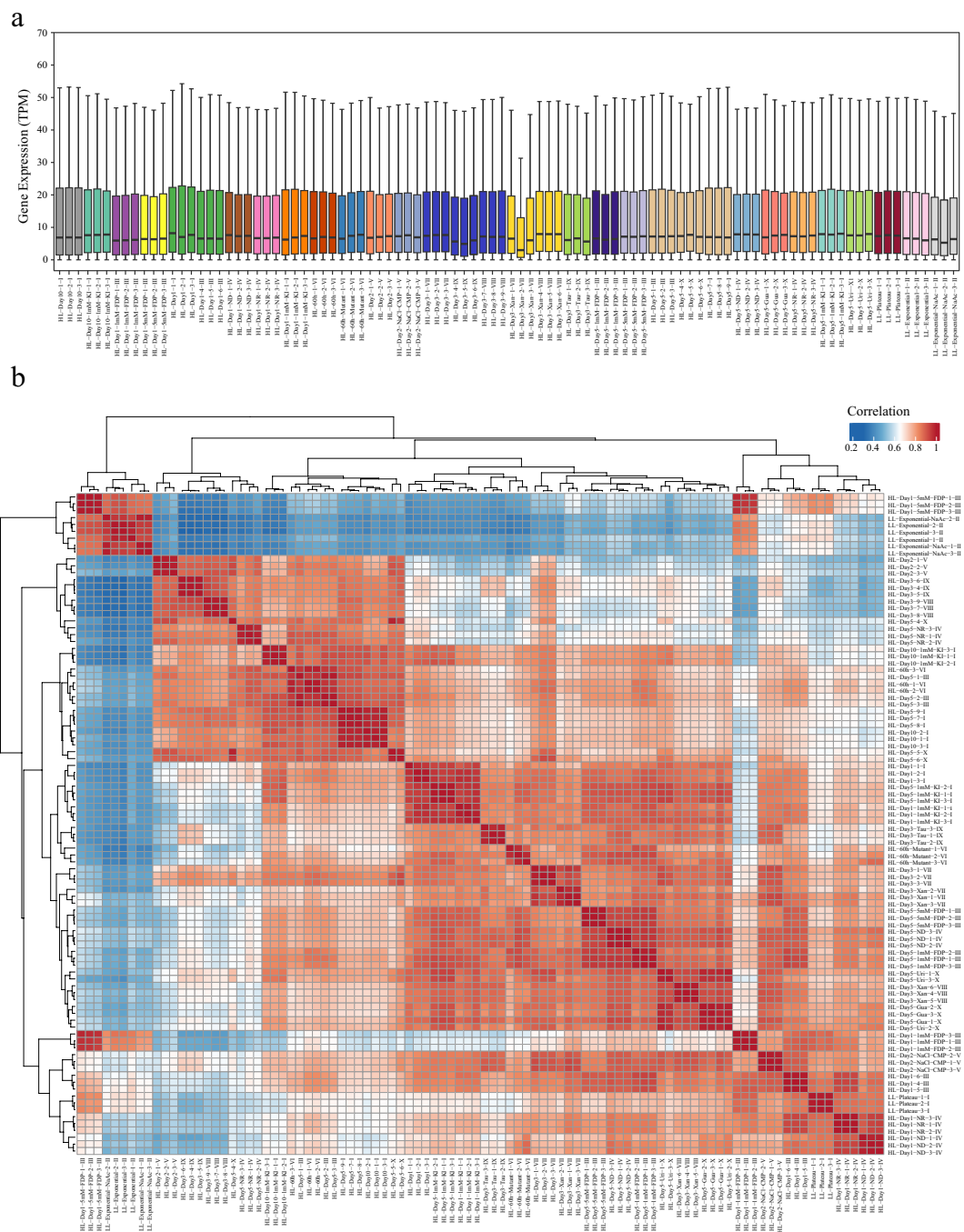


**Fig. 3** Summary of the sample clustering. (**a**) Gene expression levels of the 96 samples. (**b**) The person correlation between the 96 samples. The Roman numerals at the end of the sample name are the batch of the sample.

Under favourable environmental conditions, *H. pluvialis* exhibits rapid vegetative growth without astaxanthin accumulation or SCW synthesis. Conversely, environmental stress triggers growth cessation accompanied by simultaneous astaxanthin biosynthesis and SCW formation. However, our experimental data revealed a decoupling phenomenon between astaxanthin accumulation and SCW synthesis, as evidenced by multiple samples demonstrating astaxanthin production without concomitant SCW formation (Supplementary Table S1). This observation suggests that modules showing negative correlations with either the bisque4 or salmon4 modules may be directly associated with growth regulation. Specifically, the bisque4 module demonstrated significant negative correlations with seven modules (antiquewhite2, brown, coral2, greenyellow, mediumpurple1, mediumpurple3, and palevioletred2), while the salmon4 module showed negative correlations with four modules (coral2, grey60, mediumpurple1, and palevioletred2) (Supplementary Table S5 and Fig. 5b). Notably, the palevioletred2 and coral2 modules displayed expression patterns inversely correlated with both salmon4 and bisque4 modules. Functional enrichment analysis revealed that these two modules shared significant association with the "purine metabolism" pathway. This finding aligns with previous research demonstrating that enhanced
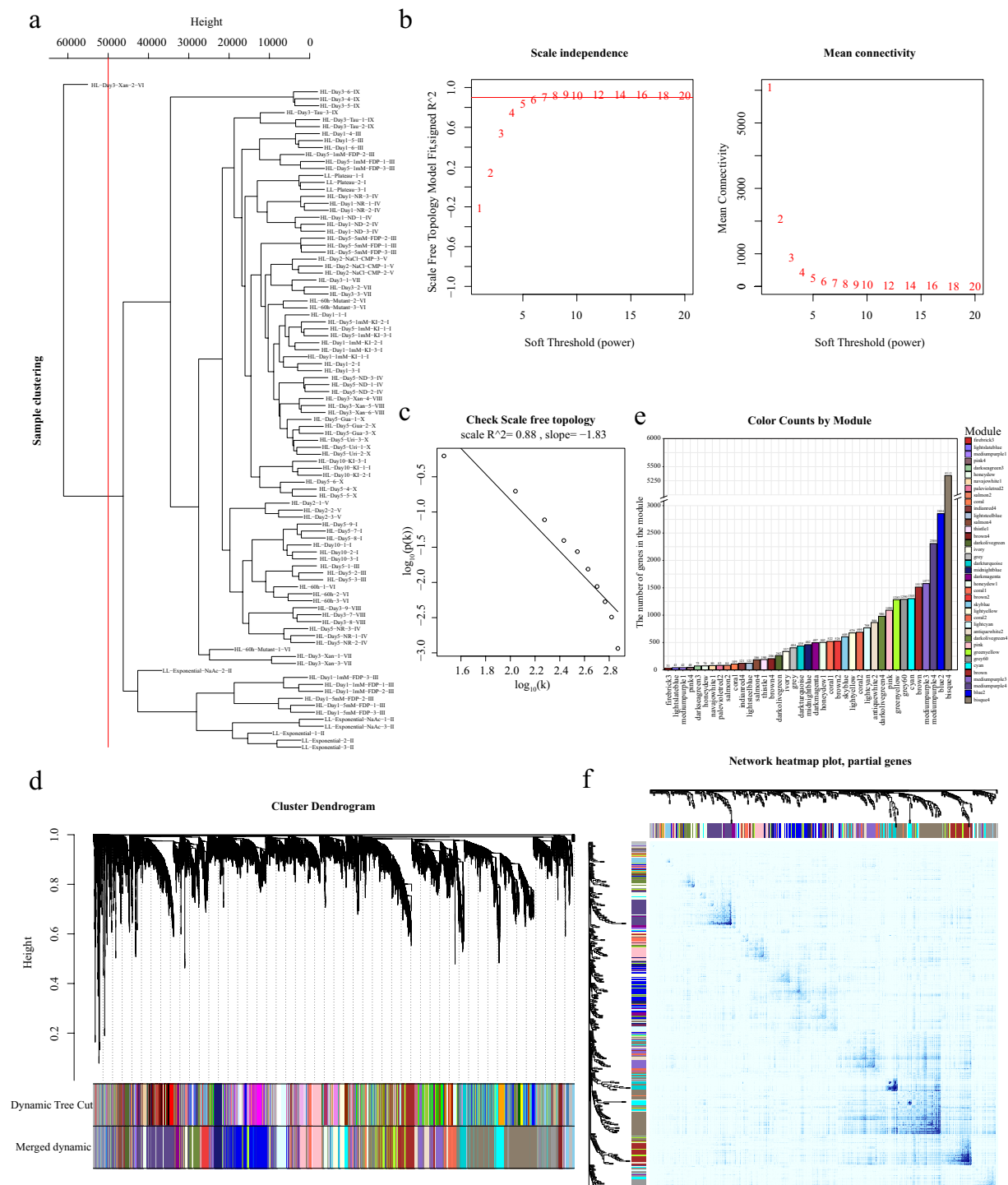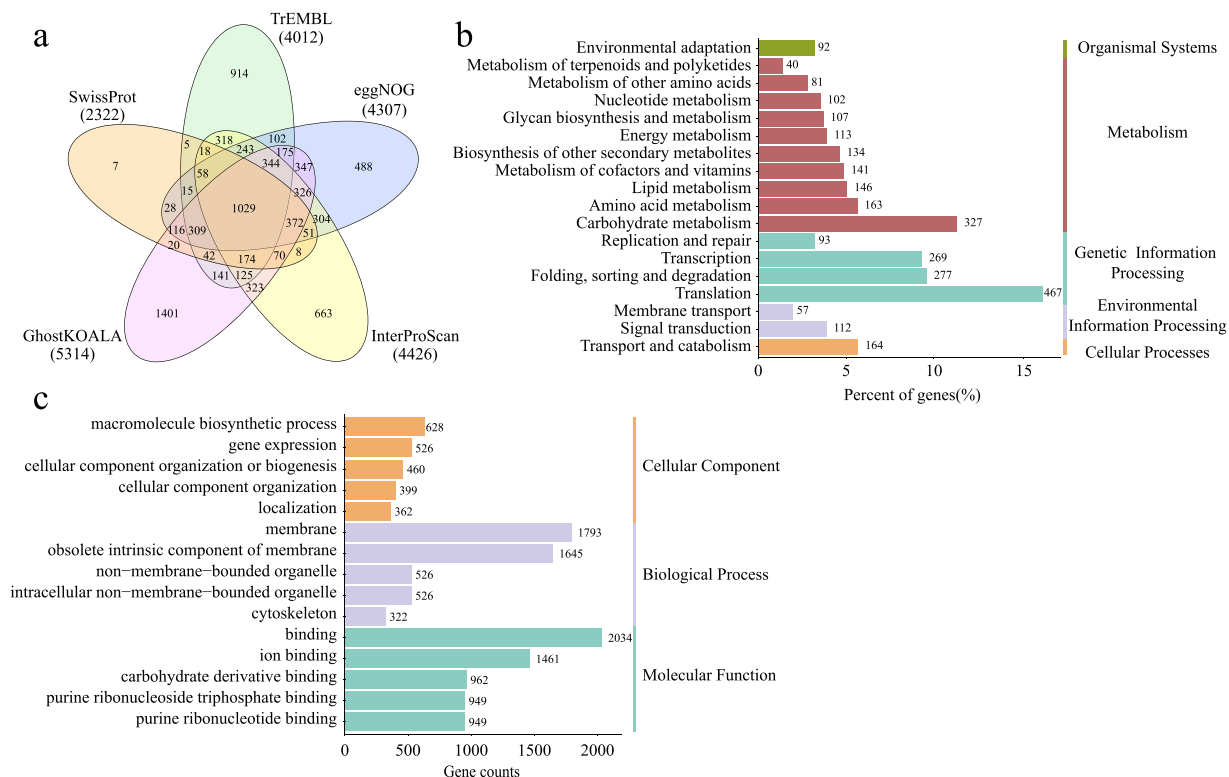
**Fig. 4** Results of weighted gene co-expression network analysis (WGCNA). (**a**) Clustering dendrogram of 96 samples. (**b**) Analysis of the scale-free index and mean connectivity for various soft-threshold powers (β). (**c**) Checking the scale free topology when β = 7; On this plot, the distribution approximately follows an approximately straight line, termed approximately scale-free topology. (**d**) Hierarchical cluster analysis was conducted to detect co-expression clusters with corresponding colour assignments. Each colour represents a module in the constructed gene co-expression network by WGCNA. (**e**) The number of genes in each module. (**f**) Heatmap depicts the Topological Overlap Matrix (TOM) of genes selected for weighted co-expression network analysis. Light colour represents low overlap and dark represents high overlap.

purine metabolism inhibits both SCW biosynthesis and astaxanthin accumulation while promoting cellular proliferation in *H. pluvialis* under high-light stress conditions[63]. These results highlight the functional specialization of distinct gene modules in regulating critical biological processes. The identified module-pathway relationships

**Fig. 5** Gene module expression and correlation analysis. (**a**) Visual module expression profile. The red line shows the expression level of β-carotene ketolase, a gene within the salmon4 module, under different treatments. The salmon4 line represents the overall expression trend of all genes in the salmon4 module across these treatments. Additionally, the palevioletred2 line represents a module with an expression trend opposite to that of the salmon4 module. (**b**) Heatmap of correlation coefficients and *p* values between modules. Blue colour indicates positive correlation, while red colour indicates negative correlation. The numbers represent the correlation values. Asterisks (*) denote statistically significant correlations ($p < 0.05$), with the number of asterisks indicating the level of significance.

provide mechanistic insights into the coordinated regulation of growth dynamics, astaxanthin biosynthesis, and SCW formation in *H. pluvialis*, establishing a foundation for further investigation of stress-responsive metabolic regulation in microalgae.

**Fig. 6** Functional annotation of the 8,598 newly annotated genes. (**a**) Venn diagram of gene annotation overlaps across for five major databases. (**b**) KEGG classification. (**c**) GO classification.

### Gene function annotation and Transcription factors identification.

In the annotation file of the reference genome utilized in this study (GenBank accession: GCA_011766145.1), 8,903 genes (31.48%) were annotated with specific functions, while 19,376 genes (68.52%) were not characterized. An extensive annotation analysis was conducted on previously uncharacterized genes by utilizing comparisons against public databases. As a result, a significant number of these genes were successfully annotated: 4,012 by TrEMBL, 4,306 by eggNOG, 2,322 by SwissProt, 5,314 by GhostKOALA, and 4,426 by InterProScan, totalling 8,598 genes (Fig. 6a). This comprehensive annotation has shed light on a substantial portion of the previously unknown gene functions. To gain a deeper insight into the roles of these newly annotated genes, they were categorized into KEGG pathways (Supplementary Table S6) and GO annotation classification (Supplementary Table S7). The KEGG pathway classification analysis revealed that "Metabolism" was the largest category (1,354 genes), followed by "Genetic Information Processing" (1,106 genes), "Environmental Information Processing" (169 genes), "Cellular Processes" (164 genes), and "Organic Systems" (92 genes) (Fig. 6b). In the second-level of classification, 18 subclass pathways were identified, including translation process, folding, sorting and degradation process, and carbohydrate metabolism (Fig. 6b). The GO classification analysis revealed these newly annotated genes were distributed in 101 GO terms of molecular function, 42 GO terms of cellular components, and 155 GO terms of biological process. The top 5 GO terms of each class with the highest number of genes were presented in Fig. 6c. In addition, 245 genes were annotated as transcription factors belonging to 45 transcription factor families. The main transcription factor families identified were GNAT (33 genes), SNF2 (28 genes), MYB-related (14 genes), SET (20 genes), C2H2 (13 genes), *etc*. and the details are provided in Supplementary Table S8. Gene functional annotation and the prediction of transcription factor are pivotal in elucidating the molecular mechanisms that govern cell growth, astaxanthin accumulation, and the biosynthesis of SCW in *H. pluvialis*.

### Code availability

The software used for quality and expression analyses of the dataset in this study is described in detail in the Methods section, and we adhered to all published bioinformatics tool manuals and protocols for executing all software and pipelines. In instances where specific parameters for the software were not detailed, default settings, as recommended by the developers, were employed. The R code used in this study is provided at figshare.

## References

1. Ariyadasa, T. U., Thevarajah, B., Anthonio, R. A. D. P., Nimarshana, P. H. V. & Wasath, W. A. J. From present to prosperity: assessing the current status and envisioning opportunities in the industrial-scale cultivation of *Haematococcus pluvialis* for astaxanthin production. *Phytochem. Rev.* **23**, 749–779 (2024).
2. Roach, T. *et al.* Acquisition of desiccation tolerance in *Haematococcus pluvialis* requires photosynthesis and coincides with lipid and astaxanthin accumulation. *Algal Res.* **64**, 102699 (2022).
3. Oslan, S. N. H. *et al. Haematococcus pluvialis* as a potential source of astaxanthin with diverse applications in industrial sectors: current research and future directions. *Molecules* **26**, 6470 (2021).
4. Kalita, N., Bhattacharjee, S., Baruah, D., Nath, S. & Baruah, P. P. Industrial and Biotechnological Applications of Algae: algal astaxanthin: genetics, synthesis, and advanced biotechnological strategies for therapeutic uses (Singapore: Springer Nature Singapore, 2025).
5. Aditi, Bhardwaj, R., Yadav, A., Swapnil, P. & Meena, M. Characterization of microalgal β-carotene and astaxanthin: exploring their health-promoting properties under the effect of salinity and light intensity. *Biotechnol. biofuels bioprod.* **18**, 1–21 (2025).
6. Long, X. W., Wu, X. G., Zhao, L., Liu, J. G. & Cheng, Y. X. Effects of dietary supplementation with *Haematococcus pluvialis* cell powder on coloration, ovarian development and antioxidation capacity of adult female Chinese mitten crab, *Eriocheir sinensis*. *Aquaculture* **473**, 545–553 (2017).
7. Xie, S. *et al.* Dietary supplementation of *Haematococcus pluvialis* improved the immune capacity and low salinity tolerance ability of post-larval white shrimp, *Litopenaeus vannamei*. *Fish Shellfish Immunol.* **80**, 452–457 (2018).
8. Ma, N. *et al.* Defatted *Haematococcus pluvialis* meal can enhance the coloration of adult Chinese mitten crab *Eriocheir sinensis*. *Aquaculture.* **510**, 371–379 (2019).
9. Long, X. W., Wang, L., Li, Y. P., Sun, W. H. & Wu, X. G. Effects of long-term *Haematococcus pluvialis* astaxanthin feeding on the growth, coloration, and antioxidant capacity of commercial-sized *Oncorhynchus mykiss*. *Aquac. Rep.* **30**, 101603 (2023).
10. Le-Feuvre, R. *et al.* Biotechnology applied to *Haematococcus pluvialis* Fotow: challenges and prospects for the enhancement of astaxanthin accumulation. *J. Appl. Phycol.* **32**, 3831–3852 (2020).
11. Patel, A. K. *et al.* Recent advancements in astaxanthin production from microalgae: A review. *Bioresour. Technol.* **364**, 128030 (2022).
12. Yu, X. L., Ye, X., Hu, C. Y., Xu, N. J. & Sun, X. Sodium acetate can promote the growth and astaxanthin accumulation in the unicellular green alga *Haematococcus pluvialis* as revealed by a proteomics approac. *J. Oceanol. Limnol.* **40**, 2052–2067 (2022).
13. Zhang, L. Q. *et al.* Enhancement of astaxanthin production, recovery, and bio-accessibility in *Haematococcus pluvialis* through taurine-mediated inhibition of secondary cell wall formation under high light conditions. *Bioresour. Technol.* **389**, 129802 (2023).
14. Christoph, H., Stefan, S. & Wolfram, B. Ultrastructural and chemical changes in the cell wall of *Haematococcus pluvialis* (Volvocales, Chlorophyta) during aplanospore formatio. *Eur. J. Phycol.* **37**, 217–226 (2011).
15. Boussiba, S. Carotenogenesis in the green alga *Haematococcus pluvialis*: Cellular physiology and stress response. *Physiol. Plant.* **108**, 111–117 (2010).
16. Park, J. C., Choi, S. P., Hong, M. E. & Sim, S. J. Enhanced astaxanthin production from microalga *Haematococcus pluvialis* by two-stage perfusion culture with stepwise light irradiation. *Bioprocess Biosyst Eng.* **37**, 2039–2047 (2014).
17. Hu, C. Y., Cui, D. D., Sun, X., Shi, J. X. & Xu, N. J. Primary metabolism is associated with the astaxanthin biosynthesis in the green alga *Haematococcus pluvialis* under light stress. *Algal Res.* **46**, 101768 (2020).
18. Ren, Y. Y., Sun, H., Deng, J. Q., Huang, J. C. & Chen, F. Carotenoid production from microalgae: biosynthesis, salinity responses and novel biotechnologies. *Mar Drugs.* **19**, 713 (2021).
19. Imamoglu, E., Fazilet, V. & Dalay, M. Effect of different culture media and light intensities on growth of *Haematococcus pluvialis*. *Int. J. Nat. Eng. Sci.* **1**, 5–9 (2007).
20. Du, F. C., Hu, C. Y., Sun, X., Zhang, L. Q. & Xu, N. J. Transcriptome analysis reveals pathways responsible for the promoting effect of sucrose on astaxanthin accumulation in *Haematococcus pluvialis* under high light condition. *Aquaculture.* **530**, 735757 (2021).
21. Cui, D. D. *et al.* Comparative transcriptome analysis unveils mechanisms underlying the promoting effect of potassium iodide on astaxanthin accumulation in *Haematococcus pluvialis* under high light stress. *Aquaculture.* **525**, 735279 (2020).
22. Du, F. C., Hu, C. Y., Sun, X., Zhang, L. Q. & Xu, N. J. Transcriptome analysis reveals the promoting effect of trisodium citrate on astaxanthin accumulation in *Haematococcus pluvialis* under high light condition. *Aquaculture.* **543**, 736978 (2021).
23. Xing, H. L. *et al.* The combination of uridine and nitrogen-deprivation promotes the efficient formation of astaxanthin-rich motile cells in *Haematococcus pluvialis*. *Bioresour. Technol.* **393**, 130150 (2024).
24. Li, Q. Q., You, J. K., Qiao, T. S., Zhong, D. B. & Yu, X. Y. Sodium chloride stimulates the biomass and astaxanthin production by *Haematococcus pluvialis* via a two-stage cultivation strategy. *Bioresour. Technol.* **344**, 126214 (2022).
25. Damiani, M. C., Leonardi, P. I., Pieroni, O. I. & Cáceres, E. J. Ultrastructure of the cyst wall of *Haematococcus pluvialis* (Chlorophyceae): Wall development and behaviour during cyst germination. *Phycologia.* **45**, 616–623 (2006).
26. Praveenkumar, R., Lee, K., Lee, J. & Oh, Y. K. Breaking dormancy: an energy-efficient means of recovering astaxanthin from microalgae. *Green Chem.* **17**, 1226–1234 (2015).
27. Irshad, M., Myint, A. A., Hong, M. E., Kim, J. & Sim, S. J. One-Pot, Simultaneous Cell Wall Disruption and Complete Extraction of Astaxanthin from *Haematococcus pluvialis* at Room Temperature. *ACS Sustainable Chem. Eng.* **7**, 13898–13910 (2019).
28. Li, D. *et al.* An evaluation of RNA-seq differential analysis methods. *PLOS ONE.* **17**, e264246 (2022).
29. Tyagi, P., Singh, D., Mathur, S., Singh, A. & Ranjan, R. Upcoming progress of transcriptomics studies on plants: An overview. *Front. Plant Sci.* **13**, 1030890 (2022).
30. Li, Y. G. *et al.* A new approach to promote astaxanthin accumulation via Na$_2$WO$_4$ in *Haematococcus pluvialis*. *J. Oceanol. Limnol.* **37**, 176–185 (2019).
31. Hu, C. Y. *et al.* Transcriptomic analysis unveils survival strategies of autotrophic *Haematococcus pluvialis* against high light stress. *Aquaculture.* **513**, 734430 (2019).
32. Chen, S. F., Zhou, Y. Q., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**, i884–i890 (2018).
33. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* **32**, 3047–3048 (2016).
34. Brandine, G. D. S. & Smith, A. D. Falco: high-speed FastQC emulation for quality control of sequencing data. *F1000Res.* **8**, 1874 (2019).
35. Morimoto, D., Yoshida, T. & Sawayama, S. Draft Genome Sequence of the Astaxanthin-Producing Microalga *Haematococcus lacustris* Strain NIES-144. *Microbiol. Resour. Announc.* **9**, 10–1128 (2020).
36. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25**, 2078–2079 (2009).
38. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* **30**, 923–930 (2014).
39. Düz, E. & Çakır, T. Effect of RNA-Seq data normalization on protein interactome mapping for Alzheimer' s disease. *Comput. Biol. Chem.* **109**, 108028 (2024).
40. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
41. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* **16**, 284–287 (2012).

42. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* **9**, 559 (2008).
43. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. EggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
44. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
45. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **27**, 49–54 (1999).
46. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
47. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.* **12**, 59–60 (2015).
48. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics.* **30**, 1236–1240 (2014).
49. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
50. Mistry, J. & Finn, R. Pfam: a domain-centric method for analyzing proteins and proteomes. *Methods Mol. Biol.* **396**, 43–58 (2007).
51. Attwood, T. K., Beck, M. E., Bleasby, A. J. & Parry-Smith, D. J. PRINTS–a database of protein motif fingerprints. *Nucleic Acids Res.* **22**, 3590–3596 (1994).
52. Hulo, N. *et al.* The PROSITE database. *Nucleic Acids Res.* **34**, D227–D230 (2006).
53. Bru, C. *et al.* The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.* **33**, D212–D215 (2005).
54. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA.* **95**, 5857–5864 (1998).
55. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
56. Zheng, Y. *et al.* ITAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Mol. Plant.* **9**, 1667–1670 (2016).
57. Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP521178 (2025).
59. *NCBI Gene Expression Omnibus* https://identifiers.org/geo/GSE289633 (2025).
60. Hu, B. *et al.* Transcriptome profiles and co-expression network analysis of 96 *Haematococcus pluvialis* samples. *figshare* https://doi.org/10.6084/m9.figshare.26409430.v2 (2025).
61. Tian, Z. L. *et al.* Identification of important modules and biomarkers in breast cancer based on WGCNA. *Onco Targets Ther.* **13**, 6805–6817 (2020).
62. Luo, Q. L. *et al.* Genome and transcriptome sequencing of the astaxanthin-producing green microalga, *Haematococcus pluvialis*. *Genome Biology and Evolution.* **11**, 166–173 (2019).
63. Liu, K. *et al.* Xanthosine-mediated modulation of *Haematococcus pluvialis* growth and stress response under high light: Insights from physiological and transcriptome analysis. *Aquaculture.* **599**, 742156 (2025).

## Acknowledgements

## Author contributions

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05533-4.

**Correspondence** and requests for materials should be addressed to N.X. or C.H.

**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.