phycology

MDPI

Article

Comparative Genomics of Stress-Associated Gene Family Copy Number Variation in Chlorophyte Microalgae

Prabhaharan Renganathan

Department of Bioecology and Biological Education, M. Akmullah Bashkir State Pedagogical University, 450000 Ufa, Russia; prabhaharan06@gmail.com

Abstract

Abiotic stresses severely limit agricultural productivity and have stimulated growing interest in microalgae as sources of stress-resilient traits and biostimulants. However, comparative genomic assessments integrating multiple stress-associated gene families across chlorophyte genome assemblies are limited. Here, we analyzed 19 chlorophyte genome assemblies to investigate the distribution and copy-number variation of 11 gene families associated with antioxidant defense, osmoprotection, and carotenoid biosynthesis. Candidate genes were identified using standardized InterPro annotations and manually curated to ensure consistent gene copy-number estimation. Hierarchical clustering, principal component analysis (PCA), descriptive statistics, and Pearson's correlation analysis were performed to characterize gene-family copy-number patterns and multivariate similarities among genome assemblies. Multiple tests in the correlation analysis were controlled using the Benjamini–Hochberg false discovery rate procedure. The total functional family assignments ranged from 28 to 51 across the analyzed genome assemblies. Thioredoxin (*TRX*) exhibited the highest mean copy number (17.26 copies per genome) and the lowest coefficient of variation (15.43%), whereas catalase (*CAT*) showed the greatest variability (CV = 59.28%). The first two principal components explained 50.18% of the total variation, with PC1 accounting for 30.86% and PC2 for 19.32%, and differentiated genome assemblies according to their stress-associated gene copy-number profiles. Several moderate-to-strong pairwise correlations were observed, but none were significant after FDR correction. Overall, this study provides a curated comparative genomic framework for identifying stress-associated gene-family copy-number patterns across chlorophyte genome assemblies and generating testable hypotheses for subsequent functional studies.

Keywords: abiotic stress; antioxidant defense; osmoprotection; carotenoid biosynthesis; thioredoxin; reactive oxygen species; InterPro; manual curation; plant biostimulants; stress tolerance

Check for updates

Academic Editor: Jiangxin Wang

Received: 21 July 2026
Revised: 19 August 2026
Accepted: 31 August 2026
Published: 2 September 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Climate change is intensifying abiotic stresses, such as drought, salinity, and heat, posing major challenges to agricultural productivity and global food security [1]. These stresses disrupt cellular homeostasis by promoting reactive oxygen species (ROS) accumulation, impairing photosynthetic efficiency, and disturbing osmotic balance, leading to oxidative damage, metabolic dysfunction, and growth inhibition [2]. Therefore, improving plant resilience to abiotic stress has become a major objective in sustainable agriculture and climate-adaptive crop management [3]. In this context, microalgae have attracted increasing attention because of their ability to tolerate diverse environmental conditions

and their capacity to synthesize a wide range of bioactive compounds, including antioxidants, compatible solutes, carotenoids, and phytohormones [4–7]. These characteristics have stimulated growing interest in the development of microalgae-based biostimulants as environmentally sustainable alternatives to conventional agrochemical inputs [8].

Stress responses in microalgae involve an integrated network of physiological, biochemical, and molecular stress response mechanisms that contribute to cellular homeostasis under adverse environmental conditions [2,3]. Antioxidant defense systems regulate intracellular ROS levels, osmoprotective pathways synthesize compatible solutes that preserve cellular integrity, and carotenoid biosynthesis protects the photosynthetic apparatus against photo-oxidative damage while contributing to light harvesting [9–12]. These interconnected pathways represent the major components of cellular responses to diverse abiotic stresses in photosynthetic organisms.

Substantial progress has been made in characterizing these stress-response mechanisms through physiological, transcriptomic, and functional studies, leading to the identification of numerous genes involved in regulating oxidative stress, osmotic adjustment, and carotenoid metabolism [13–15]. In addition, advances in genome sequencing have enabled comparative genomic studies investigating chlorophyte genome evolution, gene family variation, and metabolic diversification by exploiting the increasing availability of high-quality genome assemblies and standardized annotation resources [16,17]. However, these studies have primarily focused on genome evolution, phylogenetic relationships, or specific biological pathways, rather than providing a standardized cross-genome comparison of copy-number variation across multiple stress-associated functional modules in chlorophytes. Moreover, differences in annotation strategies and genome curation have limited direct comparisons between the genome datasets. Consequently, manually curated comparative assessments using consistent identification and counting criteria across multiple stress-associated gene families are limited.

Comparative genomics provides an efficient framework for examining the organization and copy number variation in stress-associated gene families by enabling systematic comparisons of gene content, copy number, and distribution across diverse genomes [18–20]. When combined with standardized genome annotation and consistent manual curation, comparative genomics facilitates reliable cross-species comparisons and improves the interpretation of gene family copy number patterns while minimizing methodological biases [20–22]. However, gene copy number does not directly indicate gene expression, protein abundance, biochemical activity, or physiological stress tolerance. Therefore, comparative copy-number patterns are most appropriately used to identify genomic variations and generate hypotheses for subsequent functional testing.

This study focused on 11 gene families representing three stress-associated functional modules: antioxidant defense (*SOD*, *CAT*, *PRX*, and *TRX*), osmoprotection (*GPD1*, *TPS/OtsA*, and *OtsB/TPP*), and carotenoid biosynthesis (*crtB*, *crtP*, *crtQ*, and *crtL*). These families were selected because they represent well-characterized components of the three pathways and can be identified consistently using InterPro-supported annotations and manual curation across the analyzed genomes. The selected families do not represent the complete chlorophyte stress response repertoire, which also includes additional signaling, regulatory, metabolic, and phytohormone-associated pathways.

In this study, we performed a comparative genomic analysis of 19 chlorophyte genome assemblies to investigate the copy-number variation of 11 stress-associated gene families involved in antioxidant defense, osmoprotection, and carotenoid biosynthesis. Candidate genes were identified using standardized InterPro annotations and manually curated to ensure consistent gene copy-number estimation across all genomes. The curated dataset was analyzed using descriptive statistics, hierarchical clustering, principal component

analysis (PCA), and Pearson correlation analysis to characterize gene-family copy-number variation, multivariate similarities among genome-level profiles, and pairwise associations among gene families. Cluster robustness was further assessed by bootstrap resampling, and multiple testing in the correlation analysis was controlled for using the Benjamini–Hochberg false discovery rate procedure. This study provides a standardized comparative genomic framework for investigating stress-associated gene-family copy-number variation and a curated genomic resource for future functional, evolutionary, and phylogenetically informed studies, including the experimental evaluation of candidate microalgal lineages relevant to biostimulant research.

2. Materials and Methods

2.1. Selection of Chlorophyte Genomes

Representative genomes were selected to provide wide taxonomic coverage for the comparative analysis of stress-associated gene repertoires across the phylum Chlorophyta. Genome selection was based on the availability of publicly accessible genome assemblies with predicted protein datasets and comprehensive functional annotations. To achieve analytical consistency and minimize annotation bias, only genomes annotated using a standardized pipeline within a single genomic resource were included. Genome sequences, predicted protein datasets, and functional annotation files were obtained from the Joint Genome Institute (JGI) PhycoCosm platform [23]. All datasets were retrieved from JGI PhycoCosm v20 in March 2026, and the most recent genome assembly and annotation available at that time were used for all analyses.

A total of 19 chlorophyte genome assemblies representing 16 nominal taxa, diverse taxonomic groups, and ecological lifestyles were selected, including model organisms and taxa of ecological or biotechnological importance. Multiple genome assemblies were retained for some species to capture the available strain-level genomic variation; therefore, all statistical analyses were performed at the genome assembly level rather than treating the dataset as 19 independent species. Genome assemblies lacking complete predicted protein datasets or standardized InterPro functional annotations were excluded to maintain consistency in the comparative analyses. The final dataset was used to identify, manually curate, and compare the antioxidant defense, osmoprotection, and carotenoid biosynthesis gene families. Although several genomes were available as unpublished JGI PhycoCosm assemblies, all were annotated using the same standardized JGI pipeline, ensuring annotation consistency across datasets.

Independent genome completeness assessment using BUSCO or an equivalent metric was not applied as an additional inclusion criterion because the present analysis relied on genome assemblies and standardized annotations available through JGI PhycoCosm. Consequently, variations in assembly completeness may influence gene copy number estimates, particularly through the underrepresentation of genes located in incomplete or poorly assembled genomic regions. This limitation was considered when interpreting the differences in the gene copy number among the genome assemblies.

The scientific names, genome assembly versions, genome sizes, predicted protein-coding gene numbers, and genome references are listed in Table 1.

Table 1. Genome characteristics of the 19 chlorophyte genome assemblies included in this comparative genomic study. Genome assembly versions, assembly sizes (Mb), predicted protein-coding genes, and genome references were obtained from the Joint Genome Institute (JGI) PhycoCosm v20.

Species	Genome Assembly	Genome Size (Mb)	Predicted Protein-Coding Genes	Genome Reference
Chlamydomonas reinhardtii	CC-4532 v6.1	114.04	16,090	[24]
Chlorella sorokiniana	UTEX 1230	58.53	12,209	[25]
Dunaliella salina	CCAP19/18	343.70	18,801	[26]
Desmodesmus armatus	SE107 v1.0	116.30	17,675	[27]
Tetradesmus obliquus	UTEX 72 v1.1	107.16	16,985	JGI PhycoCosm unpublished genome assembly
Monoraphidium neglectum	SAG 48.87	69.71	16,755	[28]
Haematococcus lacustris	Liv1 v1.2	291.51	21,373	[29]
Coccomyxa subellipsoidea	C-169 v3.0	48.85	10,893	[30]
Micractinium conductrix	SAG 241.80	60.77	8739	[31]
Flechtneria rotunda	SEV3VF49	76.78	16,599	JGI PhycoCosm unpublished genome assembly
Enallax costatus	CCAP 276/31	127.56	19,308	JGI PhycoCosm unpublished genome assembly
Scenedesmus obliquus	SAG 2326	104.23	17,007	JGI PhycoCosm unpublished genome assembly
Scenedesmus dimorphus	UTEX 1237	102.57	17,373	JGI PhycoCosm unpublished genome assembly
Scenedesmus sp.	WC-1	102.32	17,012	JGI PhycoCosm unpublished genome assembly
S. obliquus	UTEX 3031	205.01	13,349	[32]
Tetradesmus deserticola	SNI-2	198.16	17,197	JGI PhycoCosm unpublished genome assembly
H. lacustris	NIES-144	171.79	28,279	[33]
Coccomyxa pringsheimii	SAG 216-7	50.91	9499	[34]
S. obliquus	UTEX 393 v2.0	107.86	19,017	[35]

2.2. Identification of Stress-Associated Gene Families

Eleven stress-associated gene families were selected based on their established roles in oxidative stress defense, osmotic stress adaptation, and carotenoid biosynthesis. These gene families were selected because they represent well-characterized components of the three selected stress-associated functional modules that are widely represented in chlorophyte genomes and are supported by consistent functional annotation in JGI PhycoCosm. Therefore, the analysis was intended as a targeted comparison of these selected modules rather than a comprehensive survey of all stress-response genes.

The selected gene families were grouped into three functional modules: antioxidant defense [superoxide dismutase (*SOD*), catalase (*CAT*), peroxiredoxin (*PRX*), and thioredoxin (*TRX*)], osmoprotection [glycerol-3-phosphate dehydrogenase (*GPD1*), trehalose-6-phosphate synthase (*TPS*/*OtsA*), and trehalose-6-phosphate phosphatase (*OtsB/TPP*)], and carotenoid biosynthesis [phytoene synthase (*crtB*), phytoene desaturase (*crtP*), ζ-carotene desaturase (*crtQ*), and lycopene cyclase (*crtL*)].

TPS and OtsA were treated as a single functional gene family (*TPS/OtsA*) because both annotations correspond to the trehalose-6-phosphate synthase function and, in the curated chlorophyte datasets, frequently referred to the same synthase domain within a TPS–TPP fusion protein. *OtsB/TPP* was retained as a separate functional family because it represents a distinct trehalose-6-phosphate phosphatase domain. Their biological functions are summarized in Table 2, and the corresponding InterPro identifiers used for candidate retrieval are provided in Table S1.

Table 2. Stress-associated gene families analyzed in this study, their corresponding functional modules, and primary biological roles.

Module	Gene Family	Full Name	Biological Function	Stress Role
Antioxidant	*SOD*	Superoxide dismutase	Detoxification of superoxide radicals	ROS scavenging
Antioxidant	*CAT*	Catalase	Hydrogen peroxide degradation	ROS scavenging
Antioxidant	*PRX*	Peroxiredoxin	Peroxide reduction	ROS scavenging
Antioxidant	*TRX*	Thioredoxin	Cellular redox regulation	Redox homeostasis
Osmoprotection	*GPD1*	Glycerol-3-phosphate dehydrogenase	Glycerol biosynthesis	Osmotic adjustment
Osmoprotection	*TPS*/*OtsA*	Trehalose-6-phosphate synthase	Trehalose biosynthesis	Osmoprotection
Osmoprotection	*OtsB*/*TPP*	Trehalose-6-phosphate phosphatase	Trehalose synthesis	Osmoprotection
Carotenoid	*crtB*	Phytoene synthase	Carotenoid biosynthesis	Photoprotection
Carotenoid	*crtP*	Phytoene desaturase	Carotenoid biosynthesis	Photoprotection
Carotenoid	*crtQ*	ζ-carotene desaturase	Carotenoid biosynthesis	Photoprotection
Carotenoid	*crtL*	Lycopene cyclase	Carotenoid biosynthesis	Photoprotection

Candidate genes were identified using functional annotation resources available through the JGI PhycoCosm platform. Annotated protein datasets and the corresponding InterPro annotation files were screened for proteins assigned to the target gene families. Predefined InterPro identifiers and associated functional descriptions were used as the primary criteria for candidate retrieval. When multiple InterPro entries corresponded to a single gene family, all relevant annotations were included during the initial screening

to maximize candidate identification. The resulting candidate gene sets were manually curated and verified, as described in Section 2.3.

2.3. Manual Curation of Candidate Genes

All gene families underwent standardized manual curation to maintain accurate and comparable gene copy number estimates across the 19 chlorophyte genome assemblies. Automated annotations were used only for preliminary candidate identification, whereas the final protein inclusion was determined through manual verification using predefined criteria. Each candidate protein was evaluated by examining its annotated functional description and the associated InterPro domain architecture. Proteins were retained only when both the annotated function and conserved domains were consistent with the target gene family, ensuring that gene identification was supported by functional evidence rather than by annotation alone.

Identical inclusion and exclusion criteria were applied to all the genome assemblies. Proteins were included only if they (i) matched the predefined InterPro identifiers for the target gene family, (ii) possessed a functional annotation consistent with the expected biological role, and (iii) contained conserved domains that supported the assigned function. Proteins that did not satisfy all of these criteria were excluded. Redundant entries arising from duplicate annotations, alternative gene models, or repeated records were identified during manual inspection and counted only once. Proteins originating from distinct genomic loci were retained as independent gene copies, whereas multiple annotations referring to the same protein or alternative gene models were considered redundant and counted only once. Candidate proteins associated with multiple InterPro entries were evaluated individually to distinguish redundant annotations from distinct functional domain assignments.

For multidomain or fusion proteins, the copy number was defined at the functional gene family level. Each distinct functional family with a confirmed diagnostic domain was counted once. Thus, the bifunctional TPS–TPP fusion protein could contribute one assignment to *TPS/OtsA* and one assignment to *OtsB/TPP* because the two domains represent distinct enzymatic functions. In contrast, TPS and OtsA were not counted separately when they referred to the same trehalose-6-phosphate synthase domain. Multiple InterPro hits to the same functional domain within a single protein did not increase the gene-family copy number.

Formal orthology relationships were not established in this study. Proteins were assigned to functional gene families based on annotation and conserved domain evidence, and distinct proteins from separate genomic loci were counted as independent copies. Consequently, multiple copies within a genome may include paralogs resulting from lineage-specific duplication, and similar domain architectures among genomes should not be interpreted as evidence of one-to-one orthologs.

Proteins with ambiguous functional assignments, incomplete annotations, or insufficient domain evidence were excluded to minimize false-positive assignments. Although the use of standardized InterPro annotations together with manual review of InterPro domain architecture improved consistency among genome assemblies, this approach may fail to recover highly divergent homologs lacking recognizable diagnostic domains or proteins affected by incomplete or inaccurate annotations. Therefore, the reported copy numbers should be interpreted as curated, annotation-supported functional gene-family assignments rather than an exhaustive inventory of all possible homologs.

The final functional gene family copy numbers were obtained using the counting rules described above and used to construct a comparative genomic matrix for downstream analyses. Detailed information on curated protein identifiers, supporting InterPro annotations,

proteins excluded during manual curation, and corresponding curation decisions, is provided in Table S1.

2.4. Construction of the Comparative Genomic Matrix

The verified gene copy numbers were compiled into a comparative genomic matrix. The final dataset comprised 19 chlorophyte genome assemblies and 11 stress-associated gene families, with each entry representing the curated copy number of a specific gene family in an individual genome. This matrix was used as the primary dataset for all subsequent analyses. The complete manually curated gene copy number matrix is presented in Table S2.

The 11 gene families were organized into three functional modules, including antioxidant defense, osmoprotection, and carotenoid biosynthesis, to facilitate gene family and module-level comparisons across the genome assemblies. Before statistical analyses, the matrix underwent quality control procedures to verify dataset integrity. Matrix dimensions, data types, missing values, duplicate genome assembly entries, duplicate gene family records, and consistency between curated gene counts and matrix entries were systematically examined. No inconsistencies were identified during these quality control procedures, confirming the accuracy and completeness of the final comparative genomic matrix. Because all gene counts originated from the standardized manual curation workflow described in Section 2.3, no additional data imputation or correction was required in this study.

2.5. Comparative Genomic and Statistical Analyses

Descriptive statistics, including the mean, median, minimum, maximum, range, standard deviation, and coefficient of variation (CV), were calculated for each gene family to summarize the copy number variation across the genome assemblies. The CV was used as a normalized measure of the relative variability among gene families (Table S3; Figure S1). Because gene families differed substantially in their mean copy numbers, the CV enabled a direct comparison of relative variability and copy number stability across gene families, independent of absolute copy number magnitude. Module-level summaries were calculated for the antioxidant defense, osmoprotection, and carotenoid biosynthesis gene sets by determining the total functional family copy number for each module in every genome assembly (Tables 3 and S4). For Table 3, module totals were first calculated separately for each of the 19 genome assemblies, and the mean, median, minimum, maximum, SD, and CV were calculated across these 19 genome-level module totals. Individual gene-family values were not pooled across modules. Functional-module composition (%) was calculated as the proportion of the total functional-family assignments represented by each module within each genome assembly.

Table 3. Summary statistics of genome-level copy-number totals for three functional modules across 19 chlorophyte genome assemblies. For each genome, the copy numbers were summed within the antioxidant defense, osmoprotection, and carotenoid biosynthesis modules before calculating the mean, median, minimum, maximum, standard deviation (SD), and coefficient of variation (CV) across the 19 genomes.

Module	Gene Families	Mean	Median	Minimum	Maximum	SD	CV (%)
Antioxidant	4	28.00	28.0	22	34	3.38	12.08
Osmoprotection	3	7.32	8.0	1	10	2.00	27.36
Carotenoid	4	7.63	8.0	5	9	1.30	17.03

Genome-wise totals of stress-associated gene copy numbers were summarized as total functional-family assignments across the 11 analyzed families for each genome assembly (Table S5). Because multidomain *TPS–TPP* proteins can contribute to both *TPS/OtsA* and *OtsB/TPP* functional assignments, these totals should not be interpreted as counts of unique protein loci.

A presence–absence matrix was generated by converting gene copy numbers into binary data, where 0 indicates gene absence and values ≥ 1 indicate gene presence (Figure S2). Pairwise relationships among gene families were evaluated using Pearson's correlation analysis based on the copy number values across the 19 genome assemblies. Two-sided p values were calculated for all 55 unique pairwise comparisons among the 11 gene families. To account for multiple tests, p values were adjusted using the Benjamini–Hochberg false discovery rate (BH-FDR) procedure, with $q < 0.05$ considered statistically significant. The Pearson correlation coefficients, raw p-values, and BH-FDR-adjusted p-values are presented in Table S6.

Hierarchical clustering was performed using Z-score-normalized gene copy numbers, Euclidean distance, and complete linkage to generate clustered heat maps of the genome assemblies and gene families. The stability of genome-level clusters was evaluated by bootstrap resampling of gene family variables with replacement for 1000 iterations. For each internal cluster in the original dendrogram, bootstrap support was calculated as the percentage of resampled datasets in which the same set of genome assemblies was recovered. Bootstrap support values are provided in Table S7, and the clustering was interpreted as exploratory because support varied among nodes.

PCA was performed using the centered and scaled (Z-score standardized) gene copy number matrix to examine the multivariate patterns of gene family copy number. PCA scores and loading plots were generated to visualize the similarities among genome-level copy number profiles and the contributions of individual gene families to the principal components. The first two principal components were used for visualization, and their explained variance was reported to indicate the proportion of the total variation represented in the two-dimensional PCA space.

2.6. Data Visualization

All statistical analyses and graphical visualizations were performed using Python (version 3.12) in the Jupyter Notebook v7 environment. Data processing was conducted using Pandas, numerical computations using NumPy, statistical analyses and hierarchical clustering using SciPy, principal component analysis using scikit-learn, and figure generation using Matplotlib 3.10.0 and Seaborn 0.13.2. Multiple testing correction was performed using the statsmodels implementation of the Benjamini–Hochberg FDR procedure. Bootstrap cluster support analysis was implemented in Python using SciPy hierarchical clustering with 1000 resampling iterations.

Figures were generated from the manually curated gene copy number matrix using standardized graphical parameters, including consistent font styles, color palettes, axis formatting, and image resolution with identical plotting settings applied across comparable figure types to ensure visual consistency and facilitate direct comparison between analyses.

3. Results

3.1. Comparative Genomics Workflow

The comparative genomics workflow integrated genome mining, standardized manual curation, comparative genomic analyses, and cautious biological interpretation to characterize stress-associated gene repertoires across 19 chlorophyte genome assemblies (Figure 1). The application of this workflow generated a curated comparative genomic

dataset comprising 19 chlorophyte genome assemblies and 11 manually verified stress-associated gene families, which served as the basis for subsequent analyses. The resulting curated dataset established the foundation for subsequent analyses of gene copy number variation, multivariate similarities among genome-level copy-number profiles, pairwise gene-family associations, and patterns of gene family copy number variation across the analyzed chlorophyte genome assemblies (Figure 1).

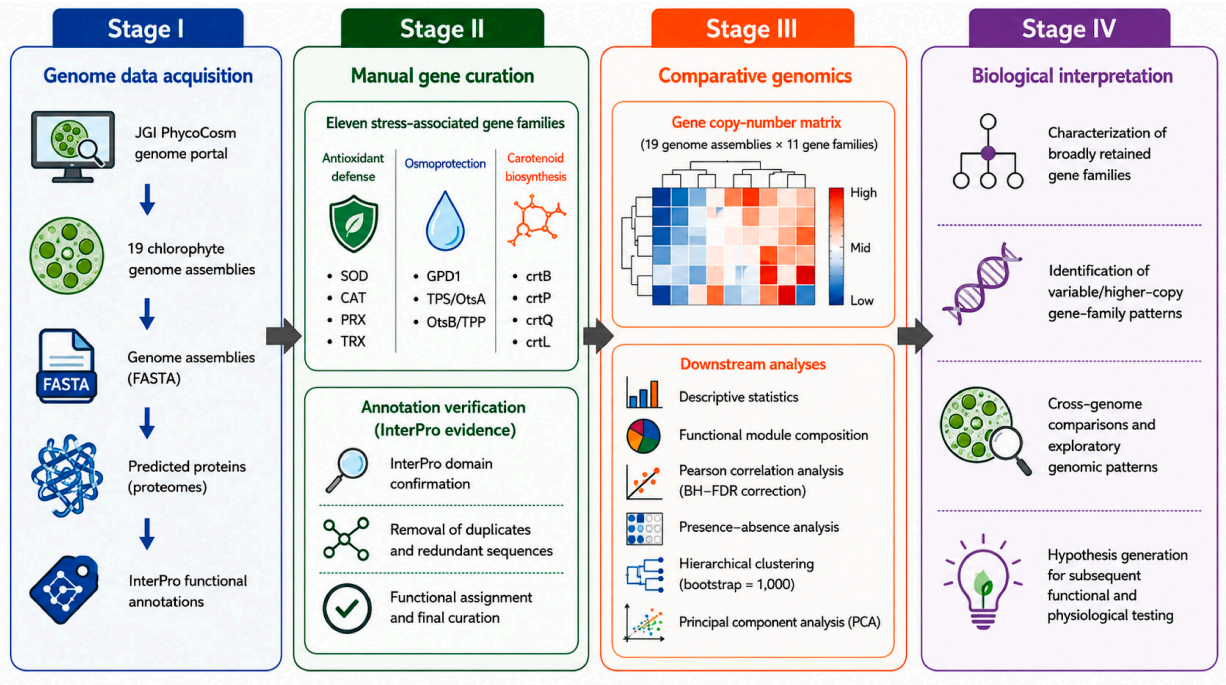

Figure 1. Comparative genomics workflow for the identification and analysis of stress-associated gene families in 19 chlorophyte genome assemblies. Genome assemblies, predicted proteomes, and InterPro-based functional annotations were obtained from the JGI PhycoCosm Database. Following standardized manual curation of 11 stress-associated gene families, a 19 × 11 gene copy number matrix was generated for comparative genomic analysis and biological interpretation. The figure was created using BioRender.com.

3.2. Gene Copy-Number Variation Among Chlorophyte Genome Assemblies

Comparative analysis of the curated stress-associated gene repertoire identified substantial variation in gene copy number distribution across the 19 chlorophyte genome assemblies, whereas most stress-associated gene families were retained across the three functional modules (Figures 2 and S2; Table 2). Hierarchical clustering of Z-score-standardized copy numbers grouped genome assemblies according to the relative abundance of individual gene families (Figure 2). Bootstrap analysis indicated variable support among dendrogram nodes, and clustering was therefore interpreted as exploratory rather than as evidence of robust evolutionary relationships (Table S7). Despite the extensive copy number variation, most gene families were present across the analyzed genome assemblies, indicating a wide representation of the investigated stress-associated gene repertoire.

Gene-family variability differed among the analyzed functional groups (Figure S1; Table S3). Within the antioxidant module, *TRX* showed consistently high copy numbers across all genomes (12–23 copies) and the lowest CV (15.43%), whereas *CAT* showed the greatest variability (59.28%). *SOD* also showed comparatively low variability (21.80%). Within the osmoprotection module, *OtsB/TPP* displayed moderate variation (CV = 28.58%), whereas *GPD1* (36.32%) and *TPS/OtsA* (53.05%) showed greater variability across genome assemblies. Among the carotenoid biosynthesis genes, *crtB* (21.08%) and *crtL* (23.24%)

showed comparatively low copy number variability, whereas *crtP* (32.13%) and *crtQ* (55.34%) showed greater variability.

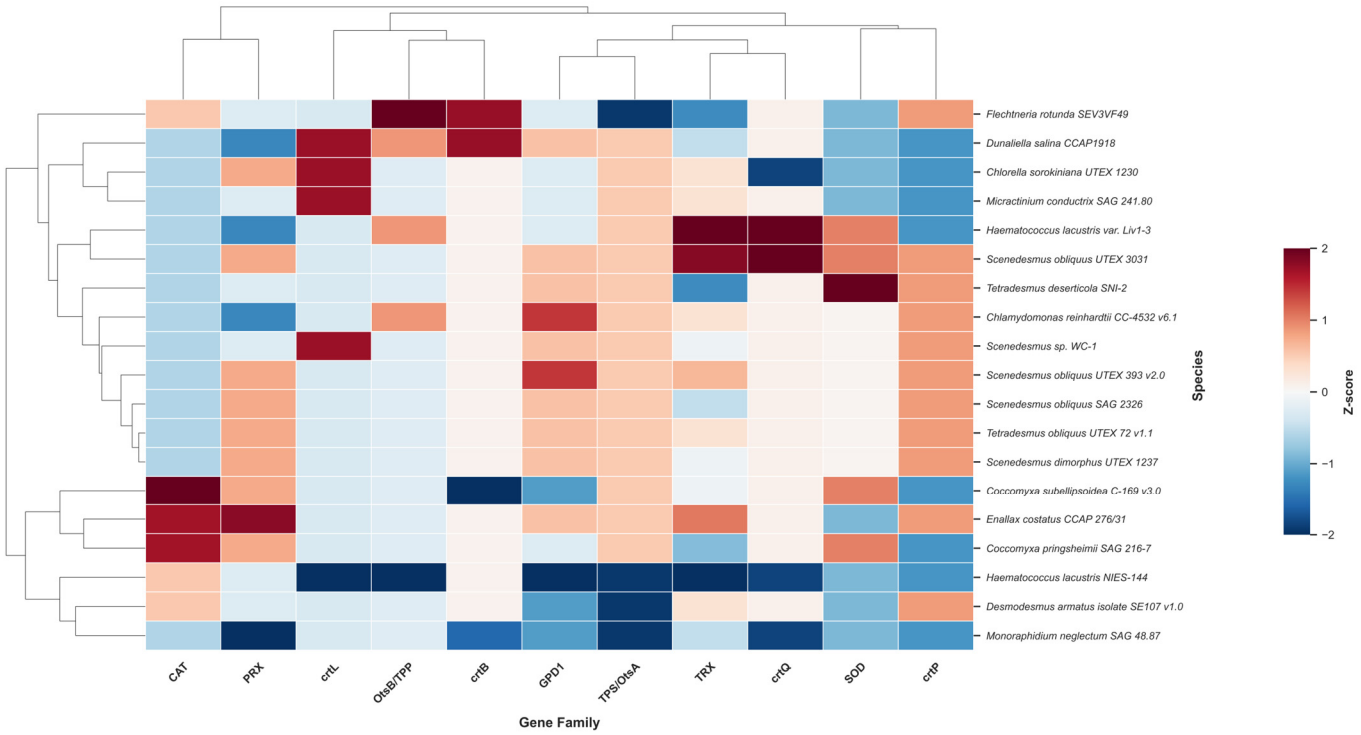

Figure 2. Hierarchically clustered heatmap showing the relative abundance of 11 stress-associated gene families across 19 chlorophyte genome assemblies. Gene copy numbers were standardized using Z-score transformation to facilitate comparisons among genomes. Positive Z-scores (red) indicate higher than average gene copy numbers, whereas negative Z-scores (blue) indicate lower than average gene copy numbers. Hierarchical clustering of both genome assemblies and gene families was performed using Euclidean distance and complete linkage to identify similarities in stress-associated gene copy number profiles.

Distinct gene copy number profiles were also evident among the genome assemblies (Figure 2; Table S5). *Scenedesmus obliquus* UTEX 3031 showed the highest total functional-family assignment count (51), followed by *Haematococcus lacustris* var. Liv1-3 and *Enallax costatus* CCAP 276/31 (49 each), whereas *H. lacustris* NIES-144 and *Monoraphidium neglectum* SAG 48.87 showed the lowest totals, with 28 and 33 assignments, respectively. The individual genomes differed in terms of the relative abundance of specific gene families. Elevated *TRX* copy numbers were observed in *H. lacustris* var. Liv1-3, *S. obliquus* UTEX 3031, and *E. costatus* CCAP 276/31, and *Tetradesmus deserticola* SNI-2 showed the highest *SOD* copy numbers, and both *Coccomyxa* genomes showed comparatively high *CAT* copy numbers. *H. lacustris* NIES-144 lacked *GPD1*, *TPS/OtsA*, and *crtQ* and showed relatively low copy numbers for several other analyzed families (Figure S2; Table S2).

At the functional module level, antioxidant defense constituted the largest component of the analyzed stress-associated repertoire, with a mean genome-level total of 28.00 functional family assignments, compared with 7.32 for osmoprotection and 7.63 for carotenoid biosynthesis (Table 3). This pattern was largely influenced by the consistently high copy number of *TRX*. Osmoprotection showed the highest relative variability (CV = 27.36%), followed by carotenoid biosynthesis (17.03%), whereas antioxidant defense showed the lowest relative variability (12.08%). Overall, the analysis showed substantial differences in stress-associated gene copy numbers among chlorophyte genome assemblies at both the individual gene family and functional module levels (Figures 2, S1 and S2; Table 3).

3.3. Multivariate Gene Copy-Number Patterns Revealed by Principal Component Analysis

PCA summarized the overall variation in stress-associated gene copy numbers across the 19 chlorophyte genome assemblies and revealed differences in their multivariate copy-number profiles (Figure 3). The first two principal components explained 50.18% of the total variance, with PC1 accounting for 30.86% and PC2 accounting for 19.32% (Figure 3). Most genome assemblies occupied the positive side of PC1, whereas several genomes were separated from the main distribution.

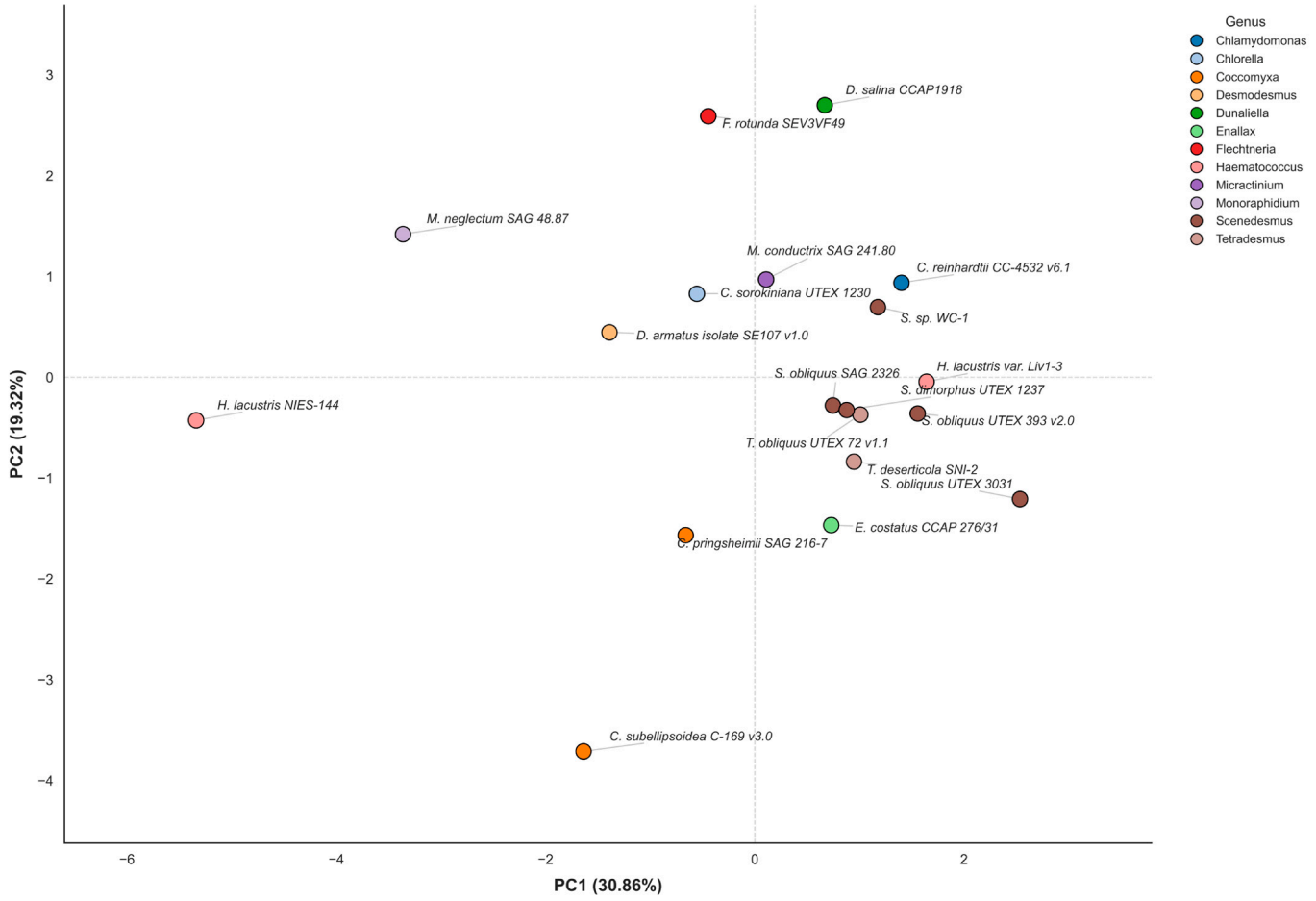

Figure 3. Principal component analysis (PCA) score plot illustrating the distribution of 19 chlorophyte genome assemblies based on the copy numbers of 11 stress-associated gene families. Each point represents an individual genome assembly and is colored according to the genus. The first two principal components explained 30.86% (PC1) and 19.32% (PC2) of the total variance. Genome assemblies positioned closer together exhibited more similar standardized stress-associated gene copy-number profiles, whereas greater separation indicated greater differences in these profiles.

Several genome assemblies were separated from the main cluster, showing distinct combinations of stress-associated gene copy numbers captured by the first two principal components. *H. lacustris* NIES-144 was positioned at the negative extreme of PC1 (Figures 3 and S2; Table S2), consistent with its comparatively low copy numbers for several analyzed families. *M. neglectum* SAG 48.87 and *Desmodesmus armatus* isolate SE107 also occurred on the negative side of PC1, whereas *Flechtneria rotunda* SEV3VF49 and *Dunaliella salina* CCAP1918 were distinguished primarily along PC2. Members of the *Scenedesmus–Tetradesmus* group occupied a relatively compact region on the positive side of PC1, although some separation was evident among the individual genome assemblies.

The PCA loading plot identified the contributions of individual stress-associated gene families to PC1 and PC2 (Figure 4). Positive loadings on PC1 were primarily associated with all analyzed gene families except *CAT*, with particularly strong positive contributions from *GPD1*, *TPS/OtsA*, and *crtQ*, whereas *CAT* showed a negative loading on PC1. Along PC2, *crtB*, *OtsB/TPP*, and *crtL* showed the strongest positive loadings, whereas *CAT*, *PRX*, *SOD*, and *TPS/OtsA* showed negative loadings (Figure 4). Based on the combined loading magnitudes across PC1 and PC2, *GPD1*, *TPS/OtsA*, *crtQ*, *crtB*, and *CAT* showed comparatively strong contributions to the variation represented in two-dimensional PCA space.

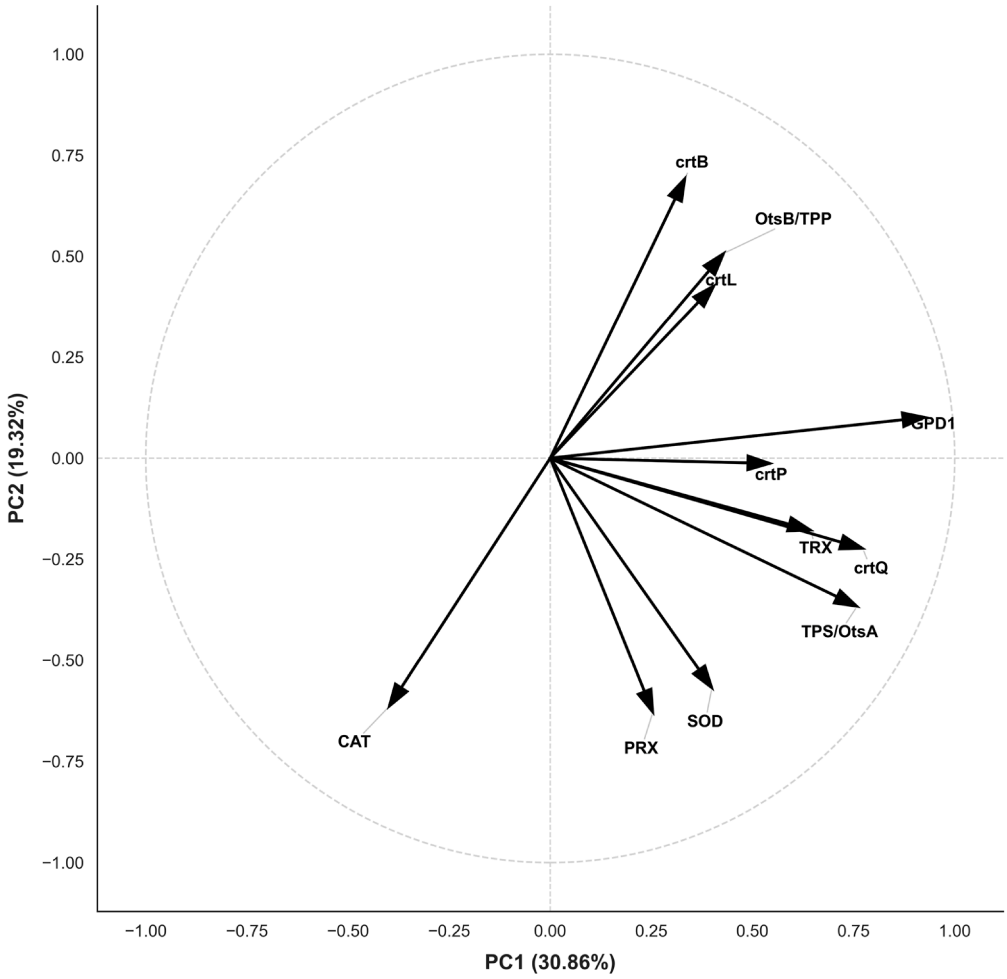

Figure 4. Principal component analysis (PCA) loading plot showing the contribution of the 11 stress-associated gene families to the first two principal components. The direction of the arrows indicates the association of each gene family with the principal component axes, whereas the length of the arrow represents the magnitude of its contribution to the variation represented by PC1 and PC2. Gene families with similar vector orientations contributed in similar directions within the PCA space, whereas oppositely oriented vectors indicated contrasting contributions. PC1 and PC2 explained 30.86% and 19.32% of the total variances, respectively.

PCA revealed broad patterns consistent with the hierarchical clustering of Z-score-standardized gene copy numbers (Figures 2 and 3). Genome assemblies positioned close together in the ordination generally showed similar standardized copy number profiles, whereas more isolated genomes displayed comparatively distinct combinations of gene family copy numbers (Figures 2 and 3; Table S2). Because the first two PCs represented approximately half of the total variance and the hierarchical cluster support varied among nodes, these multivariate patterns were interpreted as exploratory similarities in gene copy number profiles rather than functional or evolutionary relationships.

3.4. Correlation Among Stress-Associated Gene Families

Pearson's correlation analysis revealed several moderate-to-strong pairwise associations among the 11 stress-associated gene families across the 19 chlorophyte genome assemblies; however, none remained statistically significant after Benjamini–Hochberg false discovery rate (BH-FDR) correction for 55 pairwise comparisons (Figure 5; Table S6). Most pairwise correlations were positive, whereas the negative correlations were generally weak or moderate.

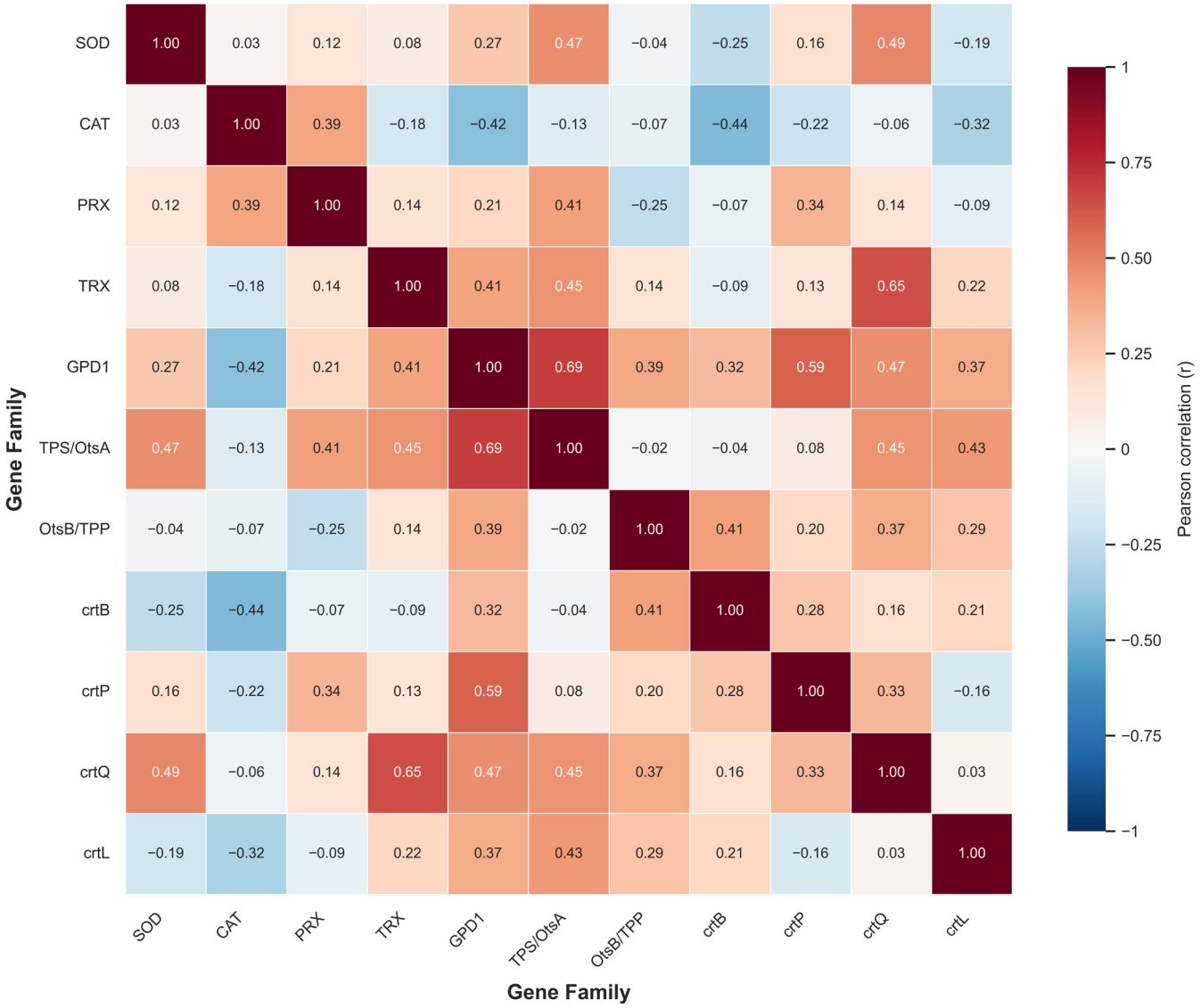

Figure 5. Pearson correlation matrix showing pairwise relationships among the copy numbers of 11 stress-associated gene families across 19 chlorophyte genome assemblies. Correlation coefficients (r) are displayed within each cell. Statistical significance was evaluated using two-sided Pearson correlation tests, followed by BH-FDR correction for 55 pairwise comparisons. No pairwise association remained significant at FDR-adjusted $q < 0.05$. Positive and negative correlations are represented by red and blue colors, respectively.

The strongest positive association was observed within the osmoprotection module between *TPS/OtsA* and *GPD1* ($r = 0.69$). In contrast, *OtsB/TPP* showed weaker associations with *GPD1* ($r = 0.39$) and *TPS/OtsA* ($r = -0.02$).

Several notable positive associations were observed between the functional modules. *TRX* showed the strongest cross-module association with *crtQ* ($r = 0.65$) and moderate

positive associations with *TPS/OtsA* ($r = 0.45$) and *GPD1* ($r = 0.41$). *GPD1* was positively associated with *crtP* ($r = 0.59$) and *crtQ* ($r = 0.47$), whereas *SOD* showed moderate positive associations with *crtQ* ($r = 0.49$) and *TPS/OtsA* ($r = 0.47$).

Negative correlations were comparatively less pronounced in this study. The strongest negative relationships were observed between *CAT* and *crtB* ($r = -0.44$) and between *CAT* and *GPD1* ($r = -0.42$). Other negative associations were generally weak, including *CAT–crtL* ($r = -0.32$), *SOD–crtB* ($r = -0.25$), and *PRX–OtsB/TPP* ($r = -0.25$).

Overall, the correlation matrix was dominated by positive associations, with the largest coefficients involving *TPS/OtsA–GPD1* and *TRX–crtQ*. However, because none of the 55 pairwise correlations remained significant after BH-FDR correction, these relationships were interpreted as exploratory patterns rather than statistically supported gene-family associations (Figure 5; Table S6).

3.5. Mean Copy Number and Observed Ranges of Stress-Associated Gene Families

The gene copy number variation differed among the 11 stress-associated gene families across the 19 chlorophyte genome assemblies (Figures 6 and S1; Table S3).

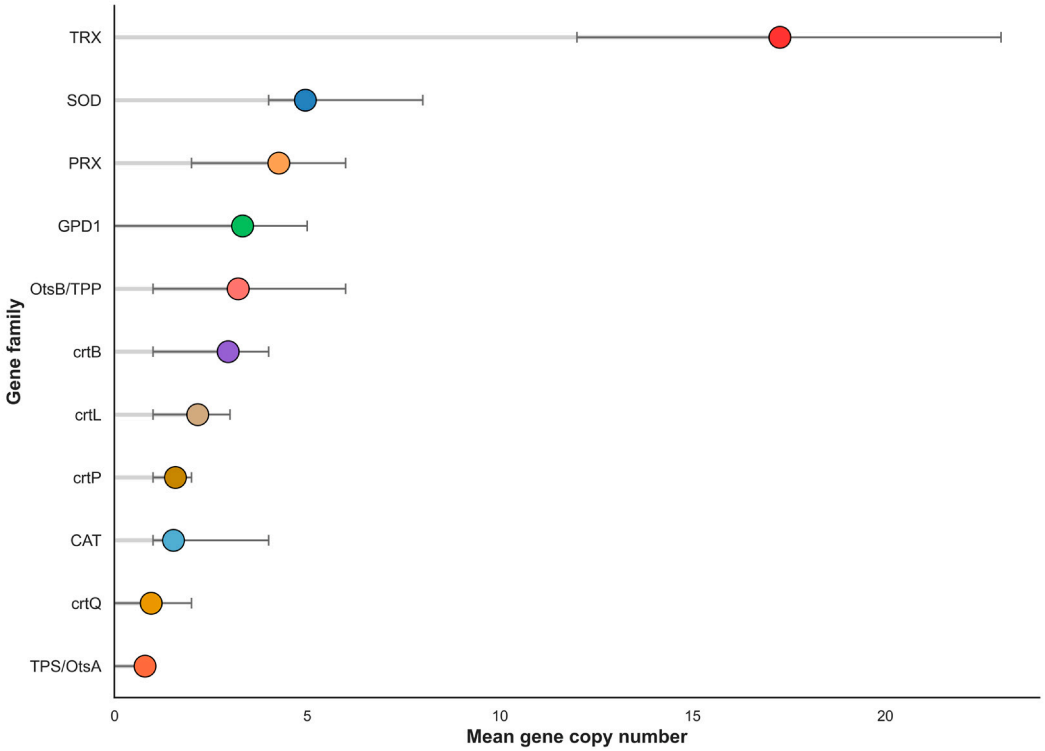

Figure 6. Horizontal lollipop plot summarizing the mean copy number of 11 stress-associated gene families across 19 chlorophyte genome assemblies. Error bars represent the observed range (minimum–maximum copy numbers) of the genomes. Higher mean copy numbers indicate greater average gene-family abundance across the analyzed genomes, whereas wider ranges indicate greater variation in copy number among genome assemblies.

TRX exhibited the highest mean copy number (17.26 copies per genome; range 12–23), followed by *SOD* (4.95; range 4–8) and *PRX* (4.26; range 2–6). Intermediate mean copy numbers were observed for *GPD1* (3.32; range 0–5), *OtsB/TPP* (3.21; range 1–6), and *crtB* (2.95; range 1–4). Lower mean copy numbers were observed for *crtL* (2.16), *crtP* (1.58), *CAT* (1.53), *crtQ* (0.95), and *TPS/OtsA* (0.79) (Figure 6; Table S3). The relative variability, expressed as CV, ranged from 15.43% for *TRX* to 59.28% for *CAT*, with the detailed gene family CV patterns presented in Section 3.2 and Figure S1.

Overall, the gene families differed markedly in both average copy number and observed copy-number range, with *TRX* showing the highest abundance and *TPS/OtsA* the lowest mean abundance across the analyzed genome assemblies (Figures 6 and S1; Table S3).

4. Discussion

4.1. Comparative Genomic Patterns of Antioxidant Defense Systems

Photosynthetic organisms rely on interconnected antioxidant and redox-regulatory systems to manage the reactive oxygen species (ROS) generated during photosynthetic metabolism and environmental stress [36–39]. Among these systems, *TRX* participate broadly in cellular redox regulation through reversible thiol–disulfide exchange and interact with numerous processes associated with photosynthesis, metabolism, protein maintenance, and stress responses [40,41].

TRX showed the highest mean copy number (17.26 copies per genome) and the lowest CV (15.43%) among the 11 analyzed gene families, indicating a comparatively limited relative copy-number variation across the 19 chlorophyte genome assemblies (Figures 6 and S1; Table S3). Unlike antioxidant enzymes that directly detoxify specific ROS molecules, thioredoxins regulate a wide spectrum of downstream proteins involved in photosynthesis, metabolism, and stress acclimation [40–42]. The comparatively high abundance and low CV of *TRX* in the present dataset are therefore consistent with the broad representation of this redox-regulatory family among the analyzed genomes. However, copy-number stability alone does not demonstrate functional constraint or conserved expression, and the observed pattern may also reflect lineage history and the retention of multiple paralogous *TRX* genes.

In contrast, the remaining antioxidant genes displayed different levels of copy number variability. *SOD* and *PRX* showed relatively low copy number variability (CV = 21.80% and 23.25%, respectively), indicating comparatively stable copy number patterns across the genome assemblies (Figure 6). Superoxide dismutases constitute the primary enzymatic defense against superoxide radicals, whereas peroxiredoxins reduce hydrogen peroxide and organic peroxides, often through thioredoxin-dependent catalytic cycles [38,43]. Their relatively low CVs indicate less variation in copy number than that observed for several other analyzed stress-associated families, although these genomic patterns do not by themselves establish equivalent antioxidant activity among taxa.

In comparison, *CAT* showed the highest copy number variability among all analyzed gene families (CV = 59.28%) (Figures 6 and S1; Table S3). This indicates a greater relative variation in *CAT* copy number among the analyzed genome assemblies than among the other antioxidant families. One possible explanation is lineage-specific duplication or loss, whereas differences in overall gene content and genome history may also contribute. Complementary peroxide detoxification systems, including peroxiredoxin- and thioredoxin-associated pathways, could reduce dependence on *CAT* copy number alone; however, this interpretation remains a hypothesis that requires functional validation [36,38,44].

Overall, the antioxidant gene families exhibited contrasting copy number patterns across the analyzed chlorophyte genome assemblies. *TRX* showed the highest mean abundance and lowest relative variability, whereas *CAT* showed the greatest relative variability. These differences provide evidence for the heterogeneous genomic organization of antioxidant-related gene families; however, they should not be interpreted as direct measures of antioxidant capacity, stress tolerance, or adaptation. Such functional consequences would require complementary transcriptomic, proteomic, biochemical, and physiological validation.

4.2. Comparative Genomic Variation in Osmoprotection Pathways

Osmoprotection in photosynthetic organisms involves several compatible solute pathways, the relative importance of which can differ among taxa and environmental conditions [45,46]. Among chlorophytes, osmotic adjustment can involve the synthesis and accumulation of low-molecular-weight compatible solutes, particularly glycerol and trehalose, which contribute to protein and membrane stabilization and cellular osmotic balance [47,48].

Within the osmoprotection module, *OtsB/TPP* showed the lowest relative copy-number variability (CV = 28.58%), followed by *GPD1* (36.32%), whereas *TPS/OtsA* showed the greatest variability (53.05%) across the 19 chlorophyte genome assemblies (Figures 6 and S1; Table S3). At the module level, osmoprotection also showed the highest relative variability among the three analyzed functional modules (CV = 27.36%; Table 3), indicating greater heterogeneity in genome-level copy number totals than those observed for antioxidant defense or carotenoid biosynthesis.

Physiological studies have shown that glycerol is an important compatible solute in several microalgae exposed to osmotic stress. In *Dunaliella* and several *Chlamydomonas* species, intracellular glycerol concentrations increase with external salinity, contributing to osmotic adjustment [47,48]. In the present genomic dataset, *GPD1* had a mean copy number of 3.32 but was not universally retained, emphasizing that gene family copy number alone cannot establish the physiological importance of glycerol metabolism across all chlorophytes. The observed variation may reflect lineage-specific duplication, gene loss, differences in genome content, and alternative osmolyte strategies.

The greater variability of *TPS/OtsA* indicates that the trehalose-6-phosphate synthase family is less uniformly represented in the analyzed genomes than *GPD1* or *OtsB/TPP*. Trehalose functions as a compatible solute and participates in metabolic and stress-related signaling processes [45,47]. Therefore, the *TPS/OtsA* copy number may reflect differences in the genomic organization and retention of trehalose-related genes among lineages, although the present analysis did not demonstrate differences in trehalose production or osmotic tolerance. *TPS/OtsA* showed a positive correlation with *GPD1* (r = 0.69); however, this association did not remain significant after BH-FDR correction and should therefore be regarded as exploratory rather than evidence of coordinated evolutionary expansion.

Overall, the osmoprotection module exhibited greater genome-level relative variability than the antioxidant and carotenoid modules, with a particularly high variation in *TPS/OtsA*. These patterns demonstrate heterogeneity in osmoprotection-associated gene copy numbers among chlorophyte genome assemblies but do not establish habitat-specific adaptation or physiological osmotic tolerance. Such interpretations require phylogenetically informed comparisons, as well as expression, metabolite, and physiological measurements under defined osmotic stress conditions.

4.3. Comparative Genomic Patterns of Carotenoid Biosynthesis Genes

Carotenoids contribute to light harvesting and photoprotection by participating in energy dissipation, stabilization of photosynthetic complexes, and protection against photo-oxidative stress [49,50]. Therefore, the carotenoid biosynthetic pathway represents an important component of photosynthetic stress physiology, although variations in gene copy number do not directly indicate pigment abundance or photoprotective capacity [51].

The comparative genomic patterns observed in this study showed different levels of copy number variability among the four analyzed carotenoid biosynthesis gene families. *crtB* and *crtL* showed relatively low copy number variability (CV = 21.08% and 23.24%, respectively), whereas *crtP* showed moderate variation (32.13%), and *crtQ* displayed substantially greater variability (55.34%) (Figures 6 and S1; Table S3). *crtB* encodes phytoene

synthase, which catalyzes an early committed step in carotenoid biosynthesis, whereas *crtL* encodes lycopene cyclase, which is involved in the formation of cyclic carotenoids. Their comparatively low CVs indicate relatively limited copy number variation among the analyzed genome assemblies, but these patterns alone cannot be taken as evidence of a strong evolutionary constraint. Previous comparative studies have reported the broad conservation of core carotenoid biosynthetic functions across oxygenic phototrophs [51].

In contrast, the higher variability of *crtQ* indicates greater heterogeneity in the genomic representation of this desaturase family among the analyzed chlorophyte genomes (Figure 6). Evolutionary studies have shown that carotenoid desaturases can undergo gene duplication, replacement, and functional diversification across photosynthetic lineages [51]. Accordingly, the variation observed for *crtP* and *crtQ* may reflect lineage-specific duplication, loss, or differences in gene retention. Possible relationships with pigment composition or photoprotective requirements remain hypotheses because environmental conditions and carotenoid profiles were not measured in this study.

Differences in copy number among pathway components may arise through several evolutionary processes, including lineage-specific gene duplication, gene loss, and differential retention [51,52]. Consequently, the present data are most appropriately interpreted as variations in the genomic organization of carotenoid-associated gene families rather than direct evidence for differences in pathway activity or pigment composition.

Overall, carotenoid biosynthesis genes showed heterogeneous copy number patterns across the 19 chlorophyte genome assemblies. *crtB* and *crtL* exhibited comparatively low relative variability, whereas *crtQ* showed a substantially greater variation. These genomic differences provide candidate patterns for future investigation; however, determining whether they influence carotenoid composition, photosynthetic performance, or stress tolerance will require transcriptomic, biochemical, pigment profile, and physiological validation.

4.4. Genome-Level Variation in Stress-Associated Gene Copy-Number Profiles

Comparative genomics has demonstrated that closely related taxa can retain broadly similar metabolic functions while differing in gene duplication, loss, and retention patterns [53,54]. Chlorophytes occupy diverse ecological environments and are exposed to different combinations of osmotic, oxidative, irradiative, and nutrient stresses [55,56]. However, the present analysis did not directly test the association between habitat characteristics and gene copy number.

Hierarchical clustering, PCA, and genome-level copy-number summaries showed substantial differences among the 19 chlorophyte genome assemblies, despite the broad retention of the analyzed stress-associated gene families (Figures 2 and 3; Table S5). Genome assemblies positioned near one another in the PCA generally showed similar standardized copy-number profiles, whereas more isolated genomes displayed distinct combinations of antioxidant, osmoprotection, and carotenoid-associated gene-family counts. These patterns represent similarities in the analyzed copy-number profiles rather than direct evidence of functional or evolutionary relationships.

Several genome assemblies illustrated this variation (Figures 2 and 3; Table S5). *S. obliquus* UTEX 3031, *H. lacustris* var. Liv1-3, and *E. costatus* CCAP 276/31 showed the highest summed functional-family assignment counts, with 51, 49, and 49 assignments, respectively. In contrast, *H. lacustris* NIES-144 showed the lowest total (28 assignments) and occupied an isolated position in the PCA, reflecting relatively low copy numbers for several analyzed gene families together with the absence of *GPD1*, *TPS/OtsA*, and *crtQ* (Figure S2). Because multidomain proteins can contribute to more than one functional family assignment, these totals should not be interpreted as the number of unique protein loci.

Several evolutionary and genomic processes could contribute to the observed differences, including lineage-specific gene duplication and loss, differential retention of paralogs, whole-genome duplication, overall genome or gene content differences, and shared ancestry [53,54]. Environmental selection may also contribute, but the present dataset does not allow for the distinction of these alternatives. Because the relative contributions of these evolutionary mechanisms have not been explicitly tested, they should be regarded as plausible explanations rather than demonstrated causes of the observed copy number variation. This distinction is illustrated by *D. salina*, whose tolerance to hypersaline environments has been associated with specialized osmotic adjustment and photoprotective mechanisms rather than simply with a generalized increase in stress-associated gene copy number [57]. Accordingly, stress tolerance should not be inferred directly from the total number of stress-associated gene-family copies.

Overall, the analyzed chlorophyte genome assemblies shared broad representation of the selected stress-associated gene families but differed substantially in their quantitative copy-number profiles. These genome-level patterns provide a comparative framework for generating hypotheses regarding lineage-specific gene duplication, loss, paralog retention, and other genome-scale evolutionary processes. Establishing whether these genomic differences contribute to physiological stress tolerance will require phylogenetically informed analyses, together with transcriptomic, proteomic, metabolomic, and experimental stress response data.

4.5. Implications for Microalgal Biostimulant Discovery and Future Comparative Genomics

The development of commercial microalgal biostimulants has largely been driven by biomass productivity, ease of cultivation, and demonstrated agronomic performance rather than by a systematic evaluation of comparative genomic variation in stress-associated pathways [7]. Consequently, only a limited number of chlorophyte taxa have been extensively studied, while many phylogenetically diverse taxa remain largely unexplored, despite potentially possessing valuable stress-associated genetic resources. Recent advances in whole-genome sequencing and comparative genomics provide an opportunity to complement conventional strain screening with genome-informed approaches that identify candidate lineages and pathways for subsequent experimental evaluations [58,59]. Such genomic information may help prioritize experimental resources, but cannot replace physiological or agronomic validation [60,61].

The comparative genomic framework established in this study illustrates how stress-associated gene copy number profiles can contribute to the preliminary screening of microalgal genomic diversity before detailed physiological and agronomic evaluation (Figures 2–6). Rather than identifying a single "superior" chlorophyte, our analyses showed that different genome assemblies possess distinct combinations of widely retained and high-copy-number stress-associated gene families. These differences can be used to generate testable hypotheses regarding candidate pathways or lineages of potential interest; however, they should not be interpreted as evidence that genomes with higher copy numbers exhibit greater stress tolerance or biostimulant efficacy.

Combined clustering, PCA, correlation, and copy number analyses provide a framework for identifying genome assemblies with similar or contrasting stress-associated gene family profiles and for selecting particular gene families for subsequent functional investigation. However, the low-to-moderate bootstrap support for many clusters, the absence of FDR-significant pairwise correlations, and the limited number of genome assemblies emphasize that these patterns are exploratory. In particular, none of the 55 pairwise gene-family correlations remained statistically significant after Benjamini–Hochberg FDR correction; therefore, the observed correlation coefficients were interpreted as descriptive

and hypothesis-generating rather than as evidence of robust functional or evolutionary associations. As genomic resources for microalgae continue to expand, comparative genomics may provide a useful complementary component for biostimulant discovery and algal biotechnology when integrated with experimental phenotyping [20,62–64].

However, gene copy number alone cannot predict biological function or biostimulant efficacy. Higher gene copy numbers do not necessarily correspond to increased transcription, protein abundance, metabolite accumulation, or physiological performance under field conditions. Consequently, the genomic patterns identified in this study should be considered hypotheses requiring functional validation rather than direct indicators of stress tolerance. Recent studies have further demonstrated that plant responses to algal biostimulants can vary among microalgal/cyanobacterial strains and algal-derived formulations, reinforcing the need for direct experimental validation rather than prediction of efficacy from genomic characteristics alone [65–67]. The present comparisons are also potentially influenced by shared phylogenetic ancestry, lineage-specific gene duplication and loss, whole-genome duplications, differences in genome size and total gene content, and variations in genome assembly and annotation completeness. Accordingly, similarities among genome-level copy number profiles cannot be assumed to represent independent adaptive responses because part of the observed similarity may reflect a common evolutionary history. These factors should be explicitly incorporated into future comparative analyses. Future investigations should integrate comparative genomics with transcriptomics, proteomics, metabolomics, and controlled physiological assays to establish direct relationships between gene copy number variation, stress-responsive molecular mechanisms, and crop performance following the application of microalgal biostimulants [53,64].

Future comparative studies should also expand taxonomic sampling to include additional chlorophyte lineages from extreme and contrasting environments and apply phylogenetically informed approaches, such as phylogenetic generalized least squares (PGLS) or phylogenetic independent contrasts (PIC), using a well-resolved phylogeny to explicitly account for evolutionary non-independence among taxa. Because no pairwise correlation remained significant after FDR correction in the present dataset, the current associations were retained as exploratory patterns rather than being subjected to additional phylogenetic correlation modeling. Future analyses with broader taxonomic sampling will be better suited to evaluate whether gene family associations persist after correction for phylogenetic relatedness. Normalization of gene family copy numbers by genome-wide gene content would provide an additional sensitivity analysis for distinguishing family-specific expansion from broader differences in genome complexity. Moreover, extending the genomic framework beyond the antioxidant, osmoprotection, and carotenoid modules to include other biostimulant-relevant pathways, including phytohormone biosynthesis and signaling, would provide a more comprehensive basis for evaluating candidate strains.

Overall, this study provides a curated comparative genomic framework for characterizing variations in selected stress-associated gene families across chlorophyte genome assemblies. The framework can support the prioritization of candidate lineages and molecular pathways for subsequent functional testing; however, its application to microalgal biostimulant discovery should remain hypothesis-generating until genomic patterns are experimentally linked to stress physiology and plant response outcomes.

5. Conclusions

This study presents a comparative genomic assessment of stress-associated gene families across 19 chlorophyte genome assemblies, providing insights into copy-number variation among 11 selected gene families involved in antioxidant defense, osmoprotection, and carotenoid biosynthesis. Comparative analyses revealed that the antioxidant module

exhibited the lowest genome-level relative variability, whereas the osmoprotection module exhibited the highest variability. Carotenoid biosynthesis genes displayed intermediate levels of copy-number variability, with relatively stable and more variable gene families. At the individual gene family level, *TRX* showed the highest mean copy number and lowest relative variability, whereas *CAT*, *crtQ*, and *TPS/OtsA* showed comparatively greater variability. These findings indicate that chlorophyte genome assemblies contain both relatively stable and more variable stress-associated gene families, highlighting the heterogeneity in their genomic copy number organization rather than direct differences in stress tolerance or adaptive capacity.

In addition to describing these genomic patterns, this study established a standardized comparative genomic framework for examining copy number variations in selected stress-associated gene families across chlorophyte lineages. The curated dataset and analytical approach provide a resource for future comparative and functional studies and may assist in prioritizing candidate lineages and pathways for subsequent experimental investigations relevant to microalgal biotechnology and biostimulant research. These genomic characteristics should be regarded as hypothesis-generating patterns that require experimental validation through transcriptomic, proteomic, metabolomic, physiological, and agronomic studies before any relationship with stress tolerance or plant biostimulant performance can be established. Future studies should incorporate broader taxonomic sampling, phylogenetically informed analyses, and normalization of genome-wide gene content to distinguish lineage-specific gene family variation from broader genomic effects. As high-quality algal genome resources continue to expand, integrating comparative genomics with functional and physiological investigations will further clarify the biological significance of stress-associated gene copy-number variation and its potential applications in sustainable agriculture.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/phycology6030097/s1, Figure S1. Coefficient of variation (CV) of 11 stress-associated gene families across 19 chlorophyte genome assemblies; Figure S2. Presence–absence heatmap of 11 stress-associated gene families across 19 chlorophyte genome assemblies; Table S1. Curated protein IDs, supporting InterPro accessions, and manual curation rationale for the stress-associated gene families; Table S2. Manually curated copy-number matrix of 11 stress-associated gene families involved in antioxidant defense, osmoprotection, and carotenoid biosynthesis across 19 chlorophyte genome assemblies; Table S3. Descriptive statistics of 11 stress-associated gene families; Table S4. Functional module composition (%) of stress-associated gene families across 19 chlorophyte genome assemblies; Table S5. Genome-wise summary of stress-associated functional-family assignments; Table S6. Pearson correlation coefficients, raw p values, and Benjamini–Hochberg FDR-adjusted p values for 11 stress-associated gene families; Table S7. Bootstrap support for hierarchical clustering of 19 chlorophyte genome assemblies based on 11 stress-associated gene families.

Funding: This research received no external funding.

Data Availability Statement: All data generated or analyzed during this study are included in this published article and its Supplementary Materials.

Acknowledgments: The authors used Paperpal (Version 2025.3; Cactus Communications, Mumbai, India), an AI-assisted language editing tool, to improve the grammar, language, and clarity of this manuscript. The authors reviewed and edited all suggested changes and took full responsibility for the final content of the manuscript. All scientific analyses, interpretations, conclusions, and intellectual contributions were performed by the authors.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Di Sario, L.; Boeri, P.; Matus, J.T.; Pizzio, G.A. Plant biostimulants to enhance abiotic stress resilience in crops. *Int. J. Mol. Sci.* **2025**, *26*, 1129. [CrossRef] [PubMed]
2. Zhu, J.; Cai, Y.; Wakisaka, M.; Yang, Z.; Yin, Y.; Fang, W.; Xu, Y.; Omura, T.; Yu, R.; Zheng, A.L. Mitigation of oxidative stress damage caused by abiotic stress to improve biomass yield of microalgae: A review. *Sci. Total Environ.* **2023**, *896*, 165200. [CrossRef] [PubMed]
3. Brito-Lopez, C.; van der Wielen, N.; Barbosa, M.; Karlova, R. Plant growth-promoting microbes and microalgae-based biostimulants: Sustainable strategy for agriculture and abiotic stress resilience. *Philos. Trans. R. Soc. B Biol. Sci.* **2025**, *380*, 20240251. [CrossRef]
4. Renganathan, P.; Gaysina, L.A.; Rueda-Puente, E.O. Microbial biofertilizers for salinity stress mitigation in hydroponic systems. *Curr. Issues Mol. Biol.* **2025**, *47*, 1029. [CrossRef] [PubMed]
5. Renganathan, P.; Puente, E.O.; Sukhanova, N.V.; Gaysina, L.A. Hydroponics with microalgae and cyanobacteria: Emerging trends and opportunities in modern agriculture. *BioTech* **2024**, *13*, 27. [CrossRef]
6. Renganathan, P.; Gaysina, L.A.; Holguín-Peña, R.J.; Sainz-Hernández, J.C.; Ortega-García, J.; Rueda-Puente, E.O. Phycoremediated microalgae and cyanobacteria biomass as biofertilizer for sustainable agriculture: A holistic biorefinery approach to promote circular bioeconomy. *Biomass* **2024**, *4*, 1047–1077. [CrossRef]
7. Renganathan, P.; Mahendhiran, M.; Ríos-Ruiz, W.F.; Gaysina, L.; Puente, E.O.R. Microalgae–microbe consortia for agricultural wastewater remediation and biofertilizer valorization: A circular bioeconomy perspective. In *Algal Bioremediation: Sustainable Solutions, Challenges, and Future Prospects*; Springer Nature: Singapore, 2026; pp. 319–342.
8. Vangenechten, B.; De Coninck, B.; Ceusters, J. How to improve the potential of microalgal biostimulants for abiotic stress mitigation in plants? *Front. Plant Sci.* **2025**, *16*, 1568423. [CrossRef]
9. Tamaki, S.; Mochida, K.; Suzuki, K. Diverse biosynthetic pathways and protective functions against environmental stress of antioxidants in microalgae. *Plants* **2021**, *10*, 1250. [CrossRef]
10. Rezayian, M.; Niknam, V.; Ebrahimzadeh, H. Oxidative damage and antioxidative system in algae. *Toxicol. Rep.* **2019**, *6*, 1309–1313. [CrossRef] [PubMed]
11. Sun, X.M.; Ren, L.J.; Zhao, Q.Y.; Ji, X.J.; Huang, H. Microalgae for the production of lipid and carotenoids: A review with focus on stress regulation and adaptation. *Biotechnol. Biofuels* **2018**, *11*, 272. [CrossRef]
12. Generalić Mekinić, I.; Šimat, V.; Rathod, N.B.; Hamed, I.; Čagalj, M. Algal carotenoids: Chemistry, sources, and application. *Foods* **2023**, *12*, 2768. [CrossRef]
13. Stavridou, E.; Karapetsi, L.; Nteve, G.M.; Tsintzou, G.; Chatzikonstantinou, M.; Tsaousi, M.; Martinez, A.; Flores, P.; Merino, M.; Dobrovic, L.; et al. Landscape of microalgae omics and metabolic engineering research for strain improvement: An overview. *Aquaculture* **2024**, *587*, 740803. [CrossRef]
14. Vilarrasa-Blasi, J.; Vellosillo, T.; Jinkerson, R.E.; Fauser, F.; Xiang, T.; Minkoff, B.B.; Wang, L.; Kniazev, K.; Guzman, M.; Osaki, J.; et al. Multi-omics analysis of green lineage osmotic stress pathways unveils crucial roles of different cellular compartments. *Nat. Commun.* **2024**, *15*, 5988. [CrossRef]
15. Goold, H.D.; Moseley, J.L.; Lauersen, K.J. The synthetic future of algal genomes. *Cell Genom.* **2024**, *4*, 100505. [CrossRef] [PubMed]
16. Grigoriev, I.V.; Hayes, R.D.; Calhoun, S.; Kamel, B.; Wang, A.; Ahrendt, S.; Dusheyko, S.; Nikitin, R.; Mondo, S.J.; Salamov, A.; et al. PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res.* **2021**, *49*, D1004–D1011. [CrossRef]
17. Dewar, A.E.; Belcher, L.J.; West, S.A. A phylogenetic approach to comparative genomics. *Nat. Rev. Genet.* **2025**, *26*, 395–405. [CrossRef] [PubMed]
18. Weisman, C.M.; Murray, A.W.; Eddy, S.R. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Curr. Biol.* **2022**, *32*, 2632–2639. [CrossRef] [PubMed]
19. Bachy, C.; Wittmers, F.; Muschiol, J.; Hamilton, M.; Henrissat, B.; Worden, A.Z. The land–sea connection: Insights into the plant lineage from a green algal perspective. *Annu. Rev. Plant Biol.* **2022**, *73*, 585–616. [CrossRef] [PubMed]
20. Blaby-Haas, C.E.; Merchant, S.S. Comparative and functional algal genomics. *Annu. Rev. Plant Biol.* **2019**, *70*, 605–638. [CrossRef]
21. Clark, J.W. Genome evolution in plants and the origins of innovation. *New Phytol.* **2023**, *240*, 2204–2209. [CrossRef]
22. Fang, Y.; Jiang, J.; Hou, X.; Guo, J.; Li, X.; Zhao, D.; Xie, X. Plant protein-coding gene families: Their origin and evolution. *Front. Plant Sci.* **2022**, *13*, 995746. [CrossRef] [PubMed]
23. Joint Genome Institute. PhycoCosm Platform. Available online: https://phycocosm.jgi.doe.gov/ (accessed on 20 June 2026).
24. Craig, R.J.; Gallaher, S.D.; Shu, S.; Salomé, P.A.; Jenkins, J.W.; Blaby-Haas, C.E.; Purvine, S.O.; O'Donnell, S.; Barry, K.; Grimwood, J.; et al. The *Chlamydomonas* Genome Project, version 6: Reference assemblies for mating-type plus and minus strains reveal extensive structural mutation in the laboratory. *Plant Cell* **2023**, *35*, 644–672. [CrossRef] [PubMed]
25. Hovde, B.T.; Hanschen, E.R.; Tyler, C.R.; Lo, C.C.; Kunde, Y.; Davenport, K.; Daligault, H.; Msanne, J.; Canny, S.; Eyun, S.I.; et al. Genomic characterization reveals significant divergence within *Chlorella sorokiniana* (Chlorellales, Trebouxiophyceae). *Algal Res.* **2018**, *35*, 449–461. [CrossRef]

26. Polle, J.E.; Barry, K.; Cushman, J.; Schmutz, J.; Tran, D.; Hathwaik, L.T.; Yim, W.C.; Jenkins, J.; McKie-Krisberg, Z.; Prochnik, S.; et al. Draft nuclear genome sequence of the halophilic and β-carotene-accumulating green alga *Dunaliella salina* strain CCAP19/18. *Genome Announc.* **2017**, *5*, e01287-17. [CrossRef] [PubMed]
27. Knoshaug, E.P.; Nag, A.; Astling, D.P.; Douchi, D.; Laurens, L.M. Draft genome sequence of the biofuel-relevant microalga *Desmodesmus armatus*. *Microbiol. Resour. Announc.* **2020**, *9*, e01280-19. [CrossRef] [PubMed]
28. Bogen, C.; Al-Dilaimi, A.; Albersmeier, A.; Wichmann, J.; Grundmann, M.; Rupp, O.; Lauersen, K.J.; Blifernez-Klassen, O.; Kalinowski, J.; Goesmann, A.; et al. Reconstruction of the lipid metabolism for the microalga *Monoraphidium neglectum* from its genome sequence reveals characteristics suitable for biofuel production. *BMC Genom.* **2013**, *14*, 926. [CrossRef] [PubMed]
29. Lashbrook, M.J.; Adesanya, J.A.; Talag, J.; Lail, K.; Lee, R.; Lipzen, A.M.; Guo, J.; Jenkins, J.; Kuo, A.; Pennacchio, C.; et al. Draft genome assembly and annotation of *Haematococcus lacustris* strain Liv1, an industrial astaxanthin-producing microalga. *Microbiol. Resour. Announc.* **2025**, *14*, e00591-25. [CrossRef] [PubMed]
30. Blanc, G.; Agarkova, I.; Grimwood, J.; Kuo, A.; Brueggeman, A.; Dunigan, D.D.; Gurnon, J.; Ladunga, I.; Lindquist, E.; Lucas, S.; et al. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* **2012**, *13*, R39. [CrossRef]
31. Arriola, M.B.; Velmurugan, N.; Zhang, Y.; Plunkett, M.H.; Hondzo, H.; Barney, B.M. Genome sequences of *Chlorella sorokiniana* UTEX 1602 and *Micractinium conductrix* SAG 241.80: Implications to maltose excretion by a green alga. *Plant J.* **2018**, *93*, 566–586. [CrossRef] [PubMed]
32. Biondi, T.C.; Kruse, C.P.; Koehler, S.I.; Kwon, T.; Davis, A.K.; Eng, W.; Kunde, Y.; Gleasner, C.D.; Mak, K.T.; Polle, J.; et al. The telomere-to-telomere, gapless, phased diploid genome and methylome of the green alga *Scenedesmus obliquus* UTEX 3031 reveals significant heterozygosity and genetic divergence of the haplotypes. *Algal Res.* **2024**, *79*, 103431. [CrossRef]
33. Morimoto, D.; Yoshida, T.; Sawayama, S. Draft genome sequence of the astaxanthin-producing microalga *Haematococcus lacustris* strain NIES-144. *Microbiol. Resour. Announc.* **2020**, *9*, e00128-20. [CrossRef] [PubMed]
34. Puginier, C.; Libourel, C.; Otte, J.; Skaloud, P.; Haon, M.; Grisel, S.; Petersen, M.; Berrin, J.G.; Delaux, P.M.; Dal Grande, F.; et al. Phylogenomics reveals the evolutionary origins of lichenization in chlorophyte algae. *Nat. Commun.* **2024**, *15*, 4452. [CrossRef]
35. Kruse, C.P.S.; Calhoun, S.; Kamel, B.; Edmundson, S.J.; Holguin, O.; Mach, P.M.; McKie-Krisberg, Z.; Baumgart, L.; Blaby, I.K.; Bowen, B.; et al. A multi-omic characterization of the physiological responses to salt stress in *Scenedesmus obliquus* UTEX393. *Commun. Biol.* **2025**, *8*, 1632. [CrossRef] [PubMed]
36. Imlay, J.A.; Sethu, R.; Rohaun, S.K. Evolutionary adaptations that enable enzymes to tolerate oxidative stress. *Free Radic. Biol. Med.* **2019**, *140*, 4–13. [CrossRef] [PubMed]
37. Balsera, M.; Uberegui, E.; Schuermann, P.; Buchanan, B.B. Evolutionary development of redox regulation in chloroplasts. *Antioxid. Redox Signal.* **2014**, *21*, 1327–1355. [CrossRef] [PubMed]
38. Cejudo, F.J.; González, M.C.; Pérez-Ruiz, J.M. Redox regulation of chloroplast metabolism. *Plant Physiol.* **2021**, *186*, 9–21. [CrossRef] [PubMed]
39. Dietz, K.J.; Hell, R. Thiol switches in redox regulation of chloroplasts: Balancing redox state, metabolism and oxidative stress. *Biol. Chem.* **2015**, *396*, 483–494. [CrossRef] [PubMed]
40. Buchanan, B.B. The path to thioredoxin and redox regulation in chloroplasts. *Annu. Rev. Plant Biol.* **2016**, *67*, 1–24. [CrossRef] [PubMed]
41. Nikkanen, L.; Rintamäki, E. Thioredoxin-dependent regulatory networks in chloroplasts under fluctuating light conditions. *Philos. Trans. R. Soc. B Biol. Sci.* **2014**, *369*, 20130224. [CrossRef] [PubMed]
42. Balsera, M.; Buchanan, B.B. Evolution of the thioredoxin system as a step enabling adaptation to oxidative stress. *Free Radic. Biol. Med.* **2019**, *140*, 28–35. [CrossRef] [PubMed]
43. Vaseghi, M.J.; Chibani, K.; Telman, W.; Liebthal, M.F.; Gerken, M.; Schnitzer, H.; Mueller, S.M.; Dietz, K.J. The chloroplast 2-cysteine peroxiredoxin functions as thioredoxin oxidase in redox regulation of chloroplast metabolism. *eLife* **2018**, *7*, e38194. [CrossRef] [PubMed]
44. Vogelsang, L.; Dietz, K.J. Regulatory thiol oxidation in chloroplast metabolism, oxidative stress response and environmental signaling in plants. *Biochem. J.* **2020**, *477*, 1865–1878. [CrossRef] [PubMed]
45. Iturriaga, G.; Suárez, R.; Nova-Franco, B. Trehalose metabolism: From osmoprotection to signaling. *Int. J. Mol. Sci.* **2009**, *10*, 3793–3810. [CrossRef] [PubMed]
46. Fürst-Jansen, J.M.R.; de Vries, S.; de Vries, J. Evo-physio: On stress responses and the earliest land plants. *J. Exp. Bot.* **2020**, *71*, 3254–3269. [CrossRef] [PubMed]
47. Colina, F.J.; Carbó, M.; Cañal, M.J.; Valledor, L. A complex metabolic rearrangement towards the accumulation of glycerol and sugars consequence of a proteome remodeling is required for the survival of *Chlamydomonas reinhardtii* growing under osmotic stress. *Environ. Exp. Bot.* **2020**, *180*, 104261. [CrossRef]
48. Raymond, J.A.; Morgan-Kiss, R.; Stahl-Rommel, S. Glycerol is an osmoprotectant in two Antarctic *Chlamydomonas* species from an ice-covered saline lake and is synthesized by an unusual bidomain enzyme. *Front. Plant Sci.* **2020**, *11*, 1259. [CrossRef] [PubMed]

49. Sun, T.; Rao, S.; Zhou, X.; Li, L. Plant carotenoids: Recent advances and future perspectives. *Mol. Hortic.* **2022**, *2*, 3. [CrossRef] [PubMed]
50. Caferri, R.; Guardini, Z.; Bassi, R.; Dall'Osto, L. Assessing photoprotective functions of carotenoids in photosynthetic systems of plants and green algae. In *Methods in Enzymology*; Academic Press: Cambridge, MA, USA, 2022; Volume 674, pp. 53–84.
51. Sandmann, G. Diversity and origin of carotenoid biosynthesis: Its history of coevolution towards plant photosynthesis. *New Phytol.* **2021**, *232*, 479–493. [CrossRef] [PubMed]
52. Takaichi, S. Distribution, biosynthesis, and function of carotenoids in oxygenic phototrophic algae. *Mar. Drugs* **2025**, *23*, 62. [CrossRef] [PubMed]
53. De Clerck, O.; Kao, S.M.; Bogaert, K.A.; Blomme, J.; Foflonker, F.; Kwantes, M.; Vancaester, E.; Vanderstraeten, L.; Aydogdu, E.; Boesger, J.; et al. Insights into the evolution of multicellularity from the sea lettuce genome. *Curr. Biol.* **2018**, *28*, 2921–2933. [CrossRef] [PubMed]
54. Hanschen, E.R.; Marriage, T.N.; Ferris, P.J.; Hamaji, T.; Toyoda, A.; Fujiyama, A.; Neme, R.; Noguchi, H.; Minakuchi, Y.; Suzuki, M.; et al. The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. *Nat. Commun.* **2016**, *7*, 11370. [CrossRef] [PubMed]
55. Delaux, P.M.; Nanda, A.K.; Mathé, C.; Sejalon-Delmas, N.; Dunand, C. Molecular and biochemical aspects of plant terrestrialization. *Perspect. Plant Ecol. Evol. Syst.* **2012**, *14*, 49–59. [CrossRef]
56. de Vries, J.; Archibald, J.M. Plant evolution: Landmarks on the path to terrestrial life. *New Phytol.* **2018**, *217*, 1428–1434. [CrossRef] [PubMed]
57. Oren, A. The ecology of *Dunaliella* in high-salt environments. *J. Biol. Res. Thessalon.* **2014**, *21*, 23. [CrossRef] [PubMed]
58. Ricci, M.; Tilbury, L.; Daridon, B.; Sukalac, K. General principles to justify plant biostimulant claims. *Front. Plant Sci.* **2019**, *10*, 494. [CrossRef] [PubMed]
59. Rouphael, Y.; Colla, G. Toward a sustainable agriculture through plant biostimulants: From experimental data to practical applications. *Agronomy* **2020**, *10*, 1461. [CrossRef]
60. Calvo, P.; Nelson, L.; Kloepper, J.W. Agricultural uses of plant biostimulants. *Plant Soil* **2014**, *383*, 3–41. [CrossRef]
61. Du Jardin, P. Plant biostimulants: Definition, concept, main categories and regulation. *Sci. Hortic.* **2015**, *196*, 3–14. [CrossRef]
62. Hanschen, E.R.; Starkenburg, S.R. The state of algal genome quality and diversity. *Algal Res.* **2020**, *50*, 101968. [CrossRef]
63. Weckwerth, W. Green systems biology—From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *J. Proteom.* **2011**, *75*, 284–305. [CrossRef] [PubMed]
64. Guarnieri, M.T.; Pienkos, P.T. Algal omics: Unlocking bioproduct diversity in algae cell factories. *Photosynth. Res.* **2015**, *123*, 255–263. [CrossRef] [PubMed]
65. Renganathan, P.; Yakupova, A.; Filippov, A.; Larionova, I.; Sushchenko, R.; Mufazalova, A.; Khilazhetdinova, L.; Gaysina, K.; Gaysina, L.A. Strain-Specific Microalgal and Cyanobacterial Suspensions Modulate Germination Kinetics and Early Seedling Vigor in Cucumber. *Horticulturae* **2026**, *12*, 414. [CrossRef]
66. Renganathan, P.; Ukhatkina, K.B.; Van Erp, I.I.; Mufazalova, A.; Sukhanova, N.V.; Gaysina, L.A. Comparative Evaluations of Commercial Seaweed Extract Formulations on Germination, Biomass Accumulation and Early Seedling Growth in *Capsicum annuum*. *Horticulturae* **2026**, *12*, 799. [CrossRef]
67. Renganathan, P.; Ukhatkina, K.B.; Almaz, M.; Van Erp, I.I.; Gaysina, L.A. Dosage-Specific Effects of Commercial Seaweed Extracts on Seedling Growth, Biomass Accumulation and Root Growth Traits of *Capsicum annuum*. *Int. J. Plant Biol.* **2026**, *17*, 70. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.