International Journal of
Molecular Sciences

MDPI

Article

Identification of Two Flip-Over Genes in Grass Family as Potential Signature of C4 Photosynthesis Evolution

Chao Wu and Dianjing Guo *

State Key Laboratory of Agrobiotechnology, School of Life Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; wuchao@link.cuhk.edu.hk
* Correspondence: djguo@cuhk.edu.hk

Abstract: In flowering plants, C4 photosynthesis is superior to C3 type in carbon fixation efficiency and adaptation to extreme environmental conditions, but the mechanisms behind the assembly of C4 machinery remain elusive. This study attempts to dissect the evolutionary divergence from C3 to C4 photosynthesis in five photosynthetic model plants from the grass family, using a combined comparative transcriptomics and deep learning technology. By examining and comparing gene expression levels in bundle sheath and mesophyll cells of five model plants, we identified 16 differentially expressed signature genes showing cell-specific expression patterns in C3 and C4 plants. Among them, two showed distinctively opposite cell-specific expression patterns in C3 vs. C4 plants (named as FOGs). The in silico physicochemical analysis of the two FOGs illustrated that C3 homologous proteins of LHCA6 had low and stable pI values of ~6, while the pI values of LHCA6 homologs increased drastically in C4 plants *Setaria viridis* (7), *Zea mays* (8), and *Sorghum bicolor* (over 9), suggesting this protein may have different functions in C3 and C4 plants. Interestingly, based on pairwise protein sequence/structure similarities between each homologous FOG protein, one FOG PGRL1A showed local inconsistency between sequence similarity and structure similarity. To find more examples of the evolutionary characteristics of FOG proteins, we investigated the protein sequence/structure similarities of other FOGs (transcription factors) and found that FOG proteins have diversified incompatibility between sequence and structure similarities during grass family evolution. This raised an interesting question as to whether the sequence similarity is related to structure similarity during C4 photosynthesis evolution.

Keywords: C4 photosynthesis; comparative transcriptomics; transcriptome signature; Differentially Expressed Genes (DEGs); flip-over genes (FOGs); LHCA6; PGRL1A; AlphaFold2; sequence similarity; structure similarity; C4 traits engineering

Citation: Wu, C.; Guo, D. Identification of Two Flip-Over Genes in Grass Family as Potential Signature of C4 Photosynthesis Evolution. *Int. J. Mol. Sci.* **2023**, *24*, 14165. https://doi.org/10.3390/ijms241814165

Academic Editor: Pedro Martínez-Gómez

Received: 9 July 2023
Revised: 5 September 2023
Accepted: 13 September 2023
Published: 15 September 2023

1. Introduction

Photosynthesis is the ultimate energy source from solar power and supports most life forms on earth [1,2]. Based on the different initial carbon fixation processes, photosynthesis can be mainly classified into two subtypes: C3 photosynthesis and C4 photosynthesis [2]. Higher plants that carry out C4 photosynthesis have Kranz anatomy presented in the leaf tissue, hence the photorespiration process is prohibited. Plants therefore obtain more organic carbon and accumulate more biomass during the photosynthetic process, which allows them to adapt better to extreme environmental conditions such as heat and drought (Figure 1). This kind of adaptation has broad pleiotropic and epistatic consequences on C4 plants. For instance, they exhibit better water use efficiency and better heat tolerance [3,4]. In this case, certain unique genes have been connected to abiotic stress, implying that some key genes may serve as signatures for predicting the implications of sophisticated environmental stresses [5–11]. However, how C4 photosynthesis evolves from the ancestral C3 types remains unclear, especially whether the dynamics of gene expression is conserved among diverse higher plants. As we know, C4 photosynthesis has evolved independently in grass lineage [12]. The basis of plant evolution is the mutations of its DNA sequence.

These mutations are reflected in the amino acid sequence divergence. Throughout the evolution of C4 photosynthesis, some proteins display a divergent expression, localization, and functionality [13], while others are more conserved. For example, the C4 enzyme coding genes are differentially presented between two photosynthetic cells, namely, mesophyll and bundle sheath. Some transcription factors and metabolite-related genes are also differentially regulated and undertake pivotal functions in the photosynthetic process [14,15]. All the differentially regulated genes compose the subtle differentiation of C4 photosynthesis.

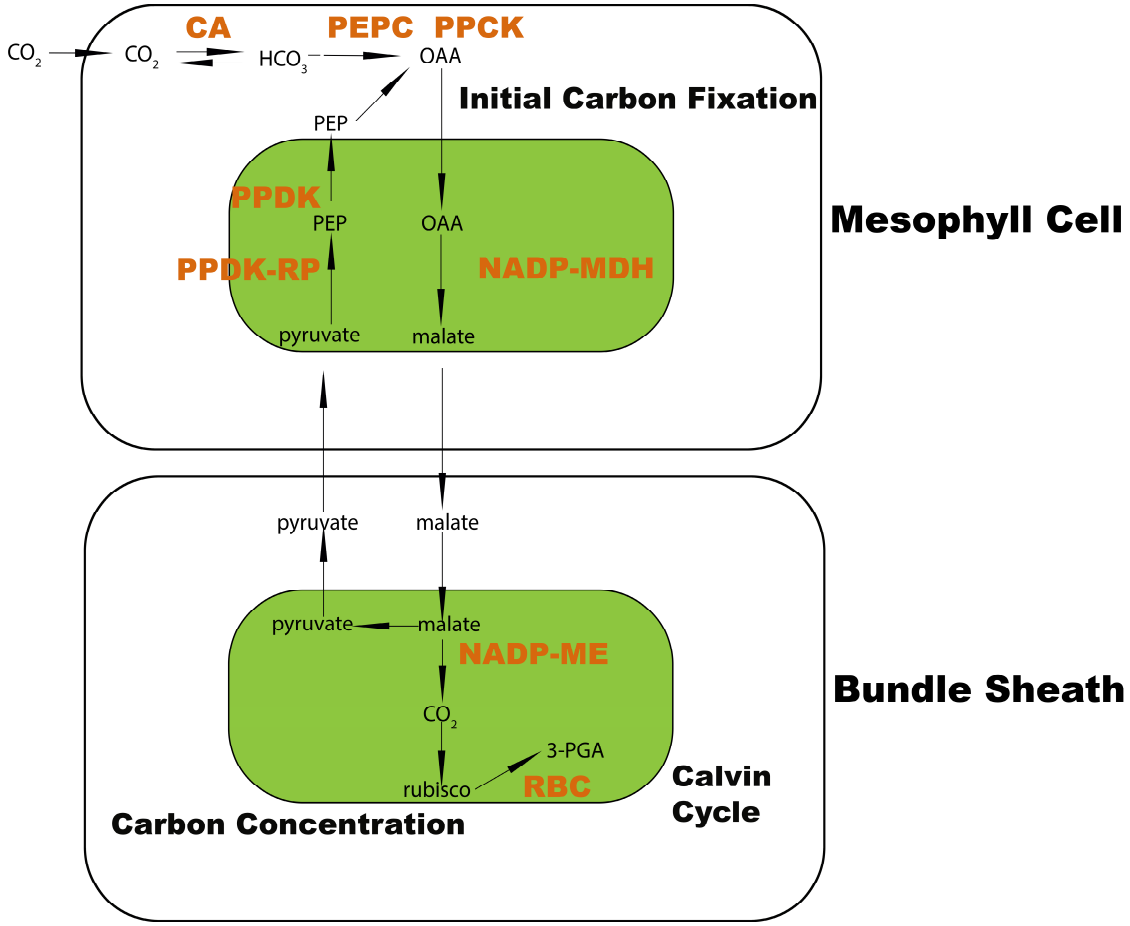

Figure 1. Typical NADP-ME subtype pathway illustration of C4 photosynthesis. Black frames represent two photosynthetic cells: bundle sheath and mesophyll. Green blocks are simplified chloroplasts. All carbon product names are in black, and all key C4 enzymes are in bold orange. Three vital processes of C4 photosynthesis are highlighted in bold black.

As the first step to uncovering the mechanisms behind C4 photosynthesis evolution, studying the Differentially Expressed Genes (DEGs) between mesophyll and bundle sheath using transcriptome data has been a challenging task [16,17]. Dating back to the initial discovery of photosynthetic DEGs, scientists identified some key C4 enzymes differentially expressed in bundle sheath cells, such as PEP carboxylase and RuBisCO [18]. To decipher the gene transcripts accumulation in maize leaf blade, scientists separated the two photosynthetic cells and measured the differential gene expression using a microarray technique [19]. The maize leaf developmental gradient DEGs and cell-specific DEGs were also identified using Illumina sequencing [20]. To compare the transcriptome between C3 and C4 plants, scientists used developing leaves from maize and *Oryza sativa* (rice) and established a statistical model to simulate the changes between C3 and C4 plants during leaf development. *Setaria viridis* has recently been adopted as a new C4 model plant due to its simple genetics. For example, scientists combined comparative transcriptomics of diverse C4 plants including

Setaria viridis and C3 rice to identify known and novel C4-related DEGs [12,21,22]. Some DEGs are photosynthetic genes that undertake key roles in various parts of photosynthesis [13,16,20,23]. To compare the C3 and C4 transcriptomes, the proper selection of the C3 model plant is pivotal. Scientists often choose panicoid grass for such study [22], whilst the consistency of cell-specific DEGs among different grass has not been discussed yet. Boosted by the subtle tissue separation technology such as laser microdissection, scientists successfully collected and sequenced the ultra-pure bundle sheath and mesophyll in two C3 plants, *Arabidopsis thaliana* and *Oryza sativa*, and identified new functions of bundle sheath from a physiology perspective [24,25]. The discoveries of known or novel DEGs in different sections such as developmental, cell-specific, or cross-species may help scientists narrow down the spectrum of C4 candidate genes for genetic engineering in C3 plants.

Among various types of DEGs, flip-over genes (FOGs) were defined as genes that display opposite cell-specific expression patterns between C3 and C4 plants. The first report about flip-over genes was from the comparison between maize (C4 NADP-ME) and *Cleome gynandra* (C4 NAD-ME), where 18 transcription factors were reported to show distinct expression preference between bundle sheath and mesophyll in maize and *Cleome gynandra* [26]. However, the authors did not reach further to investigate the structural basis of these FOGs.

Transcriptome signature discovery is a reliable approach to profiling gene expression during vital biological processes [27–29]. Generally, scientists focus on the conserved gene cascades and specifically expressed genes to meet the customized selective criteria [30]. In human biology studies, signature genes are well illustrated for further gene function studies [31]. In plant studies, however, few transcriptome signatures were identified due to limited data sets, especially for the photosynthesis study in the grass family [32].

As the end products of gene expression, proteins are often assembled as monomers or polymers to participate in diverse biological processes [33,34]. In recent years, structural scientists apply crystallography followed by X-ray or Nuclear Magnetic Resonance (NMR) [35,36] and Cryo-Electron Microscopy (Cryo-EM) to solve the structural folding of proteins of their interests [37]. Given that, the resources of universal protein folding information have been accumulated in the past decades [38]. The development of computational tools also promotes the discovery of protein three-dimensional structures [39], e.g., Swiss-model [40] and Phyre^2 [41], etc. Unfortunately, the predictive capacity and accuracy of these tools are still limited [39,42].

In recent years, artificial intelligence (AI) has been widely applied in protein structure prediction. For example, AlphaFold2 platform achieved a median score of 92.4 GDT overall on the 14th Critical Assessment of Techniques for Protein Structure Prediction (CASP) assessment [43]. Its high performance was also demonstrated in protein complex prediction and peptide–protein docking in microorganism [44,45].

The sequence and structural similarities of proteins are often regarded as being equivalenced [46]. In terms of homologous proteins, the folds of proteins with sequence homology > 50% have close tertiary structures in general [47]. It is widely believed that the structure of proteins is more conserved than their sequence during evolution [48], or at least shows a linear relationship [49,50]. However, counterexamples that facilitate our limited knowledge about the protein sequence structure relationship diversity also exist. For example, some homologous CheY-like protein pairs with low sequence similarity (partial correlation coefficients were not statistically significant) exhibited very similar structural topology (based on distance matrix analysis of the C-terminal regions in native structures of these proteins) [51]. On the other hand, due to quaternary protein–protein interactions, some proteins with high pairwise sequence similarity (sequence identity $\geq$ 50%) presented largely diverged tertiary structural geometry and occupied 22% of the total sampled protein folds. This phenomenon has been discussed in TonB and ABL proteins from *E.coli* and *Drosophila melanogaster*, respectively [52], while it has not been reported in plants so far.

Nowadays, various bioinformatic tools have been developed to predict protein structure, function, and physicochemical properties [42,43,53,54]. These tools can facilitate our

investigation of different proteins of our interests. Research about the molecular basis of the shifts from C3 to C4 photosynthesis has been well established in diverse corresponding phenotypes, such as leaf anatomy, chloroplast formation, and development of C4-specific Kranz anatomy, as well as the relevant regulatory genes that have been reported as clues to trace the trajectories of C4 evolution [14,55,56]. However, due to the uncoordinated homologous gene expression patterns or the low expression levels in our selected species, these genes cannot be utilized as reliable features to predict photosynthesis types. In addition, most machine learning-guided biomarker identifications start with a large gene expression matrix and end up with gene sets that pass the selective thresholds [57].

In this study, with the research hypothesis that cell-specific signature genes can be used to predict C4 photosynthesis type and that they have unique sequence/structure similarities from an evolutionary standpoint, we used the cutting-edge AlphaFold2 platform to predict the protein structures of homologous FOGs in five C3 and C4 model plants, and investigated the relationship between sequence and structure similarity. Our work expands our knowledge on C4 photosynthesis protein evolution, and possibly provide guidance for C4 photosynthesis engineering in C3 plants [58,59].

2. Results

2.1. Transcriptome Divergence between C3 and C4 Plants

To comprehensively compare C3 and C4 transcriptomes, we investigated the high-quality replicates of cell-specific (bundle sheath and mesophyll) transcriptomes of five grass species, including *Arabidopsis thaliana* (C3), *Oryza sativa* (C3), *Setaria viridis* (C4), *Zea mays* (C4), and *Sorghum bicolor* (C4). Pearson correlation analysis showed high consistency between each replicate, indicating good reproducibility in general, except for the first replicate of *Sorghum bicolor*. We hence removed this replicate in further analysis (Figure 2). We analyzed the cell-specific differentially expressed genes (DEGs) and found that both C3 species presented a smaller portion of DEGs, compared to the three C4 species (Figure 3), suggesting the differentiation of bundle sheath and mesophyll has a greater impact on C4 transcriptomes as the formation of Kranz anatomy recruits vast differentially expressed genes in C4 plants. Among them, *Oryza sativa* and *Zea mays* presented the DEGs' lowest and highest proportion of 10.58% and 21.72%, respectively (Figure 3A). We then analyzed the DEGs intersection among the five species and identified a group of 265 DEGs shared by both C3 and C4 species. Meanwhile, 645 DEGs shared by *Setaria viridis*, *Zea mays*, and *Sorghum bicolor* were identified as C4-specific ones (Figure 3B).

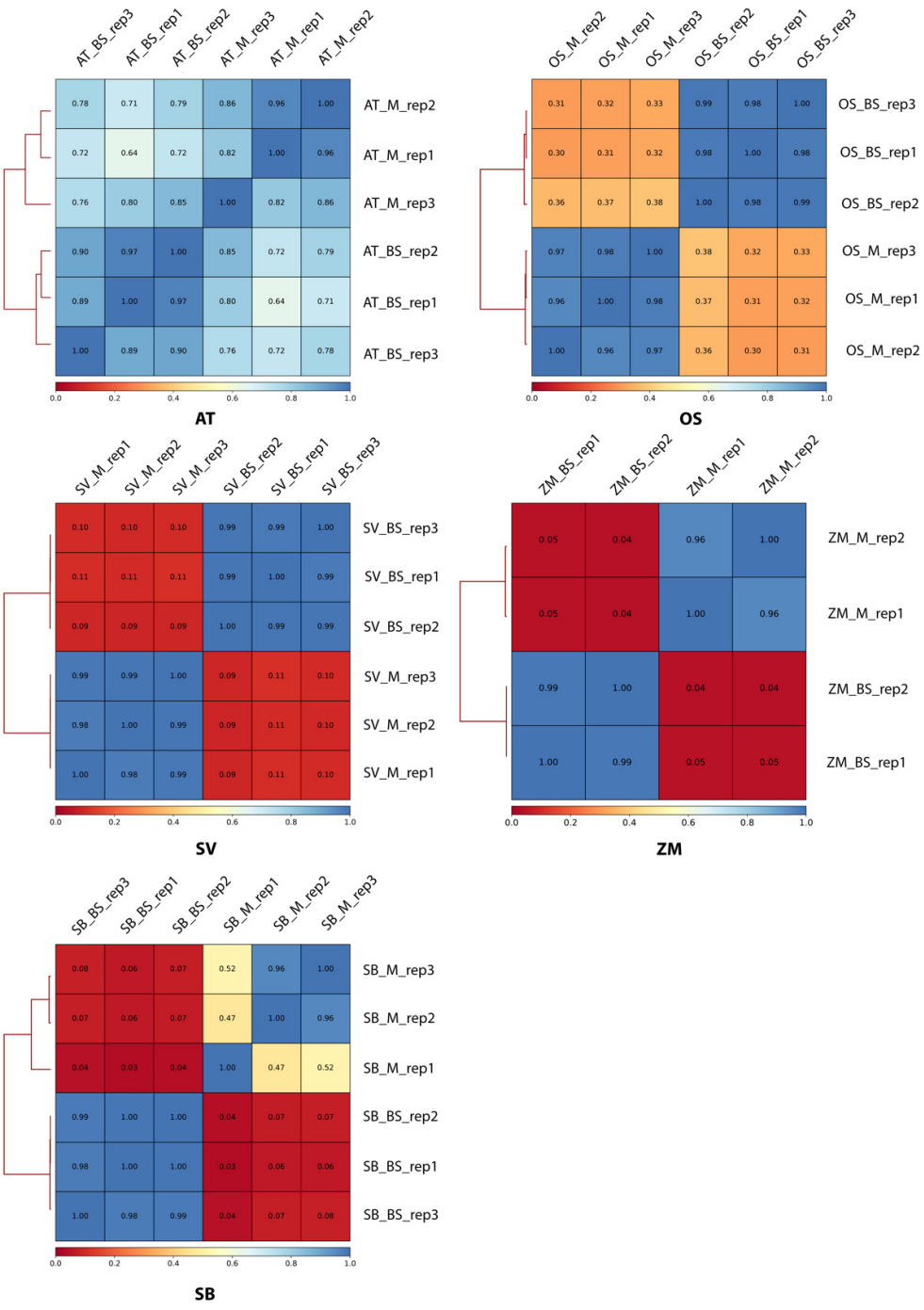

Figure 2. Internal consistency between each replicate of every selected photosynthetic model plant. Pearson correlation coefficient heatmap diagram of all collected replicates in five plants; a higher value represents higher consistency. All replicates are of high quality for reproduction except for the first replicate in SB, hence we removed it from the datasets in downstream analysis. We can see from the figure that the inter-cell-type consistency is well enough.

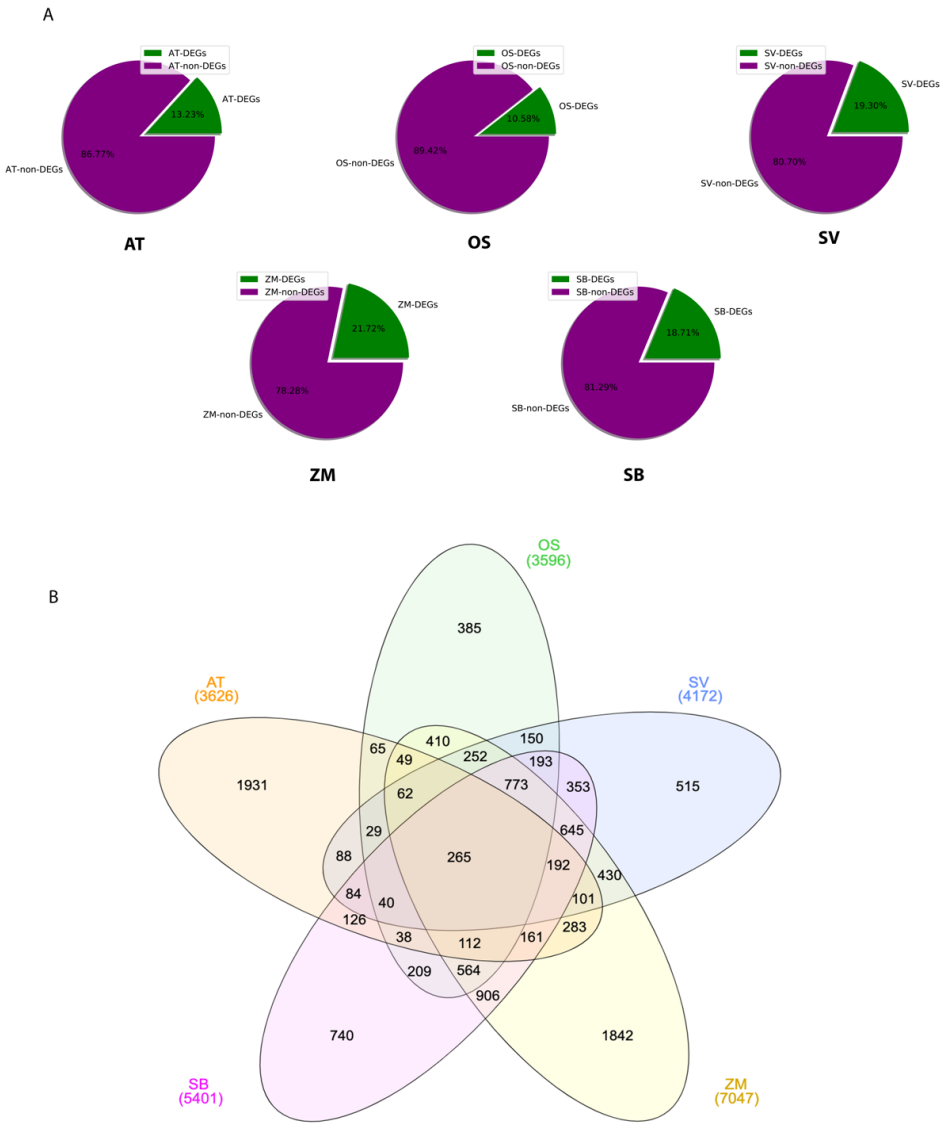

Figure 3. Comparative transcriptome analysis between C3 and C4 plants. (**A**) Cell-specific DEG proportions in each selected model plant. All data from five model plants are illustrated in pie charts. Purple represents non-DEGs, and green shows DEGs. (**B**) Multi-species Venn diagram of cell-specific DEGs from each model plant and their intersections in different sections. Among all five grass species, 265 C3 and C4 common DEGs are presented in the center of the Venn chart.

2.2. Common and Specific Pathways Enriched in C3 and C4 Plants

Based on our DEGs overlapping analysis results, we speculated that although evolutionary divergence exists between C3 and C4 plants, some genes remain active in both C3 and C4 species. The 265 common genes were mostly enriched in twelve pathways, e.g., the generation of precursor metabolites and energy, which is associated with the energy flow in photosynthesis; and hydrocarbon biosynthetic process, which is crucial for carbon transformation and delivery through photosynthesis. Apart from photosynthesis, C3 and C4 plants also have other divergent pathways in common, e.g., oxylipin biosynthetic and metabolic processes. In animals and humans, oxylipins act as pivotal precursors associated with diseases such as Alzheimer's disease. In plants, oxylipins take part in the control of plant lifespan, reproductivity, and the defense to biotic stress [60] (Figure 4A). On the other hand, six C4-specific pathways (Figure 4B) were mainly associated with nucleobase metabolic and catabolic processes.

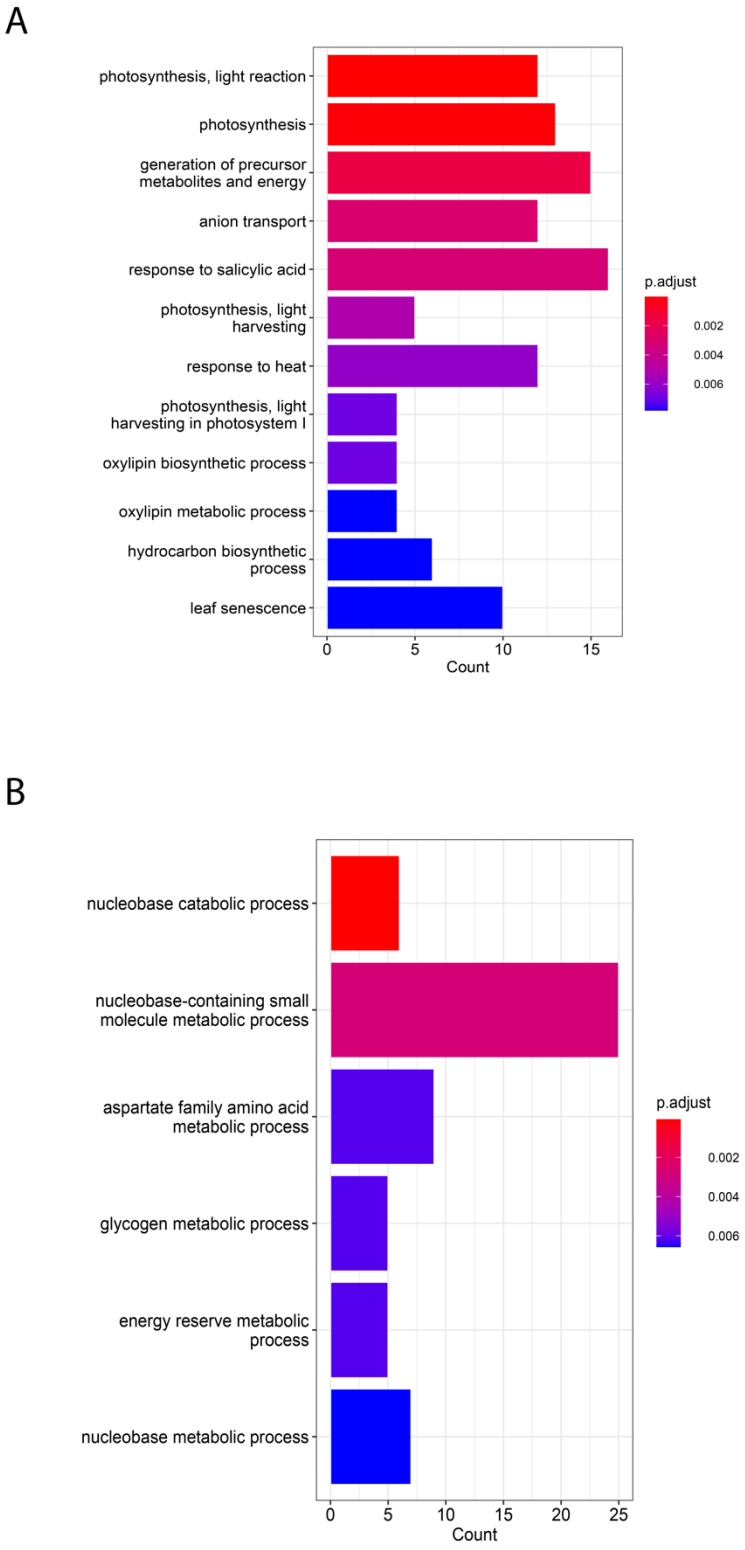

Figure 4. Gene ontology (GO) enrichment analysis of intersected DEGs. (**A**) GO enrichment analysis of C3 and C4 common DEGs. All twelve pathways are identified at the transcriptome-wide level from all five species. (**B**) GO enrichment analysis of C4-specific DEGs. All six pathways are identified at the transcriptome-wide level, and only from C4 species.

2.3. Transcriptome Signature Discovery in C3 and C4 Plants

To gain knowledge about C3 and C4 plant convergence and divergence, 16 genes with cell-specific expression patterns were identified as C3 and C4 common signature genes. Figure 5A and Figure S1 showed the log2FoldChange values of these genes between bundle sheath and mesophyll of C3 and C4 plants. Among these signature genes, two flip-over genes (FOGs) showed opposite expression patterns in mesophyll and bundle sheath. When the protein–protein interactions among the signature genes were predicted (Figure 5B), we observed two subsections of interactive genes formed two weakly connected groups. One was the light-harvesting complex (LHC) for light energy capture [61], in which the two FOGs were included. Another was the tricarboxylic acid cycle (TCA cycle) for ATP synthesis [62].

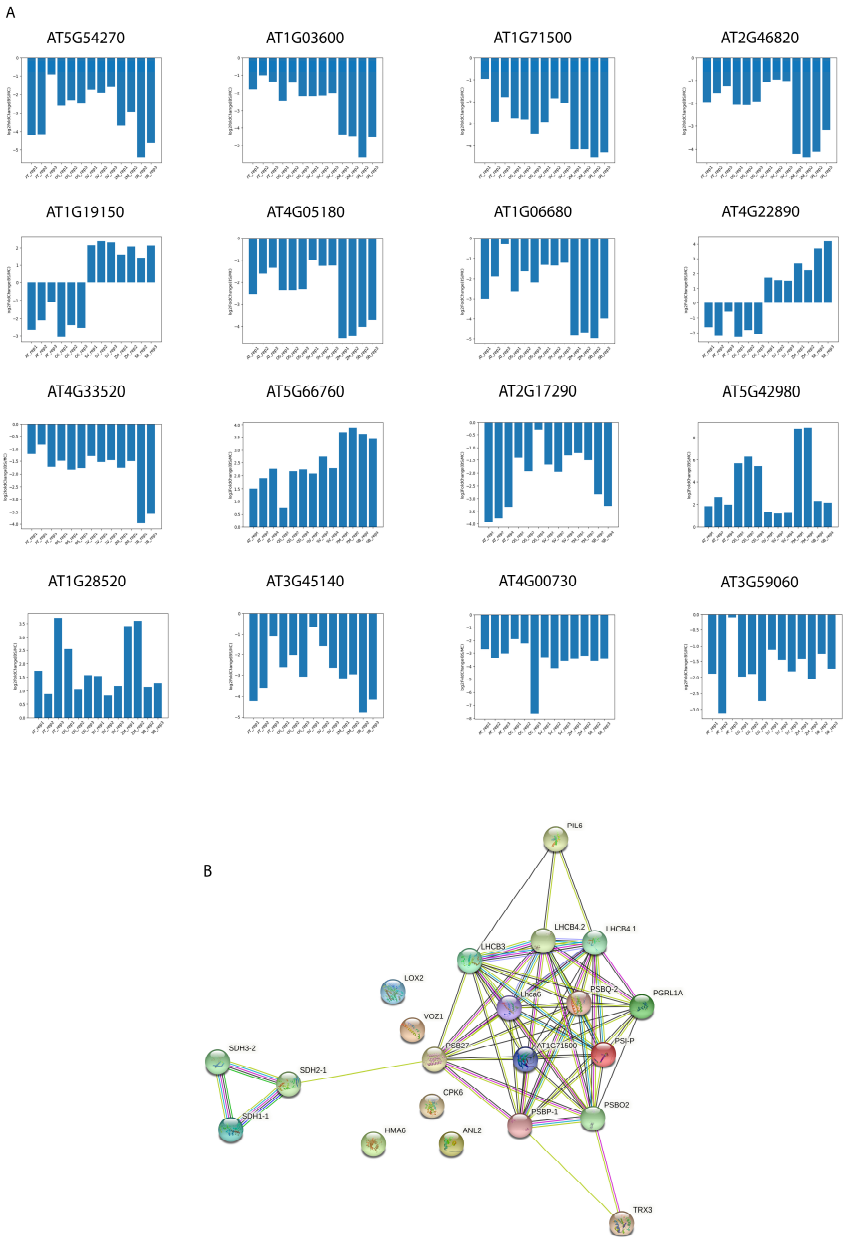

Figure 5. C3 and C4 transcriptome signature genes discovery. (**A**) Log2FoldChange values between bundle sheath and mesophyll of each replicate in five model plants. In each subplot, every column represents the log2FoldChange values. (**B**) Protein–protein interaction (PPI) prediction of 16 selected transcriptome signature genes. Each node shows a signature gene.

2.4. Domain Analysis and Protein Folding Prediction of Two FOGs

We examined the Pearson correlation coefficients against photosynthesis types (C3 or C4) and found that the cell-specific expression pattern of both FOGs were highly correlated with the photosynthetic types (over 0.9), indicating they may be used for photosynthetic type prediction (Figure 6A). We also retrieved the full sequences of FOG protein homologs in the five species and conducted protein folding prediction using AlphaFold2. To verify the performance of AlphaFold2 on plant protein structure prediction, we selected two plant proteins STP10 and ReAV, which were not included in the training dataset during the training process. The prediction accuracy was rather high as illustrated in Figure S2. Figure 6B illustrates the pLDDT values of each FOG protein structure prediction experiment. pLDDT is a metric for evaluating the prediction performance of AlphaFold2, and higher pLDDT means better prediction accuracy. AlphaFold2 achieved similar prediction performance on LHCA6 homologs. For membrane protein PGRL1A homologs, the pLDDT values were not stable. Since the sequence divergence of PGRL1A was much higher than that of LHCA6, it may also bring differences between each homolog. We then extracted the domain sequences of the two FOG protein homologs in five species and performed multiple sequence alignment (MSA) and visualized the amino acid mutations for all the selected protein domains. As shown, although the sequences were highly conserved, certain C4-specific mutations at amino acid level still existed (Figure 6C). From the sequence logo of the two FOG proteins (Figure 6D), the LHCA6 protein sequence was more conserved compared to PGRL1A. Interestingly, the consensus sequence for each protein was not continuous. We deduced that this is likely due to the genomic events that occurred during the course of evolution that break the original continuous sequence into segments. Such changes may facilitate the C4 plants' adaptation to environmental cues.

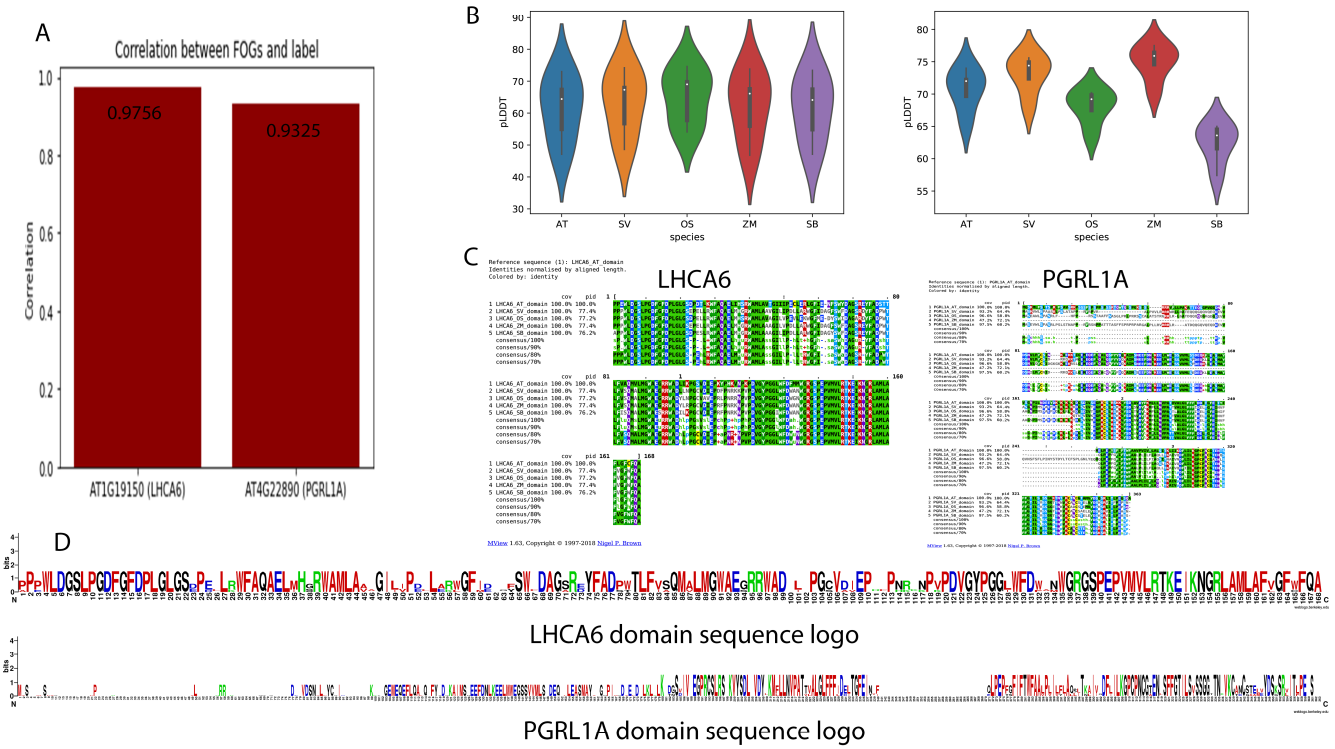

Figure 6. Domain analysis and protein three-dimension structure prediction metrics of two FOGs. (**A**) Pearson correlation between two FOGs and labels (species photosynthetic types). (**B**) pLDDT values of predicted structures from five species. (**C**) Domain sequence alignments of two FOGs in five model plants. (**D**) Domain sequence logos of two FOGs.

2.5. Amino Acids Composition Analysis and Protein Solubility Prediction of FOGs

The amino acid compositions of each FOG homolog domain were examined. The amino acids were characterized into four groups based on their physicochemical properties, including hydrophobic amino acids, amphipathic amino acids, polar amino acids, and charged amino acids. Most proteins examined consisted of hydrophobic amino acids that facilitate protein folding into a relatively stable conformation and maintain relevant functions (Figure 7A), followed by charged amino acids. In *Zea mays*, PGRL1A contained the fewest number of charged amino acids associated with the lowest pI value (Figure 7B). Polar amino acids are highly associated with protein solubility, and we found that they were correlated with protein solubility of LHCA6. This also held true for PGRL1A proteins, except for the *Zea mays* homolog. Furthermore, we found the pI values for LHCA6 protein in C3 plants AT and OS were relatively low and stable. While in C4 plants SV, ZM, and SB, the pI values increased drastically.

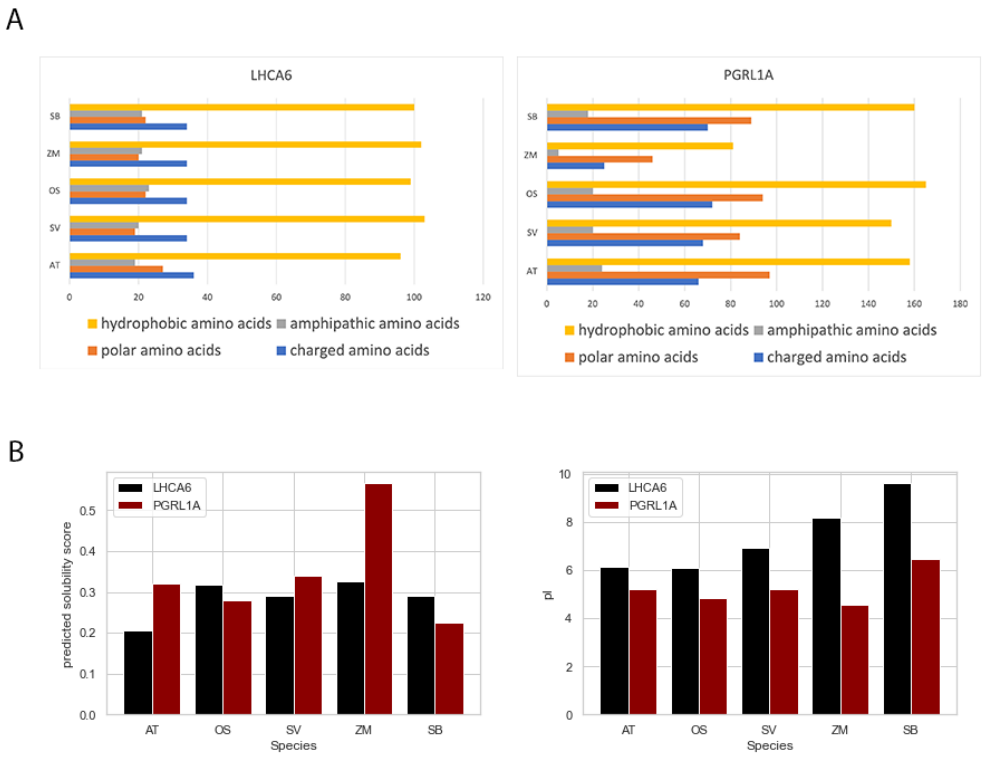

Figure 7. Amino acids composition analysis and protein solubility prediction of two FOGs. (**A**) Amino acids composition analysis based on protein domain sequences of two FOGs. (**B**) Protein solubility and pI value prediction of two FOGs.

2.6. Sequence Motif Discovery in FOGs

To investigate the changes in FOG proteins during C4 photosynthesis evolution, we conducted sequence motif analysis for each homologous protein. The top-ranked motif in *Arabidopsis thaliana* LHCA6 was annotated as a shorter consensus motif in *Oryza sativa* (Figure 8). The third consensus motif that begins with "WFD" also presented in *Oryza sativa* and *Setaria viridis* in different length and composition, but not in *Zea mays* and *Sorghum bicolor*. This indicated that *Setaria viridis* is likely more closely related to C3 plants compared to *Zea mays* and *Sorghum bicolor*. Furthermore, between *Zea mays* and *Sorghum bicolor*, the top-ranked motif "RFKERKN" appeared twice in both plants. Unlike LHCA6, the consensus motif identified in PGRL1A was less. The last-ranked motif consensus in *Arabidopsis thaliana* was also presented in *Oryza sativa*, *Setaria viridis*, and *Sorghum bicolor* with amino acid substitutions, but not in *Zea mays*. For maize, only short motif consensus was identified.

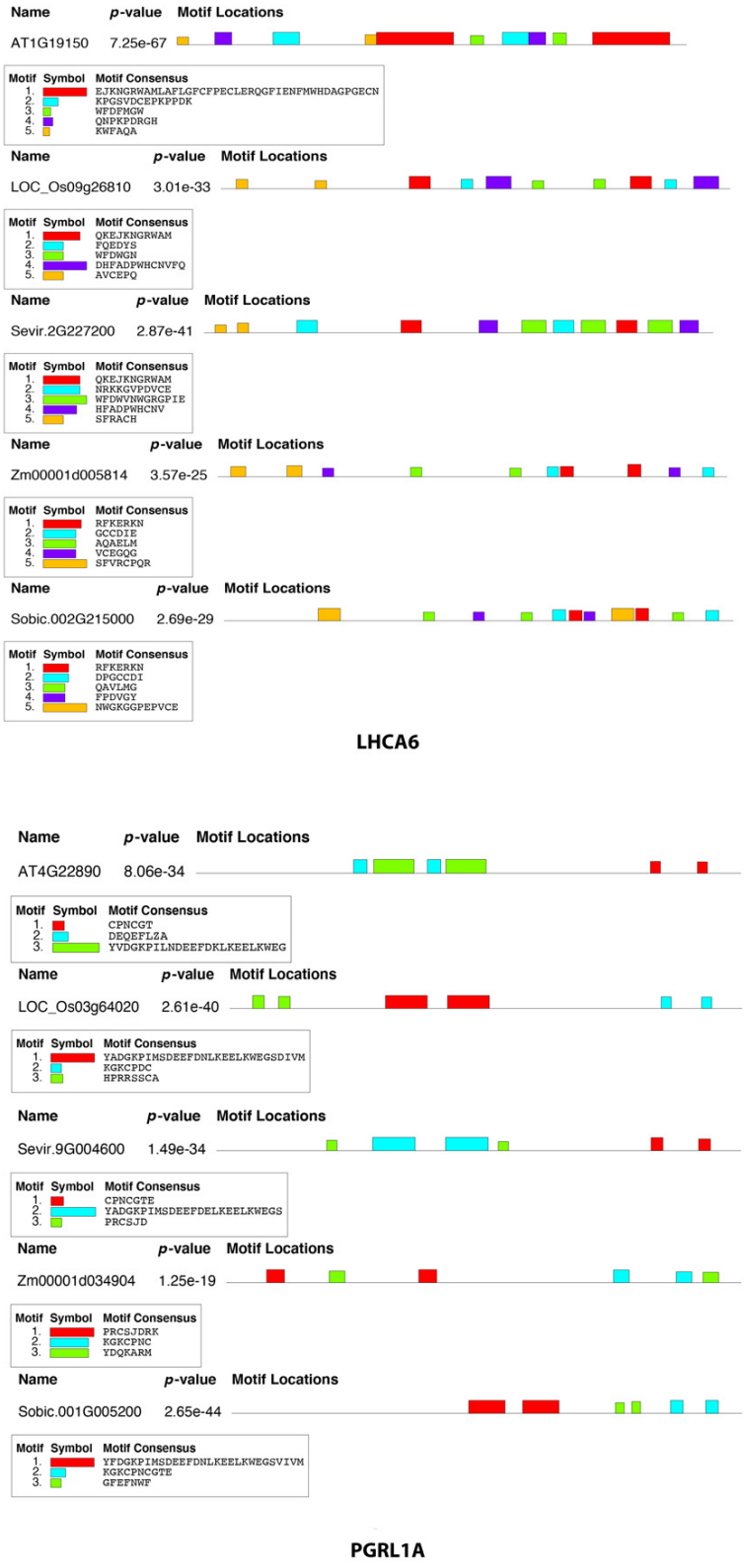

Figure 8. Sequence motif discovery of two FOGs. All five homologs of LHCA6 are listed on the left panel, and on the right panel are PGRL1A homologs. Consensus sequences are listed in the chart.

2.7. Sequence and Structure Similarities Comparison of FOG Proteins

For LHCA6, both global sequence and structure similarities were over 0.8. The consistent dual similarities showed consistency with the grass family phylogeny (Figure 9). We extracted the highest (SV-SB) and lowest (OS-ZM) structure-similarity pairs with similarity values of 0.88655 and 0.82752, respectively. Compared to the C3-C4 pair, the C4-C4 pair had closer phylogenetic relationship and hence a higher structure similarity. Moreover, we also found that the secondary structures in both pairs aligned very well, while the intrinsically disordered regions (IDRs) were differentially aligned. Precisely, the IDRs in the SV-SB pair were physically close. However, the spatial positions of the IDRs diverged in the OS-ZM pair. This finding indicated that when the global structure-similarity was high and secondary structures aligned very well, the IDRs would largely affect the overall structure similarity.

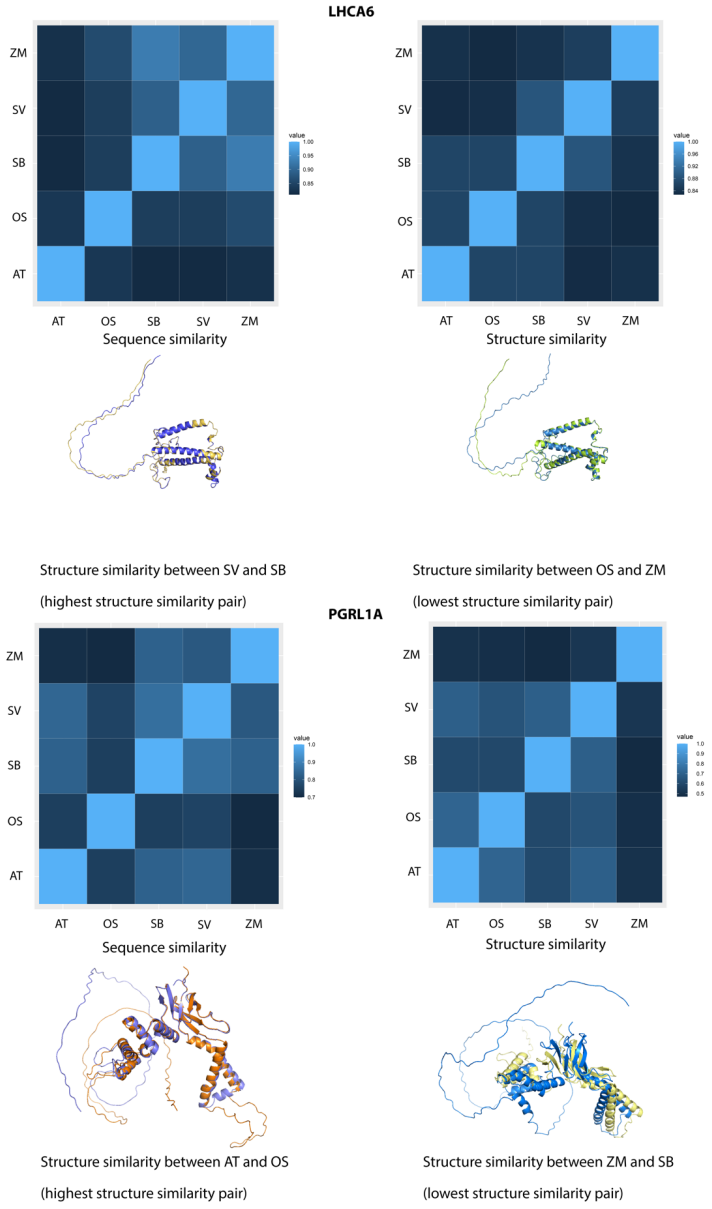

Figure 9. Sequence and structure similarities of two FOGs. The left panel is a similarity illustration of LHCA6, and another part is for PGRL1A. Each block in the heatmaps represents a pairwise comparison between homologous proteins. The brighter the color means the higher the similarity score.

On the other hand, the sequence and structure similarities of PGRL1A were over 0.7 and 0.5, respectively. In Figure 8, the AT-OS pair shared a sequence similarity value of 0.75, while for AT-SB and AT-SV pairs the values were 0.83 and 0.84, respectively. Accordingly, the structure similarity of these three pairs were 0.71106, 0.60251, and 0.69203, respectively. As we extracted the structure alignments of the AT-OS (highest structure similarity) and the ZM-SB (lowest structure similarity) pairs, we found the alignment of secondary structures between AT and OS was very good. While in the ZM and SB pair, the structural alignment of secondary structures was drastically shifted and the TM-score was only 0.47017, suggesting the protein folds were not similar and the homology relationship was weak although they shared a high sequence similarity of 0.83 (Figure 9). From the case of PGRL1A, we found that the sequence and structure similarities showed local inconsistency.

It is widely assumed that structural features are often closely related to sequence composition. As reported by previous research, protein pairs with sequence identity higher than 35–40% are very likely to be globally structurally similar as well. Our findings provided refinements of this assumption that local inconsistency may affect the sequence and structure relationship. To further validate our finding, we selected three transcription factor type FOG proteins from a previous study by Aubry et al., namely SIGC, ZFP8, and ZHD10. Unlike ZFP8, both sequence and structure similarities of SIGC and ZHD10 were over 0.5, while the local comparison in SIGC was quite interesting. Precisely, we found that even if the sequence similarities between AT-SV/ZM (over 0.5) and OS-SV/ZM (over 0.7) were different, their structure-similarities were quite similar (over 0.5), similar to the situation in PGRL1A. In addition, although the ZFP8 protein pairs showed inconsistent sequence (over 0.5) and structure (over 0.2) similarities, the global homology relationship was maintained based on the two similarity heatmaps (Figure S3). We therefore conclude that FOG proteins have diversified incompatibility between sequence and structure similarities during grass family evolution, especially for the pairs that share low/diverged sequence similarity but have high/similar local structure similarity.

We then selected the top three transcription factor type non-FOG proteins from Aubry et al. for comparison, namely EFM, RL6, and SIGB. We also observed that EFM proteins had high sequence similarity (over 0.5) while showing drastically low global structure similarity (over 0.19), which suggested that non-FOG proteins also present highly diverged three-dimensional structures due to the complexity of protein evolution (Figure S4).

3. Discussion

It is known that C4 photosynthesis evolved independently [12], while the evolutionary events that occurred during the process have not been clarified. In this study, we compared the transcriptomes between C3 and C4 plants, and identified 16 signature genes differentially expressed between mesophyll and bundle sheath cells. Using differentially expressed genes as biomarkers to predict specific diseases is a commonly used bioinformatics strategy. Such analysis can narrow down the unique genes that are closely connected with targeted biological process or treatments. When we selected the DEGs, all candidate genes should show differential expression patterns in all selected species. Based on such criteria, only two FOGs are identified. Using the cutting-edge deep learning model AlphaFold2 and other protein informatics tools, we analyzed the sequences and structures of the two FOG proteins and found that local sequence and structure similarities showed inconsistency. This finding was consistent with previous reports [34,63,64]. However, from the comparison of sequence and structure similarities, we still identified novel structural divergence between homologous proteins, especially for the rearrangements of secondary structures in PGRL1A proteins.

Prediction of protein three-dimensional structure has been a crucial biological and computational challenge for the past few decades [43]. Deep learning technology utilizes numerous protein folding data to establish prediction models based on amino acid sequences. In recent years, AlphaFold2 has played a pivotal role in new structure discovery [43] with its well-established prediction platform based on Google Colaboratory [65].

We adopted Amber relaxation and templates during the prediction process to maximize the prediction performance. Through verification of the performance on plant protein prediction, our results indicated that AlphaFold2 was robust and reliable.

Signature genes shared by C3 and C4 plants have great potential to serve as the predictive features for classification tasks from a machine-learning perspective. In this study, we successfully identified 16 C3 and C4 signature genes in the grass family, including two FOGs. Compared to LHCA6, PGRL1A seems to be more active during plant evolution. PGRL1A is the hub gene of electron transport in photosynthesis [66,67]. The highly variant structures among homologous PGRL1A proteins may indicate the complex evolution of electron transport process from C3 to C4 plants. LHCA6 is the key component of the light-harvesting complex [68,69]. Divergence in its structure may directly affect the effectivity and efficiency of solar energy capture by altering the binding affinity and specificity, and results in the biomass accumulation differences between C3 and C4 plants. Our study suggested that residue preference may occur during the folding of FOG proteins, which may contribute to the diverse functionalities. In terms of homologous species, our results investigated their evolutionary relationship in three layers, namely, gene expression level, motif occurrence, and protein structure similarity. From the first two layers, *Setaria viridis* is closer to *Arabidopsis thaliana* and *Oryza sativa*, rather than *Zea mays* and *Sorghum bicolor*, while in protein structure similarity comparison of LHCA6, a highly conserved protein, *Setaria viridis* seems to be closer to *Sorghum bicolor*. This phenomenon raised the great importance of whether point mutations will drastically affect the protein structure and the corresponding measurements and provide a good initiation to investigate with large-scale samples in the future. As reported, point mutations may cause diverged functions of proteins translated from genes that have special expression patterns [70]. And our protein structure similarity navigation coordinates well with the previous findings in mice. Based on our findings, as FOGs are differentially expressed between bundle sheath and mesophyll in C3 and C4 species, an overview of their cross-species subcellular localization may provide an important clue as to whether the shifts in their expression patterns and three-dimensional structures contribute to their functional divergence [71]. Moreover, a precise location of the point mutations would help explain the structural divergence. For proteins like the bHLH transcription factors, the key mutations will affect the structure similarity drastically if they occur in the loop region. For the FOG proteins identified in our study, both are membrane proteins. Compared to the selected transcription factor type FOG proteins, they have higher global sequence similarity. For now, we are not sure whether it is due to certain protein types or the different evolutionary backgrounds.

It must be stated that our hypothesis was based on structural data generated by computational prediction and may not always reflect the natural protein folding. Secondly, plant evolution depends on multiple evolutionary events such as point mutations, chromosomal sequence alterations and number changes. Navigating plant evolution at the genomic level (such as single nucleotide polymorphism) and at the transcriptomic level (such as differential gene expression) may anchor different gene sets for making predictions. Our study thus only gave a glimpse of photosynthesis evolution in higher plants as a convolution method.

4. Perspective

How to better integrate comparative in silico gene evolutionary analysis to assess gene diversity across species remains a great challenge in multi-omics-based systems biology. In terms of its application in plant breeding and genetic engineering of specific metabolism processes such as photosynthesis, different layers of information may serve as guidance for genetic modification. The predictive genomic approaches can boost the detection of allelic-level variants of a single gene [72–74]. In this study, we mainly focus on the expression divergence of several genes related to C4 photosynthesis. We aligned the identified FOGs to their homologous proteins, whilst the genomic variants of these genes have not been discovered. To gain a comprehensive knowledge of these signature genes, integration of genomics, transcriptomics,

and proteomics data from homologous species collected from diverse ecological regions may facilitate our understanding of protein evolution in C4 photosynthesis.

5. Materials and Methods

5.1. Comparative Transcriptome Analysis between C3 and C4 Plants

RNA-Sequencing data were retrieved from the European Nucleotide Archive under project accession PRJNA668247 (https://www.ebi.ac.uk/ena/browser/view/PRJNA668247?show=reads) (accessed on 12 September 2023) [24], PRJNA673407 (https://www.ebi.ac.uk/ena/browser/view/PRJNA673407?show=reads) (accessed on 12 September 2023) [25], and PRJEB5074 (https://www.ebi.ac.uk/ena/browser/view/PRJEB5074?show=reads) (accessed on 12 September 2023) [75] for C3 plant *Arabidopsis thaliana*, *Oryza sativa*, and C4 plant *Setaria viridis*, respectively. C4 plants *Zea mays* and *Sorghum bicolor* transcriptome data were retrieved from NCBI under accessions SRP009063 (https://www.ncbi.nlm.nih.gov/sra/?term=SRP009063) (accessed on 12 September 2023) [76] and PRJEB11652 (https://www.ncbi.nlm.nih.gov/sra/?term=PRJEB11652) (accessed on 12 September 2023) [77]. Raw reads were analyzed by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (accessed on 12 September 2023) for quality control and trimmed by Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) (accessed on 12 September 2023) [78]. The reference genomes were downloaded via JGI Phytozome as *Arabidopsis thaliana* TAIR10 (https://data.jgi.doe.gov/refine-download/phytozome?genome_id=167) (accessed on 12 September 2023) [79], *Oryza sativa* v7.0 (https://data.jgi.doe.gov/refine-download/phytozome?genome_id=323) (accessed on 12 September 2023) [80], *Setaria viridis* v2.1 (https://data.jgi.doe.gov/refine-download/phytozome?genome_id=500) (accessed on 12 September 2023) [81], *Zea mays* B73 RefGen_v4 (https://data.jgi.doe.gov/refine-download/phytozome?genome_id=493) (accessed on 12 September 2023) (https://phytozome-next.jgi.doe.gov/info/Zmays_RefGen_V4) (accessed on 12 September 2023), and *Sorghum bicolor* v3.1.1. (https://data.jgi.doe.gov/refine-download/phytozome?genome_id=454) (accessed on 12 September 2023) [82]. HISAT2 (http://daehwankimlab.github.io/hisat2/) (accessed on 12 September 2023) was used for sequencing reads alignment to reference genomes [83]. The output files conversion from SAM to BAM format was performed by Samtools (https://www.htslib.org/) (accessed on 12 September 2023) [84]. The sorted and indexed BAM files were processed by the plotcorrelation function from deepTools (https://deeptools.readthedocs.io/en/develop/) (accessed on 12 September 2023) to analyze the internal consistency between replicates [85]. The following reads count step was processed by htseq-count from HTSeq 0.11.1 (https://htseq.readthedocs.io/en/release_0.11.1/count.html) (accessed on 12 September 2023) [86], and the count tables were passed to DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) (accessed on 12 September 2023) for differentially expressed genes analysis [87]. Genes with adjusted p-value < 0.05 and the absolute value of log2FoldChange between bundle sheath and mesophyll > 1 were identified as the differentially expressed genes (DEGs) for further analysis. The overlapped DEGs between all five model plants were intersected and plotted by interactivenn (http://www.interactivenn.net/) (accessed on 12 September 2023) [88]. Among them, only C3 and C4 common DEGs and C4-specific DEGs were annotated by clusterProfiler (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) (accessed on 12 September 2023) using gene ontology terms with p-value < 0.01 and q value < 0.05 [89]. To identify transcriptome signature genes, we calculated the log2FoldChange value for each pair of expressions in bundle sheath and mesophyll of DEGs in all twelve pathways enriched as C3 and C4 common. The DEGs with null expression were removed, and the DEGs showing similar expression patterns were kept as the signature genes to identify the features of photosynthesis. Among them, two DEGs that showed opposite cell-specific expression were characterized as flip-over genes (FOGs). In total, 20 transcriptome signature genes between the C3 and C4 species were identified. The visualization of DEGs proportion and the plot of log2FoldChange values between bundle sheath and mesophyll of transcriptome signature genes was performed by Microsoft Excel, python matplotlib (https://matplotlib.org/) (accessed on 12 September 2023) and R version 4.0.4 (https://www.r-project.org/) (accessed on 12 September 2023).

5.2. Protein-Protein Interaction Prediction of C3/C4 Transcriptome Signature Genes

Using STRING, we predicted the putative protein–protein interaction relationship between the selected 16 transcriptome signature proteins (https://string-db.org/) (accessed on 12 September 2023). Each node represents a signature gene, and the edge connecting two nodes represents the interactive relationship between two genes. More edges between two nodes indicate higher confidence.

5.3. Multiple Sequence Alignment for Specific Domains and FOGs Expression Pattern

To identify the domain sequences in each FOG protein, we used Interpro (https://www.ebi.ac.uk/interpro/) (accessed on 12 September 2023) and performed multiple sequence alignment (MSA) using MEGA X [90]. The MSA results were visualized using MView (https://www.ebi.ac.uk/Tools/msa/mview/) (accessed on 12 September 2023). We used weblogo (https://weblogo.berkeley.edu/logo.cgi) (accessed on 12 September 2023) to create the sequence logo for each protein domain [91]. Pearson correlation between log2FoldChange values of FOGs and the corresponding plant photosynthetic types (C3 or C4) was calculated using python pandas (https://pandas.pydata.org/) (accessed on 12 September 2023).

5.4. Domain Amino Acids Composition Analysis, Protein Solubility Prediction, and Global Motif Discovery

We identified and extracted each FOG protein's domain sequences and analyzed their amino acid composition using ProtParam (https://web.expasy.org/protparam/) (accessed on 12 September 2023) from Expasy, aiming to compare the dynamics of amino acids with different characteristics during evolution. In addition, we utilized Protein-Sol (https://protein-sol.manchester.ac.uk/) (accessed on 12 September 2023) [54] to predict the solubility and pI of each protein. The motif occurrence in each homologous FOG protein was examined by MEME suite (https://meme-suite.org/meme/tools/meme) (accessed on 12 September 2023) [92].

5.5. Three-Dimensional Structure Prediction and Sequence Structure Similarity Comparison

To verify the prediction performance of AlphaFold2 in plant protein prediction, protein amino acid sequences and three-dimensional structures of plant protein STP10 and ReAV were retrieved from RCSB PDB with accessions 6H7D (https://www.rcsb.org/structure/6H7D) and 7OS5 (https://www.rcsb.org/structure/7OS5), (accessed on 12 September 2023) respectively. The structure predictions of the two proteins were generated by the ColabFold (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb) (accessed on 12 September 2023) with a substitution of MSA using MMseqs2 [65] and visualized by PyMol (https://pymol.org/2/) (accessed on 12 September 2023), and TM-score values were calculated by TM-align (https://zhanggroup.org/TM-align/) (accessed on 12 September 2023) [93]. Besides this, all monomers in five species of two FOG proteins were predicted using their amino acid sequences with Amber relaxation and templates on NVIDIA Tesla V100 GPU via the Google Colaboratory Pro+ platform. We collected five predicted protein folds for each input and plotted the pLDDT value using seaborn (http://seaborn.pydata.org/) (accessed on 12 September 2023). The models with the highest pLDDT were selected as the predicted model for structure comparison. The TM-score (https://zhanggroup.org/TM-score/) (accessed on 12 September 2023) and the Root Mean Square Deviation (RMSD) of superposition between predicted protein folds were calculated. TM-score is a classical measurement for pairwise protein structure topological similarity comparison, in which 1 represents the same fold, a value below 0.17 represents randomly selected unrelated structures, and 0.5 is the threshold for similar structure measurements. Generally, TM-score is regarded as a more sensitive measurement since it is length-dependent and it normalizes the distance errors, hence it can supplement RMSD [94]. We used the blastp program (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) (accessed on 12 September 2023) to

compare the sequence similarity of homologous proteins among five plant species. The dual similarities comparison method is well illustrated and utilized in protein studies [51,95,96].

6. Conclusions

In this work, we compared the cell-specific (bundle sheath and mesophyll cells) transcriptomes of five photosynthetic plant species. Our study illustrates that: (i) Compared to C3 plants, C4 plants have more cell-specific DEGs, which means that a more complex functional differentiation may occur in C4 plants. (ii) Among these cell-specific DEGs, we found 16 of them can be used as photosynthetic features for modeling. (iii) Two flip-over genes, LHCA6 and PGRL1A, are highly correlated with C3 or C4 photosynthetic types, which is due to their functional nature in the photosynthesis process. (iv) Based on protein physicochemical and structural analysis, we found the homologous proteins of these two flip-over genes are inconsistent in terms of their sequence and structure similarities, which are also found in other photosynthetic proteins, and may contribute to our understanding of the complexity of protein evolution in C4 photosynthesis.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms241814165/s1.

Author Contributions: Conceptualization, C.W. and D.G.; methodology, C.W.; formal analysis, C.W.; investigation, C.W.; resources, C.W.; data curation, C.W.; writing—original draft preparation, C.W.; writing—review and editing, D.G.; visualization, C.W.; supervision, D.G.; project administration, D.G.; funding acquisition, D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Transformation project of Hong Kong and Macao scientific and technological achievements of Guangdong province, China, grant number 6905891; and grant 8300052 from State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong SAR, China.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank John JUMPER for discussion on model performance, Pengfei DONG for assistance on plants synteny analysis, Ying AN for guiding protein sequence analysis, Biyang XU for preparing the computing resources, Chuanyang YU and Xiaojun WANG for discussion on analysis pipeline design; we cannot complete this study without your help.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sage, R.F. The evolution of C 4 photosynthesis. *New Phytol.* **2004**, *161*, 341–370. [CrossRef] [PubMed]
2. Wang, L.; Czedik-Eysenberg, A.; Mertz, R.A.; Si, Y.; Tohge, T.; Nunes-Nesi, A.; Arrivault, S.; Dedow, L.K.; Bryant, D.W.; Zhou, W.; et al. Comparative analyses of C4 and C3 photosynthesis in developing leaves of maize and rice. *Nat. Biotechnol.* **2014**, *32*, 1158–1165. [CrossRef] [PubMed]
3. Way, D.A.; Katul, G.G.; Manzoni, S.; Vico, G. Increasing water use efficiency along the C3 to C4 evolutionary pathway: A stomatal optimization perspective. *J. Exp. Bot.* **2014**, *65*, 3683–3693. [CrossRef]
4. Killi, D.; Bussotti, F.; Raschi, A.; Haworth, M. Adaptation to high temperature mitigates the impact of water deficit during combined heat and drought stress in C3 sunflower and C4 maize varieties with contrasting drought tolerance. *Physiol. Plant* **2017**, *159*, 130–147. [CrossRef]
5. Lopez-Hernandez, F.; Cortes, A.J. Last-Generation Genome-Environment Associations Reveal the Genetic Basis of Heat Tolerance in Common Bean (*Phaseolus vulgaris* L.). *Front. Genet.* **2019**, *10*, 954. [CrossRef] [PubMed]
6. Cortes, A.J.; Blair, M.W. Genotyping by Sequencing and Genome-Environment Associations in Wild Common Bean Predict Widespread Divergent Adaptation to Drought. *Front. Plant Sci.* **2018**, *9*, 128. [CrossRef]
7. Cortes, A.J.; Skeen, P.; Blair, M.W.; Chacon-Sanchez, M.I. Does the Genomic Landscape of Species Divergence in Phaseolus Beans Coerce Parallel Signatures of Adaptation and Domestication? *Front. Plant Sci.* **2018**, *9*, 1816. [CrossRef]
8. Blair, M.W.; Cortes, A.J.; This, D. Identification of an ERECTA gene and its drought adaptation associations with wild and cultivated common bean. *Plant Sci.* **2016**, *242*, 250–259. [CrossRef]

9. Cortés, A.J.; Chavarro, M.C.; Madriñán, S.; This, D.; Blair, M.W. Molecular ecology and selection in the drought-related Asr gene polymorphisms in wild and cultivated common bean (*Phaseolus vulgaris* L.). *BMC Genom. Data* **2012**, *13*, 58. [CrossRef]
10. Cortes, A.J.; This, D.; Chavarro, C.; Madrinan, S.; Blair, M.W. Nucleotide diversity patterns at the drought-related DREB2 encoding genes in wild and cultivated common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* **2012**, *125*, 1069–1085. [CrossRef]
11. Buitrago-Bitar, M.A.; Cortes, A.J.; Lopez-Hernandez, F.; Londono-Caicedo, J.M.; Munoz-Florez, J.E.; Munoz, L.C.; Blair, M.W. Allelic Diversity at Abiotic Stress Responsive Genes in Relationship to Ecological Drought Indices for Cultivated Tepary Bean, Phaseolus acutifolius A. Gray, and Its Wild Relatives. *Genes* **2021**, *12*, 556. [CrossRef] [PubMed]
12. Brutnell, T.P.; Wang, L.; Swartwood, K.; Goldschmidt, A.; Jackson, D.; Zhu, X.-G.; Kellogg, E.; Van Eck, J. Setaria viridis: A Model for C4 Photosynthesis. *Plant Cell* **2010**, *22*, 2537–2544. [CrossRef] [PubMed]
13. Li, X.; Wang, P.; Li, J.; Wei, S.; Yan, Y.; Yang, J.; Zhao, M.; Langdale, J.A.; Zhou, W. Maize GOLDEN2-LIKE genes enhance biomass and grain yields in rice by improving photosynthesis and reducing photoinhibition. *Commun. Biol.* **2020**, *3*, 151. [CrossRef]
14. Wang, P.; Fouracre, J.; Kelly, S.; Karki, S.; Gowik, U.; Aubry, S.; Shaw, M.K.; Westhoff, P.; Slamet-Loedin, I.H.; Quick, W.P.; et al. Evolution of GOLDEN2-LIKE gene function in C(3) and C(4) plants. *Planta* **2013**, *237*, 481–495. [CrossRef] [PubMed]
15. Reeves, G.; Grange-Guermente, M.J.; Hibberd, J.M. Regulatory gateways for cell-specific gene expression in C4 leaves with Kranz anatomy. *J. Exp. Bot.* **2017**, *68*, 107–116. [CrossRef]
16. Majeran, W.; Cai, Y.; Sun, Q.; van Wijk, K.J. Functional differentiation of bundle sheath and mesophyll maize chloroplasts determined by comparative proteomics. *Plant Cell* **2005**, *17*, 3111–3140. [CrossRef]
17. Leegood, R.C. Roles of the bundle sheath cells in leaves of C3 plants. *J. Exp. Bot.* **2007**, *59*, 1663–1673. [CrossRef]
18. Sinha, N.R.; Kellogg, E.A. Parallelism and diversity in multiple origins of c4photosynthesis in the grass family. *Am. J. Bot.* **1996**, *83*, 1458–1470. [CrossRef]
19. Sawers, R.J.; Liu, P.; Anufrikova, K.; Hwang, J.T.; Brutnell, T.P. A multi-treatment experimental system to examine photosynthetic differentiation in the maize leaf. *BMC Genom.* **2007**, *8*, 12. [CrossRef]
20. Li, P.; Ponnala, L.; Gandotra, N.; Wang, L.; Si, Y.; Tausta, S.L.; Kebrom, T.H.; Provart, N.; Patel, R.; Myers, C.R.; et al. The developmental dynamics of the maize leaf transcriptome. *Nat. Genet.* **2010**, *42*, 1060–1067. [CrossRef]
21. Ding, Z.; Weissmann, S.; Wang, M.; Du, B.; Huang, L.; Wang, L.; Tu, X.; Zhong, S.; Myers, C.; Brutnell, T.P.; et al. Identification of Photosynthesis-Associated C4 Candidate Genes through Comparative Leaf Gradient Transcriptome in Multiple Lineages of C3 and C4 Species. *PLoS ONE* **2015**, *10*, e0140629. [CrossRef] [PubMed]
22. Huang, P.; Brutnell, T.P. A synthesis of transcriptomic surveys to dissect the genetic basis of C4 photosynthesis. *Curr. Opin. Plant Biol.* **2016**, *31*, 91–99. [CrossRef] [PubMed]
23. Wu, C.; Guo, D. Computational Docking Reveals Co-Evolution of C4 Carbon Delivery Enzymes in Diverse Plants. *Int. J. Mol. Sci.* **2022**, *23*, 12688. [CrossRef] [PubMed]
24. Berkowitz, O.; Xu, Y.; Liew, L.C.; Wang, Y.; Zhu, Y.; Hurgobin, B.; Lewsey, M.G.; Whelan, J. RNA-seq analysis of laser microdissected Arabidopsis thaliana leaf epidermis, mesophyll and vasculature defines tissue-specific transcriptional responses to multiple stress treatments. *Plant J.* **2021**, *107*, 938–955. [CrossRef]
25. Hua, L.; Stevenson, S.R.; Reyna-Llorens, I.; Xiong, H.; Kopriva, S.; Hibberd, J.M. The bundle sheath of rice is conditioned to play an active role in water transport as well as sulfur assimilation and jasmonic acid synthesis. *Plant J.* **2021**, *107*, 268–286. [CrossRef]
26. Aubry, S.; Kelly, S.; Kumpers, B.M.; Smith-Unna, R.D.; Hibberd, J.M. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis. *PLoS Genet.* **2014**, *10*, e1004365. [CrossRef]
27. Madhok, A.; Bhat, S.A.; Philip, C.S.; Sureshbabu, S.K.; Chiplunkar, S.; Galande, S. Transcriptome Signature of Vgamma9Vdelta2 T Cells Treated With Phosphoantigens and Notch Inhibitor Reveals Interplay Between TCR and Notch Signaling Pathways. *Front. Immunol.* **2021**, *12*, 660361. [CrossRef]
28. Casella, G.; Munk, R.; Kim, K.M.; Piao, Y.; De, S.; Abdelmohsen, K.; Gorospe, M. Transcriptome signature of cellular senescence. *Nucleic Acids Res.* **2019**, *47*, 7294–7305. [CrossRef]
29. Liang, P.; Yang, W.; Chen, X.; Long, C.; Zheng, L.; Li, H.; Zuo, Y. Machine Learning of Single-Cell Transcriptome Highly Identifies mRNA Signature by Comparing F-Score Selection with DGE Analysis. *Mol. Ther. Nucleic Acids* **2020**, *20*, 155–163. [CrossRef]
30. Li, R.; Zhu, J.; Zhong, W.D.; Jia, Z. Comprehensive Evaluation of Machine Learning Models and Gene Expression Signatures for Prostate Cancer Prognosis Using Large Population Cohorts. *Cancer Res.* **2022**, *82*, 1832–1843. [CrossRef]
31. Kong, S.W.; Collins, C.D.; Shimizu-Motohashi, Y.; Holm, I.A.; Campbell, M.G.; Lee, I.H.; Brewster, S.J.; Hanson, E.; Harris, H.K.; Lowe, K.R.; et al. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS ONE* **2012**, *7*, e49475. [CrossRef] [PubMed]
32. Aubry, S.; Aresheva, O.; Reyna-Llorens, I.; Smith-Unna, R.D.; Hibberd, J.M.; Genty, B. A Specific Transcriptome Signature for Guard Cells from the C4 Plant Gynandropsis gynandra. *Plant Physiol.* **2016**, *170*, 1345–1357. [CrossRef] [PubMed]
33. McLaughlin, R.N., Jr.; Poelwijk, F.J.; Raman, A.; Gosal, W.S.; Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **2012**, *491*, 138–142. [CrossRef] [PubMed]
34. Grishin, N.V. Fold change in evolution of protein structures. *J. Struct. Biol.* **2001**, *134*, 167–185. [CrossRef]
35. Shi, Y. A glimpse of structural biology through X-ray crystallography. *Cell* **2014**, *159*, 995–1014. [CrossRef]
36. Fowler, N.J.; Sljoka, A.; Williamson, M.P. A method for validating the accuracy of NMR protein structures. *Nat. Commun.* **2020**, *11*, 6321. [CrossRef]

37. Nogales, E.; Scheres, S.H. Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Mol. Cell* **2015**, *58*, 677–689. [CrossRef]
38. Todd, A.E.; Orengo, C.A.; Thornton, J.M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **2001**, *307*, 1113–1143. [CrossRef]
39. Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 1496–1503. [CrossRef]
40. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; de Beer, T.A.P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303. [CrossRef]
41. Kelley, L.A.; Mezulis, S.; Yates, C.M.; Wass, M.N.; Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **2015**, *10*, 845–858. [CrossRef] [PubMed]
42. Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C.L.; Ma, J.; et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2016239118. [CrossRef] [PubMed]
43. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
44. Goulet, A.; Cambillau, C. Structure and Topology Prediction of Phage Adhesion Devices Using AlphaFold2: The Case of Two Oenococcus oeni Phages. *Microorganisms* **2021**, *9*, 2151. [CrossRef]
45. Tsaban, T.; Varga, J.K.; Avraham, O.; Ben-Aharon, Z.; Khramushin, A.; Schueler-Furman, O. Harnessing protein folding neural networks for peptide-protein docking. *Nat. Commun.* **2022**, *13*, 176. [CrossRef]
46. Krissinel, E. On the relationship between sequence and structure similarities in proteomics. *Bioinformatics* **2007**, *23*, 717–723. [CrossRef]
47. Chothia, C.; Lesk, A.M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*, 823–826. [CrossRef]
48. Murzin, A.G. How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **1998**, *8*, 380–387. [CrossRef]
49. Todd, C.; Wood, W.R.P. Evolution of Protein Sequences and Structures. *J. Mol. Biol.* **1999**, *291*, 977–995.
50. Koehl, P.; Levitt, M. Sequence variations within protein families are linearly related to structural variations. *J. Mol. Biol.* **2002**, *323*, 551–562. [CrossRef]
51. He, Y.; Maisuradze, G.G.; Yin, Y.; Kachlishvili, K.; Rackovsky, S.; Scheraga, H.A. Sequence-, structure-, and dynamics-based comparisons of structurally homologous CheY-like proteins. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 1578–1583. [CrossRef] [PubMed]
52. Kosloff, M.; Kolodny, R. Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins* **2008**, *71*, 891–902. [CrossRef] [PubMed]
53. Szklarczyk, D.; Gable, A.L.; Nastou, K.C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N.T.; Legeay, M.; Fang, T.; Bork, P.; et al. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **2021**, *49*, D605–D612. [CrossRef]
54. Hebditch, M.; Carballo-Amador, M.A.; Charonis, S.; Curtis, R.; Warwicker, J. Protein-Sol: A web tool for predicting protein solubility from sequence. *Bioinformatics* **2017**, *33*, 3098–3100. [CrossRef] [PubMed]
55. Wang, S.; Tholen, D.; Zhu, X.G. C(4) photosynthesis in C(3) rice: A theoretical analysis of biochemical and anatomical factors. *Plant Cell Environ.* **2017**, *40*, 80–94. [CrossRef]
56. Hughes, T.E.; Sedelnikova, O.V.; Wu, H.; Becraft, P.W.; Langdale, J.A. Redundant SCARECROW genes pattern distinct cell layers in roots and leaves of maize. *Development* **2019**, *146*, dev177543. [CrossRef]
57. di Iulio, J.; Bartha, I.; Spreafico, R.; Virgin, H.W.; Telenti, A. Transfer transcriptomic signatures for infectious diseases. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2022486118. [CrossRef]
58. Hibberd, J.M.; Sheehy, J.E.; Langdale, J.A. Using C4 photosynthesis to increase the yield of rice—Rationale and feasibility. *Curr. Opin. Plant Biol.* **2008**, *11*, 228–231. [CrossRef]
59. Ermakova, M.; Danila, F.R.; Furbank, R.T.; von Caemmerer, S. On the road to C4 rice: Advances and perspectives. *Plant J.* **2020**, *101*, 940–950. [CrossRef]
60. Eckardt, N.A. Oxylipin signaling in plant stress responses. *Plant Cell* **2008**, *20*, 495–497. [CrossRef]
61. Barros, T.; Kuhlbrandt, W. Crystallisation, structure and function of plant light-harvesting Complex II. *Biochim. Biophys. Acta* **2009**, *1787*, 753–772. [CrossRef] [PubMed]
62. Araujo, W.L.; Nunes-Nesi, A.; Nikoloski, Z.; Sweetlove, L.J.; Fernie, A.R. Metabolic control and regulation of the tricarboxylic acid cycle in photosynthetic and heterotrophic plant tissues. *Plant Cell Environ.* **2012**, *35*, 1–21. [CrossRef] [PubMed]
63. Theobald, D.L.; Wuttke, D.S. Divergent evolution within protein superfolds inferred from profile-based phylogenetics. *J. Mol. Biol.* **2005**, *354*, 722–737. [CrossRef] [PubMed]
64. Gilson, A.I.; Marshall-Christensen, A.; Choi, J.M.; Shakhnovich, E.I. The Role of Evolutionary Selection in the Dynamics of Protein Structure Evolution. *Biophys. J.* **2017**, *112*, 1350–1365. [CrossRef] [PubMed]
65. Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making protein folding accessible to all. *Nat. Methods* **2022**, *19*, 679–682. [CrossRef] [PubMed]

66. DalCorso, G.; Pesaresi, P.; Masiero, S.; Aseeva, E.; Schunemann, D.; Finazzi, G.; Joliot, P.; Barbato, R.; Leister, D. A complex containing PGRL1 and PGR5 is involved in the switch between linear and cyclic electron flow in Arabidopsis. *Cell* **2008**, *132*, 273–285. [CrossRef]
67. Hertle, A.P.; Blunder, T.; Wunder, T.; Pesaresi, P.; Pribil, M.; Armbruster, U.; Leister, D. PGRL1 is the elusive ferredoxin-plastoquinone reductase in photosynthetic cyclic electron flow. *Mol. Cell* **2013**, *49*, 511–523. [CrossRef]
68. Peng, L.; Fukao, Y.; Fujiwara, M.; Takami, T.; Shikanai, T. Efficient operation of NAD(P)H dehydrogenase requires supercomplex formation with photosystem I via minor LHCI in Arabidopsis. *Plant Cell* **2009**, *21*, 3623–3640. [CrossRef]
69. Peng, L.; Shikanai, T. Supercomplex formation with photosystem I is required for the stabilization of the chloroplast NADH dehydrogenase-like complex in Arabidopsis. *Plant Physiol.* **2011**, *155*, 1629–1639. [CrossRef]
70. Liao, B.Y.; Weng, M.P. Unraveling the association between mRNA expressions and mutant phenotypes in a genome-wide assessment of mice. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 4707–4712. [CrossRef]
71. Yerramsetty, P.; Agar, E.M.; Yim, W.C.; Cushman, J.C.; Berry, J.O. An rbcL mRNA-binding protein is associated with C3 to C4 evolution and light-induced production of Rubisco in Flaveria. *J. Exp. Bot.* **2017**, *68*, 4635–4649. [CrossRef] [PubMed]
72. Cortes, A.J.; Restrepo-Montoya, M.; Bedoya-Canas, L.E. Modern Strategies to Assess and Breed Forest Tree Adaptation to Changing Climate. *Front. Plant Sci.* **2020**, *11*, 583323. [CrossRef] [PubMed]
73. Cortes, A.J.; Lopez-Hernandez, F.; Osorio-Rodriguez, D. Predicting Thermal Adaptation by Looking Into Populations' Genomic Past. *Front. Genet.* **2020**, *11*, 564515. [CrossRef] [PubMed]
74. Cortes, A.J.; Lopez-Hernandez, F.; Blair, M.W. Genome-Environment Associations, an Innovative Tool for Studying Heritable Evolutionary Adaptation in Orphan Crops and Wild Relatives. *Front. Genet.* **2022**, *13*, 910386. [CrossRef]
75. John, C.R.; Smith-Unna, R.D.; Woodfield, H.; Covshoff, S.; Hibberd, J.M. Evolutionary convergence of cell-specific gene expression in independent lineages of C4 grasses. *Plant Physiol.* **2014**, *165*, 62–75. [CrossRef]
76. Chang, Y.M.; Liu, W.Y.; Shih, A.C.; Shen, M.N.; Lu, C.H.; Lu, M.Y.; Yang, H.W.; Wang, T.Y.; Chen, S.C.; Chen, S.M.; et al. Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. *Plant Physiol.* **2012**, *160*, 165–177. [CrossRef]
77. Doring, F.; Streubel, M.; Brautigam, A.; Gowik, U. Most photorespiratory genes are preferentially expressed in the bundle sheath cells of the C4 grass Sorghum bicolor. *J. Exp. Bot.* **2016**, *67*, 3053–3064. [CrossRef]
78. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
79. Lamesch, P.; Berardini, T.Z.; Li, D.; Swarbreck, D.; Wilks, C.; Sasidharan, R.; Muller, R.; Dreher, K.; Alexander, D.L.; Garcia-Hernandez, M.; et al. The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* **2012**, *40*, D1202–D1210. [CrossRef]
80. Ouyang, S.; Zhu, W.; Hamilton, J.; Lin, H.; Campbell, M.; Childs, K.; Thibaud-Nissen, F.; Malek, R.L.; Lee, Y.; Zheng, L.; et al. The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res.* **2007**, *35*, D883–D887. [CrossRef]
81. Mamidi, S.; Healey, A.; Huang, P.; Grimwood, J.; Jenkins, J.; Barry, K.; Sreedasyam, A.; Shu, S.; Lovell, J.T.; Feldman, M.; et al. A genome resource for green millet Setaria viridis enables discovery of agronomically valuable loci. *Nat. Biotechnol.* **2020**, *38*, 1203–1210. [CrossRef] [PubMed]
82. McCormick, R.F.; Truong, S.K.; Sreedasyam, A.; Jenkins, J.; Shu, S.; Sims, D.; Kennedy, M.; Amirebrahimi, M.; Weers, B.D.; McKinley, B.; et al. The Sorghum bicolor reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* **2018**, *93*, 338–354. [CrossRef] [PubMed]
83. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [CrossRef] [PubMed]
84. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
85. Ramirez, F.; Dundar, F.; Diehl, S.; Gruning, B.A.; Manke, T. deepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **2014**, *42*, W187–W191. [CrossRef]
86. Anders, S.; Pyl, P.T.; Huber, W. HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, 166–169. [CrossRef] [PubMed]
87. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
88. Heberle, H.; Meirelles, G.V.; da Silva, F.R.; Telles, G.P.; Minghim, R. InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinform.* **2015**, *16*, 169. [CrossRef]
89. Yu, G.; Wang, L.G.; Han, Y.; He, Q.Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **2012**, *16*, 284–287. [CrossRef]
90. Hunter, S.; Apweiler, R.; Attwood, T.K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Das, U.; Daugherty, L.; Duquenne, L.; et al. InterPro: The integrative protein signature database. *Nucleic Acids Res.* **2009**, *37*, D211–D215. [CrossRef]
91. Crooks, G.E.; Hon, G.; Chandonia, J.M.; Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res* **2004**, *14*, 1188–1190. [CrossRef] [PubMed]
92. Bailey, T.L.; Johnson, J.; Grant, C.E.; Noble, W.S. The MEME Suite. *Nucleic Acids Res.* **2015**, *43*, W39–W49. [CrossRef] [PubMed]

93. Zhang, Y.; Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302–2309. [CrossRef] [PubMed]
94. Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26*, 889–895. [CrossRef] [PubMed]
95. Tong, J.; Sadreyev, R.I.; Pei, J.; Kinch, L.N.; Grishin, N.V. Using homology relations within a database markedly boosts protein sequence similarity search. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 7003–7008. [CrossRef]
96. Sadowski, M.I.; Jones, D.T. The sequence-structure relationship and protein function prediction. *Curr. Opin. Struct. Biol.* **2009**, *19*, 357–362. [CrossRef]