Upregulation and cell specificity of C_4 genes are derived from ancestral C_3 gene regulatory networks

Pallavi Singh#, Sean R. Stevenson#, Ivan Reyna-Llorens, Gregory Reeves, Tina B. Schreier and Julian M. Hibberd*

Department of Plant Sciences, University of Cambridge, Downing street, Cambridge CB2 3EA, United Kingdom.

*Correspondence to jmh65@cam.ac.uk

#Both authors contributed equally

KEY WORDS

C_4 photosynthesis, light, de-etiolation, transcriptome, DNaseI-SEQ, *cis*-elements, light-responsive elements, *Gynandropsis gynandra*.

RUNNING TITLE: Transcriptional regulatory landscape in de-etiolating *Gynandropsis gynandra* seedlings.

ABSTRACT

The efficient C_4 pathway is based on strong up-regulation of genes found in C_3 plants, but also compartmentation of their expression into distinct cell-types such as the mesophyll and bundle sheath. Transcription factors associated with these phenomena have not been identified. To address this, we undertook genome-wide analysis of transcript accumulation, chromatin accessibility and transcription factor binding in C_4 *Gynandropsis gynandra*. From these data, two models relating to the molecular evolution of C_4 photosynthesis are proposed. First, increased expression of C_4 genes is associated with increased binding by MYB-related transcription factors. Second, mesophyll specific expression is associated with binding of homeodomain transcription factors. Overall, we conclude that during evolution of the complex C_4 trait, C_4 cycle genes gain *cis*-elements that operate in the C_3 leaf such that they become integrated into existing gene regulatory networks associated with cell specificity and photosynthesis.

INTRODUCTION

Photosynthesis fuels life on Earth and in the majority of land plants, Ribulose 1,5-bisphosphate Carboxylase Oxygenase (RuBisCO) catalyses the initial fixation of atmospheric carbon-dioxide (CO_2) to generate phosphoglyceric acid (PGA). However, oxygen (O_2) can competitively bind to the RuBisCO active site to form a toxic product 2-phosphoglycolate[1]. 2-phosphoglycolate can be metabolised at the expense of carbon and energy by photorespiration[2,3]. It is thought that to reduce rates of photorespiration, many plant lineages have evolved carbon concentrating mechanisms. C_4 photosynthesis is one such example and is characterised by compartmentation of photosynthesis, typically between mesophyll and bundle sheath cells[4]. This compartmentalisation involves cell-preferential gene expression and allows increased concentrations of CO_2 to be supplied to the RuBisCO sequestered in bundle sheath cells[5]. Rather than RuBisCO initially fixing carbon, in C_4 species fixation is initiated by phospho*enol*pyruvate carboxylase (PEPC) combining HCO_3^- to form a C_4 acid in mesophyll cells. Diffusion of C_4 acids into bundle sheath cells and subsequent decarboxylation results in elevated partial pressures of CO_2 around RuBisCO facilitating efficient carboxylation and reducing the requirement for significant rates of photorespiration.

C_4 photosynthesis results in higher water and nitrogen use efficiencies compared with the C_3 state, particularly in dry and hot climates. C_4 crops of major economic importance include maize (*Zea mays*), sugarcane (*Saccharum officinarum*), sorghum (*Sorghum bicolor*), pearl millet (*Pennisetum glaucum*) and finger millet (*Setaria italica*)[6]. Although C_4 photosynthesis is a complex trait characterized by changes in anatomy, biochemistry and gene expression[4] it has evolved convergently from C_3 ancestors in around 62 independent lineages that together account for ~8,100 species[7]. During the evolution of C_4 photosynthesis, parsimony would therefore imply that gene networks underpinning this system are derived from those that operate in C_3 ancestors.

Compared with the ancestral C_3 state, expression of genes encoding components of the C_4 pathway are restricted more precisely to either mesophyll or bundle sheath cells, and also upregulated. Our present understanding of these changes to C_4 gene regulation is mostly based on studies designed to understand the regulation of individual C_4 genes[8]. For example, a number of *cis*-regulatory motifs controlling the cell preferential expression of C_4-genes have been identified[9–15]. Whilst some *cis*-elements appear to have evolved *de novo* in C_4 genes to pattern their expression[16–20], others appear to have been recruited from pre-existing elements present in C_3 orthologs[12–15] and in these cases there is evidence that individual *cis*-elements are shared between multiple C_4 genes. In contrast to the analysis of regulators of cell specific expression, there is much less work on mechanisms that underpin the upregulation of genes important for the C_4 pathway compared with the ancestral C_3

state. One possibility is that *cis*-elements referred to as Light Responsive Elements (LREs) that have been characterized in photosynthesis genes in C_3 plants e.g. *CAB* (chlorophyll a/b binding proteins), *PC* (plastocyanin) and *RBCS* (small subunit of RuBisCO)[21,22] are acquired by genes of the core C_4 pathway. Indeed, whilst many C_4 pathway components and their orthologs in C_3 species show light-dependant induction[23], the mechanisms driving these patterns are still largely unknown.

The response of a seedling to light is the first major step towards photosynthetic maturity. The growth of seedlings in prolonged darkness leads to the development of etioplasts in place of chloroplasts[24]. Etioplasts lack chlorophyll but contain membranes composed of a paracrystalline lipid–pigment–protein structure known as the prolamellar body (PLB)[25–29]. De-etiolation of seedlings therefore marks the initiation of photosynthesis and presents a good model system to study the dynamics and regulatory mechanisms governing photosynthesis in both C_3 and C_4 species. The establishment of photosynthesis has been shown to involve two phases including an initial change in gene expression following light induction together with accompanying metabolic and structural changes to the chloroplast. A second phase involves maturation of the chloroplast and a tight coordination between chloroplast and nuclear genomes[30]. Here we used a genome-wide approach to better understand the patterns of transcript abundance and potential regulatory mechanisms responsible for these behaviours underpinning C_4 photosynthesis. By carrying out DNaseI-SEQ and coupling it with profiling of transcript abundance during de-etiolation of *Gynandropsis gynandra* seedlings, we aimed to provide insights into the transcription factor binding repertoire and the dynamics of gene expression associated with the establishment of C_4 photosynthesis. Further, we undertook comparative analysis using an analogous dataset from C_3 *Arabidopsis thaliana* (hereafter Arabidopsis) to compare the extent to which regulatory mechanisms are shared between the ancestral C_3 and derived C_4 systems.

Our data link changes in chromatin accessibility and transcription factor binding to patterns of gene expression and assembly of the photosynthetic apparatus in the C_4 species *G. gynandra*. During de-etiolation, assembly of the photosynthetic apparatus was initiated within two hours of exposure to light. Transcript profiling revealed a global remodelling of gene expression associated with the dark-to-light transition. Many genes associated with core photosynthetic processes shared by C_3 and C_4 plants, as well as those specifically encoding components of C_4 photosynthesis showed similar dynamics during this dark-to-light transition. During the first two hours of exposure to light, a relatively large restructuring of open chromatin and a shift from transcription factor binding in exons to promoters and 5’ UTRs took place. All genes encoding core proteins of the C_4 pathway were more strongly induced after exposure to light in C_4 *G. gynandra* compared with C_3 Arabidopsis. The greater induction of C_4 pathway genes in *G. gynandra* was associated with gain of light responsive elements such as EE and I-boxes, but also the C2C2-GATA box that regulates photosynthesis-associated nuclear genes (PhANGs). Moreover, binding of transcription factors *in vivo* to these sites was detected. Second, C_4 genes expressed in mesophyll cells gained homeodomain and LOB/AS2 binding sites that were also bound in the C_4 leaf. In addition to the increased number and binding to such sites in C_4 genes of *G. gynandra* compared with Arabidopsis, strong expression of *ANL2* which belongs to the homeodomain family and is preferentially expressed in mesophyll cells was detected. We conclude that the evolution of C_4 photosynthesis is associated with rewiring of photosynthesis gene regulatory networks that exist in the C_3 state such that they expand to include genes encoding C_4 enzymes.

RESULTS

De-etiolation and chloroplast development in C_4 *Gynandropsis gynandra*

The dynamics associated with unfolding of the apical hook, chlorophyll accumulation, and ultrastructural re-arrangements of chloroplasts were determined as etiolated seedlings of *Gynandropsis gynandra* were transferred from dark-to-light. The classical photomorphogenic responses of apical hook unfolding and greening of cotyledons were visible 2 hours after transfer to light (Fig. 1a). Chlorophyll quantification indicated that accumulation was detectable by 0.5 hours after exposure to light, and that an initial exponential phase was followed by a more linear increase (Fig. 1b). Little additional chlorophyll was synthesized from 12 to 24 hours after first exposure to light (Fig. 1b). Assembly of the photosynthetic membranes in chloroplasts from mesophyll and bundle sheath cells, both of which are involved in C_4 photosynthesis, was apparent over this time course (Fig. 1c-d). In the dark, prolamellar bodies dominated the internal space of chloroplasts in each cell type. After 0.5 hours of exposure to light, although prolamellar bodies were still evident, they had started to disperse. Starch grains were apparent in bundle sheath chloroplasts by 24 hours after exposure to light, and it was noticeable that thylakoids showed low stacking in this cell type (Fig. 1c-d, Supplementary Fig. 1). Overall, these data indicate that in *G. gynandra,* assembly of the photosynthetic apparatus was initiated within 0.5 hours of exposure to light and by 24 hours the apparatus appeared fully functional. To better understand the patterns of gene expression and the transcriptional regulation that underpin this induction of C_4 photosynthesis, these early time points were selected for detailed molecular analysis.

Induction of photosynthesis genes in *G. gynandra*

To investigate how transcript abundance changed during the induction of C_4 photosynthesis, mRNA from three biological replicates at 0, 0.5, 2, 4, and 24 hours after exposure to light was isolated and used for RNA-SEQ. On average, 10 million reads were obtained and ~25,000 transcripts detected per sample (Supplementary Table 1). We were primarily interested in the dynamics of gene expression throughout de-etiolation and so we analysed how transcript abundance changed relative to each previous time point. To provide a conservative estimate for the number of transcripts that were differentially expressed between consecutive time points, two independent algorithms were used and the intersect between these datasets determined (Supplementary Table 1). This showed that by far the greatest difference in transcript abundance was detected 0.5 hours after transfer from dark-to-light (number differentially expressed = 4609). At subsequent time points, the number of differentially expressed transcripts ranged from 1861 (4 hours versus 2 hours) to 2452 (24 hours versus 4 hours), and so was always less than half the number associated with the first 0.5 hours of exposure to light (Supplementary Table 1). Principle component analysis (PCA) showed that replicates from each timepoint clustered together tightly, and that 64% of the variance in transcript abundance could be explained by two main components. The first component accounted for 46% of the variance and was associated with the dark-to-light transition (Fig. 2a). To better understand the general patterns of differentially expressed genes (Supplementary Table 2), Gene Ontology (GO) terms were assessed (Fig. 2b, Supplementary Fig. 2, FDR<10^{-5}). Compared with each previous time point, upregulated genes in samples taken at 0.5 and 24 hours showed enrichment in GO terms including those related to the plastid, as well as carbohydrate, secondary, nitrogen and lipid metabolism, but also responses to light and photosynthesis. These components were also over-represented in genes down-regulated at 2 and 4-hour time points suggesting two phases of photosynthetic induction. Overall, these pairwise comparisons of transcript abundance between samples, the PCA and the GO term enrichment analysis are consistent with a major remodelling of gene expression after 0.5 hours of exposure to light, at least in part associated with establishment of photosynthesis.

To better understand the dynamics of gene expression associated with the induction of chlorophyll accumulation and remodelling of chloroplast ultrastructure in the C_4 leaf (Fig. 1), genes associated with photosynthesis were subjected to hierarchical clustering. The genes defined as such were C_4 pathway genes as well as orthologs to nuclear genes from Arabidopsis annotated with the photosynthesis-related GO term (GO:0015979). A total of 116 genes were clustered into three main groups. Cluster I (red, Fig. 2c) showed no clear induction over the time-course, but clusters II and III (yellow and green respectively, Fig. 2c) contained the majority of genes (n=76) and showed clear induction by 24 hours. Notably, C_4 pathway genes were dispersed among these clusters, with 67% of C_4 photosynthesis genes (n=20) found in clusters II and III (Fig. 2c). There are multiple paralogs of various C_4 cycle genes of which ten showed no clear induction in response to light (Cluster I). However, the majority of these non-induced members were poorly expressed, and at least one other paralog was strongly induced and so present in Cluster II and III. Overall, these data show that the majority of C_4 cycle genes populated photosynthesis gene clusters that showed increased expression during de-etiolation (Fig. 2c).

To identify candidate transcription factors that may be responsible for the induction of photosynthesis gene expression in response to light, four classes were identified on the basis of changes in their transcript abundance. Transcription factors that act as positive regulators of photosynthesis would be expected to show either a steady increase and positive correlation (>0.8 Pearson Correlation across time-course) with induced photosynthesis genes (clusters II and III from Fig. 2c), or an early burst at 0.5 hours (Fig. 2d). Repressors would be expected to show the opposite trends (Fig. 2d). This approach identified twenty-one transcription factors that were strongly and positively correlated with photosynthesis genes (Supplementary Table 3). Candidates in this class were often related to previously characterised regulators of photomorphogenesis, plastid development, light and circadian networks as well as components of cell fate determination and sucrose signalling in Arabidopsis. Twenty-six transcription factors showed a strong and specific induction at 0.5 hours (Supplementary Table 3). Again, these encoded homologs to proteins previously implicated in the circadian clock, de-etiolation and chloroplast greening components. For example, two orthologs of *ELONGATED HYPOCOTYL 5 (HY5)*, a master regulator of de-etiolation and *LATE ELONGATED HYPOCOTYL (LHY)*, a key clock component[31] were among this group. Sixty-two transcription factors were strongly negatively correlated with transcripts of induced photosynthesis genes, and these included an ortholog of *PHYTOCHROME INTERACTING FACTOR7 (PIF7)*, a negative regulator of phytochrome B-mediated seedling de-etiolation[32]. This group also contained seventeen transcription factors associated with hormone signalling and ten containing a homeodomain (Supplementary Table 3). Finally, twenty-two transcription factors were identified as showing an early and specific downregulation with the majority connected to developmental processes, often organ development (Supplementary Table 3). These four classes of transcription factors therefore contain members consistent with analysis from other systems and so appear to represent conserved *trans*-factors associated with de-etiolation in general. Taken together, the data indicate that as in C_3 plants, a global remodelling of gene expression is associated with the dark-to-light transition in C_4 *G. gynandra*. Moreover, many genes associated with core photosynthetic processes shared by C_3 and C_4 plants, as well as those specifically encoding components of C_4 photosynthesis showed similar dynamics during this transition. We conclude that the process of de-etiolation provides an attractive system to start to define gene regulatory networks that control the induction of C_4 photosynthesis. As the regulation of gene expression is highly combinatorial, to identify potential regulators of this process we opted for an unbiased genome-wide approach to assess chromatin accessibility as well as transcription factor-DNA interactions during de-etiolation.

Chromatin dynamics associated with de-etiolation in *G. gynandra*

To gain insight into how chromatin accessibility and *cis*-elements bound by transcription factors within such regions change during de-etiolation of *G. gynandra*, nuclei from three biological replicates across the five time points were treated with DNase-I and subjected to sequencing. From these time-points, a total of 1,145,530,978 reads were mapped to the *G. gynandra* genome, and 795,017 DNaseI-hypersensitive sites (DHSs) representing broad regulatory regions accessible to transcription factor binding were identified (Fig. 3a, Supplementary Fig. 3). The average length of these DHSs was ~610 base pairs, and distribution plots showed that their density was highest at the predicted transcription start sites (Fig. 3b). However, over the time-course the peak DHS density at transcription start sites altered such that compared with the dark, it more than doubled by two hours after transfer to light (Fig. 3b). This is consistent with the notion that exposure to light leads to a rapid increase in open chromatin around gene bodies[33]. To further investigate the extent to which accessible chromatin changed over the entire time-course, the proportion of DHS that were shared between time-points was examined (Fig. 3c). There was a major re-organisation of DHS by 0.5 hours, with 64% changing compared with tissue harvested from the dark. This remodelling continued until 2 hours after exposure to light when 71% of DHS had changed compared with the dark (Fig. 3c). From 4 hours after exposure to light, the extent to which DHS were modified was less striking. These data therefore support the notion that during the first two hours of exposure to light when assembly of the photosynthetic apparatus is being initiated (Fig. 1a-d), a relatively large restructuring of open chromatin takes place, but subsequent to this, and coincident with photosynthetic maturation there are fewer changes in chromatin accessibility. We conclude that this major re-patterning of DHS in the first two hours after exposure to light likely contributes to the changes to mRNA abundance detected soon after the dark-to-light transition, and thus assembly of the photosynthetic apparatus.

Next, changes in accessibility to DHS specifically associated with photosynthesis, C_4 pathway genes and the two classes of transcription factors that were either positively or negatively correlated to photosynthesis genes (Fig. 2d) were assessed (Fig 3d). To understand the extent to which accessibility in each DHS set was altered, dDHS scores[34] were computed. These dDHS scores quantify the change in normalised cut frequency in DHS shared between samples. All sets showed broadly similar patterns across the time-course with the C_3 and C_4 gene DHSs showing particularly similar patterns (Fig. 3d). This was also the case for the differentially expressed genes at 0.5 hours where there was little association between increased accessibility, as defined by a positive dDHS score, and an increase in transcript abundance (Supplementary Fig. 4). This suggests that changes to the binding of specific transcription factors in these open regions of chromatin, rather than changes in accessibility *per se*, must drive the increased in gene expression as C_4 photosynthesis is initiated in *G. gynandra*.

To better understand transcription factor binding sites that may be involved in activating photosynthesis gene expression, DHSs associated within induced C_3 and C_4 photosynthesis genes (clusters II and III Fig. 2c) were selected and scanned for binding sites[35,36]. Two complementary algorithms were used from the MEME suite. The first was FIMO[37] which finds individual motif occurrences predicted to be of high affinity. The second was AME[38], which determines the average odds scores across entire sequences and in so doing considers lower affinity sites, many of which would not be detected with FIMO. Using FIMO we did not detect a strong correlation (Pearson's correlation of 0.0014) between motif frequencies against a random background set of DHSs in C_3 and C_4 photosynthesis genes (Fig. 3e). To identify which motifs occurred more often than by chance alone, AME was run for both the C_3 and C_4 cistromes against shuffled input sequences as a control (Supplementary Table 4). Although there was little overlap between AME and FIMO outputs, a group of bZIP (G-box binding) motifs were enriched in both datasets (Fig. 3e). These motifs represent Light Responsive Elements (LREs) first defined as multipartite *cis*-elements in the *RbcS* promoter[22,39,40]. This finding is consistent with the requirement for photosynthesis genes to be responsive to light and implies that induction of C_4 pathway genes during de-etiolation may in part be due to these LREs. AME also identified motifs that were enriched in the C_4 cistrome compared with the C_3 cistrome (orange points in Fig. 3e). This was dominated by homeodomain motifs (specifically those from the HD-Zip I family) as well as some GT-box related Trihelix motifs, MYB and MADs motifs. In summary, although there was no strong correlation between the cistromes of C_3 and C_4 photosynthesis genes from *G. gynandra*, a group of bZIP (G-box) motifs were enriched in both gene sets, and homeodomain motifs were enriched in the cistrome of C_4 genes compared with that of C_3 genes.

A *cis*-regulatory atlas for de-etiolation in *G. gynandra*

Chromatin accessibility assays followed by *in silico* analysis of motifs within these regions identifies regulatory elements that could be important for gene regulation but does not indicate whether motifs are actually subject to transcription factor binding. We therefore carried out sequencing at sufficient depth to define DNA sequences that are protected from DNaseI digestion (Fig. 4a). Such sequences are diagnostic of strong and/or widespread protein binding and referred to as Digital Genomic Footprints (DGFs). Although DNaseI-SEQ has been used to predict transcription factor binding sites at base-pair resolution through DGF, the DNaseI enzyme possesses sequence bias that can lead to both type I and II errors in their identification and so to account for such bias, de-proteinated DNA was first analysed[41,42] (Supplementary Fig. 3). After this filtering, 300,091 DGFs corresponding to individual transcription factor binding sites across all time points were identified (Fig. 4a and Supplementary Fig. 3). This compares favourably with 282,030 DGFs in a publicly available dataset for de-etiolation of Arabidopsis[43] that was less conservative as the naked DNaseI filtering steps were not undertaken. To provide an overview of transcription factors likely binding these DGFs, all were scanned for 948 known Arabidopsis motifs. To be conservative, each DGF was only annotated to its top match.

The distribution of DGFs in gene features (e.g., promoters, exons, introns and UTRs) of *G. gynandra* changed during de-etiolation (Fig. 4b). Notably, in the first two hours of exposure to light, DGF density in promoter elements (defined as sequence two kilobase pairs upstream of predicted transcriptional start sites) and 5' UTRs increased (from 11 to 19% in the case of promoters and 27 to 41% for 5' UTRs). This finding is consistent with the increase in DHS density around predicted transcriptional start sites at this time (Fig. 3a). Coincident with the increase in DGFs in promoters and 5' UTRs, the density found in coding sequence was reduced by around half (Fig. 4b). In contrast, the density of DGFs associated with introns and 3' UTRs changed little during de-etiolation. These findings suggest changes to binding site distribution between genomic features may play an important role in transcriptional regulation and contribute to the induction of photosynthesis during de-etiolation. To test this further, we correlated the change in frequency of each motif with expression of the proximal gene (Supplementary Table 5). Positive and negative correlations between motif frequency and gene expression were used to classify motifs as either allowing activation or repression. Of the motifs predicted to act as activators, which included a number of Cysteine-rich polycomb-like (CPP) factors, significantly more were located in promoters and introns. In contrast motifs predicted to act as repressors were significantly more likely, roughly twice, to be found in exons (Fig. 4c, Supplementary Table 5). Motifs found to have no correlation to targets (neutral) were found to have intermediate frequencies suggesting a gradient from the two extremes (Supplementary Table 5).

To understand the dynamics of motifs during de-etiolation, after filtering out low frequency motifs, 743 were subjected to hierarchical clustering (Fig. 4d). Whilst some clusters were dominated by a small number of transcription factor families such as DOFs (II), AP2s (III), WRKYs (VI) and TCPs (XI) (Fig. 4d) others were composed of motifs associated with multiple families of transcription factors suggesting that the binding sites of these unrelated transcription factors could be involved in similar networks during de-etiolation. One of the most striking clusters was dominated by TCP motifs (Cluster XI) which was depleted in dark-grown tissue but became more enriched at thirty minutes and then two hours after exposure to light. To varying degrees, clusters I to V showed a similar pattern to the TCP cluster, peaking between thirty minutes to four hours with lower levels at zero hours. These contained more C2H2 and NAC (cluster I), bZIP (cluster I and IV), DOF (cluster II), AP2 (clusters III and IV) and bHLH (cluster IV) motifs. There were a few clusters, notably XVI and XVII, that were least represented at two hours and hence showed a broadly opposite dynamic. These large clusters are highly heterogeneous but included many MYB, MYB-related, GATA, G2-like, HSF and homeodomain motifs. Overall, the patterns indicate that certain groups of transcription factors are likely more involved at certain points during de-etiolation. In addition, many of these patterns appear to be complementary to one another (such as XI with XVI and XVII) and so antagonistic interactions appear likely between members of clusters I to XI and XII to XVII.

C_4 genes in C_3 *A thaliana* are induced but with reduced amplitude compared with orthologs in C_4 *G. gynandra*

In order to gain insight into the extent to which C_4 gene expression has altered compared with the ancestral C_3 state, we compared the RNA-SEQ data and the *cis*-regulatory atlas collected for C_4 *G. gynandra* with an equivalent dataset from Arabidopsis[43] (Fig. 5a). As in *G. gynandra* (Fig. 2c), many photosynthesis-related genes in C_3 Arabidopsis showed increased transcript abundance after the dark-to-light transition and three major behaviours were evident (Fig. 5b). Cluster I (red) showed no clear change, Cluster II (yellow) showed moderate, while Cluster III (green) showed strong induction (Fig. 5b). Of the genes orthologous to C_4 pathway genes, nineteen showed no clear induction while six were moderately and only one was strongly induced (Fig. 5b). Therefore, whilst some C_4 genes showed induction in the ancestral state, this was by no means universal. Indeed, in Arabidopsis a higher proportion occupied the non-responding cluster (19/27) compared with *G. gynandra* (10/30). Whilst in *G. gynandra* seven occupied the most strongly responding cluster, only one did in Arabidopsis. We next normalised the transcript abundance data of both species to enable direct comparison of expression of C_4 genes from *G. gynandra* with orthologous groups from Arabidopsis (Supplementary Fig. 5). This indicated that all genes encoding core proteins of the C_4 pathway were more strongly induced after exposure to light in C_4 *G. gynandra* than in C_3 Arabidopsis (Fig. 5c). This is consistent with re-analysis of publicly available data for maize and rice de-etiolation[44] (Supplementary Fig. 6). Thus, whilst many C_4 pathway orthologs were induced in response to light in Arabidopsis the amplitude of this response was larger in *G. gynandra*.

C_4 genes gain light responsive elements and motifs that regulate photosynthesis-associated nuclear genes in the C_3 state

To investigate how C_4 genes become more responsive to light in *G. gynandra* compared with Arabidopsis, we first identified *cis*-elements in DHS associated with these genes from each species (the C_4 cistromes). As well as genes encoding the core C_4 pathway, we included photosynthesis genes that showed clear induction in response to light (the C_3 cistromes, see Fig. 2d). Using these gene sets allowed us to investigate the extent to which C_4 genes in *G. gynandra* share a *cis*-code with photosynthesis genes, and whether this code is also found in the C_3 ancestral state. As the number of each motif may vary between species due to phylogenetic distance, to allow interspecies comparison we ranked motif enrichment within each species. With the exception of some AP2 and LOB/AS2 motifs that were abundant in *G. gynandra*, the *cis*-code of C_3 photosynthesis genes in both species was similar (Fig. 6a). In both species, the most enriched motifs located in DHS around C_3 genes included many bZIP (including HY5), bHLH (including PIF7) and BZR motifs (Supplementary Table 6). Interestingly, there was more similarity between the C_3 cistromes of Arabidopsis and *G. gynandra* (Fig. 6a) than between those of C_3 and C_4 genes from *G. gynandra* (Fig. 6a). The Arabidopsis C_3 and C_4 cistromes were more similar than the C_3 and C_4 cistromes from *G. gynandra* (Fig. 6b) indicating that C_4 genes from *G. gynandra* have not acquired large numbers of *cis*-elements associated with C_3 photosynthesis genes of Arabidopsis. Lastly, the *cis*-code of C_4 genes from Arabidopsis and *G. gynandra* were very different (Fig. 6c) strongly implying that as they are recruited into the C_4 pathway the regulation of these genes has diverged significantly. Collectively, these data indicate a greater divergence in motif composition of C_4 genes in these C_3 and C_4 species than between their C_3 photosynthesis genes.

We next assessed annotations associated with the top fifty motifs in the cistrome of C_4 genes from *G. gynandra* (Fig. 6d). Hierarchical clustering revealed two groups of particular interest. The first (green) contained motifs that were relatively highly ranked in all cistromes except C_4 genes from Arabidopsis. This group included YAB1 and CRC, as well as Class 1 bZIP G- and E-box motifs (Fig. 6d&e). As C_4 genes from Arabidopsis were poorly induced during de-etiolation we propose that these motifs are strong candidates for the light induced expression of C_4 genes from *G. gynandra* as well as C_3 photosynthesis genes from both species. The second group (red) contained motifs that were highly ranked only in the C_4 cistrome of *G. gynandra*. This includes a number of TCP and homeodomain motifs as well as the GT-box binding trihelix motifs and Class II bZIP TGAs (Fig. 6d&e). The remaining motifs showed intermediate patterns and included further G-box binding Class 2 bZIP motifs, GT-box binding trihelix motifs, E-box binding BZR motifs and a number of homeodomain motifs (Fig. 6d&e). Class I and Class II bZIP G-box, E-box, GT-box, EEs and I-box all represent Light Responsive Elements (LREs). Class I G-boxes are bound by bZIP transcription factors, whilst Class II G-boxes are bound by the TGA subfamily (Fig. 6e). The E-box is recognised by the bHLH family of transcription factors, whilst the GT-box is bound by members of the Trihelix family of transcription factors. Lastly, Evening Elements (EEs) are recognised by the MYB-related CCA1/LHY subfamily while the I-box is recognised by another subfamily of less well characterised MYB-related factors.

As LREs were amongst the most enriched motifs in the cistrome of C_4 genes from *G. gynandra* compared with Arabidopsis, we next tested whether this difference was statistically significant and therefore could contribute to the gain in induction observed for the twelve core pathway components. PhANGs are also known to be regulated by transcription factors such as GLK, GNC1 and CGA1[45,46] and although they were not in the top fifty most enriched motifs from C_4 genes of *G. gynandra*, we included motifs recognised by these PhANGs (Fig. 6e) in our analysis. For the LREs, there was no enrichment in Class I and Class II G-boxes or E-boxes in C_4 genes from *G. gynandra* (adjusted p-value < 0.05) (Fig. 6f), and whilst there was increase in the number of GT-boxes this was not statistically significant (Fig. 6f). However, both EEs and I-boxes were statistically more common in C_4 genes of *G. gynandra* than their orthologs in Arabidopsis (adjusted p-values of 0.00286 and 0.047 respectively) (Fig. 6f). Motifs recognised by GLK were very close to the $p < 0.05$ cut-off (0.0535), and the increase in CGA1 and GNC1 (C2C2-GATA) binding sites in C_4 genes of *G. gynandra* compared with Arabidopsis was statistically significant (adjusted p-value 0.0134) (Fig. 6f). We interrogated the DGF datasets from each species to test whether *in vivo* binding for any of the LRE or PhANG motifs was detected. Although we did not detect more binding *in vivo* for the PhANGs in C_4 genes of *G. gynandra* compared with Arabidopsis, we did for both EEs and I-boxes recognised by MYB-related transcription factors (Fig. 6g). These data support a model in which increased binding of LREs in C_4 genes from *G. gynandra* drive their increased responsiveness to light compared with orthologs from Arabidopsis. It is possible that regulators of PhANGs such as CGA1 and GNC1 have a low affinity for their cognate *cis*-element and so binding cannot be detected with the DNaseI assay.

We also found increased binding of homeodomain and LOB/AS2 transcription factors to C_4 genes of *G. gynandra* (Fig. 6g, Supplementary Fig. 7). Given that more homeodomain sites were present in the cistrome of C_4 genes from *G. gynandra* (Fig. 6d), we sought to gain further understanding of their potential roles. The C_4 genes *BASS2, PPa6* and *PPDK* contained six DGF with significant matches to homeodomain motifs. These three genes are preferentially expressed in mesophyll cells of C_4 plants, and so we hypothesized that transcription factors able to bind these motifs would be more strongly expressed in the mesophyll. To address this, we used publicly available data[47,48] and found that transcripts derived from the Homeodomain-Zip IV (HD-Zip IV) subfamily (including *GLABRA2* and others with a wide range of functions) were more abundant in mesophyll cells from Arabidopsis and *G. gynandra*, whilst those from HD-Zip III subfamily (including vascular development and leaf polarity regulators such as *CORONA* and *PHABULOSA*) were more abundant in bundle sheath strands from both species (Supplementary Fig. 8). Hits for HD-Zip IV binding were detected in *BASS2* and *PPa6*, but the strongest match was a motif recognised by ANL2 in *PPDK*. As the *G. gynandra* ortholog of ANL2 was most abundant during de-etiolation, these data are consistent with *ANL2* playing a role in driving mesophyll expressed C_4 genes in *G. gynandra*.

In summary, our data provide a genome-wide resource linking expression patterns to regulatory mechanisms during the establishment of C_4 photosynthesis. The distribution of transcription factor binding changed dramatically during de-etiolation with increased numbers of *cis*-elements in coding regions detected when genes were repressed. By combining expression, computational and *in vivo* transcription factor binding analysis, we propose two models relating to the molecular evolution of C_4 photosynthesis. First, the large increase in C_4 gene expression in *G. gynandra* compared with Arabidopsis is driven by acquisition of accessible LREs and motifs bound by the GNC and CGA1 transcription factors (Fig. 6h). For the EEs and I-box LREs, increased binding by transcription factors was detected *in planta*. Second, mesophyll preferential expression of genes such as *PPDK, BASS2* and *PPa6* is driven by a gain of motifs subject to binding by homeodomain (HD-Zip IV) transcription factors (Fig. 6h). Again, more binding to such *cis*-elements was detected *in planta*. To our knowledge, these data provides the first evidence based on chromatin accessibility and *in vivo* binding assays to link specific *cis*-elements and their cognate transcription factors with the evolution of C_4 photosynthesis.

DISCUSSION

Using de-etiolation to understand the establishment of C_4 photosynthesis and the etioplast to chloroplast transition in *G. gynandra*

Light is an important cue that triggers the onset of autotrophy after germination. Upon light perception, de-etiolation is initiated such that etioplasts found in dark-grown tissue transition to photosynthetically competent chloroplasts[49]. The dark-to-light transition is therefore considered an excellent system with which to understand assembly of the photosynthetic apparatus and as a consequence has long been used to study various aspects of the induction of photosynthesis in C_3 species[29,50–54]. Although there have also been a number of de-etiolation studies in C_4 species[44,55–57] our understanding of how genes encoding components of the C_4 cycle are integrated into the photosynthesis gene regulatory networks is limited[8]. Cotyledons of *G. gynandra* operate C_4 photosynthesis[58] and so seedlings undergoing de-etiolation represent a reasonable system with which to probe these processes. In the present study we focussed on changes associated with the induction of C_3 and C_4 photosynthesis and to investigate processes shared by C_3 Arabidopsis and C_4 *G. gynandra* we undertook a detailed analysis of de-etiolation.

Consistent with previous analysis of C_3 species[50,59–62] chloroplasts from *G. gynandra* followed a trajectory towards attaining full photosynthetic capacity over a 24-hour time course. In a recent systems biology study of tobacco, loss of the prolamellar body was detected after only 10 minutes of light, with granal stacking and chlorophyll accumulation then taking place[29]. Our observations are consistent with this, as prolamellar bodies in *G. gynandra* had started to disassemble 30 minutes after exposure to light. Furthermore, consistent with previous reports[60,63] chlorophyll content in *G. gynandra* increased linearly from 30 minutes to 12 hours after light exposure. Subsequent to this initial activity in building the photosynthetic apparatus, a period referred to as the building phase has been reported[29]. Consistent with previous reports[29], in *G. gynandra* this building phase was detected from 12 to 24 hours after light exposure and despite relatively little increase in chlorophyll content was associated with an increase in nuclear encoded photosynthesis transcripts.

Co-ordinated gene expression patterns during de-etiolation of *G. gynandra*

The induction of photosynthesis transcripts in *G. gynandra* between 0 to 0.5 hours was associated with upregulation of previously identified master regulators of de-etiolation. For example, we were able to detect upregulation of genes encoding well characterised proteins involved in circadian rhythms and light regulation. This included two paralogs of the master regulator *ELONGATED HYPOCOTYL 5 (HY5)* as well as an ortholog of *EARLY PHYTOCHROME RESPONSIVE 1*. Clock components that also showed this early light induction were two paralogs of *REVEILLE (RVE2)*, as well as one each of *RVE1* and *LATE ELONGATED HYPOCOTYL (LHY)*. Further, consistent with other species[32] the negative regulators of de-etiolation *PHYTOCHROME INTERACTING FACTOR7 (PIF7)* and *PIF3-LIKE5 PIL5* from *G. gynandra* were rapidly down-regulated in light.

GO-terms associated with chloroplast development and photosynthesis were enriched at the 0.5 and 24 hours timepoints, a finding consistent with the two-phase response reported in cell lines of Arabidopsis[30] where significant changes in expression of plastid and chlorophyll biosynthesis genes occurred soon after light exposure. In *G. gynandra*, transcript levels of most photosynthesis genes showed an increased over the de-etiolation time course. This is consistent with the trend observed during de-etiolation of Arabidopsis[43]. Furthermore, transcript levels of most C_4 cycle genes in both Arabidopsis and *G. gynandra* increased steadily over the de-etiolation time course. A similar induction of C_4 cycle and photosynthesis genes has been reported in C_3 rice and C_4 maize[44]. The most parsimonious explanation for these findings is that in C_3 plants genes encoding components of the C_4 pathway show a basal induction in response to light, and that this ancestral system becomes amplified during the evolution of C_4 photosynthesis.

Chromatin dynamics and transcription factor binding during de-etiolation of *G. gynandra*

DHS density around the predicted transcription start sites increased for the first 2 hours of light suggesting a gain in chromatin accessibility around the proximal promoters of *G. gynandra*. This finding is consistent with loci becoming more accessible in light[33]. However, for early responding genes changes in accessibility in DHS's (dDHS) was not associated with increased gene expression (Supplementary Fig. 3) suggesting that broad-scale changes at DHS are not strongly predictive of gene expression patterns. This may be because transcription factor complexes that then bind have antagonistic actions on gene expression and demonstrates the need for an integrated systems biology approach to gain more detailed insight into the regulatory mechanisms. As open chromatin dynamics around these genes were not strongly predictive of gene expression, we analysed the cistromes located in accessible regions. This suggested that there was low convergence in the regulatory elements available for binding in C_3 and C_4 genes from *G. gynandra* despite similar dynamics of expression.

Consistent with previous work[33] our data indicate that changes in the distribution of transcription factor binding was associated both with changes in the location of open chromatin but also the binding of individual transcription factors. Notably, between 0 and 2 hours of exposure to light, transcription factor binding events decreased in coding regions but increased in promoters and 5' UTRs. Transcription factor binding sites within codons are referred to as duons as they determine both gene expression and the amino acid code. Duons have been proposed to act as repressors of gene expression in humans[64]. In plants, duons from the *NAD-ME1* and *NAD-ME*2 genes placed downstream of the constitutive CaMV35S promoter restrict expression to bundle sheath cells of *G. gynandra*, consistent with them repressing gene expression in mesophyll cells[12,13]. In the current dataset, we found that transcription factor binding events predicted to be positive activators were 1.6 and 2.9 times more likely to be found in promoters and introns respectively compared to predicted negative regulators. Predicted negative regulators were twice as likely to be found in exons as predicted positive regulators. Interestingly, the relationship between gene enhancement and binding to introns was even more striking than that of promoters. In *Drosophila melanogaster* enhancers have been reported to be enriched in promoter, 5' UTRs but especially introns and depleted in exons[65]. We propose that de-etiolation offers an attractive system with which to investigate the importance of how the location of transcription factor binding impacts on gene expression.

Analysis of C_3 Arabidopsis reveals evolution has co-opted existing regulatory mechanisms to pattern C_4 gene expression

Compared with C_3 species, leaves from C_4 plants have increased expression of genes encoding the C_4 cycle, and decreased expression of those involved in photorespiration[66–68]. However, the dynamics with which these responses are established have not been fully defined. In the present study, photosynthesis as well as C_4 cycle genes showed light induction in both Arabidopsis and *G. gynandra*. Our datasets clearly demonstrate greater rates of transcript accumulation of core C_4 cycle genes in *G. gynandra* compared to orthologs from Arabidopsis during the de-etiolation time course (Fig. 5b). These findings are consistent with re-analysis of publicly available data for maize and rice[44] (Supplementary Fig. 6). We conclude that in at least two lineages that have independently evolved C_4 photosynthesis, genes associated with the C_4 cycle become part of gene regulatory networks that respond very strongly to the light-to-dark transition associated with de-etiolation.

The motifs of interest that were specifically over-represented in the *G. gynandra* C_4 cistrome included TGA bZIP and homeodomain motifs. These sequences therefore represent interesting candidates as regulators of C_4 specific processes. Homeodomain factors are documented to have a variety of roles, many related to development[69]. Homeodomain DGFs were detected in C_4 pathway components expressed in mesophyll cells including *BASS2, PPa6*, *PPDK, CA2/3, DIC1, NHD1* and *PPCK1*. Notably, *BASS2, PPa6* and *PPDK* contained DGF bound by HD-Zip IV factors with a mesophyll bias in both *G. gynandra* and Arabidopsis (Supplementary Fig. 7). Of this group of potential regulators, an ortholog to *ANL2* showed the highest expression of all HD-Zip IV factors in the *G. gynandra* de-etiolation time-course with its highest levels at 0 hours. In Arabidopsis, ANL2 is involved root development where it regulates the epidermal and cortical layers[70]. In leaves ANL2 expression is strongest in mesophyll cells[71]. This gene therefore appears to be a strong candidate for regulating mesophyll specific expression in C_4 leaves.

The regulation of gene expression by light is mediated by *cis*-regulatory elements known as LREs[21]. Promoter regions of many photosynthesis associated nuclear genes, including chlorophyll a/b binding proteins and *RBCS* contain these *cis*-elements[21,39,40,72]. LREs incorporate various G-, GT-, E-, Z-, I- and GATA-box elements. We found that many of these LREs dominated the fifty most common motifs in accessible DNA around C_4 pathway genes of *G. gynandra*. Comparison of C_3 and C_4 genes in Arabidopsis and *G. gynandra* showed that many G- and E-box related motifs were enriched in both cistromes and so is consistent with the notion that these elements are important for the basal response to light during de-etiolation. Whilst there was some evidence for increased numbers of motifs associated with GT-boxes and GLK binding, these just missed a statistical cut off of $p < 0.05$. However, statistically robust increases in EEs, I-boxes and motifs bound by CGA1 and GNC were detected in C_4 genes from *G. gynandra* compared with Arabidopsis. As EEs are bound by the CCA1/LHY proteins that are core components of the circadian clock[31], these data are consistent with evolution having made use of clock-regulation to enhance expression of C_4 genes in response to light. MYB-related I-box binding factors are not well characterised, although LeMYB1 from *Lycopersicum esculentum* (now *Solanum lycopersicum*) binds and activates the *RBCS3A* promoter[73]. The CGA1 and GNC transcription factors are known to regulate chloroplast biogenesis and photosynthesis-associated nuclear genes[74], and so represent an alternate part of the photosynthesis gene regulatory network to which C_4 genes have become connected.

Overall, we provide evidence that evolution appears to have repeatedly co-opted regulatory elements operating in C_3 species to pattern C_4 gene expression in C_4 plants. This includes regulators in both *cis* and *trans*. For example, increased numbers of *cis*-elements that respond to light and regulate PhANGs in C_3 Arabidopsis were found in C_4 genes from *G. gynandra*. In the case of EEs and I-boxes regulated by MYB-related transcription factors these motifs were more frequently bound in *G. gynandra* than in Arabidopsis. Furthermore, the mesophyll specific expression of a number of C_4 genes was associated with a gain of *cis*-elements known to be bound by homeodomain transcription factors in C_3 Arabidopsis. These transcription factors that belong to the HD-Zip IV family were preferentially expressed in mesophyll cells of both Arabidopsis and *G. gynandra* suggesting mesophyll specific expression of C_4 pathway genes is generated because they become integrated into an existing cell specific network that operates in the ancestral C_3 state. More broadly, the findings indicate that C_3 models such as Arabidopsis can provide significant insight into gene regulatory networks that operate in C_4 plants.

MATERIALS AND METHODS

Plant growth, chlorophyll quantitation and microscopy

G*ynandropsis gynandra* seeds were sown directly from intact pods and germinated on moist filter papers in the dark at 32 °C for 24 hours. Germinated seeds were then transferred to half strength Murashige and Skoog (MS) medium with 0.8 % (w/v) agar (pH 5.8) and grown for three days in a growth chamber at 26 °C. De-etiolation was induced by exposure to white light with a photon flux density (PFD) of 350 µmol m^{-2} s^{-1} and photoperiod of 16 hours. Whole seedlings were harvested at 0.5, 2, 4 and 24 hours after illumination (starting at 8:00 with light cycle 6:00 to 22:00). Tissue was flash frozen in liquid nitrogen and stored at -80 °C prior to processing.

For analysis of chlorophyll content, de-etiolating *G. gynandra* seedlings were flash frozen at 0, 0.5, 2, 4 or 24 hours post light exposure. 100 mg of tissue was suspended in 1 ml 80 % (v/v) acetone at 4 °C for 10 minutes prior to centrifugation at 15,700 g for 5 minutes and removal of the supernatant. The pellet was resuspended in 1 ml 80 % (v/v) acetone at 4 °C for 10 minutes, precipitated at 15,700 g for 5 minutes. Supernatants were pooled, and absorbance measured in a spectrophotometer at 663.8 nm and 646.6 nm. Total chlorophyll content determined as described previously[75].

For electron microscopy, *G. gynandra* cotyledons (~2 mm^2) were excised with a razor blade and fixed immediately in 2 % (v/v) glutaraldehyde and 2 % (w/v) formaldehyde in 0.05-0.1M sodium cacodylate (NaCac) buffer (pH 7.4) containing 2 mM calcium chloride. Samples were vacuum infiltrated overnight, washed five times in deionized water, and post-fixed in 1 % (v/v) aqueous osmium tetroxide, 1.5 % (w/v) potassium ferricyanide in 0.05 M NaCac buffer for 3 days at 4 °C. After osmication, samples were washed five times in deionized water and post-fixed in 0.1 % (w/v) thiocarbohydrazide in 0.05 M NaCac buffer for 20 minutes at room temperature in the dark. Samples were then washed five times in deionized water and osmicated for a second time for 1 hour in 2 % (v/v) aqueous osmium tetroxide in 0.05 M NaCac buffer at room temperature. Samples were washed five times in deionized water and subsequently stained in 2 % (w/v) uranyl acetate in 0.05 M maleate buffer (pH 5.5) for 3 days at 4 °C and washed five times afterwards in deionized water. Next, samples were dehydrated in an ethanol series, transferred to acetone, and then to acetonitrile. Samples were embedded in Quetol 651 resin mix (TAAB Laboratories Equipment Ltd). For transmission electron microscopy (TEM), ultra-thin sections were cut with a diamond knife, collected on copper grids and examined in a FEI Tecnai G2 transmission electron microscope (200 keV, 20 µm objective aperture). Images were obtained with AMT CCD camera. For scanning electron microscopy (SEM), ultrathin-sections were placed on plastic coverslips which were mounted on aluminium SEM stubs, sputter-coated with a thin layer of iridium and imaged in a FEI Verios 460 scanning electron microscope. For light microscopy, thin sections were stained with methylene blue and imaged by an Olympus BX41 light microscope with a mounted Micropublisher 3.3 RTV camera (Q Imaging).

RNA and DNaseI sequencing

Before processing, frozen samples were divided into two, the first being used for RNA-SEQ analysis and the second for DNaseI-SEQ. Samples were ground in a mortar and pestle and RNA extraction carried out with the RNeasy Plant Mini Kit (74904; QIAGEN) according to the manufacturer's instructions. RNA quality and integrity were assessed on a Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies). Library preparation was performed with 500 ng of high integrity total RNA (RNA integrity number > 8) using the QuantSeq 3' mRNA-SEQ Library Preparation Kit FWD for Illumina (Lexogen) following the manufacturer's instructions. Library quantity and quality were checked using Qubit (Life Technologies) and a Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies). Libraries were sequenced on NextSeq 500 (Illumina, Chesterford, UK) using single-end sequencing and a Mid Output 150 cycle run.

To extract nuclei, tissue was ground in liquid nitrogen and incubated for five minutes in 15mM PIPES pH 6.5, 0.3 M sucrose, 1 % (v/v) Triton X-100, 20mM NaCl, 80 mM KCl, 0.1 mM EDTA, 0.25 mM spermidine, 0.25 g Polyvinylpyrrolidone (SIGMA), EDTA-free proteinase inhibitors (ROCHE), filtered through two layers of Miracloth (Millipore) and pelleted by centrifugation at 4 °C for 15 min at 3600 g. To isolate deproteinated DNA, 100 mg of tissue from seedlings exposed to 24 hours light were harvested two hours into the light cycle, four days after germination. DNA was extracted using a QIAGEN DNeasy Plant Mini Kit (QIAGEN, UK) according to the manufacturer's instructions. 2x10^{8} nuclei were re-suspended at 4 °C in digestion buffer (15 mM Tris-HCl, 90 mM NaCl, 60 mM KCl, 6 mM $CaCl_2$, 0.5 mM spermidine, 1 mM EDTA and 0.5 mM EGTA, pH 8.0). DNAse-I (Fermentas) at 2.5 U was added to each tube and incubated at 37 °C for three minutes. Digestion was arrested by adding a 1:1 volume of stop buffer (50 mM Tris-HCl, 100 mM NaCl, 0.1 % (w/v) SDS, 100 mM EDTA, pH 8.0, 1 mM Spermidine, 0.3 mM Spermine, RNaseA40 µg/ml) and incubated at 55 °C for 15 minutes. 50 U of Proteinase K were then added and samples incubated at 55 °C for 1 h. DNA was isolated by mixing with 1 ml 25:24:1 Phenol:Chloroform:Isoamyl Alcohol (Ambion) and spun for 5 minutess at 15,700 g followed by ethanol precipitation of the aqueous phase. Samples were size-selected (50-400 bp) using agarose gel electrophoresis and quantified fluorometrically using a Qubit 3.0 Fluorometer (Life technologies), and a total of 10 ng of digested DNA (200 pg l^{-1}) used for library construction. Sequencing ready libraries were prepared using a TruSeq Nano DNA library kit according to the manufacturer's instructions. Quality of libraries was determined using a Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies) and quantified by Qubit (Life Technologies) and qPCR using an NGS Library Quantification Kit (KAPA Biosystems) prior to normalisation, and then pooled, diluted and denatured for paired-end sequencing using High Output 150 cycle run (2x 75 bp reads). Sequencing was performed using NextSeq 500 (Illumina, Chesterford UK) with 2x 75 cycles of sequencing.

RNA-SEQ data processing and quantification

Commands used are available on GitHub however an outline of steps was as follows. Raw single ended reads were trimmed using trimmomatic[76] (version 0.36). Trimmed reads were then quantified using salmon[77] (version 0.4.234) after building an index file for a modified *G. gynandra* transcriptome. The transcriptome was modified to create a pseudo 3' UTR sequence of 339 bp (the mean length of identified 3'UTRs) for *G. gynandra* gene models that lacked a 3' UTR sequence which was essentially an extension beyond the stop codon of the gDNA. Inclusion of this psuedo 3' UTR improved mapping rates. Each sample was then quantified using the salmon "quant" tool. All *.sf files had the "NumReads" columns merged into a single file (All_read_counts.txt) to allow analysis with both DEseq2[78] and edgeR[79]. The edgeR pipeline was run as the edgeR.R R script (on GitHub) on the All_read_counts.txt file to identify the significantly differentially expressed genes by comparing each time-point to the previous. A low expression filter step was also used. We then similarly analysed the data with the DEseq2 package using the DEseq2.R R script (on GitHub) on the same All_read_counts.txt file. This also included the PCA analysis. The intersection from both methods was used to identify a robust set of differentially regulated genes. For most further analysis of the RNA-SEQ data, mean TPM values for each time-points (from three biological replicates) was first quantile normalised and then each value divided by the mean such that values greater than 1 were higher than average. This processing facilitates comparisons between experiments across species in identifying changes to transcript abundance between orthologs.

GO enrichment analysis, identification of C_3 and C_4 gene lists and heatmap plotting

The agrigo-v2 web tool was used for GO analysis following the tools instructions for a custom background. The background was made by mapping all *G. gynandra* genes to their closest Arabidopsis blastp hit and inheriting all the GO terms associated with that gene from the TAIR gene annotation file (Athaliana_167_TAIR10.annotation_info.txt from Phytozome). Differentially expressed genes from each time-point were analysed and GO terms with significance < 10^{-5} in at least one DE gene set were kept (Supplementary Table 1, Supplementary Fig. 2). Representative GO terms were selected for plotting in a stacked barplot using the R script (Fig2B.R) and data file Fig2B_GO_term_data.txt (on GitHub).

In order to map orthologs between Arabidopsis and *G. gynandra*, OrthoFinder[80] was used. This allows more complex relationships than a 1:1 to be identified and placed into orthogroups. C_3 photosynthesis genes were first identified from Arabidopsis through the "photosynthesis" (GO:0015979) keyword search on the TAIR browse tool (https://www.arabidopsis.org/servlets/TairObject?type=keyword&id=6756) and gave ninety-two genes for Arabidopsis. Their orthologs were found in *G. gynandra* using the orthogroups generated between the two species and resulted in ninety-three C_3 photosynthesis genes. C_4 genes used in this study are considered the "core" pathway genes and are a manually curated set largely based on previous analysis[23]. Initially, multiple paralogs were included but non-induced transcripts were then filtered out. Orthologs between the two species were again identified from the orthogroups from OrthoFinder. The *G. gynandra* C_3 and C_4 gene normalised expression values were further processed with each value being divided by the row mean and log10 (log(x/row mean) plotted as a heatmap using the R script Fig2C.R on data file Fig2C_heatmap_data.txt (on GitHub). The heatmap for Arabidopsis C_3 and C_4 gene expression was made in the same way as for the *G. gynandra* data using the gene lists as previously described.

Identification of four expression behaviours of *G. gynandra* transcription factors

In order to identify transcription factors of interest during *G. Gynandra* de-etiolation we first found all potential transcription factors using homology (blastp) to the Arabidopsis transcription factor protein sequences found in the Plant Transcription Factor Database (http://plntfdb.bio.uni-potsdam.de/v3.0/downloads.php?sp_id=ATH). 2,481 potential *G. gynandra* transcription factors were then filtered to remove those with low expression and classified into four expression patterns of interest: i) Strongly and positively correlated with induced C_3 genes (>0.7 Pearson Correlation; n=21); ii) Strong and specific up-regulation at 0.5 hours (n=26); iii) Strongly and negatively correlated with induced C_3 genes <-0.7 Pearson Correlation; n=62); iv) Strong and specific down-regulation at 0.5 hours (n=22). These sets were plotted from using the Fig2D.R R script on the Fig2D_data.txt file (on GitHub).

DNaseI-SEQ data processing

The three biological replicates for each time-point were sequenced in multiple runs with one sample being chosen, based on initial QC scores, for deeper sequencing to provide the necessary depth for calling both DNaseI Hypersensitive Site (DHS) and Digital Genomic Footprints (DGF). For each sample, the raw reads from multiple sequencing runs were combined and trimmed for low quality reads using trimmomatic. These files were analysed with fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to ensure samples passed important QC parameters (see Fig3_MultiQC_summary.html on GitHub) and then mapped to the *G. gynandra* genome using bowtie2 (version 2.3.4.1) with the "--local" pre-set option. Following mapping, a bash script (Fig3_DNaseSEQ_tagAlign.sh on GitHub) was run on each bam file which in summary: filters low quality (MAPQ < 30) mapped reads and plots MAPQ distribution, removes duplicates, measures library complexity, fragment sizes, GC bias and finally makes tagAlign files. The three tagAlign files from each time-point were then merged before running another bash script (DHS_DGF_identification.sh on GitHub) for each time-point which in summary: uses "macs2 callpeak" to identify "narrowPeaks", finds the distance for each DHS to its closest transcriptional start site (TSS), calculates SPOT scores (see https://www.encodeproject.org/data-standards/dnase-SEQ/), plots DHS profiles using the deeptools bamCoverage, computeMatrix and plotProfile tools and finally calls DGF using the wellington_footprints.py program[81]. The footprints identified with a log(*p*-value) cut-off of < -10 were used for further analysis. DHS positions relative to TSS for each sample were plotted using an R script (Fig3B.R) on the file Fig3B_DHS_TSS_data.txt (on GitHub).

Analysis of DHS changes across the time-course

To quantify the overlap in DHS between samples, DHS (from the "narrowPeak" file) were sorted by their "-log10qvalue" column and only the top ranked DHS regions used until a total of 55,122,108 bp was reached which corresponded to the total length of DHS regions in the 4 hours sample which had the least. This allowed us to compare overlap between equal sized regions. These DHS regions for each time-point were then intersected in a pairwise fashion using bedtools intersect. The total length of intersecting regions was divided by 55,122,108 bp to obtain the proportion of overlap for each pairwise comparison generating the values in Fig. 3c. To compare the differential DHS (dDHS) scores for gene sets of interest, we defined promoter regions around each gene of interest as 1000 bp upstream from TSS.

We then identified DHS regions that intersected with each gene of interest (i.e., all DHS regions overlapping with a gene body or promoter) and merged these DHS regions (equivalent to an outer join). The dDHS tool[34] as part of the pyDNase package[82] to quantify changes in accessibility between consecutive time-points for a given region, where SAMPLEA precedes SAMPLEB in the time-course. Finally, these dDHS values were plotted in violin plots using the R script Fig3D.R on the data sets Fig3D_dDHS_data.txt (on GitHub).

Motif Analysis of DHS regions

All DHS intersecting with genes of interest were scanned for the presence of motifs from the DAPseq[35] and PBM[36] databases using the meme suite FIMO tool[37]. To identify shared behaviours while minimising noise, lists of both C_3 and C_4 pathway genes were filtered to keep only those that showed induction across the time course. As before, we found all DHS regions intersecting with genes of interest, including promoter, extracted the fasta sequences of these regions and scanned them for the motifs.

The same process was carried out on three sets of 500 random genes to generate a background frequency for random DHS region. Each motif frequency for a gene set (e.g., C_4 pathway gene DHS motif frequencies) was first normalised by the total number of motifs found in that set and then normalised again against the background value for that motif such that values greater than 0 indicated a higher than background rate and *vice versa* for values below 0. This approach allowed us to identify enrichments in high confidence, high affinity motif matches. While FIMO identifies higher confidence hits over a threshold, we also used the AME program[38] to identify statistically enriched motifs within the C_4 cistrome as compared to the C_3 cistrome using the webserver (http://meme-suite.org/tools/ame) and the DAP-SEQ motif database with default settings. We also used AME to identify DAP-SEQ motifs enriched in both the cistromes as compared to randomised sequence controls.

DNaseI bias correction

To reduce the proportion of false positive DGF calls caused by DNAseI cutting bias, DNaseI-SEQ was performed on de-proteinated gDNA and mapped to the *G. gynandra* genome. The hexamer cutting frequencies at the DNaseI cutting sites were used to generate a background signal profile that was incorporated into a mixture model to calculate the log-likehood ratio (FLR) for each footprint using the R package MixtureModel[42]. DGF with low confidence (FLR<0) were filtered out resulting in a reduction of 11.6 to 37.5 % of DGF per timepoint. Same pipeline was used in previous analysis[83]. The pipeline is illustrated in Supplementary Fig. 3.

DGF genomic feature distributions DGF motif frequencies and DGF-target correlation

To identify the distribution of DGF across genomic features we used bedtools intersect to find the frequency of intersection between the DGF with features in the genome annotation gff3 file promoter (2000 bp upstream of TSS), 5’ UTR, CDS, intron, 3’ UTR and intergenic). These frequencies were divided by the total length of each feature across the genome to determine a density of DGF per feature and these values were plotted as a pie chart for each time point.

In order to link DGF to possible functions, we scanned all DGF using the meme suite fimo tool for both DAPseq and PMB motifs as described above. To visualise how motif frequencies changed during de-etiolation, the frequency of each motif at each time-point was first normalised by the total number of motifs at that time point and then each value was mean centred across the time-course for plotting as a heatmap using the Fig4C.R R script on the Fig4C_heatmap_data.txt file (on GitHub). This hierarchical clustering was then manually grouped and word clouds generated for the motif transcription factor families using an online tool (https://www.wordclouds.com/) for each motif cluster.

Once individual DGF were annotated with potential motifs, we correlated the changes in frequency of each motif with the mean expression of all the potential targets. The changes in frequency of each motif were used as a proxy of their factor's abundance and/or activity and potential targets are identified as those genes lying closest to a DGF with a specific motif. A strong positive correlation was used to suggest positive regulation while a strong negative correlation suggests inhibitory regulation.

Arabidopsis RNA-SEQ and DNaseI-SEQ

In order to carry out comparative analysis between *G. gynandra* and Arabidopsis, an analogous de-etiolation time-course[43] was reprocessed in the same way as the *G. gynandra* data. Arabidopsis DNaseI-SEQ data was mapped to the TAIR9 genome and RNA-SEQ was mapped using Salmon to the Araport11 transcriptome, followed by the use of tximport to collapse expression values for all isomers into a single value, a step not required for *G. gynandra* as it lacks isomer information. To allow inter-species comparisons, as with *G. gynandra* the Arabidopsis RNA-SEQ data was quantile normalised and then each value divided by the samples mean expression value. Normalised expression values for the core set of C_4 pathway genes from *G. gynandra* were compared with orthologs from Arabidopsis, and when there was more than one paralog identified, the most highly expressed was selected. Line plots were generated using the Fig5B.R R script on the Fig5B_C4_pathway_data.txt file (on GitHub). To analyse and compare motifs between species, we ranked motifs by their normalised frequencies against the background for each gene sets DHS (C_4 pathway and C_3 photosynthesis from both *G. gynandra* and *A. thailiana*). These sets were filtered to remove genes that were not induced during the time-course. The top 50 motifs from each set were then plotted against their rank in other sets. As motifs that were highly ranked in *G. gynandra* C_4 genes were of particular interest, these were plotted as a heatmap using the Fig5G.R R script on the Fig5G_motif_rank_data.txt file (on GitHub).

To create cumulative line plots a number of steps were required. First, as individual motif frequencies are low for any given small set of genes (e.g. 105174 DGF found in the 24 hour time-point giving ~3 per gene loci) we grouped motifs based on motif clustering using the RSAT motif matrix clustering tool with default settings (http://rsat.sb-roscoff.fr/matrix-clustering_form.cgi). This meant all members from the same cluster were treated as one motif group (for example, all TCP motifs are found in group 10). For motif groups see the file DAPseq_PBM_Motif_Matrix_Clustering.txt (on GitHub). These values were then normalised for each time-point by dividing by the total number of motifs such that the values represented a proportion of the total. These values were then plotted in a cumulative line plot using the Fig5H.R R script on the Fig5H_data.txt (on GitHub).

The quantification of light and chloroplastic regulatory elements was carried out on the gDNA sequence of the 10 strongly induced *G. gynandra* core C_4 pathway gene loci, including a 1500 bp promoter region, and their most highly expressed Arabidopsis ortholog loci. In summary, matches to the highly conserved core sequences of each element were counted and compared between the two species gene sets (see GitHub for command example). A one-tailed t-test was used to show no significant increase in the frequencies of these elements in the *G. gynandra* genes as compared to the Arabidopsis orthologs.

Phylogeny and cell specific expression of homeodomain factors in Arabidopsis and *G. gynandra*

Homeodomain factors were identified from Arabidopsis transcription factor databases and all potential transcription factors in *G. gynandra* were identified by sequence similarity. Phygenetic trees of the protein sequences from both species were made using the ete3 tool. The tree was loaded into the iTOL web tool where the log(BS/M or Whole Leaf) ratio of each gene was added to the tree. This expression data was obtained from publicly available datasets[47,48].

Re-processing of *O. sativa* and *Z. mays* de-etiolation time-course RNA-SEQ

Data from the monocot de-etiolation study[44] was downloaded from the Short Read Archive (SRX766219). Reads for both species were quantified using Salmon quant with the *Z. mays* reads being mapped to Zm-B73-REFERENCE-NAM-5.0_Zm00001e.1.cdna.fa file available from MaizeDB while *O. sativa* reads were mapped to Osativa_323_v7.0.cds_primaryTranscriptOnly.fa available from Phytozome. TPM values were quantile normalised and then each value divided by the sample mean. *O. sativa* C_4 orthologs were identified using orthofinder to identify orthogroups with the Arabidopsis C_4 orthologs used in this study. *Z. mays* C_4 genes were identified by blasting to these same Arabidopsis genes. Line plots were then made grouping all putative orthologs.

ACCESSION NUMBERS

Raw sequencing data files are deposited in The National Center for Biotechnology Information (PRJNA640984). For full methods, commands, and scripts, see GitHub (https://github.com/hibberd-lab/Singh-Stevenson-Gynandra).

ACKNOWLEDGMENTS

We thank the Cambridge Advanced Imaging Centre, University of Cambridge for resin embedding the electron microscopy samples and providing the access to electron microscopes. The work was supported by BBSRC Grant BBP0031171 and European Research Council Grant 694733 Revolution to J.M.H.

AUTHOR CONTRIBUTIONS

PS, SRS and JMH designed the study. PS carried out the experimental work. SRS and PS analysed the data. IRL performed de-proteinated DNaseI data analysis. GR assisted in DNaseI assays and library preparations. TBS and PS carried out electron microscopy. PS, SRS and JMH wrote the article and prepared the figures.

FIGURE LEGENDS

Fig. 1: Establishment of photosynthesis in *G. gynandra.* (a) Representative images of *Gynandropsis gynandra* seedlings illustrating greening and unhooking of the cotyledons. (b) Total chlorophyll over the time-course (data shown as means from three biological replicates at each time point, ± one standard deviation from the mean). The first four hours show an exponential increase (inset). Bar along the x-axis indicates periods of light (0-14 hours), dark (14-22 hours) and light (22-24 hours). (c-d) Representative transmission electron microscope images of Mesophyll (c) and Bundle Sheath (d) chloroplasts of de-etiolating *G. gynandra* seedlings at 0, 0.5, 2, 4 and 24 hours after exposure to light. Asterisks and arrowheads indicate the prolamellar body and photosynthetic membranes respectively. Samples at each time were taken for RNA-SEQ and DNaseI-SEQ. Scale bars represent 0.5 mm for seedlings and 50 µm for cotyledons (a), and 500 nm (c-d).

Fig. 2: Changes in transcript abundance during greening of *G. gynandra*. (a) Principal component analysis of RNA-SEQ datasets. The three biological replicates from each timepoint of de-etiolating *G. gynandra* seedlings (0, 0.5, 2, 4 and 24 hours) form distinct clusters. (b) Enriched GO terms between consecutive timepoints for up- and down-regulated genes. (c) Heatmap illustrating changes in transcript abundance of photosynthesis (grey sidebar) and C_4 photosynthesis genes (black sidebar) during the time-course. Data are shown after normalisation of expression data with each gene plotted on a row and centred around the row mean. Colour-coding of the dendrograms (red, yellow and green) highlight expression clusters representing none, moderate and strong induction respectively. (d) Line graphs depicting dynamics of transcription factors positively or negatively correlated with induced photosynthesis genes, or that showed early (0.5 hours) up- or down- regulation during the de-etiolation time-course. Values shown are normalised and centred around the mean of each gene.

Fig. 3: Profiling of open chromatin during de-etiolating of *G. gynandra.* (a) Schematic illustrating DNaseI-SEQ and the total number of DNaseI-hypersensitive sites (DHSs) detected. (b) Density of open chromatin plotted relative to the nearest Transcription Start Site (TSS). Inset highlights maximum density overlapping with the TSS at each time point. (c) Percentage of DHSs non-overlapping at each timepoint. (d) Violin plots depicting changes in DHS accessibility (dDHS) associated with photosynthesis genes, C_4 photosynthesis genes, and transcription factors that were positively or negatively correlated with photosynthesis genes. Changes are relative to the previous timepoint, n values are for the number of DHS regions quantified. (e) Scatter plot of FIMO motif frequencies in C_3 and

C_4 photosynthesis gene from (log10 normalised motif frequency/normalised background frequency). Motifs annotated in orange and the associated Wordcloud highlight those enriched in the C_4 cistrome compared with the C_3 cistrome, and those in red indicate bZIP motifs enriched in both the C_3 and C_4 cistromes from both FIMO and AME analysis.

Fig. 4: Transcription factor binding atlas for de-etiolating seedlings *G. gynandra.* (a) Schematic illustrating sampling, number of Digital Genomic Footprints (DGF) identified and representative density plot of DGF positions relative to the nearest transcription start site (TSS). (b) Pie-charts summarising the density of DGF among genomic features. Promoters are defined as sequence < 2000 base pairs upstream of TSSs while intergenic represent any regions not overlapping with other features. Values indicate densities of DGFs in each feature as proportions. (c) Bar chart showing the percentage of DGFs predicted to function either as activators (coral bars) or repressors (turquoise bars) lying within gene features of target genes. Statistically significant differences were found for the promoters, CDS and intronic regions using a Chi square goodness of fit test ("*"). (d) Heatmap of motif frequencies (log10 sample normalised motif frequency/row mean) during de-etiolation. To illustrate identity and heterogeneity of motif groups clusters were annotated with Wordclouds.

Fig. 5: Comparison of transcript abundance for photosynthetic genes during de-etiolation of C_3 *Arabidopsis thaliana* and C_4 *G. gynandra*. (a) Schematic illustrating RNA-SEQ of Arabidopsis*.* (b) Expression patterns of photosynthesis genes (grey sidebar) and C_4 orthologs (black sidebar) during de-etiolation. Heatmap illustrating gene expression with each gene being represented by a row, and data centred around the row mean. Dendrograms (red, yellow and green) highlight distinct expression clusters representing no clear, moderate, or strong induction. (c) Line graphs depicting quantile normalised and mean divided expression patterns of twelve C_4 orthologs in C_3 Arabidopsis and C_4 *G. gynandra*. Where there is more than a 1:1 relationship between genes, the most abundant paralog from each orthogroup is presented.

Fig. 6: Comparative analysis of potential regulatory mechanisms for de-etiolating seedlings of C_3 *Arabidopsis thaliana* and C_4 *G. gynandra*. Scatter plots showing the most enriched motifs in each cistrome (where 1 represents the most enriched motif). (a) Top 50 motifs in photosynthesis genes of C_3 Arabidopsis (At) and C_4 *G. gynandra* (Gg), (b) C_4 and photosynthesis genes of C_3 Arabidopsis, and (c) C_4 genes from C_3 Arabidopsis and *C. gynandra*. Motifs from the cistromes of C_3 and C_4 genes that showed induction during de-etiolation. (d) Heatmap of the top 50 motifs from DHSs of C_4 genes in *G. gynandra*

compared with their ranking in C_4 genes of Arabidopsis and photosynthesis genes in both species (log2 of the motif ranks across all four cistrome sets). Two distinct groups are highlighted with green motifs being highly ranked (more enriched) in all four cistromes while the red motifs are those specifically highly ranked in the *G. gynandra* C_4 cistrome. Motifs characterised as light-regulatory elements (LREs) are labelled with symbols used in (e). (e) Sequence logos highlighting different classes of Light Responsive Elements (LREs) and regulators of photosynthesis-associated nuclear genes (PhANGs). Seqlogos were generated from DAP-SEQ and PBM consensus motifs for all members of each type. (f) Analysis of which LREs and regulators of PhANGs are statistically enriched in cistromes of C_4 genes from *G. gynandra* and Arabidopsis. AME generates likelihood score for over-representation (-1*log(adjusted p-value), y-axis), and the adjusted $p < 0.05$ is illustrated with a dashed line. (g) Transcription factor binding sites associated with EE and I-box binding as well as the homeodomain and LOB/AS2 families dis-proportionally found in C_4 genes from *G. gynandra* compared with orthologs from Arabidopsis, and photosynthesis genes in both species. Values plotted are the motif proportion of the total number of DGF at each sample over the time-course such that differences between and within experiments were normalised. (h) Model illustrating association between enhanced C_4 cycle gene expression in *G. gynandra* compared with Arabidopsis and gain of *cis*-elements bound by MYB-related and C2C2-GATA transcription factors as well as the gain of homeodomain binding sites in mesophyll expressed genes in C_4 *G. gynandra*.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Fig. 1: Representative light and scanning electron microscope (SEM) images of 0 hours (A) and 24 hours (B) de-etiolating *G. gynandra* seedlings at 0 and 24 hours after exposure to light. Scale bars represent 100 μm for light microscope images, and 500 nm for SEM.

Supplementary Fig. 2: GO term enrichment analysis for differentially expressed genes as compared to the previous time point. Significantly enriched GO terms were identified using AgriGov2 using a custom *G. gynandra* background built by mapping *G. gynandra* proteins to their closest match in Arabidopsis and inheriting their terms from the TAIR10 annotations. Values plotted are -log10(FDR) and values derived from the up-regulated gene sets are shown in red while those form the down-regulated are shown in blue. Many light and photosynthesis-related terms are enriched in the 0.5 hours up-regulated genes. Many primary and secondary metabolism terms are enriched in the 24 hours up-regulated genes suggesting that photosynthates are being produced by the end of the time course.

Supplementary Fig. 3: Pipeline for DNaseI-SEQ data processing. On the top left-hand side, pooled reads went through quality control before DHS identification using MACS2 peakcalling. The DHSs, representing accessible chromatin regions, are then searched for DGF using the pyDNase package. These DGF are prone to distorting effects due to DNaseI bias in gDNA digestion. On the top right, the pipeline for identifying this bias is shown which includes the DNaseI digestion of deproteinised ("naked") gDNA. This generates 6-mer frequencies at each cut site which is used as input for the FootPrintMixture.R tool which scores the Footprint Likelihood Ratio (FLR) of each DGF (likelihood of being a true positive). DGF with FLR < 0 were removed leaving a final set of DGF which were used for analysis. Heatmap of cut patterns are shown centred around each DGF.

Supplementary Fig. 4: Violin plots showing the distributions of dDHS scores for DHS overlapping with differentially expressed genes at 0.5 hours with both up-regulated (A) and down-regulated shown (B). Mean values are shown as line. Positive dDHS scores represent an increase in DHS accessibility and negative values represent the opposite. No clear association is observed with up-regulated genes and positive dDHS values nor down-regulated genes with negative dDHS values.

Supplementary Fig. 5: Line plots showing mean normalised expression values at each time point for both *G. gynandra* and Arabidopsis across their respective de-etiolation time-courses. Data shown for twenty orthogroups induced during de-etiolation with paralogs shown. Values are quantile normalised followed by dividing by the sample mean to facilitate expression dynamics and abundance comparisons across the species.

Supplementary Fig. 6: Line plots showing mean normalised expression values at each time point for both *O. sativa* and *Z. mays* across their respective de-etiolation time-courses. Data from a monocot de-etiolation study[44] and processed in the same way as described for the data in this study. Many orthogroups showed a similar pattern as for the *G. gynandra* and Arabidopsis comparison with one or more C_4 (*Z. mays*) paralog showing much higher abundance than other members of the orthogroup.

Supplementary Fig. 7: DGF binding in different C_4 cycle genes during de-etiolation time course in *G. gynandra*.

Supplementary Fig. 8: Unrooted phylogenetic tree of homeodomain protein sequences taken from both *G. gynandra* and Arabidopsis. Analysis was carried out using the ete3 pipeline and visualised with the iTOL web tool. Each leaf is annotated, where available, with the ratio of the bundle sheath to mesophyll expression values taken from publicly available datasets. The HD-Zip IV family is shown to be consistently mesophyll preferentially expressed in both species in contrast to the HD-Zip III family. Major family names and individual gene names of interest are shown.

REFERENCES

1. Bowes, G., Ogren, W. L. & Hageman, R. H. Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase. *Biochem. Biophys. Res. Commun.* **45**, 716–722 (1971).
2. Tolbert, N. E. & Essner, E. Microbodies: Peroxisomes and glyoxysomes. *J. Cell Biol.* **91**, 271s-283s (1981).
3. Bauwe, H., Hagemann, M. & Fernie, A. R. Photorespiration: players, partners and origin. *Trends in Plant Science* vol. 15 330–336 (2010).
4. Hatch, M. D. C4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. *BBA Reviews On Bioenergetics* vol. 895 81–106 (1987).
5. Furbank, R. T. Evolution of the C_4 photosynthetic mechanism: Are there really three C_4 acid decarboxylation types? *Journal of Experimental Botany* vol. 62 3103–3108 (2011).
6. Sage, R. F. & Zhu, X. G. Exploiting the engine of C_4 photosynthesis. *J. Exp. Bot.* **62**, 2989–3000 (2011).
7. Sage, R. F. A portrait of the C_4 photosynthetic family on the 50th anniversary of its discovery: Species number, evolutionary lineages, and Hall of Fame. *Journal of Experimental Botany* vol. 67 4039–4056 (2016).
8. Hibberd, J. M. & Covshoff, S. The Regulation of Gene Expression Required for C_4 Photosynthesis. *Annu. Rev. Plant Biol.* **61**, 181–207 (2010).
9. Burnell, J. N., Suzuki, I. & Sugiyama, T. Light Induction and the Effect of Nitrogen Status upon the Activity of Carbonic Anhydrase in Maize Leaves. *Plant Physiol.* **94**, 384–387 (1990).
10. Glackin, C. A. & Grula, J. W. Organ-specific transcripts of different size and abundance derive from the same pyruvate, orthophosphate dikinase gene in maize. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 3004–3008 (1990).
11. Gowik, U. *et al.* cis-regulatory elements for mesophyll-specific gene expression in the C_4 plant *Flaveria trinervia*, the promoter of the C_4 phosphoenolpyruvate carboxylase gene. *Plant Cell* **16**, 1077–1090 (2004).

12. Brown, N. J. *et al.* Independent and parallel recruitment of preexisting mechanisms underlying C_4 photosynthesis. *Science* **331**, 1436–1439 (2011).

13. Reyna-Llorens, I. *et al.* Ancient duons may underpin spatial patterning of gene expression in C_4 leaves. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1931–1936 (2018).

14. Williams, B. P. *et al.* An untranslated cis-element regulates the accumulation of multiple C_4 enzymes in *Gynandropsis gynandra* mesophyll cells. *Plant Cell* **28**, 454–465 (2016).

15. Kajala, K. *et al.* Multiple Arabidopsis genes primed for recruitment into C_4 photosynthesis. *Plant J.* **69**, 47–56 (2012).

16. Stockhaus, J., Poetsch, W., Steinmüller, K. & Westhoff, P. Evolution of the C_4 phosphoenolpyruvate carboxylase promoter of the C_4 dicot *Flaveria trinervia*: an expression analysis in the C_3 plant tobacco. *MGG Mol. Gen. Genet.* **245**, 286–293 (1994).

17. Stockhaus, J. *et al.* The promoter of the gene encoding the C_4 form of phosphoenolpyruvate carboxylase directs mesophyll-specific expression in transgenic C_4 *Flaveria* spp. *Plant Cell* **9**, 479–489 (1997).

18. Matsuoka, M., Kyozuka, J., Shimamoto, K. & Kano-Murakami, Y. The promoters of two carboxylases in a C_4 plant (maize) direct cell-specific, light-regulated expression in a C_3 plant (rice). *Plant J.* **6**, 311–319 (1994).

19. Nomura, M. *et al.* The evolution of C_4 plants: Acquisition of cis-regulatory sequences in the promoter of C_4-type pyruvate, orthophosphate dikinase gene. *Plant J.* **22**, 211–221 (2000).

20. Nomura, M. *et al.* The promoter for C_4-type mitochondrial aspartate aminotransferase does not direct bundle sheath-specific expression in transgenic rice plants. *Plant Cell Physiol.* **46**, 743–753 (2005).

21. Argüello-Astorga, G. R. & Herrera-Estrella, L. R. Ancestral multipartite units in light-responsive plant promoters have structural features correlating with specific phototransduction pathways. *Plant Physiol.* **112**, 1151–1166 (1996).

22. Acevedo-Hernández, G. J., León, P. & Herrera-Estrella, L. R. Sugar and ABA responsiveness of a minimal RBCS light-responsive unit is mediated by direct binding of ABI4. *Plant J.* **43**, 506–519 (2005).

23. Burgess, S. J. *et al.* Ancestral light and chloroplast regulation form the foundations for C_4 gene expression. *Nat. Plants* **2**, (2016).

24. Pribil, M., Labs, M. & Leister, D. Structure and dynamics of thylakoids in land plants. *J. Exp. Bot.* **65**, 1955–1972 (2014).

25. Bahl, J., Francke, B. & Monéger, R. Lipid composition of envelopes, prolamellar bodies and other plastid membranes in etiolated, green and greening wheat leaves.

Planta **129**, 193–201 (1976).

26. Ryberg, M. & Sundqvist, C. Characterization of prolamellar bodies and prothylakoids fractionated from wheat etioplasts. *Physiol. Plant.* **56**, 125–132 (1982).
27. Ryberg, M. & Sundqvist, C. The regular ultrastructure of isolated prolamellar bodies depends on the presence of membrane-bound NADPH-protochlorophyllide oxidoreductase. *Physiol. Plant.* **73**, 218–226 (1988).
28. Park, H., Kreunen, S. S., Cuttriss, A. J., DellaPenna, D. & Pogson, B. J. Identification of the carotenoid isomerase provides insight into carotenoid biosynthesis, prolamellar body formation, and photomorphogenesis. *Plant Cell* **14**, 321–332 (2002).
29. Armarego-Marriott, T. *et al.* Highly resolved systems biology to dissect the etioplast-to-chloroplast transition in tobacco leaves. *Plant Physiol.* **180**, 654–681 (2019).
30. Dubreuil, C. *et al.* Establishment of photosynthesis through chloroplast development is controlled by two distinct regulatory phases. *Plant Physiol.* **176**, 1199–1214 (2018).
31. Mizoguchi, T. *et al.* LHY and CCA1 are partially redundant genes required to maintain circadian rhythms in Arabidopsis. *Dev. Cell* **2**, 629–641 (2002).
32. Leivar, P. *et al.* The Arabidopsis phytochrome-interacting factor PIF7, together with PIF3 and PIF4, regulates responses to prolonged red light by modulating phyB levels. *Plant Cell* **20**, 337–352 (2008).
33. Liu, Y. *et al.* Genome-wide mapping of DNase i hypersensitive sites reveals chromatin accessibility changes in Arabidopsis euchromatin and heterochromatin regions under extended darkness. *Sci. Rep.* **7**, 1–15 (2017).
34. He, H. H. *et al.* Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res.* **22**, 1015–1025 (2012).
35. O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280–1292 (2016).
36. Franco-Zorrilla, J. M. *et al.* DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2367–2372 (2014).
37. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
38. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**, 165 (2010).
39. Green, P. J., Kay, S. A. & Chua, N. H. Sequence-specific interactions of a pea nuclear factor with light-responsive elements upstream of the rbcS-3A gene. *EMBO J.* **6**, 2543–2549 (1987).
40. Giuliano, G. *et al.* An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 7089–7093 (1988).

41. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **11**, 73–78 (2014).

42. Yardimci, G. G., Frank, C. L., Crawford, G. E. & Ohler, U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.* **42**, 11865–11878 (2014).

43. Sullivan, A. M. *et al.* Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep.* **8**, 2015–2030 (2014).

44. Xu, J., Bräutigam, A., Li, Y., Weber, A. P. M. & Zhu, X.-G. Systems analysis of cis-regulatory motifs in C_4 photosynthesis genes using maize and rice leaf transcriptomic data during a process of de-etiolation. *J. Exp. Bot.* **67**, 5105–5117 (2016).

45. Waters, M. T. *et al.* GLK transcription factors coordinate expression of the photosynthetic apparatus in Arabidopsis. *Plant Cell* **21**, 1109–1128 (2009).

46. Xu, Z., Casaretto, J. A., Bi, Y.-M. & Rothstein, S. J. Genome-wide binding analysis of AtGNC and AtCGA1 demonstrates their cross-regulation and common and specific functions. *Plant Direct* **1**, e00016 (2017).

47. Aubry, S., Smith-Unna, R. D., Boursnell, C. M., Kopriva, S. & Hibberd, J. M. Transcript residency on ribosomes reveals a key role for the *Arabidopsis thaliana* bundle sheath in sulfur and glucosinolate metabolism. *Plant J.* **78**, 659–673 (2014).

48. Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D. & Hibberd, J. M. Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C_4 Photosynthesis. *PLoS Genet.* **10**, (2014).

49. Kirk, J. T. 0. & Tilney-Bassett, R. A. E. *The Plastids*. (Elsevier/North-Holland Biomedical Press, 1967).

50. Boffey, S. A., Selldén, G. & Leech, R. M. Influence of Cell Age on Chlorophyll Formation in Light-grown and Etiolated Wheat Seedlings. *Plant Physiol.* **65**, 680–684 (1980).

51. Virgin, H. I., Kahn, A. & von Wettstein, D. The Physiology of Chlorophyll Formation in Relation to Structural Changes in Chloroplasts. *Photochem. Photobiol.* **2**, 83–91 (1963).

52. Thorne, S. W. & Boardman, N. K. Formation of Chlorophyll b , and the Fluorescence Properties and Photochemical Activities of Isolated Plastids from Greening Pea Seedlings . *Plant Physiol.* **47**, 252–261 (1971).

53. Høyer-Hansen, G. & Simpson, D. J. Changes in the polypeptide composition of internal membranes of barley plastids during greening. *Carlsberg Res. Commun.* **42**, 379–389 (1977).

54. Bennett, J. Chloroplast protein phosphorylation and the regulation of photosynthesis. *Physiol. Plant.* **60**, 583–590 (1984).

55. Demeter, S., Mustardy, L. & Machowicz, E. The development of the intense circular dichroic signal during granum formation in greening etiolated maize. *Biochem. J.* **156**, 469–472 (1976).

56. Lonosky, P. M. *et al.* A Proteomic Analysis of Maize Chloroplast Biogenesis. *Plant Physiol.* **134**, 560–574 (2004).

57. Cahoon, A. B., Takacs, E. M., Sharpe, R. M. & Stern, D. B. Nuclear, chloroplast, and mitochondrial transcript abundance along a maize leaf developmental gradient. *Plant Mol. Biol.* **66**, 33–46 (2008).

58. Koteyeva, N. K., Voznesenskaya, E. V, Roalson, E. H. & Edwards, G. E. Diversity in forms of C_4 in the genus *Cleome* (Cleomaceae). *Ann. Bot.* **107**, 269–283 (2011).

59. Bahl, J., Francke, B. & Monéger, R. Lipid composition of envelopes, prolamellar bodies and other plastid membranes in etiolated, green and greening wheat leaves. *Planta* **129**, 193–201 (1976).

60. Baker, N. R. & Butler, W. L. Development of the Primary Photochemical Apparatus of Photosynthesis during Greening of Etiolated Bean Leaves. *Plant Physiol.* **58**, 526–529 (1976).

61. Boardman, N. K. Development of Chloroplast Structure and Function. in *Photosynthesis I* 583–600 (Springer Berlin Heidelberg, 1977). doi:10.1007/978-3-642-66505-9_42.

62. Krupinska, K. & Apel, K. Light-induced transformation of etioplasts to chloroplasts of barley without transcriptional control of plastid gene expression. *MGG Mol. Gen. Genet.* **219**, 467–473 (1989).

63. Boasson, R. & Laetsch, W. M. Chloroplast replication and growth in tobacco. *Science* **166**, 749–751 (1969).

64. Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, 888–903 (2013).

65. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

66. Bräutigam, A. *et al.* An mRNA blueprint for C_4 photosynthesis derived from comparative transcriptomics of closely related C_3 and C_4 species. *Plant Physiol.* **155**, 142–156 (2011).

67. Külahoglu, C. *et al.* Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C_3 and C_4 plant species. *Plant Cell* **26**, 3243–3260 (2014).

68. Gowik, U., Bräutigam, A., Weber, K. L., Weber, A. P. M. & Westhoff, P. Evolution of C_4 photosynthesis in the genus *Flaveria*: How many and which genes does it take to make C_4? *Plant Cell* **23**, 2087–2105 (2011).

69. Ariel, F. D., Manavella, P. A., Dezar, C. A. & Chan, R. L. The true story of the HD-Zip family. *Trends in Plant Science* vol. 12 419–426 (2007).

70. Kubo, H. & Hayashi, K. Characterization of root cells of *anl2* mutant in *Arabidopsis thaliana*. *Plant Sci.* **180**, 679–685 (2011).

71. Kubo, H., Kishi, M. & Goto, K. Expression analysis of *ANTHOCYANINLESS2* gene in Arabidopsis. *Plant Sci.* **175**, 853–857 (2008).

72. Martinez, E. Multi-protein complexes in eukaryotic gene transcription. *Plant Mol. Biol.* **50**, 925–947 (2002).

73. Rose, A., Meier, I. & Wienand, U. The tomato I-box binding factor LeMYBI is a member of a novel class of Myb-like proteins. *Plant J.* **20**, 641–652 (1999).

74. Richter, R., Behringer, C., Müller, I. K. & Schwechheimer, C. The GATA-type transcription factors GNC and GNL/CGA1 repress gibberellin signaling downstream from DELLA proteins and phytochrome-interacting factors. *Genes Dev.* **24**, 2093–2104 (2010).

75. Porra, R. J., Thompson, W. A. & Kriedemann, P. E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *BBA - Bioenerg.* **975**, 384–394 (1989).

76. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

77. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

78. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

79. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

80. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

81. Piper, J. *et al.* Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41**, (2013).

82. Piper, J. *et al.* Wellington-bootstrap: Differential DNase-seq footprinting identifies cell-type determining transcription factors. *BMC Genomics* **16**, 1000 (2015).

83. Burgess, S. J. *et al.* Genome-wide transcription factor binding in leaves from C_3 and C_4 grasses. *Plant Cell* **31**, 2297–2314 (2019).

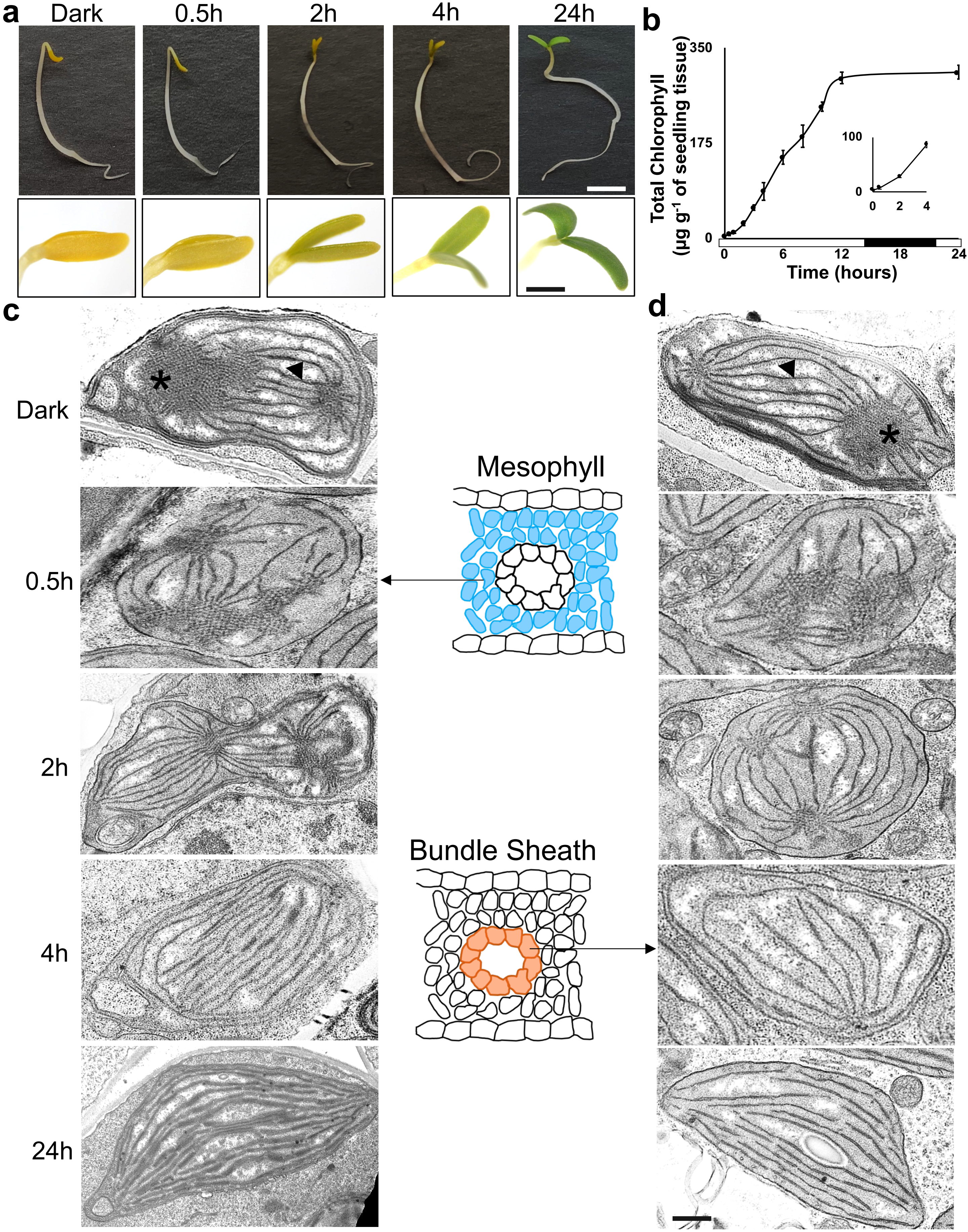


Fig. 1: Establishment of photosynthesis in *G. gynandra*. (a) Representative images of *Gynandropsis gynandra* seedlings illustrating greening and unhooking of the cotyledons. (b) Total chlorophyll over the time-course (data shown as means from three biological replicates at each time point, ± one standard deviation from the mean). The first four hours show an exponential increase (inset). Bar along the x-axis indicates periods of light (0-14 hours), dark (14-22 hours) and light (22-24 hours). (c-d) Representative transmission electron microscope images of Mesophyll (c) and Bundle Sheath (d) chloroplasts of de-etiolating *G. gynandra* seedlings at 0, 0.5, 2, 4 and 24 hours after exposure to light. Asterisks and arrowheads indicate the prolamellar body and photosynthetic membranes respectively. Samples at each time were taken for RNA-SEQ and DNaseI-SEQ. Scale bars represent 0.5 mm for seedlings and 50 µm for cotyledons (a) and 500 nm (c-d).

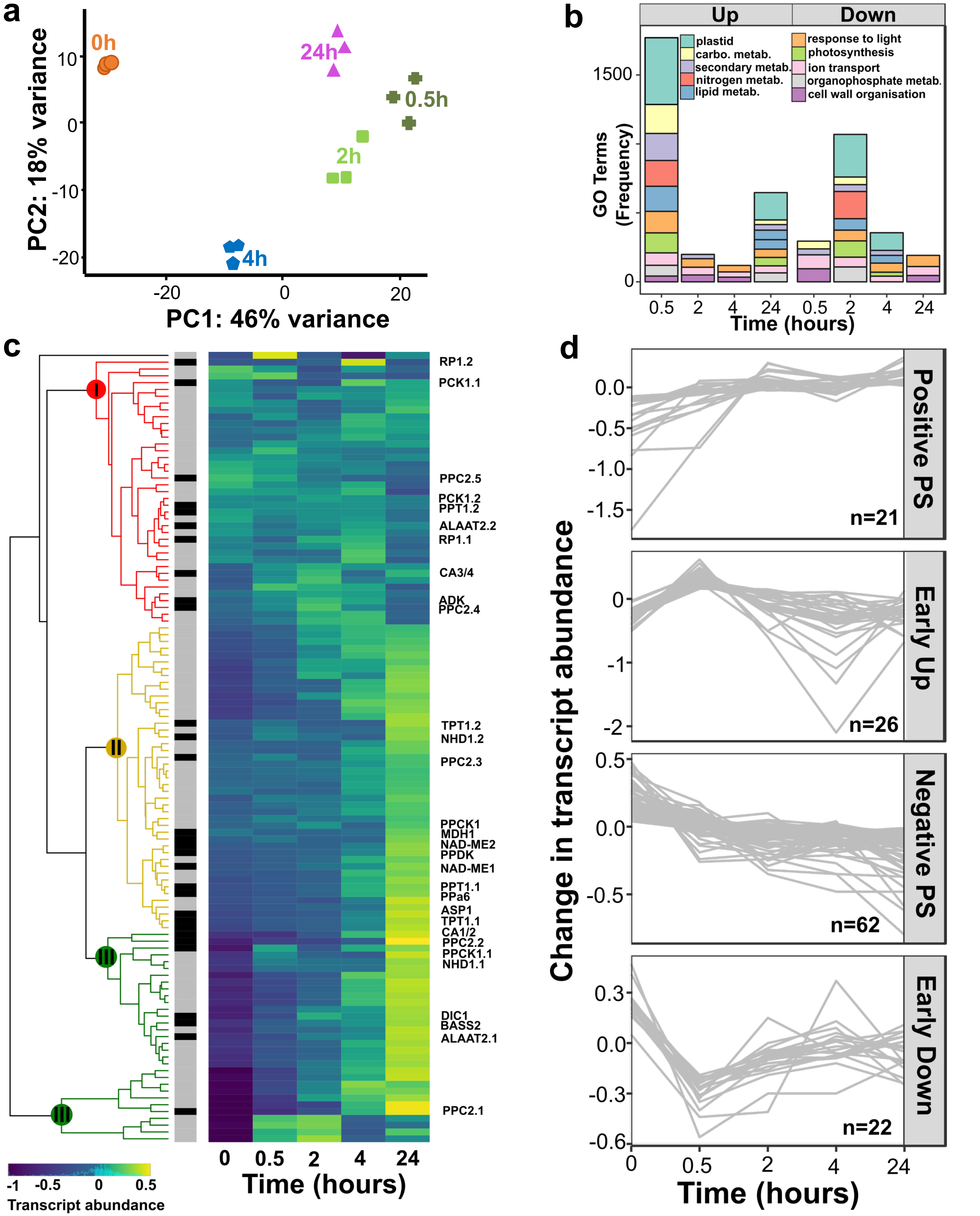
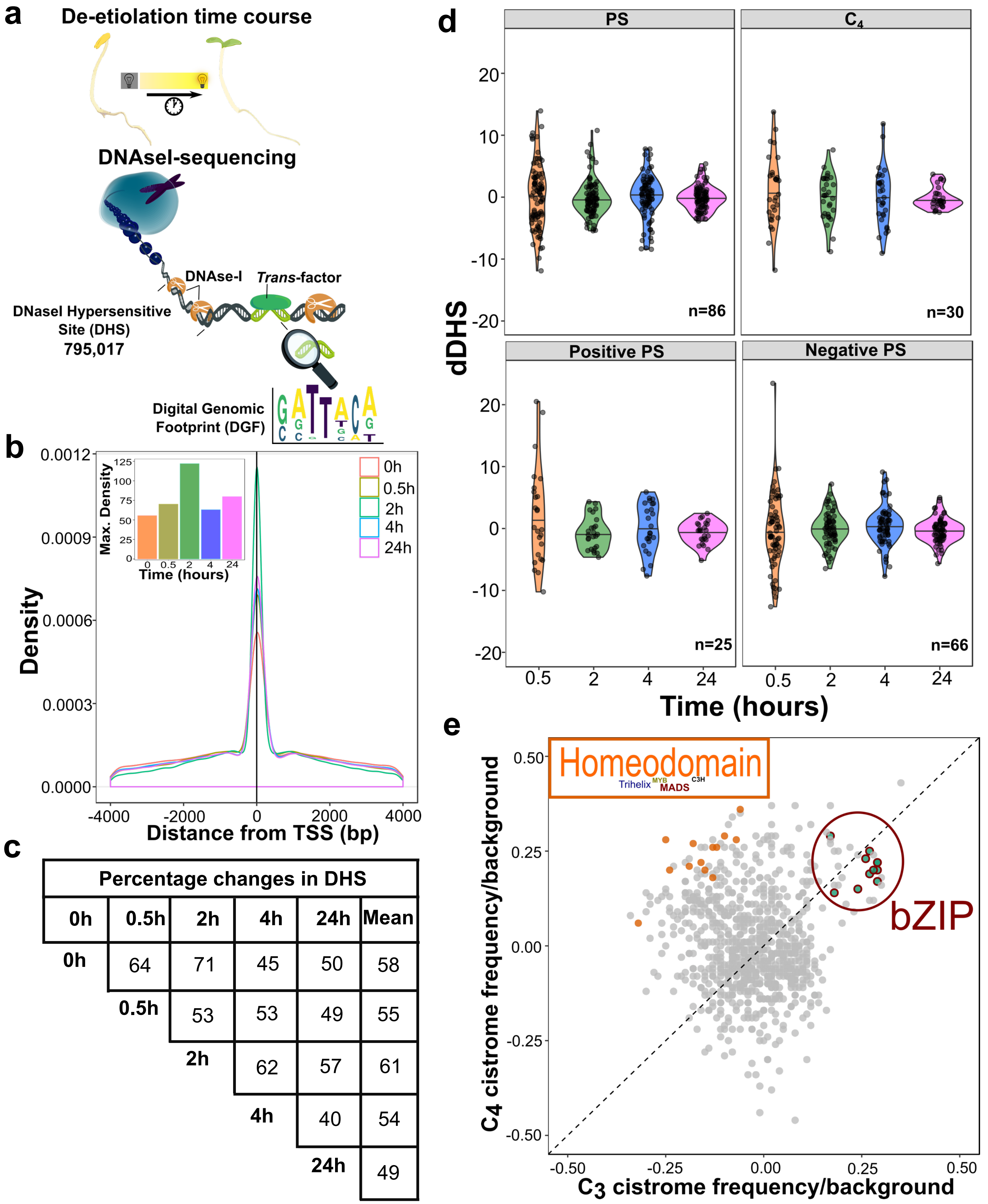


Fig. 2: Changes in transcript abundance during greening of *G. gynandra*. (a) Principal component analysis of RNA-SEQ datasets. The three biological replicates from each timepoint of de-etiolating *G. gynandra* seedlings (0, 0.5, 2, 4 and 24 hours) form distinct clusters. (b) Enriched GO terms between consecutive timepoints for up- and down-regulated genes. (c) Heatmap illustrating changes in transcript abundance of photosynthesis (grey sidebar) and C$_4$ photosynthesis genes (black sidebar) during the time-course. Data are shown after normalisation of expression data with each gene plotted on a row and centred around the row mean. Colour-coding of the dendrograms (red, yellow and green) highlight expression clusters representing none, moderate and strong induction respectively. (d) Line graphs depicting dynamics of transcription factors positively or negatively correlated with induced photosynthesis genes, or that showed early (0.5 hours) up- or down- regulation during the de-etiolation time-course. Values shown are normalised and centred around the mean of each gene.



Percentage changes in DHS					
0h	0.5h	2h	4h	24h	Mean
0h	64	71	45	50	58
	0.5h	53	53	49	55
		2h	62	57	61
			4h	40	54
				24h	49

Fig. 3: Profiling of open chromatin during de-etiolating of *G. gynandra*. (a) Schematic illustrating DNaseI-SEQ and the total number of DNaseI-hypersensitive sites (DHSs) detected. (b) Density of open chromatin plotted relative to the nearest Transcription Start Site (TSS). Inset highlights maximum density overlapping with the TSS at each time point. (c) Percentage of DHSs non-overlapping at each timepoint. (d) Violin plots depicting changes in DHS accessibility (dDHS) associated with photosynthesis genes, C_4 photosynthesis genes, and transcription factors that were positively or negatively correlated with photosynthesis genes. Changes are relative to the previous timepoint, n values are for the number of DHS regions quantified. (e) Scatter plot of FIMO motif frequencies in C_3 and C_4 photosynthesis gene from (log10 normalised motif frequency/normalised background frequency). Motifs annotated in orange and the associated Wordcloud highlight those enriched in the C_4 cistrome compared with the C_3 cistrome, and those in red indicate bZIP motifs enriched in both the C_3 and C_4 cistromes from both FIMO and AME analysis.

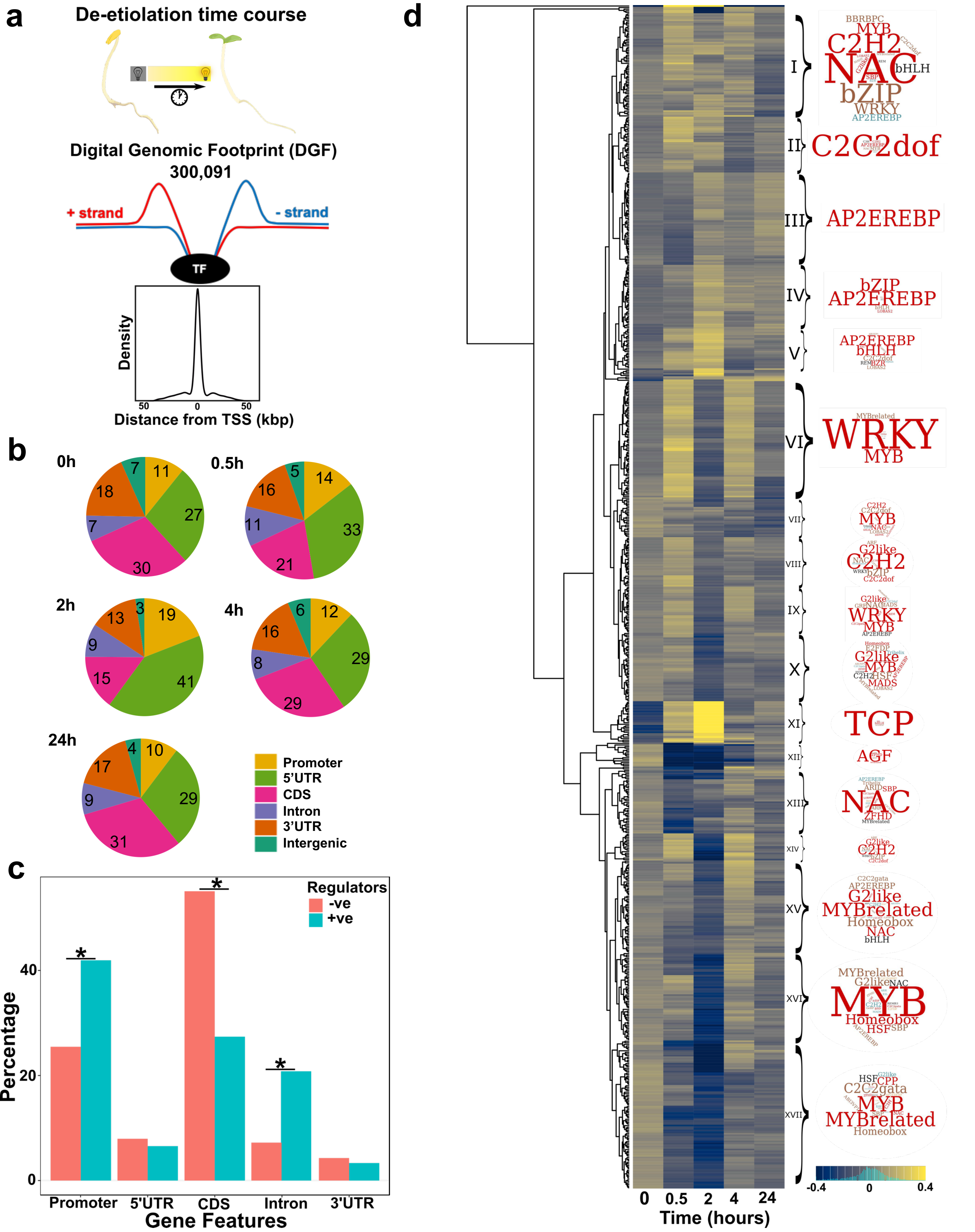


Fig. 4: Transcription factor binding atlas for de-etiolating seedlings *G. gynandra.* (a) Schematic illustrating sampling, number of Digital Genomic Footprints (DGF) identified and representative density plot of DGF positions relative to the nearest transcription start site (TSS). (b) Pie-charts summarising the density of DGF among genomic features. Promoters are defined as sequence < 2000 base pairs upstream of TSSs while intergenic represent any regions not overlapping with other features. Values indicate densities of DGFs in each feature as percentages. (c) Bar chart showing the percentage of DGFs predicted to function either as activators (coral bars) or repressors (turquoise bars) lying within gene features of target genes. Statistically significant differences were detected for promoters, coding sequence (CDS) and introns using a Chi square goodness of fit test ("*"). (d) Heatmap of motif frequencies (log10 sample normalised motif frequency/row mean) during de-etiolation. To illustrate identity and heterogeneity of motif groups clusters were annotated with Wordclouds.

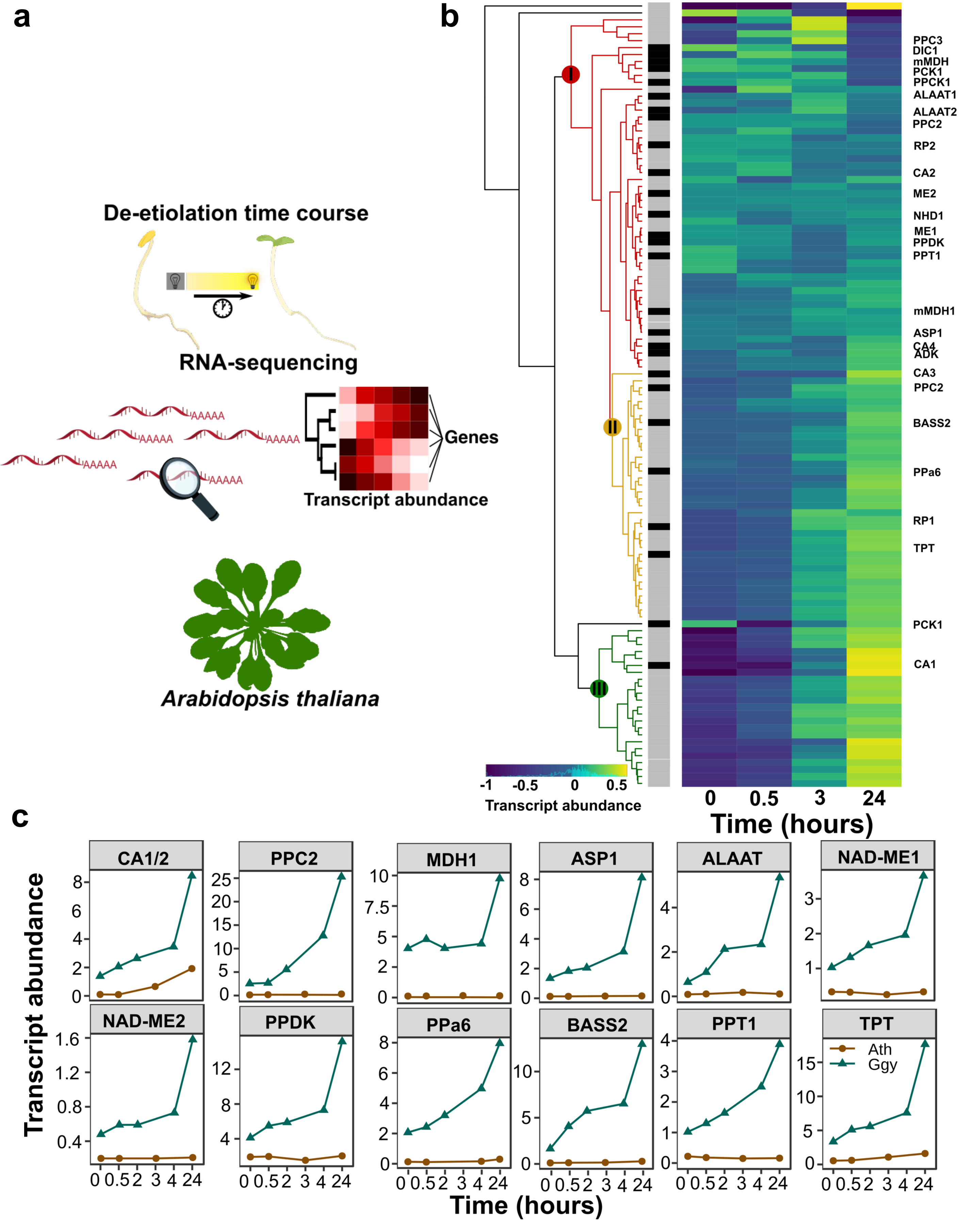


Fig. 5: Comparison of transcript abundance for photosynthetic genes during de-etiolation of C$_3$ *Arabidopsis thaliana* and C$_4$ *G. gynandra*. (a) Schematic illustrating RNA-SEQ of Arabidopsis. (b) Expression patterns of photosynthesis genes (grey sidebar) and C$_4$ orthologs (black sidebar) during de-etiolation. Heatmap illustrating gene expression with each gene being represented by a row, and data centred around the row mean. Dendrograms (red, yellow and green) highlight distinct expression clusters representing no clear, moderate, or strong induction. (c) Line graphs depicting quantile normalised and mean divided expression patterns of twelve C$_4$ orthologs in C$_3$ Arabidopsis and C$_4$ *G. gynandra*. Where there is more than a 1:1 relationship between genes, the most abundant paralog from each orthogroup is presented.

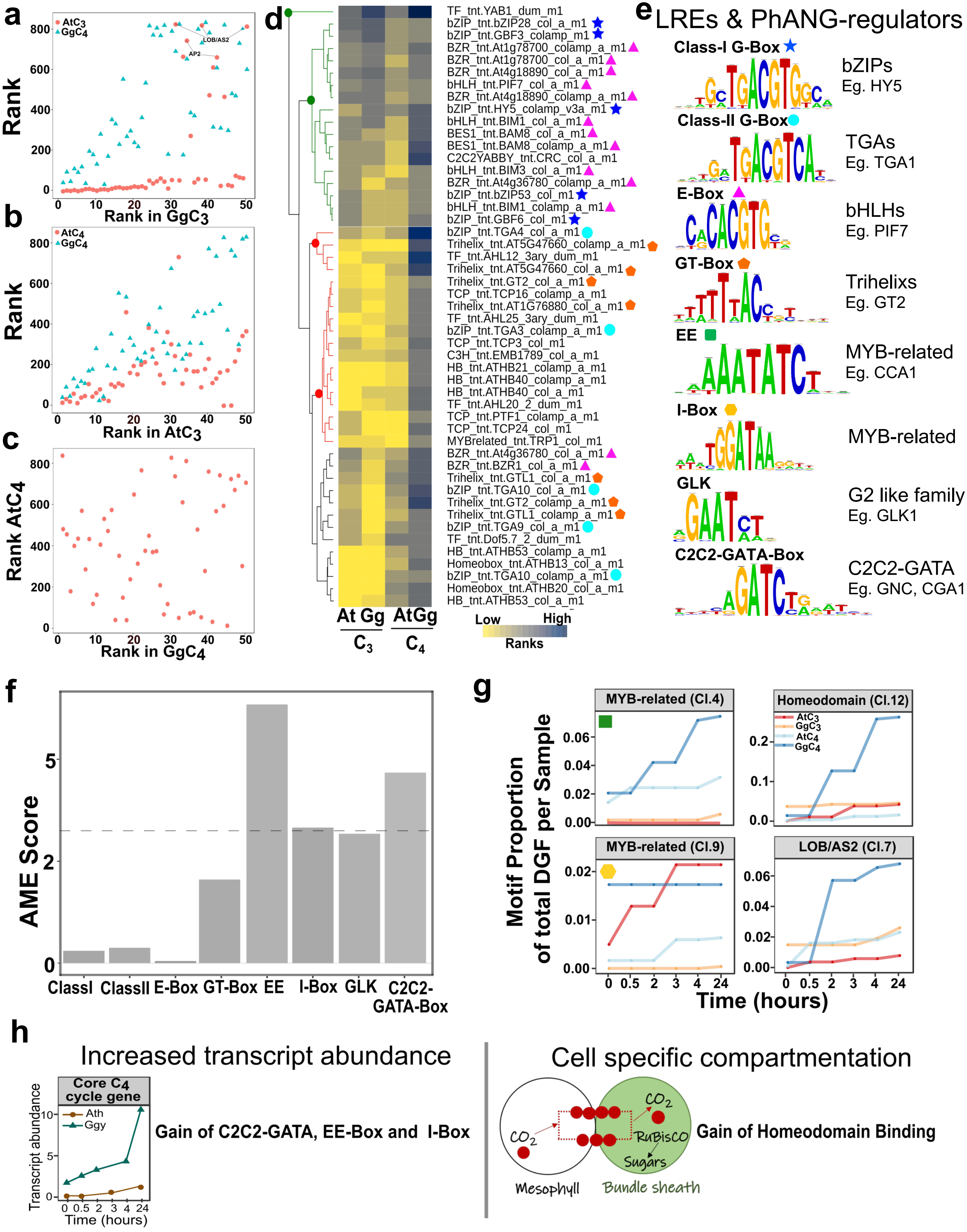


Fig. 6: Comparative analysis of potential regulatory mechanisms for de-etiolating seedlings of C_3 *Arabidopsis thaliana* and C_4 *G. gynandra*. Scatter plots showing the most enriched motifs in each cistrome (where 1 represents the most enriched motif). (a) Top 50 motifs in photosynthesis genes of C_3 Arabidopsis (At) and C_4 *G. gynandra* (Gg), (b) C_4 and photosynthesis genes of C_3 Arabidopsis, and (c) C_4 genes from C_3 Arabidopsis and *C. gynandra*. Motifs from the cistromes of C_3 and C_4 genes that showed induction during de-etiolation. (d) Heatmap of the top 50 motifs from DHSs of C_4 genes in *G. gynandra* compared with their

ranking in C_4 genes of Arabidopsis and photosynthesis genes in both species (log2 of the motif ranks across all four cistrome sets). Two distinct groups are highlighted with green motifs being highly ranked (more enriched) in all four cistromes while the red motifs are those specifically highly ranked in the *G. gynandra* C_4 cistrome. Motifs characterised as light-regulatory elements (LREs) are labelled with symbols used in (e). (e) Sequence logos highlighting different classes of Light Responsive Elements (LREs) and regulators of photosynthesis-associated nuclear genes (PhANGs). Seqlogos were generated from DAP-SEQ and PBM consensus motifs for all members of each type. (f) Analysis of which LREs and regulators of PhANGs are statistically enriched in cistromes of C_4 genes from *G. gynandra* and Arabidopsis. AME generates likelihood score for over-representation (-1*log(adjusted p-value), y-axis), and the adjusted $p < 0.05$ is illustrated with a dashed line. (g) Transcription factor binding sites associated with EE and I-box binding as well as the homeodomain and LOB/AS2 families disproportionally found in C_4 genes from *G. gynandra* compared with orthologs from Arabidopsis, and photosynthesis genes in both species. Values plotted are the motif proportion of the total number of DGF at each sample over the time-course such that differences between and within experiments were normalised. (h) Model illustrating association between enhanced C_4 cycle gene expression in *G. gynandra* compared with Arabidopsis and gain of *cis*-elements bound by MYB-related and C2C2-GATA transcription factors as well as the gain of homeodomain binding sites in mesophyll expressed genes in C_4 *G. gynandra*.