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

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

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

97 Our data link changes in chromatin accessibility and transcription factor binding to
98 patterns of gene expression and assembly of the photosynthetic apparatus in the C_4 species
99 *G. gynandra*. During de-etiolation, assembly of the photosynthetic apparatus was initiated
100 within two hours of exposure to light. Transcript profiling revealed a global remodelling of
101 gene expression associated with the dark-to-light transition. Many genes associated with
102 core photosynthetic processes shared by C_3 and C_4 plants, as well as those specifically
103 encoding components of C_4 photosynthesis showed similar dynamics during this dark-to-
104 light transition. During the first two hours of exposure to light, a relatively large restructuring
105 of open chromatin and a shift from transcription factor binding in exons to promoters and 5'
106 UTRs took place. All genes encoding core proteins of the C_4 pathway were more strongly
107 induced after exposure to light in C_4 *G. gynandra* compared with C_3 Arabidopsis. The greater
108 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)

156 showed that replicates from each timepoint clustered together tightly, and that 64% of the
157 variance in transcript abundance could be explained by two main components. The first
158 component accounted for 46% of the variance and was associated with the dark-to-light
159 transition (Fig. 2a). To better understand the general patterns of differentially expressed
160 genes (Supplementary Table 2), Gene Ontology (GO) terms were assessed (Fig. 2b,
161 Supplementary Fig. 2, FDR<10^{-5}). Compared with each previous time point, upregulated
162 genes in samples taken at 0.5 and 24 hours showed enrichment in GO terms including those
163 related to the plastid, as well as carbohydrate, secondary, nitrogen and lipid metabolism, but
164 also responses to light and photosynthesis. These components were also over-represented
165 in genes down-regulated at 2 and 4-hour time points suggesting two phases of
166 photosynthetic induction. Overall, these pairwise comparisons of transcript abundance
167 between samples, the PCA and the GO term enrichment analysis are consistent with a
168 major remodelling of gene expression after 0.5 hours of exposure to light, at least in part
169 associated with establishment of photosynthesis.
170 To better understand the dynamics of gene expression associated with the induction of
171 chlorophyll accumulation and remodelling of chloroplast ultrastructure in the C_4 leaf (Fig. 1),
172 genes associated with photosynthesis were subjected to hierarchical clustering. The genes
173 defined as such were C_4 pathway genes as well as orthologs to nuclear genes from
174 Arabidopsis annotated with the photosynthesis-related GO term (GO:0015979). A total of
175 116 genes were clustered into three main groups. Cluster I (red, Fig. 2c) showed no clear
176 induction over the time-course, but clusters II and III (yellow and green respectively, Fig. 2c)
177 contained the majority of genes (n=76) and showed clear induction by 24 hours. Notably, C_4
178 pathway genes were dispersed among these clusters, with 67% of C_4 photosynthesis genes
179 (n=20) found in clusters II and III (Fig. 2c). There are multiple paralogs of various C_4 cycle
180 genes of which ten showed no clear induction in response to light (Cluster I). However, the
181 majority of these non-induced members were poorly expressed, and at least one other
182 paralog was strongly induced and so present in Cluster II and III. Overall, these data show
183 that the majority of C_4 cycle genes populated photosynthesis gene clusters that showed
184 increased expression during de-etiolation (Fig. 2c).
185 To identify candidate transcription factors that may be responsible for the induction of
186 photosynthesis gene expression in response to light, four classes were identified on the
187 basis of changes in their transcript abundance. Transcription factors that act as positive
188 regulators of photosynthesis would be expected to show either a steady increase and
189 positive correlation (>0.8 Pearson Correlation across time-course) with induced
190 photosynthesis genes (clusters II and III from Fig. 2c), or an early burst at 0.5 hours (Fig.
191 2d). Repressors would be expected to show the opposite trends (Fig. 2d). This approach
192 identified twenty-one transcription factors that were strongly and positively correlated with

193 photosynthesis genes (Supplementary Table 3). Candidates in this class were often related
194 to previously characterised regulators of photomorphogenesis, plastid development, light
195 and circadian networks as well as components of cell fate determination and sucrose
196 signalling in Arabidopsis. Twenty-six transcription factors showed a strong and specific
197 induction at 0.5 hours (Supplementary Table 3). Again, these encoded homologs to proteins
198 previously implicated in the circadian clock, de-etiolation and chloroplast greening
199 components. For example, two orthologs of *ELONGATED HYPOCOTYL 5 (HY5)*, a master
200 regulator of de-etiolation and *LATE ELONGATED HYPOCOTYL (LHY)*, a key clock
201 component[31] were among this group. Sixty-two transcription factors were strongly negatively
202 correlated with transcripts of induced photosynthesis genes, and these included an ortholog
203 of *PHYTOCHROME INTERACTING FACTOR7 (PIF7)*, a negative regulator of phytochrome
204 B-mediated seedling de-etiolation[32]. This group also contained seventeen transcription
205 factors associated with hormone signalling and ten containing a homeodomain
206 (Supplementary Table 3). Finally, twenty-two transcription factors were identified as showing
207 an early and specific downregulation with the majority connected to developmental
208 processes, often organ development (Supplementary Table 3). These four classes of
209 transcription factors therefore contain members consistent with analysis from other systems
210 and so appear to represent conserved *trans*-factors associated with de-etiolation in general.
211 Taken together, the data indicate that as in C_3 plants, a global remodelling of gene
212 expression is associated with the dark-to-light transition in C_4 *G. gynandra*. Moreover, many
213 genes associated with core photosynthetic processes shared by C_3 and C_4 plants, as well as
214 those specifically encoding components of C_4 photosynthesis showed similar dynamics
215 during this transition. We conclude that the process of de-etiolation provides an attractive
216 system to start to define gene regulatory networks that control the induction of C_4
217 photosynthesis. As the regulation of gene expression is highly combinatorial, to identify
218 potential regulators of this process we opted for an unbiased genome-wide approach to
219 assess chromatin accessibility as well as transcription factor-DNA interactions during de-
220 etiolation.
221
222 **Chromatin dynamics associated with de-etiolation in *G. gynandra***
223 To gain insight into how chromatin accessibility and *cis*-elements bound by transcription
224 factors within such regions change during de-etiolation of *G. gynandra*, nuclei from three
225 biological replicates across the five time points were treated with DNase-I and subjected to
226 sequencing. From these time-points, a total of 1,145,530,978 reads were mapped to the *G.*
227 *gynandra* genome, and 795,017 DNaseI-hypersensitive sites (DHSs) representing broad
228 regulatory regions accessible to transcription factor binding were identified (Fig. 3a,
229 Supplementary Fig. 3). The average length of these DHSs was ~610 base pairs, and

230 distribution plots showed that their density was highest at the predicted transcription start
231 sites (Fig. 3b). However, over the time-course the peak DHS density at transcription start
232 sites altered such that compared with the dark, it more than doubled by two hours after
233 transfer to light (Fig. 3b). This is consistent with the notion that exposure to light leads to a
234 rapid increase in open chromatin around gene bodies[33]. To further investigate the extent to
235 which accessible chromatin changed over the entire time-course, the proportion of DHS that
236 were shared between time-points was examined (Fig. 3c). There was a major re-
237 organisation of DHS by 0.5 hours, with 64% changing compared with tissue harvested from
238 the dark. This remodelling continued until 2 hours after exposure to light when 71% of DHS
239 had changed compared with the dark (Fig. 3c). From 4 hours after exposure to light, the
240 extent to which DHS were modified was less striking. These data therefore support the
241 notion that during the first two hours of exposure to light when assembly of the
242 photosynthetic apparatus is being initiated (Fig. 1a-d), a relatively large restructuring of open
243 chromatin takes place, but subsequent to this, and coincident with photosynthetic maturation
244 there are fewer changes in chromatin accessibility. We conclude that this major re-patterning
245 of DHS in the first two hours after exposure to light likely contributes to the changes to
246 mRNA abundance detected soon after the dark-to-light transition, and thus assembly of the
247 photosynthetic apparatus.

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

261 To better understand transcription factor binding sites that may be involved in activating
262 photosynthesis gene expression, DHSs associated within induced C_3 and C_4 photosynthesis
263 genes (clusters II and III Fig. 2c) were selected and scanned for binding sites[35,36]. Two
264 complementary algorithms were used from the MEME suite. The first was FIMO[37] which
265 finds individual motif occurrences predicted to be of high affinity. The second was AME[38],
266 which determines the average odds scores across entire sequences and in so doing

267 considers lower affinity sites, many of which would not be detected with FIMO. Using FIMO
268 we did not detect a strong correlation (Pearson's correlation of 0.0014) between motif
269 frequencies against a random background set of DHSs in C_3 and C_4 photosynthesis genes
270 (Fig. 3e). To identify which motifs occurred more often than by chance alone, AME was run
271 for both the C_3 and C_4 cistromes against shuffled input sequences as a control
272 (Supplementary Table 4). Although there was little overlap between AME and FIMO outputs,
273 a group of bZIP (G-box binding) motifs were enriched in both datasets (Fig. 3e). These
274 motifs represent Light Responsive Elements (LREs) first defined as multipartite *cis*-elements
275 in the *RbcS* promoter[22,39,40]. This finding is consistent with the requirement for
276 photosynthesis genes to be responsive to light and implies that induction of C_4 pathway
277 genes during de-etiolation may in part be due to these LREs. AME also identified motifs that
278 were enriched in the C_4 cistrome compared with the C_3 cistrome (orange points in Fig. 3e).
279 This was dominated by homeodomain motifs (specifically those from the HD-Zip I family) as
280 well as some GT-box related Trihelix motifs, MYB and MADs motifs. In summary, although
281 there was no strong correlation between the cistromes of C_3 and C_4 photosynthesis genes
282 from *G. gynandra*, a group of bZIP (G-box) motifs were enriched in both gene sets, and
283 homeodomain motifs were enriched in the cistrome of C_4 genes compared with that of C_3
284 genes.
285
286 **A *cis*-regulatory atlas for de-etiolation in *G. gynandra***
287 Chromatin accessibility assays followed by *in silico* analysis of motifs within these regions
288 identifies regulatory elements that could be important for gene regulation but does not
289 indicate whether motifs are actually subject to transcription factor binding. We therefore
290 carried out sequencing at sufficient depth to define DNA sequences that are protected from
291 DNaseI digestion (Fig. 4a). Such sequences are diagnostic of strong and/or widespread
292 protein binding and referred to as Digital Genomic Footprints (DGFs). Although DNaseI-SEQ
293 has been used to predict transcription factor binding sites at base-pair resolution through
294 DGF, the DNaseI enzyme possesses sequence bias that can lead to both type I and II errors
295 in their identification and so to account for such bias, de-proteinated DNA was first
296 analysed[41,42] (Supplementary Fig. 3). After this filtering, 300,091 DGFs corresponding to
297 individual transcription factor binding sites across all time points were identified (Fig. 4a and
298 Supplementary Fig. 3). This compares favourably with 282,030 DGFs in a publicly available
299 dataset for de-etiolation of Arabidopsis[43] that was less conservative as the naked DNaseI
300 filtering steps were not undertaken. To provide an overview of transcription factors likely
301 binding these DGFs, all were scanned for 948 known Arabidopsis motifs. To be
302 conservative, each DGF was only annotated to its top match.

303 The distribution of DGFs in gene features (e.g., promoters, exons, introns and UTRs) of
304 *G. gynandra* changed during de-etiolation (Fig. 4b). Notably, in the first two hours of
305 exposure to light, DGF density in promoter elements (defined as sequence two kilobase
306 pairs upstream of predicted transcriptional start sites) and 5' UTRs increased (from 11 to
307 19% in the case of promoters and 27 to 41% for 5' UTRs). This finding is consistent with the
308 increase in DHS density around predicted transcriptional start sites at this time (Fig. 3a).
309 Coincident with the increase in DGFs in promoters and 5' UTRs, the density found in coding
310 sequence was reduced by around half (Fig. 4b). In contrast, the density of DGFs associated
311 with introns and 3' UTRs changed little during de-etiolation. These findings suggest changes
312 to binding site distribution between genomic features may play an important role in
313 transcriptional regulation and contribute to the induction of photosynthesis during de-
314 etiolation. To test this further, we correlated the change in frequency of each motif with
315 expression of the proximal gene (Supplementary Table 5). Positive and negative correlations
316 between motif frequency and gene expression were used to classify motifs as either allowing
317 activation or repression. Of the motifs predicted to act as activators, which included a
318 number of Cysteine-rich polycomb-like (CPP) factors, significantly more were located in
319 promoters and introns. In contrast motifs predicted to act as repressors were significantly
320 more likely, roughly twice, to be found in exons (Fig. 4c, Supplementary Table 5). Motifs
321 found to have no correlation to targets (neutral) were found to have intermediate frequencies
322 suggesting a gradient from the two extremes (Supplementary Table 5).
323 To understand the dynamics of motifs during de-etiolation, after filtering out low frequency
324 motifs, 743 were subjected to hierarchical clustering (Fig. 4d). Whilst some clusters were
325 dominated by a small number of transcription factor families such as DOFs (II), AP2s (III),
326 WRKYs (VI) and TCPs (XI) (Fig. 4d) others were composed of motifs associated with
327 multiple families of transcription factors suggesting that the binding sites of these unrelated
328 transcription factors could be involved in similar networks during de-etiolation. One of the
329 most striking clusters was dominated by TCP motifs (Cluster XI) which was depleted in dark-
330 grown tissue but became more enriched at thirty minutes and then two hours after exposure
331 to light. To varying degrees, clusters I to V showed a similar pattern to the TCP cluster,
332 peaking between thirty minutes to four hours with lower levels at zero hours. These
333 contained more C2H2 and NAC (cluster I), bZIP (cluster I and IV), DOF (cluster II), AP2
334 (clusters III and IV) and bHLH (cluster IV) motifs. There were a few clusters, notably XVI and
335 XVII, that were least represented at two hours and hence showed a broadly opposite
336 dynamic. These large clusters are highly heterogeneous but included many MYB, MYB-
337 related, GATA, G2-like, HSF and homeodomain motifs. Overall, the patterns indicate that
338 certain groups of transcription factors are likely more involved at certain points during de-
339 etiolation. In addition, many of these patterns appear to be complementary to one another

340 (such as XI with XVI and XVII) and so antagonistic interactions appear likely between
341 members of clusters I to XI and XII to XVII.
342
343 **C_4 genes in C_3 *A thaliana* are induced but with reduced amplitude compared with**
344 **orthologs in C_4 *G. gynandra***
345 In order to gain insight into the extent to which C_4 gene expression has altered compared
346 with the ancestral C_3 state, we compared the RNA-SEQ data and the *cis*-regulatory atlas
347 collected for C_4 *G. gynandra* with an equivalent dataset from Arabidopsis[43] (Fig. 5a). As in *G.*
348 *gynandra* (Fig. 2c), many photosynthesis-related genes in C_3 Arabidopsis showed increased
349 transcript abundance after the dark-to-light transition and three major behaviours were
350 evident (Fig. 5b). Cluster I (red) showed no clear change, Cluster II (yellow) showed
351 moderate, while Cluster III (green) showed strong induction (Fig. 5b). Of the genes
352 orthologous to C_4 pathway genes, nineteen showed no clear induction while six were
353 moderately and only one was strongly induced (Fig. 5b). Therefore, whilst some C_4 genes
354 showed induction in the ancestral state, this was by no means universal. Indeed, in
355 Arabidopsis a higher proportion occupied the non-responding cluster (19/27) compared with
356 *G. gynandra* (10/30). Whilst in *G. gynandra* seven occupied the most strongly responding
357 cluster, only one did in Arabidopsis. We next normalised the transcript abundance data of
358 both species to enable direct comparison of expression of C_4 genes from *G. gynandra* with
359 orthologous groups from Arabidopsis (Supplementary Fig. 5). This indicated that all genes
360 encoding core proteins of the C_4 pathway were more strongly induced after exposure to light
361 in C_4 *G. gynandra* than in C_3 Arabidopsis (Fig. 5c). This is consistent with re-analysis of
362 publicly available data for maize and rice de-etiolation[44] (Supplementary Fig. 6). Thus, whilst
363 many C_4 pathway orthologs were induced in response to light in Arabidopsis the amplitude
364 of this response was larger in *G. gynandra*.
365
366 **C_4 genes gain light responsive elements and motifs that regulate photosynthesis-**
367 **associated nuclear genes in the C_3 state**
368 To investigate how C_4 genes become more responsive to light in *G. gynandra* compared
369 with Arabidopsis, we first identified *cis*-elements in DHS associated with these genes from
370 each species (the C_4 cistromes). As well as genes encoding the core C_4 pathway, we
371 included photosynthesis genes that showed clear induction in response to light (the C_3
372 cistromes, see Fig. 2d). Using these gene sets allowed us to investigate the extent to which
373 C_4 genes in *G. gynandra* share a *cis*-code with photosynthesis genes, and whether this code
374 is also found in the C_3 ancestral state. As the number of each motif may vary between
375 species due to phylogenetic distance, to allow interspecies comparison we ranked motif
376 enrichment within each species. With the exception of some AP2 and LOB/AS2 motifs that

377 were abundant in *G. gynandra*, the *cis*-code of C_3 photosynthesis genes in both species was
378 similar (Fig. 6a). In both species, the most enriched motifs located in DHS around C_3 genes
379 included many bZIP (including HY5), bHLH (including PIF7) and BZR motifs (Supplementary
380 Table 6). Interestingly, there was more similarity between the C_3 cistromes of Arabidopsis
381 and *G. gynandra* (Fig. 6a) than between those of C_3 and C_4 genes from *G. gynandra* (Fig.
382 6a). The Arabidopsis C_3 and C_4 cistromes were more similar than the C_3 and C_4 cistromes
383 from *G. gynandra* (Fig. 6b) indicating that C_4 genes from *G. gynandra* have not acquired
384 large numbers of *cis*-elements associated with C_3 photosynthesis genes of Arabidopsis.
385 Lastly, the *cis*-code of C_4 genes from Arabidopsis and *G. gynandra* were very different (Fig.
386 6c) strongly implying that as they are recruited into the C_4 pathway the regulation of these
387 genes has diverged significantly. Collectively, these data indicate a greater divergence in
388 motif composition of C_4 genes in these C_3 and C_4 species than between their C_3
389 photosynthesis genes.

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

409 As LREs were amongst the most enriched motifs in the cistrome of C_4 genes from *G.*
410 *gynandra* compared with Arabidopsis, we next tested whether this difference was statistically
411 significant and therefore could contribute to the gain in induction observed for the twelve
412 core pathway components. PhANGs are also known to be regulated by transcription factors
413 such as GLK, GNC1 and CGA1[45,46] and although they were not in the top fifty most enriched

414 motifs from C_4 genes of *G. gynandra*, we included motifs recognised by these PhANGs (Fig.
415 6e) in our analysis. For the LREs, there was no enrichment in Class I and Class II G-boxes
416 or E-boxes in C_4 genes from *G. gynandra* (adjusted p-value < 0.05) (Fig. 6f), and whilst there
417 was increase in the number of GT-boxes this was not statistically significant (Fig. 6f).
418 However, both EEs and I-boxes were statistically more common in C_4 genes of *G. gynandra*
419 than their orthologs in Arabidopsis (adjusted p-values of 0.00286 and 0.047
420 respectively) (Fig. 6f). Motifs recognised by GLK were very close to the $p < 0.05$ cut-off
421 (0.0535), and the increase in CGA1 and GNC1 (C2C2-GATA) binding sites in C_4 genes of *G.*
422 *gynandra* compared with Arabidopsis was statistically significant (adjusted p-value 0.0134)
423 (Fig. 6f). We interrogated the DGF datasets from each species to test whether *in vivo*
424 binding for any of the LRE or PhANG motifs was detected. Although we did not detect more
425 binding *in vivo* for the PhANGs in C_4 genes of *G. gynandra* compared with Arabidopsis, we
426 did for both EEs and I-boxes recognised by MYB-related transcription factors (Fig. 6g).
427 These data support a model in which increased binding of LREs in C_4 genes from *G.*
428 *gynandra* drive their increased responsiveness to light compared with orthologs from
429 Arabidopsis. It is possible that regulators of PhANGs such as CGA1 and GNC1 have a low
430 affinity for their cognate *cis*-element and so binding cannot be detected with the DNaseI
431 assay.
432 We also found increased binding of homeodomain and LOB/AS2 transcription factors to
433 C_4 genes of *G. gynandra* (Fig. 6g, Supplementary Fig. 7). Given that more homeodomain
434 sites were present in the cistrome of C_4 genes from *G. gynandra* (Fig. 6d), we sought to gain
435 further understanding of their potential roles. The C_4 genes *BASS2, PPa6* and *PPDK*
436 contained six DGF with significant matches to homeodomain motifs. These three genes are
437 preferentially expressed in mesophyll cells of C_4 plants, and so we hypothesized that
438 transcription factors able to bind these motifs would be more strongly expressed in the
439 mesophyll. To address this, we used publicly available data[47,48] and found that transcripts
440 derived from the Homeodomain-Zip IV (HD-Zip IV) subfamily (including *GLABRA2* and
441 others with a wide range of functions) were more abundant in mesophyll cells from
442 Arabidopsis and *G. gynandra*, whilst those from HD-Zip III subfamily (including vascular
443 development and leaf polarity regulators such as *CORONA* and *PHABULOSA*) were more
444 abundant in bundle sheath strands from both species (Supplementary Fig. 8). Hits for HD-
445 Zip IV binding were detected in *BASS2* and *PPa6*, but the strongest match was a motif
446 recognised by ANL2 in *PPDK*. As the *G. gynandra* ortholog of ANL2 was most abundant
447 during de-etiolation, these data are consistent with *ANL2* playing a role in driving mesophyll
448 expressed C_4 genes in *G. gynandra*.

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

464 **DISCUSSION**

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

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

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

492

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

494 The induction of photosynthesis transcripts in *G. gynandra* between 0 to 0.5 hours was
495 associated with upregulation of previously identified master regulators of de-etiolation. For
496 example, we were able to detect upregulation of genes encoding well characterised proteins
497 involved in circadian rhythms and light regulation. This included two paralogs of the master
498 regulator *ELONGATED HYPOCOTYL 5 (HY5)* as well as an ortholog of *EARLY*
499 *PHYTOCHROME RESPONSIVE 1*. Clock components that also showed this early light

500 induction were two paralogs of *REVEILLE (RVE2)*, as well as one each of *RVE1* and *LATE*
501 *ELONGATED HYPOCOTYL (LHY)*. Further, consistent with other species[32] the negative
502 regulators of de-etiolation *PHYTOCHROME INTERACTING FACTOR7 (PIF7)* and *PIF3-*
503 *LIKE5 PIL5* from *G. gynandra* were rapidly down-regulated in light.
504 GO-terms associated with chloroplast development and photosynthesis were enriched at
505 the 0.5 and 24 hours timepoints, a finding consistent with the two-phase response reported
506 in cell lines of Arabidopsis[30] where significant changes in expression of plastid and
507 chlorophyll biosynthesis genes occurred soon after light exposure. In *G. gynandra*, transcript
508 levels of most photosynthesis genes showed an increased over the de-etiolation time
509 course. This is consistent with the trend observed during de-etiolation of Arabidopsis[43].
510 Furthermore, transcript levels of most C_4 cycle genes in both Arabidopsis and *G. gynandra*
511 increased steadily over the de-etiolation time course. A similar induction of C_4 cycle and
512 photosynthesis genes has been reported in C_3 rice and C_4 maize[44]. The most parsimonious
513 explanation for these findings is that in C_3 plants genes encoding components of the C_4
514 pathway show a basal induction in response to light, and that this ancestral system becomes
515 amplified during the evolution of C_4 photosynthesis.
516
517 **Chromatin dynamics and transcription factor binding during de-etiolation of *G.***
518 ***gynandra***
519 DHS density around the predicted transcription start sites increased for the first 2 hours of
520 light suggesting a gain in chromatin accessibility around the proximal promoters of *G.*
521 *gynandra*. This finding is consistent with loci becoming more accessible in light[33]. However,
522 for early responding genes changes in accessibility in DHS's (dDHS) was not associated
523 with increased gene expression (Supplementary Fig. 3) suggesting that broad-scale
524 changes at DHS are not strongly predictive of gene expression patterns. This may be
525 because transcription factor complexes that then bind have antagonistic actions on gene
526 expression and demonstrates the need for an integrated systems biology approach to gain
527 more detailed insight into the regulatory mechanisms. As open chromatin dynamics around
528 these genes were not strongly predictive of gene expression, we analysed the cistromes
529 located in accessible regions. This suggested that there was low convergence in the
530 regulatory elements available for binding in C_3 and C_4 genes from *G. gynandra* despite
531 similar dynamics of expression.
532 Consistent with previous work[33] our data indicate that changes in the distribution of
533 transcription factor binding was associated both with changes in the location of open
534 chromatin but also the binding of individual transcription factors. Notably, between 0 and 2
535 hours of exposure to light, transcription factor binding events decreased in coding regions
536 but increased in promoters and 5' UTRs. Transcription factor binding sites within codons are

537 referred to as duons as they determine both gene expression and the amino acid code.
538 Duons have been proposed to act as repressors of gene expression in humans[64]. In plants,
539 duons from the *NAD-ME1* and *NAD-ME2* genes placed downstream of the constitutive
540 CaMV35S promoter restrict expression to bundle sheath cells of *G. gynandra*, consistent
541 with them repressing gene expression in mesophyll cells[12,13]. In the current dataset, we
542 found that transcription factor binding events predicted to be positive activators were 1.6 and
543 2.9 times more likely to be found in promoters and introns respectively compared to
544 predicted negative regulators. Predicted negative regulators were twice as likely to be found
545 in exons as predicted positive regulators. Interestingly, the relationship between gene
546 enhancement and binding to introns was even more striking than that of promoters. In
547 *Drosophila melanogaster* enhancers have been reported to be enriched in promoter, 5'
548 UTRs but especially introns and depleted in exons[65]. We propose that de-etiolation offers an
549 attractive system with which to investigate the importance of how the location of transcription
550 factor binding impacts on gene expression.
551
552 **Analysis of C_3 Arabidopsis reveals evolution has co-opted existing regulatory**
553 **mechanisms to pattern C_4 gene expression**
554 Compared with C_3 species, leaves from C_4 plants have increased expression of genes
555 encoding the C_4 cycle, and decreased expression of those involved in photorespiration[66–68].
556 However, the dynamics with which these responses are established have not been fully
557 defined. In the present study, photosynthesis as well as C_4 cycle genes showed light
558 induction in both Arabidopsis and *G. gynandra*. Our datasets clearly demonstrate greater
559 rates of transcript accumulation of core C_4 cycle genes in *G. gynandra* compared to
560 orthologs from Arabidopsis during the de-etiolation time course (Fig. 5b). These findings are
561 consistent with re-analysis of publicly available data for maize and rice[44] (Supplementary
562 Fig. 6). We conclude that in at least two lineages that have independently evolved C_4
563 photosynthesis, genes associated with the C_4 cycle become part of gene regulatory
564 networks that respond very strongly to the light-to-dark transition associated with de-
565 etiolation.
566 The motifs of interest that were specifically over-represented in the *G. gynandra* C_4
567 cistrome included TGA bZIP and homeodomain motifs. These sequences therefore
568 represent interesting candidates as regulators of C_4 specific processes. Homeodomain
569 factors are documented to have a variety of roles, many related to development[69].
570 Homeodomain DGFs were detected in C_4 pathway components expressed in mesophyll
571 cells including *BASS2, PPa6*, *PPDK, CA2/3, DIC1, NHD1* and *PPCK1*. Notably, *BASS2,*
572 *PPa6* and *PPDK* contained DGF bound by HD-Zip IV factors with a mesophyll bias in both
573 *G. gynandra* and Arabidopsis (Supplementary Fig. 7). Of this group of potential regulators,

574 an ortholog to *ANL2* showed the highest expression of all HD-Zip IV factors in the *G.*
575 *gynandra* de-etiolation time-course with its highest levels at 0 hours. In Arabidopsis, ANL2 is
576 involved root development where it regulates the epidermal and cortical layers[70]. In leaves
577 ANL2 expression is strongest in mesophyll cells[71]. This gene therefore appears to be a
578 strong candidate for regulating mesophyll specific expression in C_4 leaves.

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

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

611 findings indicate that C_3 models such as Arabidopsis can provide significant insight into gene
612 regulatory networks that operate in C_4 plants.

613 **MATERIALS AND METHODS**

614 **Plant growth, chlorophyll quantitation and microscopy**

615 G*ynandropsis gynandra* seeds were sown directly from intact pods and germinated on
616 moist filter papers in the dark at 32 °C for 24 hours. Germinated seeds were then transferred
617 to half strength Murashige and Skoog (MS) medium with 0.8 % (w/v) agar (pH 5.8) and
618 grown for three days in a growth chamber at 26 °C. De-etiolation was induced by exposure
619 to white light with a photon flux density (PFD) of 350 µmol m^{-2} s^{-1} and photoperiod of 16
620 hours. Whole seedlings were harvested at 0.5, 2, 4 and 24 hours after illumination (starting
621 at 8:00 with light cycle 6:00 to 22:00). Tissue was flash frozen in liquid nitrogen and stored at
622 -80 °C prior to processing.

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

630 For electron microscopy, *G. gynandra* cotyledons (~2 mm^2) were excised with a razor
631 blade and fixed immediately in 2 % (v/v) glutaraldehyde and 2 % (w/v) formaldehyde in 0.05-
632 0.1M sodium cacodylate (NaCac) buffer (pH 7.4) containing 2 mM calcium chloride. Samples
633 were vacuum infiltrated overnight, washed five times in deionized water, and post-fixed in 1
634 % (v/v) aqueous osmium tetroxide, 1.5 % (w/v) potassium ferricyanide in
635 0.05 M NaCac buffer for 3 days at 4 °C. After osmication, samples were washed five times in
636 deionized water and post-fixed in 0.1 % (w/v) thiocarbohydrazide in 0.05 M NaCac buffer for
637 20 minutes at room temperature in the dark. Samples were then washed five times in
638 deionized water and osmicated for a second time for 1 hour in 2 % (v/v) aqueous osmium
639 tetroxide in 0.05 M NaCac buffer at room temperature. Samples were washed five times in
640 deionized water and subsequently stained in 2 % (w/v) uranyl acetate in 0.05 M maleate
641 buffer (pH 5.5) for 3 days at 4 °C and washed five times afterwards in deionized water. Next,
642 samples were dehydrated in an ethanol series, transferred to acetone, and then to
643 acetonitrile. Samples were embedded in Quetol 651 resin mix (TAAB
644 Laboratories Equipment Ltd). For transmission electron microscopy (TEM), ultra-thin
645 sections were cut with a diamond knife, collected on copper grids and examined in a
646 FEI Tecnai G2 transmission electron microscope (200 keV, 20 µm objective aperture).
647 Images were obtained with AMT CCD camera. For scanning electron microscopy (SEM),
648 ultrathin-sections were placed on plastic coverslips which were mounted on aluminium SEM
649 stubs, sputter-coated with a thin layer of iridium and imaged in a FEI Verios 460 scanning

650 electron microscope. For light microscopy, thin sections were stained with methylene blue
651 and imaged by an Olympus BX41 light microscope with a mounted Micropublisher 3.3 RTV
652 camera (Q Imaging).
653
654 **RNA and DNaseI sequencing**
655 Before processing, frozen samples were divided into two, the first being used for RNA-
656 SEQ analysis and the second for DNaseI-SEQ. Samples were ground in a mortar and pestle
657 and RNA extraction carried out with the RNeasy Plant Mini Kit (74904; QIAGEN) according
658 to the manufacturer's instructions. RNA quality and integrity were assessed on a Bioanalyzer
659 High Sensitivity DNA Chip (Agilent Technologies). Library preparation was performed with
660 500 ng of high integrity total RNA (RNA integrity number > 8) using the QuantSeq 3' mRNA-
661 SEQ Library Preparation Kit FWD for Illumina (Lexogen) following the manufacturer's
662 instructions. Library quantity and quality were checked using Qubit (Life Technologies) and a
663 Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies). Libraries were sequenced
664 on NextSeq 500 (Illumina, Chesterford, UK) using single-end sequencing and a Mid Output
665 150 cycle run.
666 To extract nuclei, tissue was ground in liquid nitrogen and incubated for five minutes in
667 15mM PIPES pH 6.5, 0.3 M sucrose, 1 % (v/v) Triton X-100, 20mM NaCl, 80 mM KCl, 0.1
668 mM EDTA, 0.25 mM spermidine, 0.25 g Polyvinylpyrrolidone (SIGMA), EDTA-free
669 proteinase inhibitors (ROCHE), filtered through two layers of Miracloth (Millipore) and
670 pelleted by centrifugation at 4 °C for 15 min at 3600 g. To isolate deproteinated DNA, 100
671 mg of tissue from seedlings exposed to 24 hours light were harvested two hours into the light
672 cycle, four days after germination. DNA was extracted using a QIAGEN DNeasy Plant Mini
673 Kit (QIAGEN, UK) according to the manufacturer's instructions. $2x10^8$ nuclei were re-
674 suspended at 4 °C in digestion buffer (15 mM Tris-HCl, 90 mM NaCl, 60 mM KCl, 6 mM
675 $CaCl_2$, 0.5 mM spermidine, 1 mM EDTA and 0.5 mM EGTA, pH 8.0). DNAse-I (Fermentas)
676 at 2.5 U was added to each tube and incubated at 37 °C for three minutes. Digestion was
677 arrested by adding a 1:1 volume of stop buffer (50 mM Tris-HCl, 100 mM NaCl, 0.1 % (w/v)
678 SDS, 100 mM EDTA, pH 8.0, 1 mM Spermidine, 0.3 mM Spermine, RNaseA40 µg/ml) and
679 incubated at 55 °C for 15 minutes. 50 U of Proteinase K were then added and samples
680 incubated at 55 °C for 1 h. DNA was isolated by mixing with 1 ml
681 25:24:1 Phenol:Chloroform:Isoamyl Alcohol (Ambion) and spun for 5 minutess at 15,700 g
682 followed by ethanol precipitation of the aqueous phase. Samples were size-selected (50-400
683 bp) using agarose gel electrophoresis and quantified fluorometrically using a Qubit 3.0
684 Fluorometer (Life technologies), and a total of 10 ng of digested DNA (200 pg l^{-1}) used for
685 library construction. Sequencing ready libraries were prepared using a TruSeq Nano DNA
686 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).

724 In order to map orthologs between Arabidopsis and *G. gynandra*, OrthoFinder[80] was
725 used. This allows more complex relationships than a 1:1 to be identified and placed into
726 orthogroups. C_3 photosynthesis genes were first identified from Arabidopsis through the
727 "photosynthesis" (GO:0015979) keyword search on the TAIR browse tool
728 (https://www.arabidopsis.org/servlets/TairObject?type=keyword&id=6756) and gave ninety-
729 two genes for Arabidopsis. Their orthologs were found in *G. gynandra* using the orthogroups
730 generated between the two species and resulted in ninety-three C_3 photosynthesis genes.
731 C_4 genes used in this study are considered the "core" pathway genes and are a manually
732 curated set largely based on previous analysis[23]. Initially, multiple paralogs were included
733 but non-induced transcripts were then filtered out. Orthologs between the two species were
734 again identified from the orthogroups from OrthoFinder. The *G. gynandra* C_3 and C_4 gene
735 normalised expression values were further processed with each value being divided by the
736 row mean and log10 (log(x/row mean) plotted as a heatmap using the R script Fig2C.R on
737 data file Fig2C_heatmap_data.txt (on GitHub). The heatmap for Arabidopsis C_3 and C_4 gene
738 expression was made in the same way as for the *G. gynandra* data using the gene lists as
739 previously described.
740
741 **Identification of four expression behaviours of *G. gynandra* transcription factors**
742 In order to identify transcription factors of interest during *G. Gynandra* de-etiolation we
743 first found all potential transcription factors using homology (blastp) to the Arabidopsis
744 transcription factor protein sequences found in the Plant Transcription Factor Database
745 (http://plntfdb.bio.uni-potsdam.de/v3.0/downloads.php?sp_id=ATH). 2,481 potential *G.*
746 *gynandra* transcription factors were then filtered to remove those with low expression and
747 classified into four expression patterns of interest: i) Strongly and positively correlated with
748 induced C_3 genes (>0.7 Pearson Correlation; n=21); ii) Strong and specific up-regulation at
749 0.5 hours (n=26); iii) Strongly and negatively correlated with induced C_3 genes <-0.7
750 Pearson Correlation; n=62); iv) Strong and specific down-regulation at 0.5 hours (n=22).
751 These sets were plotted from using the Fig2D.R R script on the Fig2D_data.txt file (on
752 GitHub).
753
754 **DNaseI-SEQ data processing**
755 The three biological replicates for each time-point were sequenced in multiple runs with
756 one sample being chosen, based on initial QC scores, for deeper sequencing to provide the
757 necessary depth for calling both DNaseI Hypersensitive Site (DHS) and Digital Genomic
758 Footprints (DGF). For each sample, the raw reads from multiple sequencing runs were
759 combined and trimmed for low quality reads using trimmomatic. These files were analysed
760 with fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to ensure samples

761 passed important QC parameters (see Fig3_MultiQC_summary.html on GitHub) and then
762 mapped to the *G. gynandra* genome using bowtie2 (version 2.3.4.1) with the "--local" pre-set
763 option. Following mapping, a bash script (Fig3_DNaseSEQ_tagAlign.sh on GitHub) was run
764 on each bam file which in summary: filters low quality (MAPQ < 30) mapped reads and plots
765 MAPQ distribution, removes duplicates, measures library complexity, fragment sizes, GC
766 bias and finally makes tagAlign files. The three tagAlign files from each time-point were then
767 merged before running another bash script (DHS_DGF_identification.sh on GitHub) for each
768 time-point which in summary: uses "macs2 callpeak" to identify "narrowPeaks", finds the
769 distance for each DHS to its closest transcriptional start site (TSS), calculates SPOT scores
770 (see https://www.encodeproject.org/data-standards/dnase-SEQ/), plots DHS profiles using
771 the deeptools bamCoverage, computeMatrix and plotProfile tools and finally calls DGF using
772 the wellington_footprints.py program[81]. The footprints identified with a log(*p*-value) cut-off of
773 < -10 were used for further analysis. DHS positions relative to TSS for each sample were
774 plotted using an R script (Fig3B.R) on the file Fig3B_DHS_TSS_data.txt (on GitHub).
775
776 **Analysis of DHS changes across the time-course**
777 To quantify the overlap in DHS between samples, DHS (from the "narrowPeak" file) were
778 sorted by their "-log10qvalue" column and only the top ranked DHS regions used until a total
779 of 55,122,108 bp was reached which corresponded to the total length of DHS regions in the
780 4 hours sample which had the least. This allowed us to compare overlap between equal
781 sized regions. These DHS regions for each time-point were then intersected in a pairwise
782 fashion using bedtools intersect. The total length of intersecting regions was divided by
783 55,122,108 bp to obtain the proportion of overlap for each pairwise comparison generating
784 the values in Fig. 3c. To compare the differential DHS (dDHS) scores for gene sets of
785 interest, we defined promoter regions around each gene of interest as 1000 bp upstream
786 from TSS.
787 We then identified DHS regions that intersected with each gene of interest (i.e., all DHS
788 regions overlapping with a gene body or promoter) and merged these DHS regions
789 (equivalent to an outer join). The dDHS tool[34] as part of the pyDNase package[82] to quantify
790 changes in accessibility between consecutive time-points for a given region, where
791 SAMPLEA precedes SAMPLEB in the time-course. Finally, these dDHS values were plotted
792 in violin plots using the R script Fig3D.R on the data sets Fig3D_dDHS_data.txt (on GitHub).
793
794 **Motif Analysis of DHS regions**
795 All DHS intersecting with genes of interest were scanned for the presence of motifs from
796 the DAPseq[35] and PBM[36] databases using the meme suite FIMO tool[37]. To identify shared
797 behaviours while minimising noise, lists of both C_3 and C_4 pathway genes were filtered to

798 keep only those that showed induction across the time course. As before, we found all DHS
799 regions intersecting with genes of interest, including promoter, extracted the fasta
800 sequences of these regions and scanned them for the motifs.
801 The same process was carried out on three sets of 500 random genes to generate a
802 background frequency for random DHS region. Each motif frequency for a gene set (e.g.,
803 C_4 pathway gene DHS motif frequencies) was first normalised by the total number of motifs
804 found in that set and then normalised again against the background value for that motif such
805 that values greater than 0 indicated a higher than background rate and *vice versa* for values
806 below 0. This approach allowed us to identify enrichments in high confidence, high affinity
807 motif matches. While FIMO identifies higher confidence hits over a threshold, we also used
808 the AME program[38] to identify statistically enriched motifs within the C_4 cistrome as
809 compared to the C_3 cistrome using the webserver (http://meme-suite.org/tools/ame) and the
810 DAP-SEQ motif database with default settings. We also used AME to identify DAP-SEQ
811 motifs enriched in both the cistromes as compared to randomised sequence controls.
812
813 **DNaseI bias correction**
814 To reduce the proportion of false positive DGF calls caused by DNAseI cutting bias,
815 DNaseI-SEQ was performed on de-proteinated gDNA and mapped to the *G.*
816 *gynandra* genome. The hexamer cutting frequencies at the DNaseI cutting sites were used
817 to generate a background signal profile that was incorporated into a mixture model to
818 calculate the log-likehood ratio (FLR) for each footprint using the R package MixtureModel[42].
819 DGF with low confidence (FLR<0) were filtered out resulting in a reduction of 11.6 to 37.5 %
820 of DGF per timepoint. Same pipeline was used in previous analysis[83]. The pipeline is
821 illustrated in Supplementary Fig. 3.
822
823 **DGF genomic feature distributions DGF motif frequencies and DGF-target correlation**
824 To identify the distribution of DGF across genomic features we used bedtools intersect to
825 find the frequency of intersection between the DGF with features in the genome annotation
826 gff3 file promoter (2000 bp upstream of TSS), 5' UTR, CDS, intron, 3' UTR and intergenic).
827 These frequencies were divided by the total length of each feature across the genome to
828 determine a density of DGF per feature and these values were plotted as a pie chart for
829 each time point.
830 In order to link DGF to possible functions, we scanned all DGF using the meme suite fimo
831 tool for both DAPseq and PMB motifs as described above. To visualise how motif
832 frequencies changed during de-etiolation, the frequency of each motif at each time-point was
833 first normalised by the total number of motifs at that time point and then each value was
834 mean centred across the time-course for plotting as a heatmap using the Fig4C.R R script

835 on the Fig4C_heatmap_data.txt file (on GitHub). This hierarchical clustering was then
836 manually grouped and word clouds generated for the motif transcription factor families using
837 an online tool (https://www.wordclouds.com/) for each motif cluster.
838 Once individual DGF were annotated with potential motifs, we correlated the changes in
839 frequency of each motif with the mean expression of all the potential targets. The changes in
840 frequency of each motif were used as a proxy of their factor's abundance and/or activity and
841 potential targets are identified as those genes lying closest to a DGF with a specific motif. A
842 strong positive correlation was used to suggest positive regulation while a strong negative
843 correlation suggests inhibitory regulation.
844
845 **Arabidopsis RNA-SEQ and DNaseI-SEQ**
846 In order to carry out comparative analysis between *G. gynandra* and Arabidopsis, an
847 analogous de-etiolation time-course[43] was reprocessed in the same way as the *G.*
848 *gynandra* data. Arabidopsis DNaseI-SEQ data was mapped to the TAIR9 genome and RNA-
849 SEQ was mapped using Salmon to the Araport11 transcriptome, followed by the use of
850 tximport to collapse expression values for all isomers into a single value, a step not required
851 for *G. gynandra* as it lacks isomer information. To allow inter-species comparisons, as with
852 *G. gynandra* the Arabidopsis RNA-SEQ data was quantile normalised and then each value
853 divided by the samples mean expression value. Normalised expression values for the core
854 set of C_4 pathway genes from *G. gynandra* were compared with orthologs from Arabidopsis,
855 and when there was more than one paralog identified, the most highly expressed was
856 selected. Line plots were generated using the Fig5B.R R script on the
857 Fig5B_C4_pathway_data.txt file (on GitHub). To analyse and compare motifs between
858 species, we ranked motifs by their normalised frequencies against the background for each
859 gene sets DHS (C_4 pathway and C_3 photosynthesis from both *G. gynandra* and *A. thailiana*).
860 These sets were filtered to remove genes that were not induced during the time-course. The
861 top 50 motifs from each set were then plotted against their rank in other sets. As motifs that
862 were highly ranked in *G. gynandra* C_4 genes were of particular interest, these were plotted
863 as a heatmap using the Fig5G.R R script on the Fig5G_motif_rank_data.txt file (on GitHub).
864 To create cumulative line plots a number of steps were required. First, as individual motif
865 frequencies are low for any given small set of genes (e.g. 105174 DGF found in the 24 hour
866 time-point giving ~3 per gene loci) we grouped motifs based on motif clustering using the
867 RSAT motif matrix clustering tool with default settings (http://rsat.sb-roscoff.fr/matrix-
868 clustering_form.cgi). This meant all members from the same cluster were treated as one
869 motif group (for example, all TCP motifs are found in group 10). For motif groups see the file
870 DAPseq_PBM_Motif_Matrix_Clustering.txt (on GitHub). These values were then normalised
871 for each time-point by dividing by the total number of motifs such that the values represented

872 a proportion of the total. These values were then plotted in a cumulative line plot using the
873 Fig5H.R R script on the Fig5H_data.txt (on GitHub).
874 The quantification of light and chloroplastic regulatory elements was carried out on the
875 gDNA sequence of the 10 strongly induced *G. gynandra* core C_4 pathway gene loci,
876 including a 1500 bp promoter region, and their most highly expressed Arabidopsis ortholog
877 loci. In summary, matches to the highly conserved core sequences of each element were
878 counted and compared between the two species gene sets (see GitHub for command
879 example). A one-tailed t-test was used to show no significant increase in the frequencies of
880 these elements in the *G. gynandra* genes as compared to the Arabidopsis orthologs.
881
882 **Phylogeny and cell specific expression of homeodomain factors in Arabidopsis and**
883 ***G. gynandra***
884 Homeodomain factors were identified from Arabidopsis transcription factor databases and
885 all potential transcription factors in *G. gynandra* were identified by sequence similarity.
886 Phygenetic trees of the protein sequences from both species were made using the ete3 tool.
887 The tree was loaded into the iTOL web tool where the log(BS/M or Whole Leaf) ratio of each
888 gene was added to the tree. This expression data was obtained from publicly available
889 datasets[47,48].
890
891 **Re-processing of *O. sativa* and *Z. mays* de-etiolation time-course RNA-SEQ**
892 Data from the monocot de-etiolation study[44] was downloaded from the Short Read
893 Archive (SRX766219). Reads for both species were quantified using Salmon quant with the
894 *Z. mays* reads being mapped to Zm-B73-REFERENCE-NAM-5.0_Zm00001e.1.cdna.fa file
895 available from MaizeDB while *O. sativa* reads were mapped to
896 Osativa_323_v7.0.cds_primaryTranscriptOnly.fa available from Phytozome. TPM values
897 were quantile normalised and then each value divided by the sample mean. *O. sativa* C_4
898 orthologs were identified using orthofinder to identify orthogroups with the Arabidopsis C_4
899 orthologs used in this study. *Z. mays* C_4 genes were identified by blasting to these same
900 Arabidopsis genes. Line plots were then made grouping all putative orthologs.
901
902 **ACCESSION NUMBERS**
903 Raw sequencing data files are deposited in The National Center for Biotechnology
904 Information (PRJNA640984). For full methods, commands, and scripts, see GitHub
905 (https://github.com/hibberd-lab/Singh-Stevenson-Gynandra).
906
907 **ACKNOWLEDGMENTS**

908 We thank the Cambridge Advanced Imaging Centre, University of Cambridge for resin
909 embedding the electron microscopy samples and providing the access to electron
910 microscopes. The work was supported by BBSRC Grant BBP0031171 and European
911 Research Council Grant 694733 Revolution to J.M.H.
912

913 **AUTHOR CONTRIBUTIONS**

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

918 **FIGURE LEGENDS**

919 **Fig. 1: Establishment of photosynthesis in *G. gynandra.*** (a) Representative images of
920 *Gynandropsis gynandra* seedlings illustrating greening and unhooking of the cotyledons. (b)
921 Total chlorophyll over the time-course (data shown as means from three biological replicates
922 at each time point, ± one standard deviation from the mean). The first four hours show an
923 exponential increase (inset). Bar along the x-axis indicates periods of light (0-14 hours), dark
924 (14-22 hours) and light (22-24 hours). (c-d) Representative transmission electron
925 microscope images of Mesophyll (c) and Bundle Sheath (d) chloroplasts of de-etiolating *G.*
926 *gynandra* seedlings at 0, 0.5, 2, 4 and 24 hours after exposure to light. Asterisks and
927 arrowheads indicate the prolamellar body and photosynthetic membranes respectively.
928 Samples at each time were taken for RNA-SEQ and DNaseI-SEQ. Scale bars represent 0.5
929 mm for seedlings and 50 µm for cotyledons (a), and 500 nm (c-d).

930

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

944

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

954 C_4 photosynthesis gene from (log10 normalised motif frequency/normalised background
955 frequency). Motifs annotated in orange and the associated Wordcloud highlight those
956 enriched in the C_4 cistrome compared with the C_3 cistrome, and those in red indicate bZIP
957 motifs enriched in both the C_3 and C_4 cistromes from both FIMO and AME analysis.
958
959 **Fig. 4: Transcription factor binding atlas for de-etiolating seedlings *G. gynandra.*** (a)
960 Schematic illustrating sampling, number of Digital Genomic Footprints (DGF) identified and
961 representative density plot of DGF positions relative to the nearest transcription start site
962 (TSS). (b) Pie-charts summarising the density of DGF among genomic features. Promoters
963 are defined as sequence < 2000 base pairs upstream of TSSs while intergenic represent
964 any regions not overlapping with other features. Values indicate densities of DGFs in each
965 feature as proportions. (c) Bar chart showing the percentage of DGFs predicted to function
966 either as activators (coral bars) or repressors (turquoise bars) lying within gene features of
967 target genes. Statistically significant differences were found for the promoters, CDS and
968 intronic regions using a Chi square goodness of fit test ("*"). (d) Heatmap of motif
969 frequencies (log10 sample normalised motif frequency/row mean) during de-etiolation. To
970 illustrate identity and heterogeneity of motif groups clusters were annotated with
971 Wordclouds.
972
973 **Fig. 5: Comparison of transcript abundance for photosynthetic genes during de-**
974 **etiolation of C_3 *Arabidopsis thaliana* and C_4 *G. gynandra*.** (a) Schematic illustrating RNA-
975 SEQ of Arabidopsis. (b) Expression patterns of photosynthesis genes (grey sidebar) and C_4
976 orthologs (black sidebar) during de-etiolation. Heatmap illustrating gene expression with
977 each gene being represented by a row, and data centred around the row mean.
978 Dendrograms (red, yellow and green) highlight distinct expression clusters representing no
979 clear, moderate, or strong induction. (c) Line graphs depicting quantile normalised and mean
980 divided expression patterns of twelve C_4 orthologs in C_3 Arabidopsis and C_4 *G. gynandra*.
981 Where there is more than a 1:1 relationship between genes, the most abundant paralog from
982 each orthogroup is presented.
983
984 **Fig. 6: Comparative analysis of potential regulatory mechanisms for de-etiolating**
985 **seedlings of C_3 *Arabidopsis thaliana* and C_4 *G. gynandra*.** Scatter plots showing the most
986 enriched motifs in each cistrome (where 1 represents the most enriched motif). (a) Top 50
987 motifs in photosynthesis genes of C_3 Arabidopsis (At) and C_4 *G. gynandra* (Gg), (b) C_4 and
988 photosynthesis genes of C_3 Arabidopsis, and (c) C_4 genes from C_3 Arabidopsis and *C.*
989 *gynandra*. Motifs from the cistromes of C_3 and C_4 genes that showed induction during de-
990 etiolation. (d) Heatmap of the top 50 motifs from DHSs of C_4 genes in *G. gynandra*

991 compared with their ranking in C_4 genes of Arabidopsis and photosynthesis genes in both
992 species (log2 of the motif ranks across all four cistrome sets). Two distinct groups are
993 highlighted with green motifs being highly ranked (more enriched) in all four cistromes while
994 the red motifs are those specifically highly ranked in the *G. gynandra* C_4 cistrome. Motifs
995 characterised as light-regulatory elements (LREs) are labelled with symbols used in (e). (e)
996 Sequence logos highlighting different classes of Light Responsive Elements (LREs) and
997 regulators of photosynthesis-associated nuclear genes (PhANGs). Seqlogos were generated
998 from DAP-SEQ and PBM consensus motifs for all members of each type. (f) Analysis of
999 which LREs and regulators of PhANGs are statistically enriched in cistromes of C_4 genes
1000 from *G. gynandra* and Arabidopsis. AME generates likelihood score for over-representation
1001 (-1*log(adjusted p-value), y-axis), and the adjusted $p < 0.05$ is illustrated with a dashed line.
1002 (g) Transcription factor binding sites associated with EE and I-box binding as well as the
1003 homeodomain and LOB/AS2 families dis-proportionally found in C_4 genes from *G. gynandra*
1004 compared with orthologs from Arabidopsis, and photosynthesis genes in both species.
1005 Values plotted are the motif proportion of the total number of DGF at each sample over the
1006 time-course such that differences between and within experiments were normalised. (h)
1007 Model illustrating association between enhanced C_4 cycle gene expression in *G. gynandra*
1008 compared with Arabidopsis and gain of *cis*-elements bound by MYB-related and C2C2-
1009 GATA transcription factors as well as the gain of homeodomain binding sites in mesophyll
1010 expressed genes in C_4 *G. gynandra*.
1011
1012 **SUPPLEMENTARY FIGURE LEGENDS**
1013 **Supplementary Fig. 1:** Representative light and scanning electron microscope (SEM)
1014 images of 0 hours (A) and 24 hours (B) de-etiolating *G. gynandra* seedlings at 0 and 24
1015 hours after exposure to light. Scale bars represent 100 µm for light microscope images, and
1016 500 nm for SEM.
1017
1018 **Supplementary Fig. 2:** GO term enrichment analysis for differentially expressed genes as
1019 compared to the previous time point. Significantly enriched GO terms were identified using
1020 AgriGov2 using a custom *G. gynandra* background built by mapping *G. gynandra* proteins to
1021 their closest match in Arabidopsis and inheriting their terms from the TAIR10 annotations.
1022 Values plotted are -log10(FDR) and values derived from the up-regulated gene sets are
1023 shown in red while those form the down-regulated are shown in blue. Many light and
1024 photosynthesis-related terms are enriched in the 0.5 hours up-regulated genes. Many
1025 primary and secondary metabolism terms are enriched in the 24 hours up-regulated genes
1026 suggesting that photosynthates are being produced by the end of the time course.

1027

1028 **Supplementary Fig. 3:** Pipeline for DNaseI-SEQ data processing. On the top left-hand side,
1029 pooled reads went through quality control before DHS identification using
1030 MACS2 peakcalling. The DHSs, representing accessible chromatin regions, are then
1031 searched for DGF using the pyDNase package. These DGF are prone to distorting effects
1032 due to DNaseI bias in gDNA digestion. On the top right, the pipeline for identifying this bias
1033 is shown which includes the DNaseI digestion of deproteinised ("naked") gDNA. This
1034 generates 6-mer frequencies at each cut site which is used as input for
1035 the FootPrintMixture.R tool which scores the Footprint Likelihood Ratio (FLR) of each DGF
1036 (likelihood of being a true positive). DGF with FLR < 0 were removed leaving a final set of
1037 DGF which were used for analysis. Heatmap of cut patterns are shown centred around each
1038 DGF.

1039

1040 **Supplementary Fig. 4:** Violin plots showing the distributions of dDHS scores for DHS
1041 overlapping with differentially expressed genes at 0.5 hours with both up-regulated (A) and
1042 down-regulated shown (B). Mean values are shown as line. Positive dDHS scores represent
1043 an increase in DHS accessibility and negative values represent the opposite. No clear
1044 association is observed with up-regulated genes and positive dDHS values nor down-
1045 regulated genes with negative dDHS values.

1046

1047 **Supplementary Fig. 5:** Line plots showing mean normalised expression values at each time
1048 point for both *G. gynandra* and Arabidopsis across their respective de-etiolation time-
1049 courses. Data shown for twenty orthogroups induced during de-etiolation with paralogs
1050 shown. Values are quantile normalised followed by dividing by the sample mean to facilitate
1051 expression dynamics and abundance comparisons across the species.

1052

1053 **Supplementary Fig. 6:** Line plots showing mean normalised expression values at each time
1054 point for both *O. sativa* and *Z. mays* across their respective de-etiolation time-courses. Data
1055 from a monocot de-etiolation study[44] and processed in the same way as described for the
1056 data in this study. Many orthogroups showed a similar pattern as for the *G. gynandra* and
1057 Arabidopsis comparison with one or more C_4 (*Z. mays*) paralog showing much higher
1058 abundance than other members of the orthogroup.

1059

1060 **Supplementary Fig. 7:** DGF binding in different C_4 cycle genes during de-etiolation time
1061 course in *G. gynandra*.

1062

1063 **Supplementary Fig. 8:** Unrooted phylogenetic tree of homeodomain protein sequences
1064 taken from both *G. gynandra* and Arabidopsis. Analysis was carried out using the ete3
1065 pipeline and visualised with the iTOL web tool. Each leaf is annotated, where available, with
1066 the ratio of the bundle sheath to mesophyll expression values taken from publicly available
1067 datasets. The HD-Zip IV family is shown to be consistently mesophyll preferentially
1068 expressed in both species in contrast to the HD-Zip III family. Major family names
1069 and individual gene names of interest are shown.
1070

1071 **REFERENCES**

1072 1. Bowes, G., Ogren, W. L. & Hageman, R. H. Phosphoglycolate production catalyzed
1073 by ribulose diphosphate carboxylase. *Biochem. Biophys. Res. Commun.* **45**, 716–722
1074 (1971).
1075 2. Tolbert, N. E. & Essner, E. Microbodies: Peroxisomes and glyoxysomes. *J. Cell Biol.*
1076 **91**, 271s-283s (1981).
1077 3. Bauwe, H., Hagemann, M. & Fernie, A. R. Photorespiration: players, partners and
1078 origin. *Trends in Plant Science* vol. 15 330–336 (2010).
1079 4. Hatch, M. D. C4 photosynthesis: a unique blend of modified biochemistry, anatomy
1080 and ultrastructure. *BBA Reviews On Bioenergetics* vol. 895 81–106 (1987).
1081 5. Furbank, R. T. Evolution of the C_4 photosynthetic mechanism: Are there really three
1082 C_4 acid decarboxylation types? *Journal of Experimental Botany* vol. 62 3103–3108
1083 (2011).
1084 6. Sage, R. F. & Zhu, X. G. Exploiting the engine of C_4 photosynthesis. *J. Exp. Bot.* **62**,
1085 2989–3000 (2011).
1086 7. Sage, R. F. A portrait of the C_4 photosynthetic family on the 50th anniversary of its
1087 discovery: Species number, evolutionary lineages, and Hall of Fame. *Journal of*
1088 *Experimental Botany* vol. 67 4039–4056 (2016).
1089 8. Hibberd, J. M. & Covshoff, S. The Regulation of Gene Expression Required for C_4
1090 Photosynthesis. *Annu. Rev. Plant Biol.* **61**, 181–207 (2010).
1091 9. Burnell, J. N., Suzuki, I. & Sugiyama, T. Light Induction and the Effect of Nitrogen
1092 Status upon the Activity of Carbonic Anhydrase in Maize Leaves. *Plant Physiol.* **94**,
1093 384–387 (1990).
1094 10. Glackin, C. A. & Grula, J. W. Organ-specific transcripts of different size and
1095 abundance derive from the same pyruvate, orthophosphate dikinase gene in maize.
1096 *Proc. Natl. Acad. Sci. U. S. A.* **87**, 3004–3008 (1990).
1097 11. Gowik, U. *et al.* cis-regulatory elements for mesophyll-specific gene expression in the
1098 C_4 plant *Flaveria trinervia*, the promoter of the C_4 phosphoenolpyruvate carboxylase
1099 gene. *Plant Cell* **16**, 1077–1090 (2004).

1100 12. Brown, N. J. *et al.* Independent and parallel recruitment of preexisting mechanisms
1101 underlying C_4 photosynthesis. *Science* **331**, 1436–1439 (2011).
1102 13. Reyna-Llorens, I. *et al.* Ancient duons may underpin spatial patterning of gene
1103 expression in C_4 leaves. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1931–1936 (2018).
1104 14. Williams, B. P. *et al.* An untranslated cis-element regulates the accumulation of
1105 multiple C_4 enzymes in *Gynandropsis gynandra* mesophyll cells. *Plant Cell* **28**, 454–
1106 465 (2016).
1107 15. Kajala, K. *et al.* Multiple Arabidopsis genes primed for recruitment into C_4
1108 photosynthesis. *Plant J.* **69**, 47–56 (2012).
1109 16. Stockhaus, J., Poetsch, W., Steinmüller, K. & Westhoff, P. Evolution of the C_4
1110 phosphoenolpyruvate carboxylase promoter of the C_4 dicot *Flaveria trinervia*: an
1111 expression analysis in the C_3 plant tobacco. *MGG Mol. Gen. Genet.* **245**, 286–293
1112 (1994).
1113 17. Stockhaus, J. *et al.* The promoter of the gene encoding the C_4 form of
1114 phosphoenolpyruvate carboxylase directs mesophyll-specific expression in transgenic
1115 C_4 *Flaveria* spp. *Plant Cell* **9**, 479–489 (1997).
1116 18. Matsuoka, M., Kyozuka, J., Shimamoto, K. & Kano-Murakami, Y. The promoters of
1117 two carboxylases in a C_4 plant (maize) direct cell-specific, light-regulated expression
1118 in a C_3 plant (rice). *Plant J.* **6**, 311–319 (1994).
1119 19. Nomura, M. *et al.* The evolution of C_4 plants: Acquisition of cis-regulatory sequences
1120 in the promoter of C_4-type pyruvate, orthophosphate dikinase gene. *Plant J.* **22**, 211–
1121 221 (2000).
1122 20. Nomura, M. *et al.* The promoter for C_4-type mitochondrial aspartate aminotransferase
1123 does not direct bundle sheath-specific expression in transgenic rice plants. *Plant Cell*
1124 *Physiol.* **46**, 743–753 (2005).
1125 21. Argüello-Astorga, G. R. & Herrera-Estrella, L. R. Ancestral multipartite units in light-
1126 responsive plant promoters have structural features correlating with specific
1127 phototransduction pathways. *Plant Physiol.* **112**, 1151–1166 (1996).
1128 22. Acevedo-Hernández, G. J., León, P. & Herrera-Estrella, L. R. Sugar and ABA
1129 responsiveness of a minimal RBCS light-responsive unit is mediated by direct binding
1130 of ABI4. *Plant J.* **43**, 506–519 (2005).
1131 23. Burgess, S. J. *et al.* Ancestral light and chloroplast regulation form the foundations for
1132 C_4 gene expression. *Nat. Plants* **2**, (2016).
1133 24. Pribil, M., Labs, M. & Leister, D. Structure and dynamics of thylakoids in land plants.
1134 *J. Exp. Bot.* **65**, 1955–1972 (2014).
1135 25. Bahl, J., Francke, B. & Monéger, R. Lipid composition of envelopes, prolamellar
1136 bodies and other plastid membranes in etiolated, green and greening wheat leaves.

1137 *Planta* **129**, 193–201 (1976).
1138 26. Ryberg, M. & Sundqvist, C. Characterization of prolamellar bodies and prothylakoids
1139 fractionated from wheat etioplasts. *Physiol. Plant.* **56**, 125–132 (1982).
1140 27. Ryberg, M. & Sundqvist, C. The regular ultrastructure of isolated prolamellar bodies
1141 depends on the presence of membrane-bound NADPH-protochlorophyllide
1142 oxidoreductase. *Physiol. Plant.* **73**, 218–226 (1988).
1143 28. Park, H., Kreunen, S. S., Cuttriss, A. J., DellaPenna, D. & Pogson, B. J. Identification
1144 of the carotenoid isomerase provides insight into carotenoid biosynthesis, prolamellar
1145 body formation, and photomorphogenesis. *Plant Cell* **14**, 321–332 (2002).
1146 29. Armarego-Marriott, T. *et al.* Highly resolved systems biology to dissect the etioplast-
1147 to-chloroplast transition in tobacco leaves. *Plant Physiol.* **180**, 654–681 (2019).
1148 30. Dubreuil, C. *et al.* Establishment of photosynthesis through chloroplast development
1149 is controlled by two distinct regulatory phases. *Plant Physiol.* **176**, 1199–1214 (2018).
1150 31. Mizoguchi, T. *et al.* LHY and CCA1 are partially redundant genes required to maintain
1151 circadian rhythms in Arabidopsis. *Dev. Cell* **2**, 629–641 (2002).
1152 32. Leivar, P. *et al.* The Arabidopsis phytochrome-interacting factor PIF7, together with
1153 PIF3 and PIF4, regulates responses to prolonged red light by modulating phyB levels.
1154 *Plant Cell* **20**, 337–352 (2008).
1155 33. Liu, Y. *et al.* Genome-wide mapping of DNase i hypersensitive sites reveals chromatin
1156 accessibility changes in Arabidopsis euchromatin and heterochromatin regions under
1157 extended darkness. *Sci. Rep.* **7**, 1–15 (2017).
1158 34. He, H. H. *et al.* Differential DNase I hypersensitivity reveals factor-dependent
1159 chromatin dynamics. *Genome Res.* **22**, 1015–1025 (2012).
1160 35. O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA
1161 Landscape. *Cell* **165**, 1280–1292 (2016).
1162 36. Franco-Zorrilla, J. M. *et al.* DNA-binding specificities of plant transcription factors and
1163 their potential to define target genes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2367–2372
1164 (2014).
1165 37. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given
1166 motif. *Bioinformatics* **27**, 1017–1018 (2011).
1167 38. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: A unified framework and an
1168 evaluation on ChIP data. *BMC Bioinformatics* **11**, 165 (2010).
1169 39. Green, P. J., Kay, S. A. & Chua, N. H. Sequence-specific interactions of a pea nuclear
1170 factor with light-responsive elements upstream of the rbcS-3A gene. *EMBO J.* **6**,
1171 2543–2549 (1987).
1172 40. Giuliano, G. *et al.* An evolutionarily conserved protein binding sequence upstream of a
1173 plant light-regulated gene. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 7089–7093 (1988).

1174 41. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in
1175 transcription factor footprint identification. *Nat. Methods* **11**, 73–78 (2014).
1176 42. Yardimci, G. G., Frank, C. L., Crawford, G. E. & Ohler, U. Explicit DNase sequence
1177 bias modeling enables high-resolution transcription factor footprint detection. *Nucleic*
1178 *Acids Res.* **42**, 11865–11878 (2014).
1179 43. Sullivan, A. M. *et al.* Mapping and dynamics of regulatory DNA and transcription factor
1180 networks in *A. thaliana*. *Cell Rep.* **8**, 2015–2030 (2014).
1181 44. Xu, J., Bräutigam, A., Li, Y., Weber, A. P. M. & Zhu, X.-G. Systems analysis of cis-
1182 regulatory motifs in C_4 photosynthesis genes using maize and rice leaf transcriptomic
1183 data during a process of de-etiolation. *J. Exp. Bot.* **67**, 5105–5117 (2016).
1184 45. Waters, M. T. *et al.* GLK transcription factors coordinate expression of the
1185 photosynthetic apparatus in Arabidopsis. *Plant Cell* **21**, 1109–1128 (2009).
1186 46. Xu, Z., Casaretto, J. A., Bi, Y.-M. & Rothstein, S. J. Genome-wide binding analysis of
1187 AtGNC and AtCGA1 demonstrates their cross-regulation and common and specific
1188 functions. *Plant Direct* **1**, e00016 (2017).
1189 47. Aubry, S., Smith-Unna, R. D., Boursnell, C. M., Kopriva, S. & Hibberd, J. M. Transcript
1190 residency on ribosomes reveals a key role for the *Arabidopsis thaliana* bundle sheath
1191 in sulfur and glucosinolate metabolism. *Plant J.* **78**, 659–673 (2014).
1192 48. Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D. & Hibberd, J. M. Deep
1193 Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-
1194 Factors in Two Independent Origins of C_4 Photosynthesis. *PLoS Genet.* **10**, (2014).
1195 49. Kirk, J. T. 0. & Tilney-Bassett, R. A. E. *The Plastids*. (Elsevier/North-Holland
1196 Biomedical Press, 1967).
1197 50. Boffey, S. A., Selldén, G. & Leech, R. M. Influence of Cell Age on Chlorophyll
1198 Formation in Light-grown and Etiolated Wheat Seedlings. *Plant Physiol.* **65**, 680–684
1199 (1980).
1200 51. Virgin, H. I., Kahn, A. & von Wettstein, D. The Physiology of Chlorophyll Formation in
1201 Relation to Structural Changes in Chloroplasts. *Photochem. Photobiol.* **2**, 83–91
1202 (1963).
1203 52. Thorne, S. W. & Boardman, N. K. Formation of Chlorophyll b , and the Fluorescence
1204 Properties and Photochemical Activities of Isolated Plastids from Greening Pea
1205 Seedlings . *Plant Physiol.* **47**, 252–261 (1971).
1206 53. Høyer-Hansen, G. & Simpson, D. J. Changes in the polypeptide composition of
1207 internal membranes of barley plastids during greening. *Carlsberg Res. Commun.* **42**,
1208 379–389 (1977).
1209 54. Bennett, J. Chloroplast protein phosphorylation and the regulation of photosynthesis.
1210 *Physiol. Plant.* **60**, 583–590 (1984).

1211 55. Demeter, S., Mustardy, L. & Machowicz, E. The development of the intense circular
1212 dichroic signal during granum formation in greening etiolated maize. *Biochem. J.* **156**,
1213 469–472 (1976).
1214 56. Lonosky, P. M. *et al.* A Proteomic Analysis of Maize Chloroplast Biogenesis. *Plant*
1215 *Physiol.* **134**, 560–574 (2004).
1216 57. Cahoon, A. B., Takacs, E. M., Sharpe, R. M. & Stern, D. B. Nuclear, chloroplast, and
1217 mitochondrial transcript abundance along a maize leaf developmental gradient. *Plant*
1218 *Mol. Biol.* **66**, 33–46 (2008).
1219 58. Koteyeva, N. K., Voznesenskaya, E. V, Roalson, E. H. & Edwards, G. E. Diversity in
1220 forms of C_4 in the genus *Cleome* (Cleomaceae). *Ann. Bot.* **107**, 269–283 (2011).
1221 59. Bahl, J., Francke, B. & Monéger, R. Lipid composition of envelopes, prolamellar
1222 bodies and other plastid membranes in etiolated, green and greening wheat leaves.
1223 *Planta* **129**, 193–201 (1976).
1224 60. Baker, N. R. & Butler, W. L. Development of the Primary Photochemical Apparatus of
1225 Photosynthesis during Greening of Etiolated Bean Leaves. *Plant Physiol.* **58**, 526–529
1226 (1976).
1227 61. Boardman, N. K. Development of Chloroplast Structure and Function. in
1228 *Photosynthesis I* 583–600 (Springer Berlin Heidelberg, 1977). doi:10.1007/978-3-642-
1229 66505-9_42.
1230 62. Krupinska, K. & Apel, K. Light-induced transformation of etioplasts to chloroplasts of
1231 barley without transcriptional control of plastid gene expression. *MGG Mol. Gen.*
1232 *Genet.* **219**, 467–473 (1989).
1233 63. Boasson, R. & Laetsch, W. M. Chloroplast replication and growth in tobacco. *Science*
1234 **166**, 749–751 (1969).
1235 64. Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human
1236 regulatory DNA landscapes. *Cell* **154**, 888–903 (2013).
1237 65. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by
1238 STARR-seq. *Science* **339**, 1074–1077 (2013).
1239 66. Bräutigam, A. *et al.* An mRNA blueprint for C_4 photosynthesis derived from
1240 comparative transcriptomics of closely related C_3 and C_4 species. *Plant Physiol.* **155**,
1241 142–156 (2011).
1242 67. Külahoglu, C. *et al.* Comparative transcriptome atlases reveal altered gene
1243 expression modules between two Cleomaceae C_3 and C_4 plant species. *Plant Cell* **26**,
1244 3243–3260 (2014).
1245 68. Gowik, U., Bräutigam, A., Weber, K. L., Weber, A. P. M. & Westhoff, P. Evolution of
1246 C_4 photosynthesis in the genus *Flaveria*: How many and which genes does it take to
1247 make C_4? *Plant Cell* **23**, 2087–2105 (2011).

1248 69. Ariel, F. D., Manavella, P. A., Dezar, C. A. & Chan, R. L. The true story of the HD-Zip
1249 family. *Trends in Plant Science* vol. 12 419–426 (2007).
1250 70. Kubo, H. & Hayashi, K. Characterization of root cells of *anl2* mutant in *Arabidopsis*
1251 *thaliana*. *Plant Sci.* **180**, 679–685 (2011).
1252 71. Kubo, H., Kishi, M. & Goto, K. Expression analysis of *ANTHOCYANINLESS2* gene in
1253 Arabidopsis. *Plant Sci.* **175**, 853–857 (2008).
1254 72. Martinez, E. Multi-protein complexes in eukaryotic gene transcription. *Plant Mol. Biol.*
1255 **50**, 925–947 (2002).
1256 73. Rose, A., Meier, I. & Wienand, U. The tomato I-box binding factor LeMYBI is a
1257 member of a novel class of Myb-like proteins. *Plant J.* **20**, 641–652 (1999).
1258 74. Richter, R., Behringer, C., Müller, I. K. & Schwechheimer, C. The GATA-type
1259 transcription factors GNC and GNL/CGA1 repress gibberellin signaling downstream
1260 from DELLA proteins and phytochrome-interacting factors. *Genes Dev.* **24**, 2093–
1261 2104 (2010).
1262 75. Porra, R. J., Thompson, W. A. & Kriedemann, P. E. Determination of accurate
1263 extinction coefficients and simultaneous equations for assaying chlorophylls a and b
1264 extracted with four different solvents: verification of the concentration of chlorophyll
1265 standards by atomic absorption spectroscopy. *BBA - Bioenerg.* **975**, 384–394 (1989).
1266 76. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina
1267 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
1268 77. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast
1269 and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419
1270 (2017).
1271 78. Anders, S. & Huber, W. Differential expression analysis for sequence count data.
1272 *Genome Biol.* **11**, R106 (2010).
1273 79. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of
1274 multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids*
1275 *Res.* **40**, 4288–4297 (2012).
1276 80. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for
1277 comparative genomics. *Genome Biol.* **20**, 238 (2019).
1278 81. Piper, J. *et al.* Wellington: A novel method for the accurate identification of digital
1279 genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41**, (2013).
1280 82. Piper, J. *et al.* Wellington-bootstrap: Differential DNase-seq footprinting identifies cell-
1281 type determining transcription factors. *BMC Genomics* **16**, 1000 (2015).
1282 83. Burgess, S. J. *et al.* Genome-wide transcription factor binding in leaves from C_3 and
1283 C_4 grasses. *Plant Cell* **31**, 2297–2314 (2019).
1284

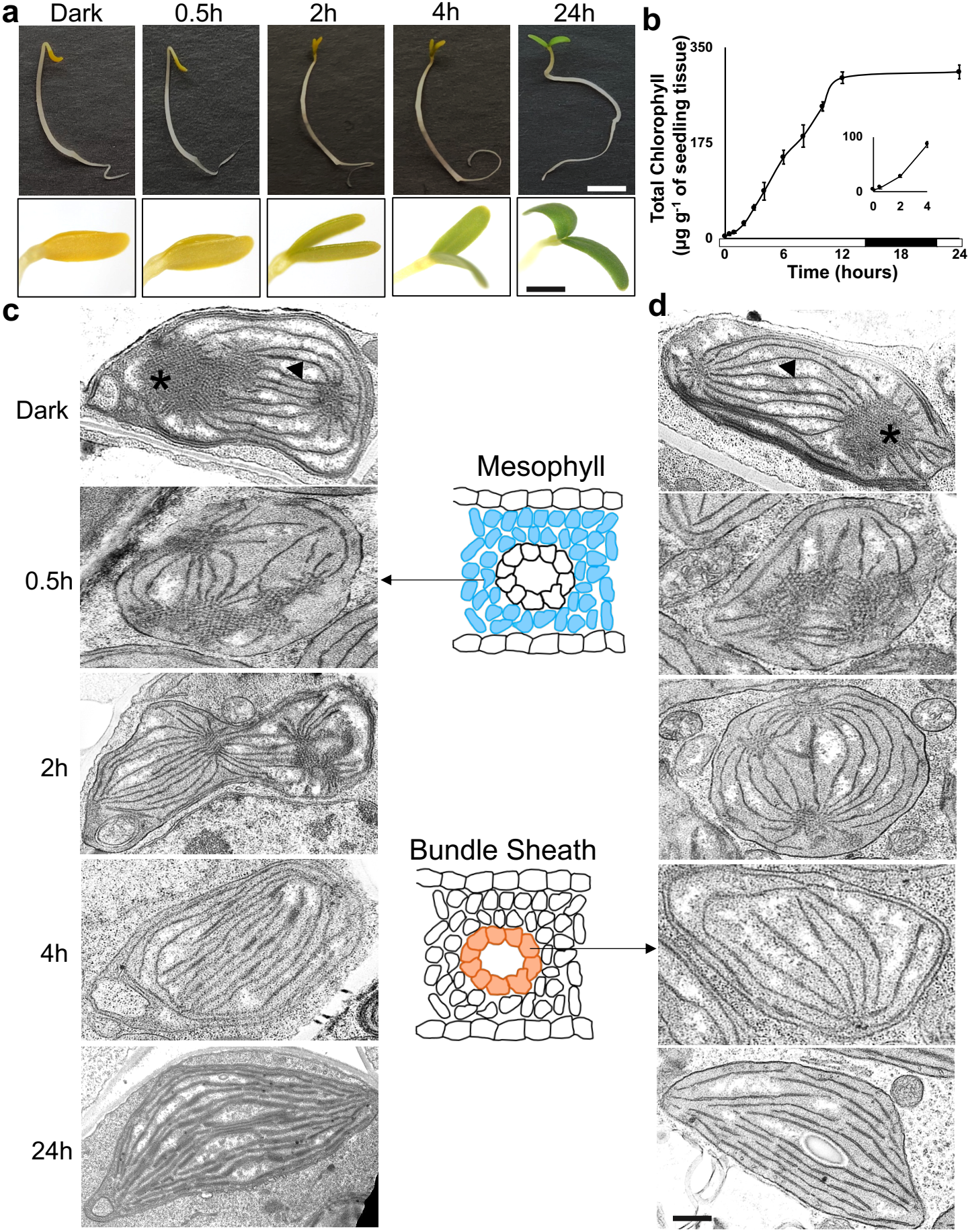


Fig. 1: Establishment of photosynthesis in *G. gynandra*. (a) Representative images of *Gynandropsis gynandra* seedlings illustrating greening and unhooking of the cotyledons. (b) Total chlorophyll over the time-course (data shown as means from three biological replicates at each time point, ± one standard 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).

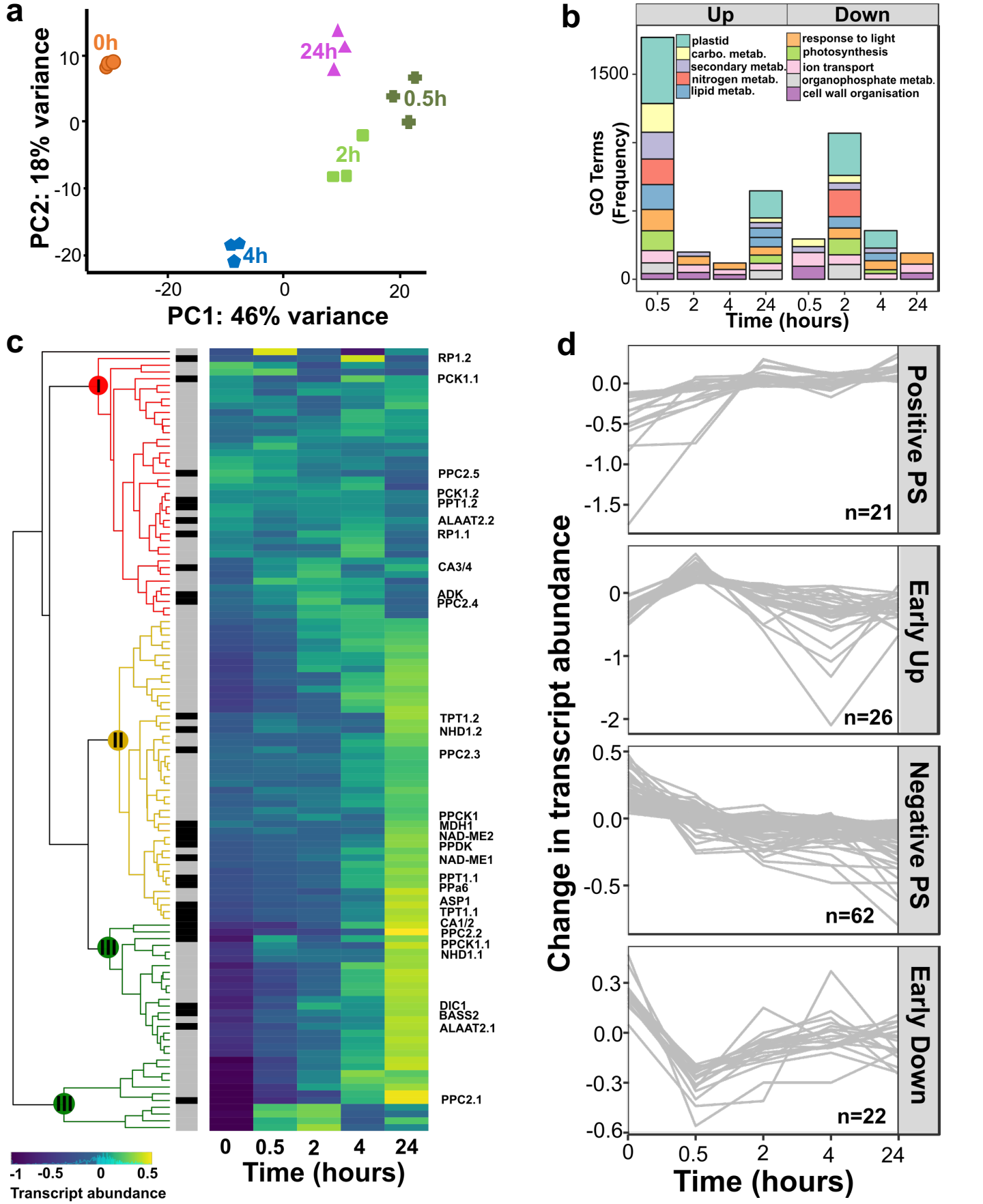
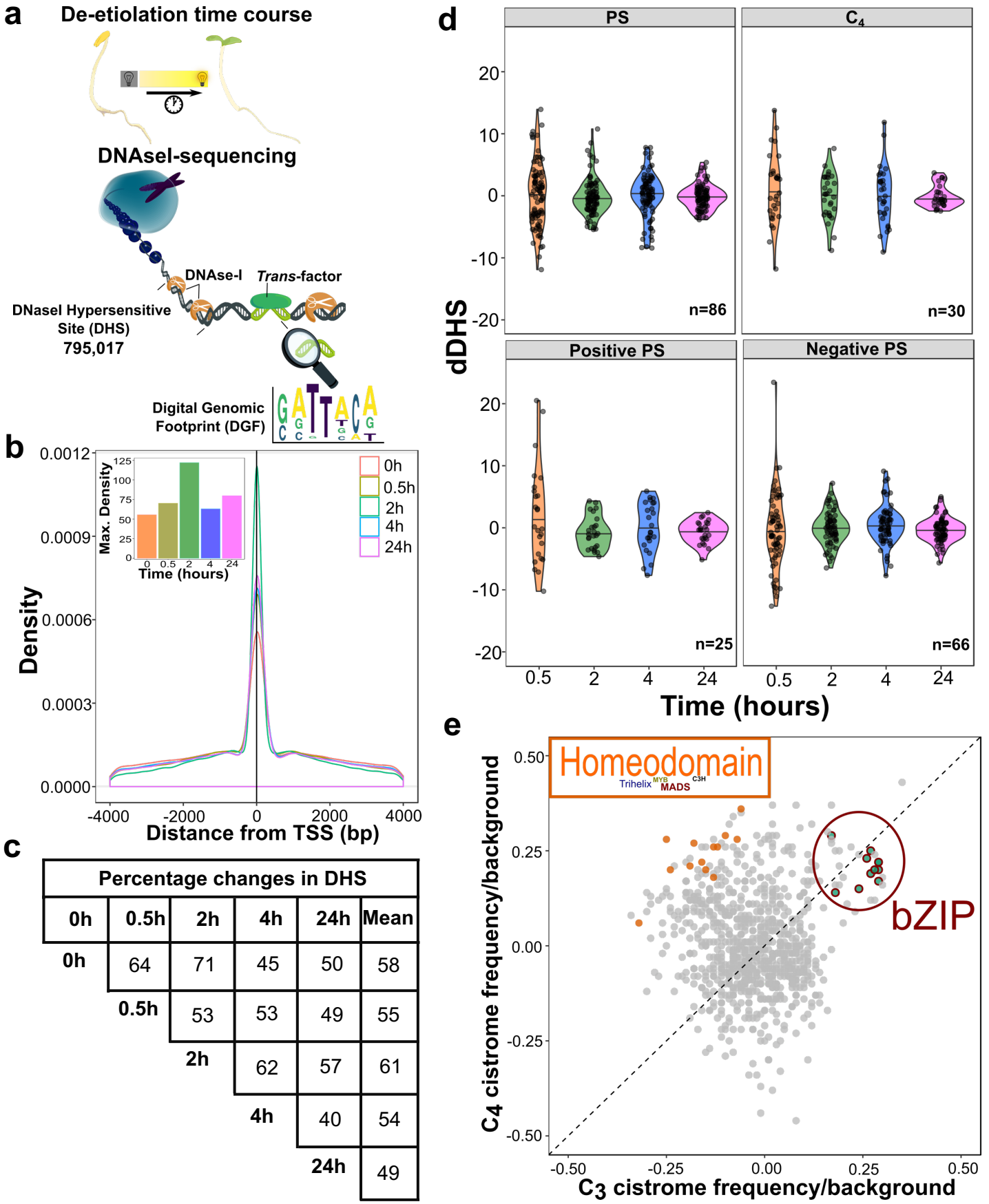


Fig. 2: Changes in transcript abundance during greening of *G. gynandra*. (a) Principal component analysis of RNA-SEQ datasets. The three biological replicates from each timepoint of de-etiolating *G. gynandra* seedlings (0, 0.5, 2, 4 and 24 hours) form distinct clusters. (b) Enriched GO terms between 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.

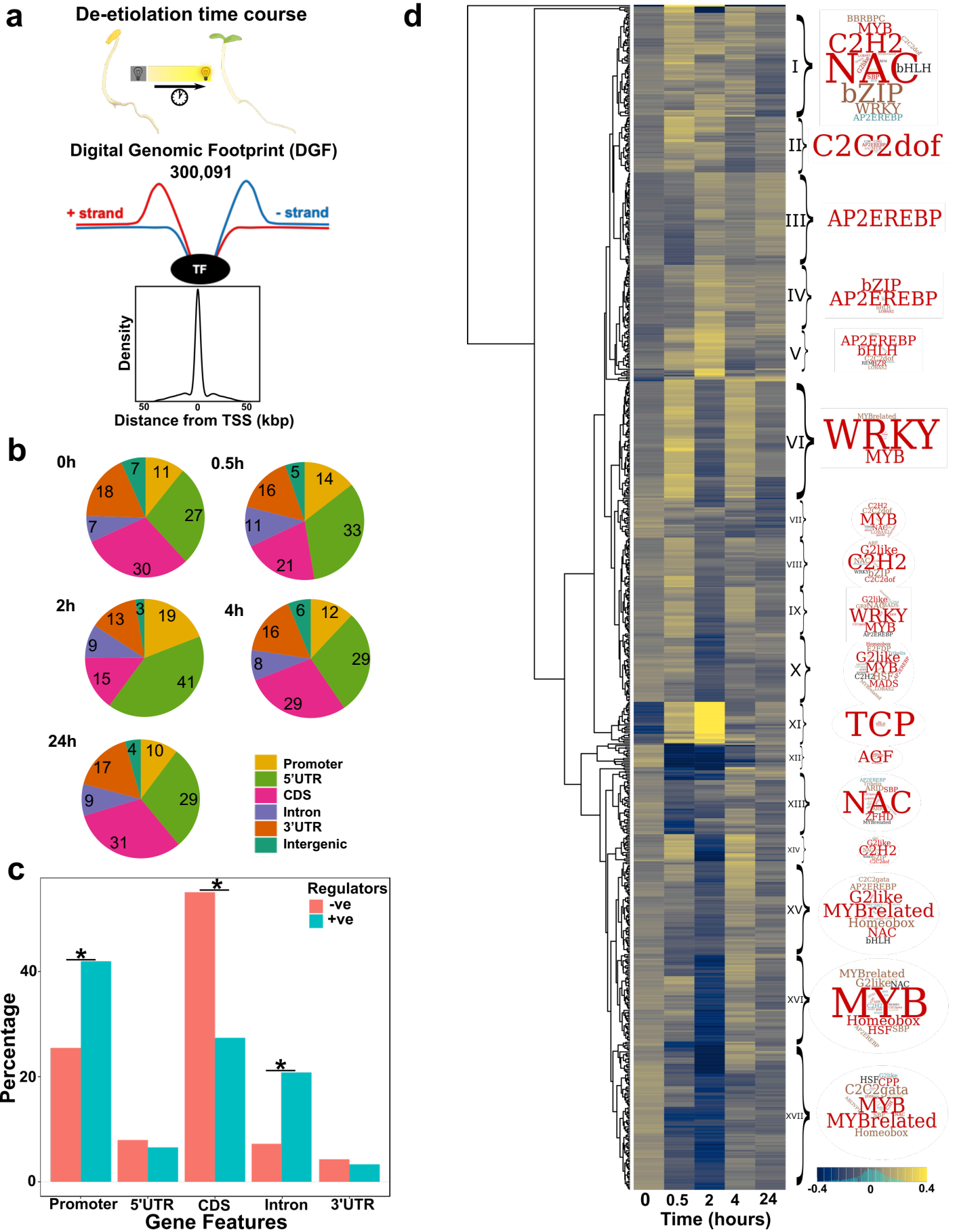


Fig. 4: Transcription factor binding atlas for de-etiolating seedlings *G. gynandra*. (a) Schematic illustrating sampling, number of Digital Genomic Footprints (DGF) identified and representative density plot of DGF positions relative to the nearest transcription start site (TSS). (b) Pie-charts summarising 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.

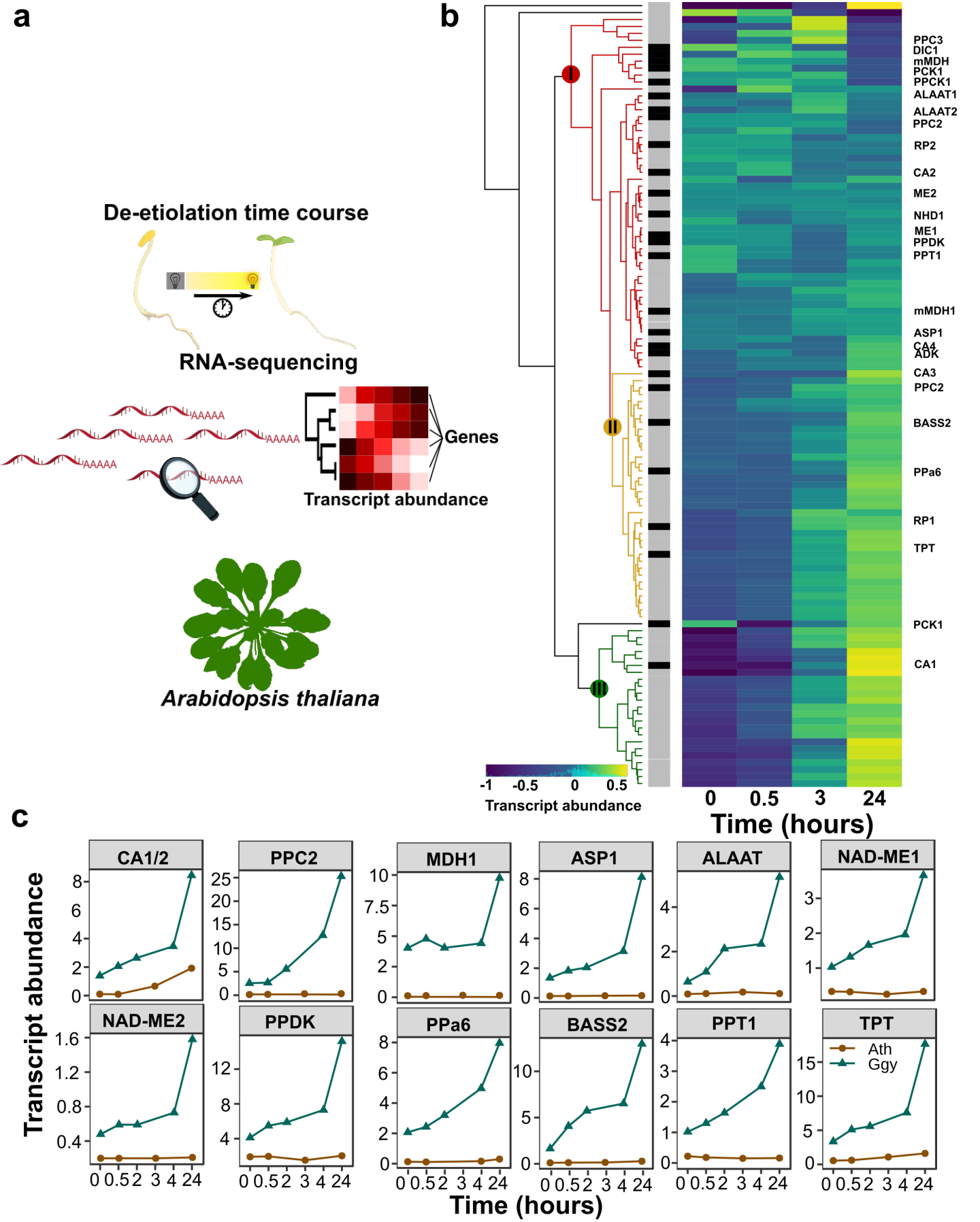


Fig. 5: Comparison of transcript abundance for photosynthetic genes during de-etiolation of C_3 *Arabidopsis thaliana* and C_4 *G. gynandra*. (a) Schematic illustrating RNA-SEQ of Arabidopsis. (b) Expression patterns of photosynthesis genes (grey sidebar) and C_4 orthologs (black sidebar) during de-etiolation. 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.

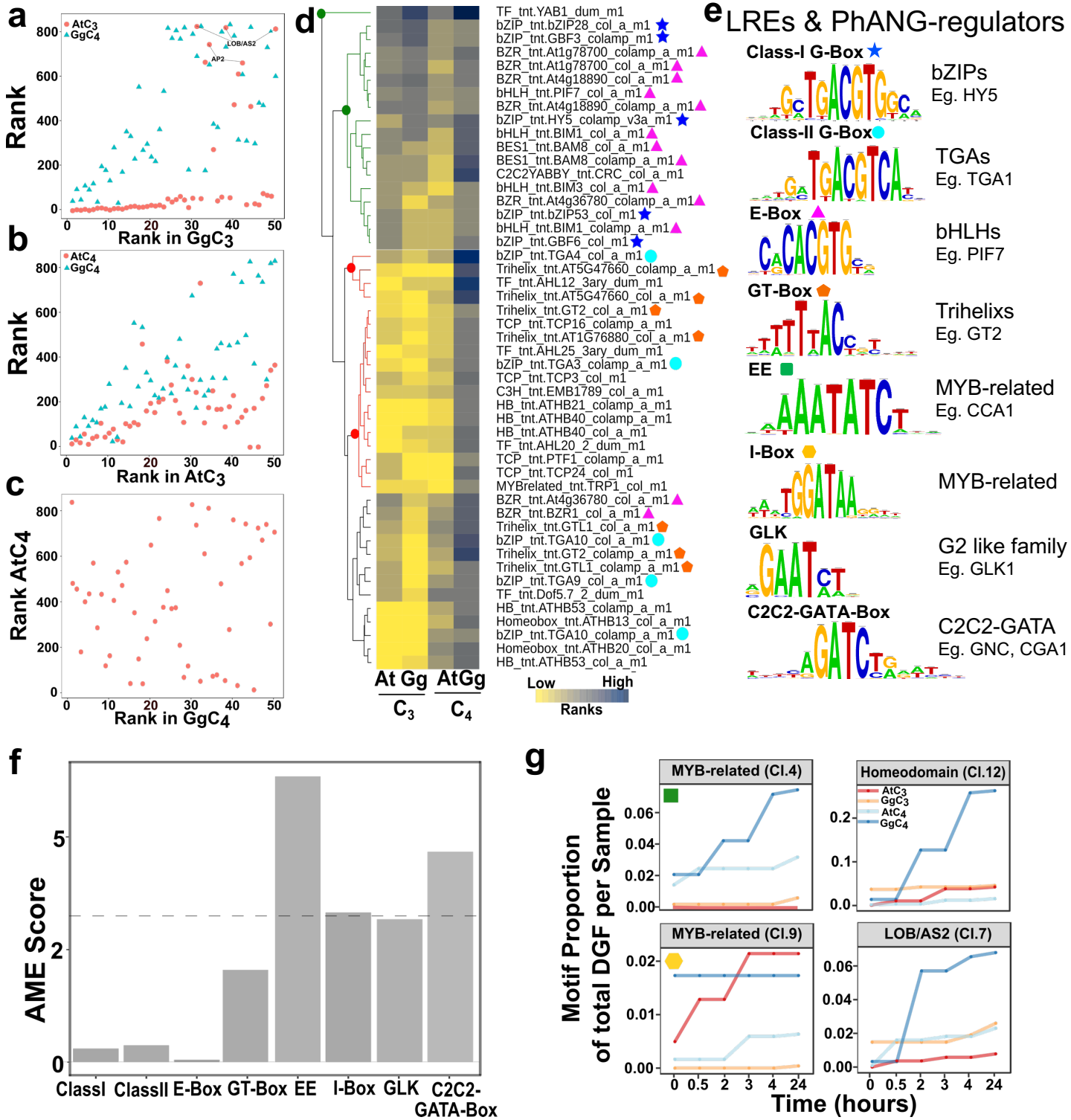


Fig. 6: Comparative analysis of potential regulatory mechanisms for de-etiolating seedlings of C_3 *Arabidopsis thaliana* and C_4 *G. gynandra*. Scatter plots showing the most enriched motifs in each cistrome (where 1 represents the most enriched motif). (a) Top 50 motifs in photosynthesis genes of C_3 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*.