Ancient duons may underpin spatial patterning of gene expression in C_4 leaves

Ancient duons may underpin spatial patterning of gene expression in C_4 leaves

Ivan Reyna-Llorens, Steven J. Burgess, Gregory Reeves, Pallavi Singh, Sean R. Stevenson, Ben P. Williams, Susan Stanley and Julian M. Hibberd

Department of Plant Sciences, Downing Street, University of Cambridge, Cambridge CB2 3EA, UK.

I-RL - suallorens@gmail.com

SJB - sburgess011@gmail.com

GR - gr360@cam.ac.uk

PS - ps753@cam.ac.uk

SRS - srs62@cam.ac.uk

BPM - bp.williams0@gmail.com

SS - ss383@cam.ac.uk

JMH (corresponding) - jmh65@cam.ac.uk. Tel - 44(0) 1223 766547

Abstract

If the highly efficient C_4 photosynthesis pathway could be transferred to crops with the C_3 pathway there could be yield gains of up to 50%. It has been proposed that the multiple metabolic and developmental modifications associated with C_4 photosynthesis are underpinned by relatively few master regulators that have allowed the evolution of C_4 photosynthesis more than sixty times in flowering plants. Here we identify a component of one such regulator that consists of a pair of *cis*-elements located in coding sequence of multiple genes that are preferentially expressed in bundle sheath cells of C_4 leaves. These motifs represent duons as they play a dual role in coding for amino acids as well as controlling the spatial patterning of gene expression associated with the C_4 leaf. They act to repress transcription of C_4 photosynthesis genes in mesophyll cells. These duons are also present in the C_3 model *Arabidopsis thaliana*, and in fact, are conserved in all land plants and even some algae that use C_3 photosynthesis. C_4 photosynthesis therefore appears to have co-opted an ancient regulatory code to generate the spatial patterning of gene expression that is a hallmark of C_4 photosynthesis. This novel intragenic transcriptional regulatory sequence could be exploited in the engineering of efficient photosynthesis of crops.

Significance statement

We describe an ancient regulatory code that patterns the gene expression required for the efficient C_4 pathway. This code is based on two novel regulatory elements located in exonic sequence that act co-operatively to repress transcription in specific cells. As these regulators are located in gene bodies they determine amino acid sequence as well as gene expression and so are known as duons. They are found in multiple genes, are not restricted to C_4 species, but rather found in all land plants and even some algae. The prevalence of these motifs has likely facilitated the repeated evolution of the complex C_4 system, and they also provide a mechanism that could be used to engineer photosynthesis.

/body

Introduction

Photosynthesis forms the basis of life on earth. When plants moved onto land they inherited a photosynthetic system developed by bacteria in which Ribulose Bisphosphate Carboxylase Oxygenase (RuBisCO) generates phosphoglyceric acid (PGA)[1]. As PGA contains three carbon atoms, this is referred to as C_3 photosynthesis. A side-reaction of RuBisCO fixes O_2 rather than CO_2 and generates the toxic compound phosphoglycolate. Although plants use photorespiration to remove phosphoglycolate, both carbon and energy are lost in the process[2]. Around 30 million years ago, some species evolved a system in which CO_2 is concentrated around RuBisCO such that oxygenation is minimised, and photosynthetic efficiency increases by around 50%[3,4]. These species now represent the most productive vegetation on the planet[5,6], and because they initially generate a C_4 acid in the photosynthetic process, are known as C_4 plants.

The C_4 mechanism depends on spatial separation of photosynthetic reactions, most commonly between mesophyll (M) and bundle sheath (BS) cells. How this complex process has evolved in over 66 independent lineages remains a mystery. The finding that expression of multiple *NAD-dependent MALIC ENZYME (NAD-ME)* genes in the BS of C_4 *Gynandropsis gynandra* is dependent on coding sequence[7] hinted at the importance of exons in this process, which are more highly conserved than intragenic sequences and therefore a potential hotspot underpinning repeated evolution. Although genome-wide studies commonly report transcription factor binding within genes[8,9], there is little functional validation to support these findings and their general importance remains unclear.

Results

To better understand mechanisms responsible for generating preferential expression in BS cells of C_4 plants analysis was focussed on the coding region of *GgNAD-ME1*. Although a 240 nucleotide fragment under control of the constitutive CaMV35S promoter confers BS accumulation in *G. gynandra* (Figure 1A&B; Figure S1,2) it was unclear if this was due to transcriptional and post-

transcriptional mechanisms. These can be distinguished using an antisense construct that maintains DNA sequence, but when transcribed generates a complementary mRNA. An antisense construct under the constitutive CaMV35S promoter maintained preferential accumulation in the BS (Figure 1A&B) indicating that DNA sequence is recognised by *trans*-acting factors regardless of the orientation of the sequence. The preferential rather than exclusive accumulation of GUS in BS cells reported here is consistent with quantification of transcripts encoding components of the C_4 cycle in BS and M cells of C_4 leaves[7,10], and also previous analysis of elements controlling gene expression[11–13]. Thus, the coding sequence of *GgNAD-ME1* controls transcription, and as it supresses activity of the constitutive 35SCaMV promoter in M cells, the simplest explanation is that this region interacts with a repressive transcription factor.

Without definition of the motifs within *NAD-ME* genes specifying expression in the BS, it was not possible to determine if they control expression of additional genes, nor to understand if the same elements are used in other species. To identify specific nucleotides responsible for BS expression a deletion series was generated (Figure 1A). Deletion of 24 or 78 nucleotides from the 3' end did not affect preferential accumulation in BS cells (Figure 1A-B) but removal of 90 nucleotides did (Figure 1A-C). Similarly, deletion of the first 63 nucleotides from the 5' end did not abolish preferential accumulation of GUS in BS cells, but removing 78 nucleotides did (Figure 1A-C). A fragment incorporating bases 64 to 162 was sufficient for preferential accumulation in the BS after microprojectile bombardment (Figure 1B) and stable transformation (Figure 1C; Figure S2). We conclude that one region composed of nucleotides TTGGGTGAA (64 to 79 downstream of the translational start codon) and another of GATCCTTG (141 to 162 nucleotides downstream of the translational start codon) are necessary for preferential accumulation of *GgNAD-ME1* in BS cells of C_4 *G. gynandra*. These two regions are separated by 75 nucleotides and will hereafter be referred to as Bundle Sheath Motif 1a (BSM1a) and Bundle Sheath Motif 1b (BSM1b).

In order to refine efforts to identify the presence of this *cis*-regulatory logic in other genes, each motif was subjected to site-directed mutagenesis and the spacer separating BSM1a and BSM1b was replaced with exogenous sequence (Figure S3). This showed that both regions are necessary but

that the intervening sequence is not required for preferential accumulation of GUS in BS cells (Figure S3). The fact that BSM1a and BSM1b alone are sufficient to repress transcription in M cells indicates that they do not function as promoters for long antisense mRNAs. Although the exact sequence separating BSM1a and BSM1b does not impact on their function, the distance separating them does. BSM1a and BSM1b do not generate preferential accumulation in BS cells when fused together directly, or when separated by 999 nucleotides (Figure S4). However, when the intervening sequence was between 240 and 550 nucleotides preferential accumulation in BS cells occurred (Figure S4). We conclude that in the coding region of *NAD-ME1,* two sequences separated by a spacer are necessary and sufficient to generate strong expression in BS cells.

Although thousands of genes are differentially expressed between M and BS cells of C_4 plants[14,13,12,11], to our knowledge no DNA motifs that determine the patterning of more than one gene in BS cells have been identified. To test whether BSM1a and BSM1b operate more widely to generate preferential expression in BS cells, coding sequence of other genes relevant to C_4 photosynthesis was scanned. Sequences similar to BSM1a and BSM1b were identified in two such genes encoding mitochondrial MALATE DEHYDROGENASE (mMDH) and GLYCOLATE OXIDASE 1 (GOX1). Fragments from *mMDH* and *GOX1* containing each motif were sufficient to drive BS accumulation of GUS (Figure 2A), and when either was deleted preferential accumulation in BS cells was lost (Figure 2A). The identification of BSM1a and BSM1b in these additional genes allowed consensus sequences to be defined (Figure 2B). These data suggest that multiple gene families involved in C_4 photosynthesis and photorespiration have been recruited into BS expression using a regulatory network based on these two motifs.

Sequences similar to BSM1a and BSM1b were identified near the predicted translational start sites of *GgNAD-ME2,* but also in the orthologues *AtNAD-ME1* and *AtNAD-ME2* from C_3 *Arabidopsis thaliana* (Figure 3A). In these additional genes BSM1a is located in the mitochondrial transit peptide and its position varies relative to the translational start site. BSM1b is located in the mature processed protein and its position appears invariant (Figure 3A). When either motif was removed preferential accumulation in BS cells was lost (Figure 3B). Thus, sequences defined by BSM1a and

BSM1b from eight genes (Figure 3C, Figure S6A) are necessary and sufficient to generate BS expression in the C_4 leaf. Although synonymous substitutions to the BSM1a and BSM1b sequences from *GgNAD-ME1* failed to abolish preferential expression in BS cells (Figure S5) amino acid sequence encoded by BSM1a and BSM1b vary considerably for two reasons. First, because BSM1a is found on either DNA strand (Figure S6). Second, in the eight genes studied amino acids encoded by each motif differ because codons are not in identical frames (Figure S6). Combined with the fact that both BSM1a and BSM1b are functional when present in antisense orientation, this variation in amino acid sequence supports the notion that these motifs do not act post-translationally through amino acid sequence, but rather function transcriptionally via transcription factor binding.

To provide orthogonal evidence that BSM1a and BSM1b are indeed the targets for transcription factors binding *in vivo* two additional approaches were pursued. First, the consensus motifs for BSM1a and BSM1b were used to search databases that document transcription factor binding specificities[15,16]. Six members of the MYOBLASTOMA (MYB) family have been reported to bind sequences that are similar but not identical to the BSM1a motif ($p < e^{-3}$, Supplementary Table IV) but there were no clear matches for BSM1b. It is therefore possible that a member of the MYB family binds BSM1a, but the cognate factor binding BSM1b remains unclear. Second, DNaseI sequencing of *G. gynandra* was used to define regions of chromatin that are accessible for transcription factor binding. Genome-wide DNaseI sequencing has previously been used to define the landscape of DNase I hypersensitive sites (DHS) as well as digital genomic footprints (DGF) that mark sequence motifs subject to transcription factor binding[9,16,18,19]. Consistent with the reporter analysis described above, DNaseI analysis showed that BSM1a and BSM1b from *GgNAD-ME1, GgNAD-ME2, GgmMDH* and *GgGOX1* were associated with DHS and so accessible to *trans*-acting factors (Figure S7). For *GgNAD-ME1*, *GgmMDH* and *GgGOX1* these DHS were present in germinating seedlings grown in the dark, whilst for *GgNAD-ME2* the DHS appeared within 4 hours of transfer from dark to light. The four DHS remained in adult leaf tissue (Figure S7). Moreover, BSM1a and BSM1b either overlapped with, or were adjacent to DGF present in these DHS (Figure S7). Whilst DNA sequence bound by a transcription factor can be found in the centre of a DGF, variation in how individual

152 transcription factors bind and contort DNA and thus make sequence available for DNaseI digestion
153 result in different cleavage profiles[20–23]. The similarity of BSM1a to a MYB binding site, the location
154 of both BSM1a and BSM1b in DNA accessible for transcription factor binding, and the presence of
155 DGF in close proximity to each sequence supports the contention that these motifs function as duons
156 and recruit transcription factors that lead to preferential expression in the BS of C_4 *G. gynandra*.
157 In contrast, analysis of publicly available DNAseI data for Arabidopsis[9,20] indicated that for *AtNAD-*
158 *ME1, AtNAD-ME2* or *AtGOX1* BSM1a and BSM1b are not located within DHS (Figure S8). These
159 data are consistent with a model in which BSM1a and BSM1b are present in the ancestral C_3 state,
160 but their relative inaccessibility inhibits binding by cognate transcription factors, and so cell specific
161 gene expression is not achieved. In contrast, in the C_4 leaf BSM1a and BSM1b are accessible to
162 transcription factor binding and this leads to preferential gene expression in BS compared with M
163 cells.
164 To determine the extent to which BSM1a and BSM1b could control gene expression more widely,
165 their distribution across the genome was quantified. More than 4000 genes contained both motifs in
166 *G. gynandra*, but fewer than half of these genes were separated by 35-550 nucleotides required to
167 repress expression in M cells (S Table III). Compared with randomly generated hits of the same size
168 both BSM1a and BSM1b were over-represented in coding sequence DHS (Figure S9), supporting
169 the notion that they carry out a specialised role in transcription factor binding within gene bodies. Of
170 the genes containing BSM1a and BSM1b separated by 35-550 nucleotides, 250 are found in genes
171 that are preferentially expressed in BS cells of *G. gynandra*. We therefore estimate that these duons
172 control BS expression of between four and 250 genes in the C_4 leaf. It was noticeable that BSM1a
173 and BSM1b were also widespread in C_3 *A. thaliana* (S Table III). The simplest explanation for the
174 presence of BSM1a and BSM1b in genes that are not preferentially expressed in BS cells may be
175 that they contain additional *cis*-elements that over-ride their function. The highly combinatorial nature
176 of transcription factor binding and impact that this has on cell-specific gene expression in the root
177 has previously been documented[25]. It is also possible that chromatin environment also influences
178 the function of these motifs and that this can repress their role in restricting expression to BS cells.

We next investigated the extent to which BSM1a and BSM1b are conserved across 1135 wild inbred *A. thaliana* accessions[26]. No Single Nucleotide Polymorphisms (SNP) were detected within either BSM1a or BSM1b (Figure 4A&B). Despite the lack of chromatin accessibility observed in the orthologues from *A. thaliana* (Figure S7) the BSM1a and BSM1b motifs and sequence around them are highly conserved suggesting a role for these duons in C_3 *A. thaliana* that is independent cell preferential expression in the leaf. To determine whether these motifs are also found more widely in *NAD-ME* genes, homologues were retrieved from all sequenced land plants (Figure S10). All dicotyledons contained at least one *NAD-ME* gene carrying sequences that define BSM1a and BSM1b (Figure 4C). In monocotyledons, BSM1a was completely conserved in rice, *Brachypodium* and *Panicum*. Although BSM1b showed one nucleotide substitution in all monocotyledenous genomes available, its conservation in spikemoss and moss argues for it being ancient (Figure 4C, S10). Both BSM1a and BSM1b are highly conserved in *GOX1* and *MDH* genes of land plants, but BSM1b appears more ancient as it is found in *GOX1* genes from all land plants and even chlorophyte algae.

In view of the importance of exonic sequences in regulating BS accumulation of *NAD-ME, MDH* and *GOX1,* the wider importance of exonic regulation in controlling C_4 gene expression was investigated. Seven of twelve core C_4 cycle genes contained regulatory elements located in transcribed sequence that were sufficient to generate preferential accumulation in M or BS cells (Figure 5). Except for *PPCk1*, which encodes the kinase that post-translationally modifies PEPC in M cells of C_4 plants, coding sequences from orthologues of these C_4 genes in C_3 *A. thaliana* led to preferential expression in either M or BS cells of the C_4 leaf (Figure 5). Overall, these data indicate that spatial patterning of gene expression in the C_4 leaf is largely derived from *cis*-regulatory elements present in genes found in the ancestral C_3 state.

Discussion

The data presented here, combined with previous reports[7,27,28] portray an overview of the contribution that untranslated regions (UTRs) and coding sequences make to the generation of cell-specific gene

206 expression in leaves of C_4 *G. gynandra*. Seven of the twelve core C_4 cycle genes possess regulatory
207 elements in their transcript sequences that are sufficient for preferential accumulation in either M or
208 BS cells of the C_4 leaf. These data strongly imply that, in addition to promoters being involved in
209 generating cell-specificity in C_4 leaves[29,30], coding sequences and UTRs play a widespread role in
210 the preferential accumulation of C_4 transcripts to either M or BS cells. It remains to be seen whether
211 this high degree of regulation from genic sequence is critical for the spatial control of gene expression
212 in other tissues and other species. However, as genome-wide studies of transcription factor
213 recognition sites in organisms as diverse as *A. thaliana* and human cells[18,31] have reported that
214 significant binding occurs in genic sequence, we anticipate many more examples of spatial regulation
215 of gene expression being associated with *cis*-elements outside of promoter sequences.

216 The accumulation of *GgNADME1*, *GgNADME2, mMDH* and *GOX1* transcripts in BS cells is
217 dependent on the co-operative function of two *cis*-elements that are separated by a spacer
218 sequence, all of which are located in the first exon of these genes. There are a number of options
219 relating to how such regulatory sequence found within genes could operate. For example, they could
220 impact on transcription by either interacting with a transcription factor or by acting as promoters to
221 drive expression of an antisense RNA. In addition, they could act post-transcriptionally via the RNA
222 molecule once the gene is transcribed. As the BSM1a and BSM1b motifs without additional
223 sequence are sufficient to restrict gene expression to BS, this excludes them acting as promoters to
224 antisense RNA molecules. Furthermore, two pieces of evidence indicate that they do not function
225 post-transcriptionally via the transcripts generated from the endogenous gene. First, in different
226 genes BSM1a is found on each strand of DNA so that it would not be present in all the mRNAs that
227 accumulate preferentially in BS cells. Second, an antisense construct inserted as a transgene into
228 *G. gynandra* still generates BS expression again indicating that the transcribed RNA is not involved.
229 The simplest hypothesis is that BSM1a and BSM1b therefore act transcriptionally to suppress activity
230 of the constitutive 35SCaMV promoter in M cells. The proposal that BSM1a and BSM1b act as duons
231 does not exclude additional levels of regulation contributing to the cell specific accumulation of
232 proteins in the C_4 leaf. Indeed, it is notable that these *cis*-elements alone do not completely restrict

accumulation of GUS to BS cells, and so post-transcriptional and/or post-translational regulation may well act to reinforce their function.

BSM1a and BSM1b are conserved both in their sequences, but also in the number of nucleotides that separates them. In orthologous *NAD-ME* genes from C_3 *A. thaliana,* which diverged from the Cleomaceae ~38 million years ago[32,33], although these motifs are present, they are not sufficient to generate cell preferential expression in the C_3 leaf[7]. This finding indicates that for *NAD-ME* genes to be preferentially expressed in BS cells of C_4 plants, a change in the behaviour of one or more *trans*-factors was a fundamental event. At least in *G. gynandra*, evolution appears to have repeatedly made use of *cis*-elements that exist in genes of C_3 species that are orthologous to those recruited into C_4 photosynthesis[7,27,28,34]. The alteration in *trans*-factors such that they recognise ancestral elements in *cis* in the M or BS therefore appears to be an important and common mechanism associated with evolution of the highly complex C_4 system.

The dual role of exons in protein coding as well as the regulation of gene expression has received significant attention in vertebrates[18,35–38]. Although, 11% of transcription factor binding sites are located in exonic sequence in *A. thaliana*[31], to our knowledge, the identification of BSM1a and BSM1b represents the first functional evidence for *cis*-elements in plant exons. The fact that these motifs are present in C_3 *A. thaliana*, and in fact, also found in the genomes of many land plants and some chlorophyte algae, indicates that these duons play ancient and conserved roles in photosynthetic organisms. The role of such regulatory elements within coding sequences has previously been proposed to be associated with constraints on both protein coding function and codon bias. For example, mutation to these *cis*- elements could be deleterious to both the correct function of the protein, but also to codon usage and so translational efficiency[39–41]. If this is the case, BSM1a and BSM1b could be highly conserved across deep phylogeny because of strong selection pressure on these elements that impact on translation, and this conservation is then co-opted to also regulate transcription during the evolution of C_4 photosynthesis to generate cell-specific gene expression. Establishing the role of BSM1a and BSM1b in C_3 plants would provide insight into the extent to which their role has altered during the transition from C_3 to C_4 photosynthesis.

Duons under strong selection pressure may represent a rich resource of *cis*-elements upon which the C_4 pathway has evolved. Although C_4 photosynthesis is a complex trait that requires multiple changes to gene expression, the repeated evolution of C_4 species across multiple plant lineages suggests that a relatively low number of changes may be required to acquire the C_4 syndrome[42–44]. A single C_4 master switch has been proposed[44] but despite multiple comparative transcriptomic studies[13,45–47], there is as yet no evidence for it. The only transcription factor associated with photosynthesis in C_4 species is Golden-like1[48], which controls expression of genes associated with chlorophyll biosynthesis and the light harvesting complexes in both C_3 and C_4 species[49]. Given the repeated and highly convergent evolution of the C_4 pathway, as well as evidence that separate lineages can arrive at the C_4 state via different routes[50], it appears more plausible that C_4 photosynthesis made use of a number of gene sub-networks. This is now supported by a number of findings. First, just as core photosynthesis genes encoding the light harvesting complexes and Calvin-Benson-Bassham cycle are regulated by light, the vast majority of genes that encode proteins of the C_4 cycle in C_3 *A. thaliana* are also regulated by light signalling, yet, during the evolution of C_4 photosynthesis there was a significant gain of responsiveness to chloroplast signalling[51]. Second, it has been suggested that evolution of the C_4 pathway is associated with the recruitment of developmental motifs into leaves that in C_3 species operate in roots[47]. Lastly, the identification of the *cis*-element MEM2[28], which controls preferential expression of multiple genes in C_4 M cells, and now BSM1a and BSM1b in four different genes that are strongly expressed in BS cells, indicates that that C_4 evolution has made use of small-scale recruitment of gene sub-networks in both cell-types.

In summary, the data indicate regulatory elements located in transcript sequences are of central importance in patterning gene expression in the C_4 leaf. Expression of multiple genes in BS cells is regulated by highly conserved duons that appear to play ancient roles in photosynthetic organisms. Our data also indicate that evolution of the highly C_4 complex trait is built upon ancient *cis*-regulatory architecture that is common to all land plants and some green algae.

Materials and Methods

G. gynandra seed were germinated in the dark at 30°C for 24 h. For microprojectile bombardment seedlings were then transferred to Murashige and Skoog (MS) medium with 1% (w/v) sucrose and 0.8% (w/v) agar (pH 5.8) and grown for a further 13 days in a growth room at 22°C and 200 µmol m^{-2} s^{-1} photon flux density (PFD) with a photoperiod of 16 h light. For DNAseI-seq, germinating seeds were placed in the dark to promote hypocotyl extension for 3 days or transferred a growth cabinet maintained at 25°C±0.5°C, 60% relative humidity, ambient CO_2 and 300 µmol m^{-2} s^{-1} PFD with photoperiod of 16 h light. For DNaseI analysis tissue was flash frozen in liquid nitrogen and stored at -80°C before processing. After bombardment or stable transformation, plant tissue was GUS stained and then chlorophyll removed in 70% (v/v) ethanol.

295 **Figure legends**

296 **Figure 1: Two regions within the coding sequence of *GgNAD-ME1* are necessary for**
297 **preferential gene expression in the bundle sheath.** An antisense construct, as well as a deletion
298 series from the 5' and 3' ends of *GgNAD-ME1* coding sequence were translationally fused to the
299 *uidA* reporter under the control of the CaMV35S promoter **(A)**. Percentage of cells containing GUS
300 in the bundle sheath (BS) after microprojectile bombardment of *G. gynandra* leaves. Bars represent
301 the percentage of stained cells in BS cells, error bars denote the standard error. * represents
302 statistically significant differences with P-values <0.05 and CI = 95% determined by a one-tailed t
303 test (**B**). GUS in *G. gynandra* transformants containing *uidA* fused to 1-240, 1-141, 79-240 and 64-
304 162 base pairs from the translational start site of *GgNAD-ME1* **(C).** Scale bars, 100 µm.
305
306 **Figure 2: BSM1a and BSM1b drive the expression of additional genes in C_4 photosynthesis**
307 **and photorespiration.** Sequences similar to BSM1a and BSM1b were identified in coding
308 sequences of *mMDH* and *GOX1* genes of *G. gynandra*. Deleting the motifs resulted in the loss of
309 preferential accumulation of GUS in the bundle sheath (BS) (**A**). Consensus sequences were defined
310 from motifs operational in *NAD-ME1, mMDH* and *GOX1* (**B**). Error bars denote the standard error. *
311 represents statistically significant differences with P-values <0.05 and CI = 95% determined by a
312 one-tailed t test.
313
314 **Figure 3: Functional versions of BSM1a and BSM1b are present in additional *NAD-MEs.***
315 BSM1a and BSM1b are found in *GgNAD-ME2* and in orthologs of *GgNAD-ME1&2* from the C_3
316 species *A. thaliana* (**A**). Translational fusions carrying these fragments confer bundle sheath (BS)
317 preferential expression in *G. gynandra* leaves. When BSM1a or BSM1b were removed this pattern
318 of GUS was lost (**B**). Consensus sequences derived from all versions of BSM1a and BSM1b tested
319 experimentally (**C**). Error bars denote the standard error. * represents statistically significant
320 differences with P-values <0.05 and CI = 95% determined by a one-tailed t test.
321

Figure 4: BSM1a and BSM1b are highly conserved in land plants. Single nucleotide polymorphisms (SNP) in *AtNAD-ME1* (**A**) and *AtNAD-ME2* (**B**) genes from 1135 wild inbred *A. thaliana* accessions. On the left, the position of BSM1a and BSM1b are highlighted by dashed blue lines, UTRs, exons and introns are denoted by black, grey and white bars respectively on the X-axis. To the right an expanded area representing exon 1, intron 1 and exon 2 is shown, with BSM1a and BSM1b marked within the blue dashed lines. For both genes, no SNP were detected in either motif. The presence of each motifs was investigated in gene sequences of *NAD-ME1, mMDH* and *GOX1* retrieved from 44 species in Phytozome (v10.1)[52]. Each Pie-chart shows the percentage of motif instances that were identical (green), or had 1 base pair (yellow), 2 base pair (orange) substitutions or no similarity (white) detected.

Figure 5: Intragenic regulatory sequences play a major role controlling C_4 photosynthesis genes. Coding sequences encoding core proteins of the C_4 pathway from *G. gynandra* together with orthologs from *A. thaliana* were translationally fused to *uidA* and placed under control of the CaMV35S promoter. After introduction into *G. gynandra* leaves by microprojectile bombardment mesophyll preferential expression of *CA2, CA4, PPDK* and *PPCk1*, together with bundle sheath (BS) preferential expression of *mMDH, NAD-ME1* and *NAD-ME2* were observed (**A**). With the exception of *PPCk* these regulatory elements are conserved in orthogues from *A. thaliana* (**B**). The contribution of intragenic sequences controlling gene regulation of the C_4 pathway is summarized in (**C**), *CA2, CA4, PPDK* and *PPCk1* (blue) and *mMDH, NAD-ME1&2* (red) denote genes where intragenic sequences control cell preferential gene expression. Error bars denote the standard error. * represents statistically significant differences with P-values <0.05 and CI = 95% determined by a one-tailed t test.

Acknowledgements

I-RL was supported by CONACyT and BBSRC grant BB/L014130; SJB by the *3to4* grant from the EU and BB/I002243 from the BBSRC; GR by a Gates Cambridge Trust PhD Fellowship; PS and

SRS by Advanced ERC grant 694733 Revolution to JMH; and BPW by a BBSRC PhD Studentship. We thank Chris Boursnell for help with data retrieval. I-RL designed and undertook experiments to identify BSM1a and BSM1b; I-RL and SJB identified these motifs in MDH and GOX1; SJB undertook analysis of the additional coding regions; GR, PS and SRS performed DNaseI analysis and BPW designed and undertook the antisense experiment. SS made and maintained transgenic plants. I-RL, SJB and JMH wrote the manuscript and prepared the figures.

Competing Interests

The authors have no competing interests.

References

1. Anbar, A. D. *et al.* A Whiff of Oxygen before the Great Oxidation Event? *Source Sci. New Ser. Nat. Sci. Philos. Trans. R. Soc London Ser. B Sci. Rapid Commun. Mass Spectrom S. Ono ai Earth Planet Sei Lett. N. J. Beukes, Sediment Geol Econ. Geol* **317,** 1903–1906 (2007).
2. Bauwe, H., Hagemann, M. & Fernie, A. R. Photorespiration: players, partners and origin. *Trends Plant Sci.* **15,** 330–6 (2010).
3. Hatch, M. D. & Slack, C. R. Photosynthesis by sugar-cane leaves. A new carboxylation reaction and the pathway of sugar formation. *Biochem. J.* **101,** 103–11 (1966).
4. Sage, R. F., Wedin, D. a & Li, M. The Biogeography of C_4 Photosynthesis: Patterns and Controlling Factors. *C_4 Plant Biol.* 313–373 (1999). doi:10.1016/B978-012614440-6/50011-2
5. Sage, R. F. Tansley review: The evolution of C_4 photosynthesis. *New Phytol* **161,** 30 (2004).
6. Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C. & Foley, J. A. Recent patterns of crop yield growth and stagnation. *Nat. Commun.* **3,** 1293 (2012).
7. Brown, N. J. *et al.* Independent and parallel recruitment of preexisting mechanisms underlying C_4 photosynthesis. *Science* **331,** 1436–1439 (2011).
8. Stergachis, A. B. *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342,** 1367–72 (2013).
9. Sullivan, A. M. *et al.* Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell Rep.* **8,** 2015–2030 (2014).
10. Xu, T., Purcell, M., Zucchi, P., Helentjaris, T. & Bogorad, L. TRM1, a YY1-like suppressor of *RbcS-m3* expression in maize mesophyll cells. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 2295–300 (2001).
11. John, C. R., Smith-Unna, R. D., Woodfield, H., Covshoff, S. & Hibberd, J. M. Evolutionary convergence of cell-specific gene expression in independent lineages of C_4 grasses. *Plant Physiol.* **165,** 62–75 (2014).
12. Chang, Y.-M. *et al.* Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. *Plant Physiol.* **160,** 165–77 (2012).
13. Li, P. *et al.* The developmental dynamics of the maize leaf transcriptome. *Nat. Genet.* **42,** 1060–1067 (2010).
14. Aubry, S., Kelly, S., Kümpers, B. & Smith-Unna, R. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C_4 photosynthesis. *PLoS* (2014).
15. Franco-Zorrilla, J. M. *et al.* DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 2367–72 (2014).
16. O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165,** 1280–1292 (2016).
17. Song, L. & Crawford, G. E. DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. doi:10.1101/pdb.prot5384
18. Stergachis, A. B. *et al.* Conservation of *trans*-acting circuitry during mammalian regulatory evolution. *Nature* **515,** 365–370 (2014).
19. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83 (2012).
20. Sung, M.-H., Guertin, M. J., Baek, S. & Hager, G. L. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell* **56,** 275–85 (2014).
21. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41,** e201 (2013).
22. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6,** 283–289 (2009).
23. Kulakovskiy, I. V., Favorov, A. V. & Makeev, V. J. Motif discovery and motif finding from

genome-mapped DNase footprint data. *Bioinformatics* **25,** 2318–2325 (2009).
24. Zhang, T., Marand, A. P. & Jiang, J. PlantDHS: a database for DNase I hypersensitive sites in plants. *Nucleic Acids Res.* **44,** D1148–D1153 (2016).
25. Sparks, E. E. *et al.* Establishment of Expression in the SHORTROOT-SCARECROW Transcriptional Cascade through Opposing Activities of Both Activators and Repressors. *Dev. Cell* **39,** 585–596 (2016).
26. Consortium, 1001 Genomes. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* (2016).
27. Kajala, K. *et al.* Multiple Arabidopsis genes primed for recruitment into C_4 photosynthesis. *Plant J* **69,** 47–56 (2012).
28. Williams, B. P. *et al.* An untranslated *cis*-element regulates the accumulation of multiple C_4 enzymes in *Gynandropsis gynandra* mesophyll cells. *Plant Cell* (2016).
29. Gowik, U. *et al.* cis-Regulatory elements for mesophyll-specific gene expression in the C_4 plant *Flaveria trinervia*, the promoter of the C_4 phospho*enol*pyruvate carboxylase gene. *Plant Cell* **16,** 1077–1090 (2004).
30. Sheen, J. C_4 gene expression. *Ann. Rev Plant Physiol. Plant Mol Biol* **50,** 187–217 (1999).
31. Sullivan, A. M. *et al.* Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. *Cell Rep.* 1–16 (2014). doi:10.1016/j.celrep.2014.08.019
32. Schranz, M. E. & Mitchell-Olds, T. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell* **18,** 1152–1165 (2006).
33. Couvreur, T. L. P. *et al.* Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Mol. Biol. Evol.* **27,** 55–71 (2010).
34. Reyna-Llorens, I. & Hibberd, J. M. Recruitment of pre-existing networks during the evolution of C_4 photosynthesis. *Philos. Trans. R. Soc. London B Biol. Sci.* **372,** (2017).
35. Lang, G., Gombert, W. & Gould, H. A transcriptional regulatory element in the coding 529 sequence of the human *Bcl-2* gene. *Immunology* **114,** 25–36 (2005).
36. Goren, A. *et al.* Comparative Analysis Identifies Exonic Splicing Regulatory Sequences—The Complex Definition of Enhancers and Silencers. *Mol. Cell* **22,** 769–781 (2006).
37. Tümpel, S., Cambronero, F., Sims, C., Krumlauf, R. & Wiedemann, L. M. A regulatory module embedded in the coding region of *Hoxa2* controls expression in rhombomere 2. *Proc. Natl. Acad. Sci.* **105,** 20077–20082 (2008).
38. Dong, X. *et al.* Exonic remnants of whole-genome duplication reveal *cis*-regulatory function of coding exons. *Nucleic Acids Res.* **38,** 1071–1085 (2009).
39. Robinson, M. *et al.* Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic Acids Res.* **12,** 6663–6671 (1984).
40. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 3645–50 (2010).
41. Nakahigashi, K. *et al.* Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo. *BMC Genomics* **15,** 1115 (2014).
42. Sinha, N. R. & Kellogg, E. A. Parallelism and diversity in multiple origins of C_4 photosynthesis in grasses. *Am. J. Bot.* **83,** 1458–1470 (1996).
43. Hibberd, J. M., Sheehy, J. E. & Langdale, J. A. Using C_4 photosynthesis to increase the yield of rice - rationale and feasibility. *Curr. Opin. Plant Biol.* **11,** 228–231 (2008).
44. Westhoff, P. & Gowik, U. Evolution of C_4 Photosynthesis—Looking for the Master Switch. *Plant Physiol.* **154,** 598–601 (2010).
45. Brautigam, A. *et al.* An mRNA blueprint for C_4 photosynthesis derived from comparative transcriptomics of closely related C_3 and C_4 species. *Plant Physiol* (2010).
46. Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D. & Hibberd, J. M. Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C_4 Photosynthesis. *PLoS Genet* **10,** e1004365 (2014).
47. Kulahoglu, C. *et al.* Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C_3 and C_4 Plant Species. *Plant Cell* **26,** 3243–3260

(2014).
48. Rossini, L., Cribb, L., Martin, D. J. & Langdale, J. A. The maize *GOLDEN2* gene defines a novel class of transcriptional regulators in plants. *Plant Cell* **13,** 1231–1244 (2001).
49. Waters, M. T. *et al.* GLK transcription factors coordinate expression of the photosynthetic apparatus in Arabidopsis. *Plant Cell* **21,** 1109–1128 (2009).
50. Williams, B. P., Johnston, I. G., Covshoff, S. & Hibberd, J. M. Phenotypic landscape inference reveals multiple evolutionary paths to C_4 photosynthesis. *Elife* **2,** e00961 (2013).
51. Burgess, S. J. *et al.* Ancestral light and chloroplast regulation form the foundations for C_4 gene expression. *Nat. Plants* **2,** (2016).
52. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40,** D1178-86 (2012).
53. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–5 (2009).
54. Newell, C. A. *et al. Agrobacterium tumefaciens* mediated transformation of *Cleome gynandra* L ., a C_4 dicotyledon that is closely related to *Arabidopsis thaliana*. *J. Exp. Bot.* **61,** 1311–1319 (2010).
55. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9,** 676–682 (2012).
56. Bailey, T. L. *et al.* MEME S UITE : tools for motif discovery and searching. *Conflict* **37,** 202–208 (2009).
57. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. Quantifying similarity between motifs. *Genome Biol.* **8,** R24 (2007).