Evolutionary convergence of cell specific gene expression in independent lineages of C₄ grasses

Christopher R. John, Richard D. Smith-Unna, Helen Woodfield and Julian M. Hibberd§

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

§Corresponding author

One sentence summary: Maize and Setaria have independently recruited syntenic orthologues into the C₄ pathway and transcript abundance in the mesophyll and bundle sheath cells of these species is highly convergent.
Summary

Leaves of almost all C₄ lineages separate the reactions of photosynthesis into the mesophyll (M) and bundle sheath (BS). The extent to which mRNA profiles of M and BS cells from independent C₄ lineages resemble each other is not known. To address this we conducted deep sequencing of RNA isolated from the M and BS of *Setaria viridis* and compared these data with publically available information from maize. This revealed a high correlation ($r = 0.89$) between the relative abundance of transcripts encoding proteins of the core C₄ pathway in M and BS cells in these species, indicating significant convergence in transcript accumulation in these evolutionarily independent C₄ lineages. We also found that the vast majority of genes encoding proteins of the C₄ cycle in *S. viridis* are syntenic to homologues used by maize. In both lineages, 122 and 212 homologous transcription factors were preferentially expressed in the M and BS respectively. Sixteen shared regulators of chloroplast biogenesis were identified, fourteen of which were syntenic homologues in maize and *S. viridis*. In *Sorghum bicolor*, a third C₄ grass, we found that 82% of these trans-factors were also differentially expressed in either M or BS cells. Taken together these data provide the first quantification of convergence in transcript abundance in the M and BS cells from independent lineages of C₄ grasses. Furthermore, the repeated recruitment of syntenic homologues from large gene families strongly implies that parallel evolution of both structural genes and trans-factors underpins the polyphyletic evolution of this highly complex trait in the monocotyledons.
Introduction

C₄ species represent many of the world’s most productive crops (Edwards et al., 2010), and in the tropics and subtropics the C₄ pathway allows increased productivity compared with ancestral C₃ photosynthesis. The increased productivity of C₄ plants is due to their ability to concentrate CO₂ around Ribulose Bisphosphate Carboxylase Oxygenase (RuBisCO) (Hatch et al., 1967), and in the majority of C₄ species this is achieved through spatial compartmentation of the photosynthetic apparatus into mesophyll (M) and bundle sheath (BS) cells (Langdale, 2011). Compared with ancestral C₃ plants, leaf anatomy of C₄ species is altered such that the BS and M compartments are increased and decreased in size respectively. Despite this complexity, C₄ plants are now documented in more than 60 independent lineages of angiosperm (Sage, 2011).

In all C₄ lineages carbonic anhydrase (CA) catalyses the conversion of CO₂ to HCO₃⁻ in M cells. Subsequently, fixation of HCO₃⁻ by phosphoenolpyruvate carboxylase (PEPC) allows C₄ acids to accumulate in the M. In combination with the drawdown of C₄ acids caused by decarboxylation and subsequent reactions of the C₄ cycle this drives the diffusion of C₄ acids from the M to the adjacent BS cells. This increase in CO₂ concentration minimises the oxygenation reaction of RuBisCO and therefore reduces photorespiration. At least three C₄ acid decarboxylases have been recruited in different C₄ lineages to release CO₂ around RuBisCO in BS cells: NAD-dependent malic enzyme (NAD-ME), NADP-dependent malic enzyme (NADP-ME), and phosphoenolpyruvate carboxykinase (PEPCK). To complete the C₄ cycle, phosphoenolpyruvate is regenerated by pyruvate,orthophosphate dikinase (PPDK) in M chloroplasts.

For the two-celled C₄ cycle to operate, enzymes and transporters must be specifically localised in either M or BS cells. This preferential accumulation of proteins is underpinned by transcriptional and post-transcriptional regulation of gene expression, as well as post-translational modification (Hibberd and Covshoff, 2010). While there is considerable diversity in the mechanisms responsible for cell-specific gene expression in C₄ leaves, and modelling predicts that the evolution of C₄ photosynthesis has occurred via distinct routes in various taxa (Williams et al., 2013), it is now clear that there are also examples where the same mechanism has been used by independent C₄ lineages to generate expression in either M or BS cells. For example, the accumulation of NAD-ME
in *Cleome gynandra* and NADP-ME in maize is mediated by a conserved element found within the coding sequence of these genes (Brown et al., 2011). Because this sequence element is present in orthologous genes of C₃ *A. thaliana* and rice, the most parsimonious explanation is that these elements have repeatedly been co-opted from an unknown ancestral function into generating BS specificity in C₄ leaves. Furthermore, the same chromatin marks have been documented on *PEP* and *NADP-ME* genes in M and BS cells of both maize and *Setaria italic* (Heimann et al., 2013), showing that separate C₄ lineages regulate expression of genes in either the M or BS cells via the same mechanisms in *cis*. However, the extent to which the expression of other genes and pathways in these cells resemble each other is not clear, nor is it known whether the same mechanisms in *trans* have been co-opted by multiple C₄ lineages.

Deep sequencing now allows an unbiased analysis of complete mRNA populations. This approach has been used to investigate transcript abundance in M and BS cells from maize (Li et al., 2010; Chang et al., 2012), and led to estimates that in the M between 53 and 78, while in the BS between 102 and 214 transcription factors accumulated preferentially. However, maize represents only one lineage of C₄ plant. We sought to determine the patterns of gene expression in M and BS cells that underpin the photosynthetic reactions in a mature photosynthetic leaf of an independent lineage of C₄ plant, and to define the extent to which they overlap with those in maize. For this, we chose *Setaria viridis*, a weedy relative of domesticated *Setaria italic*. *S. viridis* has an annotated genome sequence and is increasingly being used as a model C₄ grass (Bennetzen et al., 2012; Zhang et al., 2012. While *S. viridis* represents an independent origin of C₄ photosynthesis from maize (Brutnell et al., 2010), both species use NADP-ME as the primary C₄ acid decarboxylase in the BS cells.

In this study, we first defined the mRNA profiles of M and BS cells from *S. viridis*. We combined this information with publicly available databases to examine the extent to which patterns of transcript abundance are convergent in *S. viridis* and maize. We quantify convergence at the mRNA level in M and BS cells of these independent C₄ lineages. This includes structural genes that are known to be required for the C₄ pathway to operate, but also *trans*-factors that previously have not been implicated in the function of M or BS cells.
Results

Rapid isolation of RNA and protein from M cells and BS strands of *Setaria viridis*

To enable quantification of transcript abundance in M and BS cells from fully photosynthetic leaves of *S. viridis* we used leaf rolling to extract M cell contents (Covshoff et al., 2013) followed by mechanical isolation of BS strands (Markelz et al., 2003). Leaf rolling reduced the chlorophyll content of *S. viridis* leaves (Figure 1A&B). After mechanical blending of rolled leaves, BS strands surrounding the veins were visible and very few M cells remained (Figure 1C). Separation of soluble protein from these M and BS extracts followed by immunoblotting indicated that proteins characteristic of the C₄ cycle were partitioned as expected between these two cell types. For example, CA, PEPC, and NADP-MDH were enriched in M cell extracts, while NADP-ME and the LSU of RuBisCO (Figure 1D) were enriched in the BS. As has been observed previously for maize (Majeran et al., 2005) LSU was detectable in M cells of *S. viridis*, although at much lower abundance than in the BS. We extracted RNA from analogous M and BS samples. Electropherograms confirmed that this RNA was good quality (RNA Integrity Numbers ≥ 7.4 and the ratio of 28S rRNA to 18S rRNA was ≥ 1; Supplementary Table I) and quantitative polymerase chain reactions (qPCR) were carried out to determine abundance of transcripts encoding proteins characteristic of the C₄ cycle (Figure 1E).

Deep sequencing of transcripts from M and BS cells of *S. viridis*

To quantify transcript abundance in the M and BS of *S. viridis* we undertook deep sequencing of triplicate M rolled samples and BS strands. Approximately 200 million 91bp paired end reads were obtained (Table I) and after cleaning 97.8% aligned to the genome. Of the 27,045 genes for which transcripts were detected, 9,680 were differentially expressed (p adjusted ≤ 0.05) between M and BS cells (Supplementary File I). Based on homology to known maize C₄ proteins (Li et al., 2010; Chang et al., 2012) and strong cell specific expression we identified putative *S. viridis* C₄ gene families, from which 31 genes were selected for further analysis (Table II and Supplementary File II).

Core C₄ transcripts encoding CA, PEPC, NADP-MDH and PPDK were all at least twenty-fold
more abundant in the M than the BS, while transcripts encoding NADP-ME, PCK, RuBisCO activase, and the SSU of RuBisCO were at least twelve-fold more abundant in BS cells compared with the M (Table II). Consistent with photosynthesis requiring significant amounts of C₄ cycle proteins, for 23 of the 31 C₄ genes, transcript abundance was very high in either the M or BS (Transcripts Per Million - TPM ≥ 1000, Supplementary File II). We conclude that the M rolled samples and BS strands showed patterns of transcript abundance consistent with the C₄ pathway.

We next determined the extent to which *S. viridis* and maize, which represent independent lineages of C₄ grasses, show convergent patterns of transcript abundance in M and BS cells.

**Convergence in C₄ transcript abundance in two independent grass lineages**

We compared the *S. viridis* M and BS mRNA datasets with analogous publicly available data from maize (Chang et al., 2012; Li et al., 2010). Due to higher similarity in both sampling and sequencing procedures, unless explicitly stated, all comparisons were made with data from Chang et al. (2012). Transcripts encoding proteins known to be involved in the C₄ cycle showed very similar compartmentalisation between M and BS cells of these two C₄ grass lineages (Figure 2A). This included enzymes of the core C₄ pathway, transporters that allow flux of metabolites across organelle membranes, and proteins of the Calvin-Benson cycle, which are known to be compartmentalised between the two cell types (Figure 2A). Notably, in addition to the accumulation of transcripts encoding the A subunit of GAPDH (Si010261m.g) in the BS of *S. viridis*, we also detected Si035707m.g encoding the B subunit of GAPDH in the M. It is thought that the BS specific isoform of GAPDH forms the GAPDH-CP12-PRK super-complex (Majeran et al., 2005) that regulates PRK activity (Howard et al., 2011). Consistent with this, and with maize RNA-seq data (Chang et al., 2013; Li et al., 2010), we also found that in *S. viridis* CP12 (Si003343m.g) was BS specific.

Each pair of homologues recruited into the C₄ pathway in *S. viridis* and maize were ranked in terms of mean fold change in transcript abundance within M or BS cells (Figure 2B, Supplementary Table II), but also by the extent to which transcript abundances were convergent (Figure 2C, Supplementary Table III). In the BS, transcripts encoding GDC and FBA were highly enriched...
(Supplementary Table II) and highly convergent in their cell specificity (Figure 2C & Supplementary Table III). However, while transcripts encoding PCK were highly enriched in the BS they were far less convergent, and TKL transcripts were less abundant in BS cells, but highly convergent (Figure 2B&C). CA and NADP-MDH transcripts were the most enriched in the M of both species (Figure 2B). It was noticeable that OMT1 transcripts were highly enriched in M to very similar extents in both maize and S. viridis (Figure 2B&C).

To quantify the convergence in patterns of gene expression between S. viridis and maize, we compared the enrichment in BS versus M cells for pairs of homologous genes in the two species (Figure 3A). For all differentially expressed genes in M and BS cells the Pearson's correlation coefficient (r) was 0.58, while for genes important for the C₄ cycle the correlation coefficient was 0.89. These data indicate a high degree of convergence in the relative abundance of transcripts encoding C₄ cycle proteins between M and BS cells of these species. We also investigated the correlation between S. viridis RNA-seq and maize chloroplast proteomic data (Figure 3B) and found that C₄ genes were more highly correlated than the background. This finding extends previous analysis indicating that the abundance of transcripts encoding components of the core C₄ cycle in M and BS cells of maize were highly correlated (r = 0.95) with their cognate proteins (Li et al., 2010).

Two genes encoding proteins of the Calvin-Benson cycle showed opposite patterns of expression in S. viridis and maize (Figure 2B&C). First, in S. viridis RPI transcripts were strongly preferential to the BS and highly expressed, and this is consistent with the maize RPI protein preferentially accumulating in the BS (Majeran et al., 2005; Friso et al., 2010). However, transcripts from the only strongly expressed (TPM ≥ 100) maize RPI gene (GRMZM5G874903), were weakly M specific. An independent study using laser capture microdissection to isolate RNA also detected RPI transcripts in M cells (Li et al., 2010), and so it would appear likely that in maize strong post-transcriptional or translational control leads to accumulation of RPI in the BS. Second, while in maize both RNA-seq (Figure 2D) and proteomics (Majeran et al., 2005; Friso et al., 2010) indicated enrichment of phosphoglycerate kinase (PGK) in the M, in S. viridis a chloroplast targeted PGK (Si021917m.g) was highly expressed but weakly preferential to the BS (Table II & Figure 2B&C).
As the presence of PGK, GAPDH, and TPI in M cells of C\textsubscript{4} leaves allows balancing of reducing equivalents between M and BS (Majeran et al., 2005) we propose that PGK is under strong post-transcriptional or post-translational control in M cells of \textit{S. viridis}.

Although C\textsubscript{4} plants have been classified into three subtypes depending on the major C\textsubscript{4} acid decarboxylase that they use to provide CO\textsubscript{2} to RuBisCO, it is clear that mixed and flexible decarboxylase systems are common (Furbank et al., 2011). In maize, although NADP-ME is considered the primary C\textsubscript{4} acid decarboxylase, a simultaneous decarboxylation reaction mediated by phosphoenolpyruvate carboxykinase (PCK) (Wingler et al., 1999; Leegood and Walker, 2003; Furbank, 2011) also takes place. Of the two PCK genes in maize, transcripts derived from GRMZM2G001696 (TPM = 11,202) are much more abundant than those from GRMZM5G870932 (TPM = 68). There is only one PCK gene in \textit{S. viridis} (Si034404m.g), and although it is syntenic to the strongly expressed maize isoform, its expression was relatively low (BS TPM = 26). We therefore conclude that \textit{S. viridis} likely operates a minimal PCK decarboxylase pathway.

The light-dependent reactions of photosynthesis can also be compartmented between M and BS cells in C\textsubscript{4} leaves. For example, in NADP-ME subtypes linear and cyclic electron transport occur preferentially in the M and BS respectively (Takabayashi et al., 2005). This is because Photosystem II (PSII) is enriched in the M, while Photosystem I (PSI) and the NDH complex are enriched in the BS (Majeran et al., 2005; Friso et al., 2010; Li et al., 2010). Consistent with these maize proteomics data, but in contrast to RNAseq (Chang et al., 2013), in \textit{S. viridis} transcripts encoding components of PSII were more abundant in M cells, but transcripts encoding components of PSI or proteins allowing cyclic electron transport were more abundant in the BS (Table III). We suggest that differences between the \textit{S. viridis} and maize RNAseq datasets are because the leaf rolling approach is faster than the protoplasting that has previously been used to release M cells (Chang et al. 2013), although it is also possible that differences in growth conditions are responsible.

Many proteins used in the C\textsubscript{4} pathway are co-opted from multi-gene families in C\textsubscript{3} species (Aubry et al., 2011). For proteins that become more abundant in C\textsubscript{4} compared with C\textsubscript{3} leaves, these multi-gene families provide evolution with a rich resource for natural selection to up-regulate
a gene. To investigate the extent to which the same members of multi-gene families have been
recruited into the C₄ pathway in maize and S. viridis we searched a database of syntenic orthologs
within the grasses for genes of the core C₄ cycle, the Calvin-Benson cycle, and associated
transporters (Schnable et al., 2012). All ten of the genes defined as enzymes of the core C₄ cycle
were syntenic orthologs, while three of the six metabolite transporters were syntenic
(Supplementary Table IV). These data imply that for many of the key enzymes that are upregulated
in M or BS cells of these C₄ grasses a specific member of each gene family is repeatedly recruited.
As almost all of these genes belong to sizeable gene families (Supplementary Table IV), these
data indicate strong selective pressure to recruit particular isoforms into C₄ photosynthesis.

Global comparisons of the M and BS transcriptomes from S. viridis and maize

To investigate the extent to which global patterns of transcript abundance were similar in M and
BS cells of S. viridis and maize, we used protein alignments to link differentially expressed genes
in S. viridis to homologues in maize. This resulted in the annotation of 92% of genes whose
transcripts accumulated differentially between M and BS cells of S. viridis with maize homologues.
Of the 9,680 transcripts that accumulated differentially in S. viridis, 5,049 were preferential to the M
while 4,631 were preferential to the BS. In maize, 14,338 transcripts accumulated differentially
between the two cell types, with 6,691 being more abundant in the M and 7,647 upregulated in the
BS. The highly duplicated and complex nature of the maize genome (Schnable et al., 2009) likely
contributes to higher numbers in this species. Of the transcripts that were more abundant in the M
and BS of S. viridis, 1,848 and 1,825 respectively shared homologues that were M or BS specific in
maize (Figure 4A and Supplementary File 3). We therefore detected a higher degree of similarity in
M and BS mRNA profiles between maize and S. viridis than that estimated by two separate studies
on maize (Chang et al., 2012; Li et al., 2010).

We investigated functional enrichment of gene ontology terms within the M and BS cells of both
maize and S. viridis. In maize, of 201 categories, 44 were functionally enriched (FDR = 10%),
whilst in S. viridis, of 197 gene categories, 20 were enriched (FDR = 10%) (Figure 4B and
Supplementary File 4). Although we detected many differences between the two species (Figure 4),
we also found convergence in a small number of functional categories. Ten categories were enriched in both species, and seven of these were enriched within the same cell type (Table IV). In the M we detected convergent enrichment of the secondary metabolism of isoprenoids, protein targeting to the chloroplast, protein synthesis, and RNA processing categories (Figure 4B). In the BS, the TCA cycle, transcription factor, and carbohydrate metabolism categories were over-represented in both species (Figure 4B).

Our data also imply compartmentation of protein-degrading enzymes between the two cell types. In both species the metalloproteases category was enriched in the M (Table V). In maize serine proteases were enriched in the M, while subtilases and AAA-type proteases were more abundant in the BS (Figure 4B). Although not statistically significant, we also note that categories defined as the Calvin-Benson cycle, development, fatty acid lipid degradation, ABC transporters, and amino acid transporters were more represented in the BS, while lipid metabolism, PSII and RNA binding were upregulated in the M of both species (Table V). In summary, the analysis of functional categories indicated shared patterns in transcript abundance within M and BS of two C₄ monocot lineages, but we also note a significant amount of functional diversification (Figure 4B). It is not clear to what extent the similarities in transcript abundance in these cells is associated with their ancestral roles in C₃ species, or specialisations associated with evolution of the C₄ pathway.

Transcription factors underpinning M and BS gene expression in maize, S. viridis and Sorghum bicolor

We identified transcription factors (TFs) using the Mapman (Thimm et al., 2004) and Plant TFDB (Riaño-Pachón et al., 2007) databases. Transcripts encoding four hundred TFs were more abundant in the M than BS of S. viridis, and 122 of these had a direct homologue that accumulated in the M of maize. In addition, of 474 TFs whose transcripts were more abundant in the BS compared with M cells of S. viridis, 212 had a BS specific homologue in maize (Figure 4C).

Within these transcriptional regulators we assessed those with known roles in the regulation of either nuclear or chloroplast encoded photosynthesis genes to identify the extent to which they have been recruited into cell specific roles in independent lineages of C₄ plants. This included
GLK1 and GLK2 transcripts that preferentially accumulated in the M and BS respectively (Table VI). Many of the genes induced by GLK1&2 in A. thaliana (Waters et al., 2009) were differentially expressed between the M and BS of S. viridis (Supplementary Table V). In both S. viridis and maize, transcripts encoding two sigma factors, SIG2 (Si026193m.g and GRMZM2G143392) and SIG3 (Si021619m.g and GRMZM5G830932) were enriched in M and BS cells respectively (Figure 5).

To identify regulatory factors involved in C₄ chloroplast organisation we selected genes that were differentially expressed between M and BS cells of S. viridis and maize that were associated with an A. thaliana GO term for plastid organisation defined by co-expression analysis as well as physical and genetic interactions (GO:0009657, Obayashi et al., 2011). This identified 183 transcripts that encode proteins involved in PSII assembly, maintenance and repair, as well as others involved in the Calvin-Benson Cycle, photorespiration, and cyclic electron transport (Supplementary File 5). A relatively small number of transcriptional regulators were present in this dataset. Because of two apparent gene duplications within S. viridis, we identified fourteen genes from S. viridis that were homologous to twelve loci from maize (Table VI). This included GLK1, GLK2, and SIG2, but also pTAC12, a regulator of plastid transcription (Gao et al., 2011). In S. viridis four homologues of IDD5, an uncharacterised transcription factor in A. thaliana, were highly expressed and M specific, as were two uncharacterised SMAD/FHA transcriptional regulators (Si026826m.g & Si011021m.g). Both the IDD5 and SMAD/FHA regulators were predicted by TargetP to be chloroplast localised (Emanuelsson et al., 2000). We therefore propose that these proteins, along with GLK1&2 contribute to cell specific C₄ plastid differentiation in S. viridis and maize.

Gene expression in the plastid is also under post-transcriptional control (Del Campo et al., 2009). Within the plastid organisation GO term four RNA-binding proteins known or predicted to be chloroplast targeted were identified (Table VI). While two are currently unnamed, the others correspond to ORRM1 (Sun et al., 2013) and CRB (Qi et al., 2012). CRB (Si022373m.g) transcripts preferentially accumulated in BS cells and the protein is known to bind and stabilise chloroplast transcripts including RbcL, PsaA/B, and PsbC/D (Qi et al., 2012).
Many of these transcription factors whose mRNAs accumulated preferentially in either M or BS cells of *S. viridis* and maize belong to large gene families. To investigate the extent to which evolution co-opted genes derived from a common ancestor independently into C₄ photosynthesis we used synteny to determine orthology. Remarkably, this showed that the vast majority of these transcriptional and post-transcriptional regulators in *S. viridis* and maize were syntenic (Supplementary Table VI). We then used this information to define syntenic orthologues to these regulators in a third C₄ grass *Sorghum bicolor*, and using Q-RTPCR investigated the extent to which each was preferentially expressed in M or BS cells of this species. Of the eleven syntenic orthologues in *S. bicolor*, all but two were preferentially expressed in the same cell type as in maize and *S. viridis* (Supplementary Table VII). Overall, these data imply that different lineages of C₄ grass have repeatedly recruited the same trans-factors during evolution.
Discussion

Quantifying \( \text{C}_4 \) transcript convergence between \textit{S. viridis} and maize

\( \text{C}_4 \) photosynthesis is thought to have evolved independently in at least 62 lineages of angiosperms (Sage et al., 2011), and in almost all cases this requires modifications to gene expression in M and BS cells (Hibberd and Covshoff, 2010). Twenty six of these lineages are found in the monocotyledons, with a large cluster being restricted to the Panicoideae (Sage et al., 2011). However, to date patterns of transcript abundance in M and BS cells have only been investigated in the maize lineage (Li et al., 2010; Chang et al., 2013). Using deep sequencing we aimed to initiate an understanding of the extent to which the mRNA profiles of M and BS cells are the same in separate \( \text{C}_4 \) grass lineages. While genes recruited into the \( \text{C}_4 \) pathway showed a very high degree of convergence (Pearson’s \( r = 0.89 \)) in terms of relative transcript abundance in M or BS cells, for all genes with direct homologues in the two species the correlation was relatively low (\( r = 0.58 \)). Of thirty-one \( \text{C}_4 \) genes, only two were expressed in opposite cell types in maize and \textit{S. viridis}. We were also able to quantify the degree to which the abundance of particular transcripts converged in M and BS cells, and this indicated that \textit{GDC} and \textit{CA} were the most convergent in BS and M cells respectively. This is presumably because strong selection pressure leads to very similar levels of gene expression in both cell types of these two species. The extent to which this conservation is ubiquitous in \( \text{C}_4 \) plants will require this type of analysis in many more lineages.

There was also a strong correlation between abundance of transcripts encoding \( \text{C}_4 \) cycle proteins in \textit{S. viridis} and the abundance of those proteins in maize (\( r = 0.81 \)). This implies that these cell specific patterns of transcript accumulation make an important contribution to compartmentation of \( \text{C}_4 \) cycle proteins in both species.

Quantifying convergence in M and BS transcriptomes of \( \text{C}_4 \) grasses

We estimate that of the genes for which transcripts were compartmentalised between M and BS of \textit{S. viridis}, 37% and 39% respectively shared this distribution with a direct homologue in maize. This may represent an upper estimate of cell specific convergence in independent \( \text{C}_4 \) lineages as these species are both Panicoid monocotyledons and belong to the NADP-ME sub-type (Brutnell
et al., 2010; Sage et al., 2011). Previous work has shown that the M of maize plays important roles in protein synthesis, chloroplast protein targeting, secondary metabolism, and RNA processing, while the BS is critical for transport and carbon metabolism (Majeran et al., 2005; Friso et al., 2010; Chang et al., 2012; Li et al., 2010). Many gene categories showed divergent patterns of transcript abundance in *S. viridis* and maize, but transcripts encoding proteins required for chloroplast targeting, isoprenoid metabolism, RNA processing and protein synthesis were up-regulated in the M of both maize and *S. viridis*. Increased representation of transcripts encoding components of the protein synthesis machinery in M cells was largely associated with structural components of the chloroplast ribosomes, which may facilitate synthesis of chloroplast encoded components of PSII with high turnover rates (Majeran et al., 2005). The chloroplast targeting term was associated with up-regulation of translocon components in the M, which is surprising given that majority of the Calvin-Benson cycle is found in the BS in C₄ leaves. It is possible that photosynthesis proteins in the M have faster turn-over rates, and so increased import is required. Over-representation of genes encoding proteins required for isoprenoid synthesis in the M may be related to the presence of PSII in the M. For example PIGMENT DEFECTIVE EMBRYO 181, an enzyme involved in the synthesis of xanthophyll pigments (Josse et al., 2000) was preferentially expressed in M cells. Transcripts derived from eighteen genes encoding RNA-binding proteins targeted to the chloroplast were more abundant in the M, while only two were more abundant in the BS. These genes include members of the RRM/RBD/RNP family involved in RNA-stabilisation, editing, and splicing (Sun et al., 2013).

Previous analysis indicated that transcripts encoding enzymes associated with proteolysis were over-represented in the BS of maize (Li et al., 2010; Chang et al., 2012). Our finding that the TCA cycle and transcription factor categories were over-represented in both species therefore extends our understanding of the C₄ BS. We also report compartmentation of transcripts encoding specific classes of protein degrading enzymes between the two cell types, with serine and metallo-proteases enriched in the M, while subtilases and AAA-type proteases were enriched in the BS. This specialisation in protein degradation machinery between the two cell types may be important for post-translational control of C₄ proteins (Meierhoff et al., 1993; Roth et al., 1996; Brutnell et al.,
Regulators underlying patterns of convergence in C₄ M and BS cells

In maize, the GOLDEN2 (G2) protein regulates accumulation of photosynthesis proteins in the BS (Roth et al. 1996). This role in photosynthesis gene expression of BS cells of C₄ plants appears to represent a spatially more confined version of its function in all photosynthetic cells of C₃ leaves (Waters et al., 2009). For example, in C₃ A. thaliana GOLDEN-LIKE1 and 2 (GLK1 and GLK2) redundantly regulate nuclear encoded photosystem genes including PSAD-O; PSBO-Z; LHCA1-5 and LHCB1-6 (Waters et al., 2009), and the pale leaf phenotype of glk mutants in maize and rice suggests this function is maintained in the monocotyledons (Langdale et al., 1994; Wang et al., 2013). The fact that transcripts homologous to GLK1 and GLK2 in S. viridis accumulate in the M and BS respectively indicates that they likely perform analogous roles to GLK and G2 in maize.

The GLK family of transcription factors is the only one with a confirmed role in maintaining C₄ photosynthesis. Our data indicate that 122 and 212 additional trans-factors were preferentially expressed in the M and BS respectively of maize and S. viridis. We also note that of these trans-factors, 8% and 2% were annotated as fulfilling roles in chloroplast function in the M and BS respectively. Our analysis of S. viridis and maize therefore identifies a small number of regulators whose transcripts preferentially accumulate in either M or BS cells of two independent C₄ grass lineages. Compared with C₃ species it appears that for two of these trans-factors, their targets may have diverged. For example, in A. thaliana SIG3 regulates PsbN (Zghidi et al., 2007), and while SIG3 transcripts in S. viridis and maize were both BS specific we did not detect any cell specificity of PsbN transcripts. Furthermore, ORRM1 is involved in editing transcripts encoding the NDH complex of A. thaliana (Sun et al., 2013). While transcripts encoding the NDH complex accumulated in the BS of S. viridis and maize, ORRM1 transcripts were more abundant in M of all three species, implying that its targets may have altered.

In contrast, of transcripts from S. viridis and maize that accumulate in the same cell types as their known targets in other species, one was a transcription factor, while two are involved in post-transcriptional regulation. Nuclear-encoded sigma factors control chloroplast-encoded genes,
including components of the photosystems (Noordally et al., 2013; Puthiyaveetil et al., 2013; Tsunoyama et al., 2004). SIG2 is thought to regulate PsbA in C₃ A. thaliana (Woodson et al., 2013) and since both transcripts derived from the SIG2 and PsbA genes are enriched in the M of S. viridis and maize we infer that SIG2 drives enrichment of PsbA in both species.

OTP82 and CRB likely play roles in RNA-binding in the BS of maize and S. viridis. In A. thaliana OTP82 edits NDH-B (Hammani et al., 2009), and so the accumulation of NDH transcripts and the OTP82 homologue (Si006246m.g) in the BS of maize and S. viridis is consistent with this function. CRB stabilises rbcL transcripts in C₃ A. thaliana (Qi et al., 2012) and CRB transcripts accumulate in the BS of maize and S. viridis where rbcL transcripts are abundant. The fact that transcripts encoding three of the four RNA-binding proteins implicated in plastid gene regulation were M preferential reflects a trend in S. viridis and maize (Li et al., 2010; Chang et al., 2012) for increased accumulation of mRNAs encoding RNA-binding and RNA-processing proteins in M cells.

Recruitment of syntenic orthologues in C₄ grasses

Phylogenetic reconstructions have led to the inference that specific members of multi-gene families have repeatedly been co-opted into the C₄ cycle and therefore that parallel evolution underlies their recruitment into the C₄ pathway (Christin et al., 2013). Using information on the relative abundance of transcripts in M and BS cells, which is a hallmark of C₄ photosynthesis, as well as synteny (Schnable et al., 2012), we show that a high proportion of genes recruited into the C₄ pathway are syntenic. For example, all ten structural genes of the C₄ cycle and half of the metabolite transporters that are upregulated in either M or BS cells of maize and S. viridis are syntenic. Our analysis supports the proposals of Christin et al. (2013), but we also find that syntenic homologues from the OMT1 and RCA gene families have been recruited into C₄ photosynthesis. We excluded genes encoding alanine aminotransferase and pyrophosphorylase from our analysis because the former is not associated with the NADP-ME pathway used by maize and S. viridis (Furbank et al., 2011), and the latter was not differentially expressed between M and BS cells. As genes are recruited into the C₄ cycle they are up-regulated, but their expression is
also restricted to M or BS cells (Hibberd & Covshoff, 2010). The extent to which parallel evolution
underlies both of these alterations in gene expression (Christin et al., 2013) may differ for each
gene. The ancestral localisation of each protein in M and BS cells of C₃ species will need to be
determined to provide insight into this phenomenon. The high proportion of syntenic orthologues
that are recruited into the C₄ cycle is remarkable and indicates that specific members of multi-gene
copies are more likely to be co-opted into the C₄ pathway than others. The simplest explanation for
repeated recruitment of syntenic orthologues is presumably that they are part of existing gene
regulatory networks in C₃ species that are altered in the same way in C₄ leaves. It is also possible
that the ancestral characteristics of these specific isoforms are more appropriate for a role in C₄
photosynthesis than others (Christin et al., 2013).

Notably, in addition to these structural genes, we also detect strong cell-specific expression of
transcriptional regulators that are both homologous and syntenic in the maize, Setaria, and S. bicolor genomes. The fact that some of these transcription factors belong to families that contain
more than ten genes makes this result compelling. The repeated recruitment of GLK genes from
redundant and constitutive expression in C₃ leaves (Waters et al., 2009) into cell specific function
in C₄ plants indicates parallel evolution of trans-factors. Analysis of expression patterns in the
leaves of ancestral C₃ species will be required to confirm whether additional trans-factors have
undergone parallel evolution as they are recruited into cell specific roles in C₄ plants. If C₄ species
have repeatedly used homologous transcription factors to underpin the patterns of gene
expression required for the C₄ pathway, comparative analysis of multiple C₄ and C₃ lineages
provides an alternative approach to mutant screens and reverse genetics to identify key regulators
of this highly complex trait.

Conclusions
We report highly convergent patterns of transcript abundance in independent lineages of C₄
grasses. The data strongly implicate the recruitment of homologous trans-factors into cell specific
roles in independent groups of C₄ plant, and also provide the first quantitative insight into the
extent of convergence of transcript accumulation in M and BS cells of C₄ leaves. Specific members
of large gene families have been repeatedly recruited into M or BS roles in the C₄ leaf. It will be
interesting to determine the extent to which other C₄ plants have recruited syntenic orthologs into
the pathway, and converged on very similar levels of transcript compartmentation between M and
BS cells.
Materials and Methods

Plant growth, mesophyll and bundle sheath separation, and RNA and protein isolation

*Setaria viridis* was grown in a mixture of 3:1 medium compost to fine vermiculite in a growth chamber. The light cycle was set at 12 hours light, 12 hours dark, with a photon flux density of 200 µmol photons m$^{-1}$ s$^{-1}$, relative humidity at 75% and temperature of 23°C. Seeds were placed directly into soil and after germination, plants watered 1-2 times per week. Seventeen days after sowing, plants were watered and 8 hours into the photoperiod mesophyll cell contents and bundle sheath cells isolated from third leaves. The top and bottom 0.5 cm of these 8 cm leaves were discarded and the mid-rib removed to generate two leaf segments that were subsequently divided into two sections for rolling. Leaf rolling was performed after Covshoff et al. (2013) with the following modification; a glass rod was rolled twice over the surface of each leaf to release the mesophyll cell contents. These were then rapidly collected with a pipette filled with mirVana lysis/binding buffer (Ambion).

To isolate bundle sheath (BS) strands leaves were cut into 2 mm$^2$ segments and placed in isolation buffer (0.33 M Sorbitol, 0.3 M NaCl, 0.01 M NaCl, 0.01 M EGTA, 0.01 M DTT, 0.2 M Tris pH 9.0, 5 mM diethylthiophosphoryl chloride) and then pulsed for 10 seconds three times in a Waring blender on low speed. The suspension was then filtered through a 60 µM mesh, and blending buffer (0.35 M Sorbitol, 5 mM EDTA, 0.05 M Tris pH 8, 0.1% (v/v) beta-mercaptoethanol) used to return the BS material back into the blender. Homogenisation at maximum speed for 1 minute followed by filtering was repeated three times. Purified BS cells were placed on a paper towel stack to remove excess moisture, and then snap frozen in liquid nitrogen prior to RNA isolation. BS tissue was ground to a fine powder in liquid nitrogen and resuspended in 1 ml lysis/binding buffer from the mirVana miRNA isolation kit (Ambion).

RNA was eluted in nuclease-free water and quantity and quality assessed using a 2100 Bioanalyzer (Agilent Technologies). M and BS proteins were extracted using the same mechanical methods used for RNA extraction, but 20mM Na$_2$PO$_4$ pH 7.5 plus protease inhibitors (Roche) were used for resuspension. Samples were centrifuged at 12,000g for 10 minutes prior to the supernatant being removed and then snap frozen in liquid nitrogen. Three replicate samples were
generated for sequencing, each of which was derived from ten plants to provide sufficient RNA for
analysis.

Soluble protein (5.5 µg per lane) was separated by SDS-PAGE (12% (v/v) polyacrylamide) and
transferred to a 0.2 µm nitrocellulose membrane (Biorad). After transfer the membrane was placed
in 5% (w/v) milk powder in wash buffer (0.33 M NaCl, 0.02 M TRIS, 0.3% v/v Tween 20) overnight
at 4°C. This was followed by an one hour incubation with primary antibody (1:1000 dilution, rabbit
anti CA, PEPC, NADP-MDH, NADP-ME, and LSU of RuBisCO) followed by washes, and then one
hour incubation in secondary antibody (anti rabbit IgG peroxidase 1:5000, Sigma) followed by
further washes. Antibodies were gifts from Prof R.C. Leegood (Sheffield, UK) and Prof J.C. Gray
(Cambridge). Western lightning chemiluminescent substrate (PerkinElmer) was applied to the film
to allow visualisation. Growth conditions as well as cell and RNA extractions from Sorghum were
performed as described previously (Covshoff et al. 2013).

Quantitative polymerase chain reactions, deep sequencing, and analysis of gene
expression

For quantitative PCR (Q-PCR) 400ng of RNA were treated with RNase-Free DNase (Promega)
in 10 µl at 37°C for 30 minutes. The reaction was stopped with 1 µl of RQ1 DNase Stop solution at
65°C for 10 minutes. Reverse transcription was performed with Superscript II according to the
manufacturers protocol (Invitrogen). Each reaction was diluted 15 fold upon completion. Q-PCR
was performed using SYBR Green JumpStart Taq ReadyMix (Sigma-Aldrich) with 4 µl of cDNA
and 4 µM primers in each reaction. Relative expression was normalized based on an RNA spike
(Agilent) and primer sequences provided in Supplementary File 6.

While there were two comparable maize data sets available, we compared our data with Chang
et al. (2012) as they were generated using experimental and sequencing procedures similar to
those used in this study. For example, in this study and Chang et al. (2012), whole leaf extractions
were performed and sequencing was performed in triplicate with high depth and long paired-end
reads. In contrast, Li et al. (2010) used only leaf tips, and sequenced two biological replicates with
single-end reads.
RNA-seq libraries were prepared from 1 µg of total RNA (TruSeq RNA sample preparation v2 guide, Illumina). Six libraries (three from each cell type) were sequenced by synthesis with TruSeq v3 chemistry using one lane of the HiSeq 2000 to generate approximately 202 million 91bp paired end reads. Reads for Chang et al. (2012) and Li et al. (2010) were obtained from the Short Read Archive. Reads were quality trimmed and adapters removed using Trimmomatic (Lohse et al., 2012). Latest versions of the genomes for *Setaria italica* (v2.0.18) and *Zea mays* (v3.18) were used from Ensembl plants (http://plants.ensembl.org/) with corresponding annotations. Reads were aligned with TopHat2 (default settings, set to 2 mismatches; Kim et al., 2013) and alignments were then counted to exons with HT-SEQ (Anders, 2011) with mode set to union. Read counts were used as input for DESEQ (Anders and Huber, 2010) for differential expression analysis. Multiple testing correction was by the Benjamini-Hochberg procedure and FDR set to 5%. Counts from HT-SEQ were TPM (transcripts per million) normalised following the method of Li et al. (2010). Raw and normalised data are in Supplementary File 7.

Annotation of genes with homologues was performed using alignments of *Setaria italica* (v2.0.18) and *Zea mays* (v3.18) protein sequences obtained from Ensembl Plants (http://plants.ensembl.org/) using the usearch program (Edgar et al., 2010). *Setaria* and maize C₄ gene families were identified using the Ensembl Plants Biomart service (Kinsella et al., 2011) and target peptides designated by WoLF PSORT (Horton et al., 2007). Genes were annotated using Mapman mappings (Thimm et al., 2004), except for transcription factors that were supplemented with annotations using PlantTFDB (Riaño-Pachón et al., 2007). Functional enrichment was tested using Fisher's exact test in R with Benjamini-Hochberg multiple testing correction with FDR set to 10%.

**Acknowledgements** This work was supported by BBSRC studentships to CRJ and HW, and a PhD studentship from the Millenium Seed Bank to RDSU.
References


Hatch MD, Slack CR, Johnson HS (1967) Further studies on a new pathway of photosynthetic
carbon dioxide fixation in sugar-cane and its occurrence in other plant species. The Biochemical J 102: 417-22


Slack CR, & Hatch MD (1967) Comparative studies on the activity of carboxylases and other
enzymes in relation to the new pathway of photosynthetic carbon dioxide fixation in tropical  
grasses. Biochem J 103: 660-665

RNA recognition motif-containing protein is required for plastid RNA editing in Arabidopsis and  
maize. Proceedings of the National Academy of Sciences 110: E1169-E1178

electron flows around photosystem I for driving CO$_2$-concentration mechanism in C$_4$  
photosynthesis. Proc Nat Acad Sci 102: 16898-16903

driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological  
processes. Plant J 37: 914-939

light-induced transcription of plastid-encoded psbD gene is mediated by a nuclear-encoded  
transcription initiation factor, AtSig5. Proc Nat Acad Sci 9: 3304-3309

function in C$_3$ and C$_4$ plants. Planta 237: 481-495

transcription factors coordinate expression of the photosynthetic apparatus in Arabidopsis. Plant  
Cell 21: 1109-1128

Williams BP, Johnston IG, Covshoff S, Hibberd JM (2013) Phenotypic landscape inference  
reveals multiple evolutionary paths to C$_4$ photosynthesis. eLife 2: e00961

Wingler A, Walker RP, Chen ZH, Leegood RC (1999) Phosphoenolpyruvate carboxykinase is  
involved in the decarboxylation of aspartate in the bundle sheath of maize. Plant Physiol 120: 539-  
546


Accession number for RNA-seq: PRJEB5074
**Figure legends**

**Figure 1. Extraction of protein and RNA from mesophyll and bundle sheath cells of *Setaria viridis*. (A) Representative image of leaf prior to rolling, with alternating bands of mesophyll (arrowhead) and bundle sheath cells. (B) After leaf rolling the chloroplasts within the bundle sheath cells (circle) are visible. (C) Representative image of bundle sheath preparation after blending. (D) Immunoblotting demonstrates that CA, PEPC and NADP-MDH proteins were abundant in mesophyll samples, but not detectable in BS strands. In contrast RuBisCO LSU and NADP-ME were preferentially localised in BS strands. The molecular weight of each protein is annotated to the left of each blot. (E) Quantitative PCR for CA, PEPC, NADP-MDH, RbcS and NADP-ME indicated preferential transcript accumulation in the same cell type as each protein. Scale bars represent 200µm (A&B) and 4µm (C).

**Figure 2. Abundance of transcripts encoding proteins of the C₄ cycle in mesophyll and bundle sheath cells of *S. viridis* and maize. (A) Summary of transcript quantification in mesophyll and bundle sheath cells, components of the Calvin-Benson cycle are shown in the lower panel. Transcripts that are more abundant in the mesophyll are coloured yellow, while those in the bundle sheath are red (scale shown in the heatmap to the right). For each component of the C₄ cycle, quantifications for *S. viridis* and maize are shown on the left and right respectively. (B&C) Log₂foldchange of transcript abundance in bundle sheath and mesophyll cells for all C₄ genes sorted by mean enrichment (high to low) in *S. viridis* and maize (B), or convergence between the two species (C). The top and middle panels represent transcripts that in both species were preferential to BS or M cells respectively, while the bottom panel represents transcripts that showed divergent patterns between the two species. Abbreviations are as follows. AK (adenylate kinase); ASP-AT (aspartate aminotransferase); CA (carbonic anhydrase); DCT2 (dicarboxylate transporter); DIT1 (dicarboxylate transporter); FBA (fructose-bisphosphate aldolase); FBP (fructose-1,6-bisphosphatase); GAPDH (glyceraldehyde phosphate dehydrogenase); GCH (glycine cleavage H-protein); GDC (glycine decarboxylase); GOX (glycolate oxidase); MEP3 (putative
protein/pyruvate symporter; NADP-MDH (NADP malate dehydrogenase); NADP-ME (NADP-malic enzyme); OMT1 (2-oxoglutarate/malate transporter); PCK (phosphoenolpyruvate carboxykinase); PEPC (phosphoenolpyruvate carboxylase); PGK (phosphoglycerate kinase); PPDK (pyruvate, orthophosphate dikinase); PPDK-RP (pyruvate,orthophosphate dikinase regulatory protien); PPT (phosphoenolpyruvate/phosphate translocator); PRK (phosphoribulokinase); RBCACT (ribulose-1,5-bisphosphate carboxylase oxygenase activase); RbcS (ribulose-1,5-bisphosphate carboxylase oxygenase small subunit); RPE (ribulose-phosphate 3 epimerase); RPI (ribose-5P-isomerase); SBP (sedoheptulose-1,7-bisphosphatase); SHMT (serine hydroxymethyltransferase); TLK (transketolase); TPI (triose phosphate isomerase); TPT (triose phosphate/phosphate antiporter).

Figure 3. Convergence in abundance of transcripts and proteins between S. viridis and maize. (A) Relationship between abundance of transcripts in bundle sheath and mesophyll cells of S. viridis and maize. (B) Relationship between abundance of transcripts in bundle sheath and mesophyll cells of S. viridis and chloroplast proteins in maize defined by best BlastP hits (from Friso et al., 2010). All differentially expressed genes are represented in red, while C4 genes are in black. Pearson's correlation co-efficients (r) are shown.

Figure 4. Global convergence in transcript abundance within mesophyll and bundle sheath cells of S. viridis and maize. (A) Venn diagrams show the number of homologous genes from S. viridis and maize that were differentially expressed (FDR = 5%) in the two cell types. (B) Functionally enriched gene categories in Setaria viridis and maize, defined by a Fisher exact test (FDR = 10%). (C) Venn diagrams showing the extent to which transcripts encoding homologous transcription factors accumulate in either mesophyll or bundle sheath cells of S. viridis and maize (FDR = 5%).

Figure 5. Model for the regulation of photosynthesis gene expression in mesophyll and bundle sheath cell chloroplasts of maize and Setaria viridis. Genes previously implicated in the regulation of photosynthesis genes in the nucleus (GLK1 and GLK2) and chloroplast (SIG2)
are depicted with solid arrows. The proposed regulation by SIG3 of PsaA/B genes is shown with a dashed black arrow. Purple and dark green ovals represent the nucleus and chloroplast respectively. Genes highlighted in bold encode components of the light harvesting complexes known to be differentially expressed between mesophyll and bundle sheath cells of maize (Li et al., 2010).
Table I: Summary of sequencing, read processing, mapping, and differential expression analysis.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Read length</td>
<td>91bp</td>
</tr>
<tr>
<td>Read type</td>
<td>Paired</td>
</tr>
<tr>
<td>Replicates</td>
<td>3</td>
</tr>
<tr>
<td>Reads before cleaning</td>
<td>202,048,285</td>
</tr>
<tr>
<td>Reads after cleaning</td>
<td>183,508,816</td>
</tr>
<tr>
<td>Average cleaned reads per library</td>
<td>30,584,803</td>
</tr>
<tr>
<td>Reads aligned</td>
<td>179,478,816</td>
</tr>
<tr>
<td>Read alignment percentage</td>
<td>97.8%</td>
</tr>
<tr>
<td>Average reads aligned per library</td>
<td>29,913,136</td>
</tr>
<tr>
<td>Detected genes</td>
<td>27,045</td>
</tr>
<tr>
<td>Differentially expressed genes</td>
<td>9,680</td>
</tr>
</tbody>
</table>
Table II: Abundance of transcripts encoding enzymes required for the C₄ pathway in mesophyll (M) or bundle sheath (BS) cells of *Setaria viridis*. Each gene recruited into the C₄ pathway in *S. viridis* is highly expressed and preferentially generates transcripts in either M or BS cells. Proteins are listed in groups according to whether they function in the core C₄ pathway, in metabolite transport, the Calvin-Benson cycle, or in photorespiration. Library normalised read counts, fold enrichment in each cell type, and Benjamini-Hochberg procedure adjusted p-values are shown.

<table>
<thead>
<tr>
<th>Gene</th>
<th>ID</th>
<th>Role</th>
<th>Cell type</th>
<th>M</th>
<th>BS</th>
<th>Fold change BS/M</th>
<th>Log₂ Fold change BS/M</th>
<th>P Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>Si003882m.g</td>
<td>Core pathway</td>
<td>M</td>
<td>453421.09</td>
<td>12315.93</td>
<td>0.03</td>
<td>-5.20</td>
<td>5.23E-093</td>
</tr>
<tr>
<td>PEPC</td>
<td>Si005789m.g</td>
<td>Core pathway</td>
<td>M</td>
<td>1413131.82</td>
<td>47109.98</td>
<td>0.03</td>
<td>-4.91</td>
<td>7.31E-060</td>
</tr>
<tr>
<td>NADP-MDH</td>
<td>Si013632m.g</td>
<td>Core pathway</td>
<td>M</td>
<td>108641.70</td>
<td>4588.13</td>
<td>0.04</td>
<td>-4.57</td>
<td>5.21E-076</td>
</tr>
<tr>
<td>PPDK</td>
<td>Si021174m.g</td>
<td>Core pathway</td>
<td>M</td>
<td>1169325.81</td>
<td>57155.39</td>
<td>0.05</td>
<td>-4.35</td>
<td>5.61E-071</td>
</tr>
<tr>
<td>AK</td>
<td>Si014186m.g</td>
<td>Core pathway</td>
<td>M</td>
<td>99257.32</td>
<td>12079.26</td>
<td>0.12</td>
<td>-3.04</td>
<td>1.91E-039</td>
</tr>
<tr>
<td>ASP-AT</td>
<td>Si017156m.g</td>
<td>Core pathway</td>
<td>M</td>
<td>74263.60</td>
<td>10050.17</td>
<td>0.14</td>
<td>-2.89</td>
<td>4.34E-036</td>
</tr>
<tr>
<td>NADP-ME</td>
<td>Si000645m.g</td>
<td>Core pathway</td>
<td>BS</td>
<td>4155.43</td>
<td>613075.40</td>
<td>147.54</td>
<td>7.20</td>
<td>7.82E-084</td>
</tr>
<tr>
<td>RBCS</td>
<td>Si023465m.g</td>
<td>Core pathway</td>
<td>BS</td>
<td>2717.93</td>
<td>290283.67</td>
<td>106.80</td>
<td>6.74</td>
<td>6.79E-103</td>
</tr>
<tr>
<td>PK</td>
<td>Si034040m.g</td>
<td>Core pathway</td>
<td>BS</td>
<td>240.11</td>
<td>3170.30</td>
<td>13.20</td>
<td>3.72</td>
<td>3.18E-051</td>
</tr>
<tr>
<td>RBCACT</td>
<td>Si026402m.g</td>
<td>Core pathway</td>
<td>BS</td>
<td>6245.84</td>
<td>77396.56</td>
<td>12.39</td>
<td>3.63</td>
<td>3.21E-021</td>
</tr>
<tr>
<td>OMT1</td>
<td>Si024404m.g</td>
<td>Transport</td>
<td>M</td>
<td>32208.32</td>
<td>1289.06</td>
<td>0.04</td>
<td>-4.64</td>
<td>2.17E-077</td>
</tr>
<tr>
<td>MEP3b</td>
<td>Si000451m.g</td>
<td>Transport</td>
<td>M</td>
<td>63801.82</td>
<td>10077.02</td>
<td>0.16</td>
<td>-2.66</td>
<td>2.68E-029</td>
</tr>
<tr>
<td>PPT</td>
<td>Si013874m.g</td>
<td>Transport</td>
<td>M</td>
<td>64440.52</td>
<td>17850.85</td>
<td>0.28</td>
<td>-1.85</td>
<td>3.81E-016</td>
</tr>
<tr>
<td>DIT1</td>
<td>Si029415m.g</td>
<td>Transport</td>
<td>M</td>
<td>3703.36</td>
<td>2349.68</td>
<td>0.63</td>
<td>-0.66</td>
<td>9.64E-003</td>
</tr>
<tr>
<td>DCT2</td>
<td>Si035016m.g</td>
<td>Transport</td>
<td>BS</td>
<td>650.50</td>
<td>137421.82</td>
<td>211.25</td>
<td>7.72</td>
<td>3.68E-157</td>
</tr>
<tr>
<td>MEP3a</td>
<td>Si024314m.g</td>
<td>Transport</td>
<td>BS</td>
<td>6245.84</td>
<td>77396.56</td>
<td>12.39</td>
<td>3.63</td>
<td>3.21E-021</td>
</tr>
<tr>
<td>TPI</td>
<td>Si030636m.g</td>
<td>Calvin cycle</td>
<td>M</td>
<td>52985.29</td>
<td>3153.93</td>
<td>0.06</td>
<td>-4.07</td>
<td>1.39E-063</td>
</tr>
<tr>
<td>GAPDH</td>
<td>Si035707m.g</td>
<td>Calvin cycle</td>
<td>M</td>
<td>230052.78</td>
<td>124502.17</td>
<td>0.54</td>
<td>-0.89</td>
<td>1.41E-004</td>
</tr>
<tr>
<td>PRK</td>
<td>Si017390m.g</td>
<td>Calvin cycle</td>
<td>BS</td>
<td>1348.39</td>
<td>210577.18</td>
<td>156.17</td>
<td>7.29</td>
<td>2.86E-065</td>
</tr>
<tr>
<td>FBA</td>
<td>Si010312m.g</td>
<td>Calvin cycle</td>
<td>BS</td>
<td>4611.67</td>
<td>616786.63</td>
<td>133.74</td>
<td>7.06</td>
<td>6.1E-088</td>
</tr>
<tr>
<td>SBP</td>
<td>Si001775m.g</td>
<td>Calvin cycle</td>
<td>BS</td>
<td>1174.47</td>
<td>134922.00</td>
<td>114.88</td>
<td>6.84</td>
<td>1.20E-059</td>
</tr>
<tr>
<td>RPI</td>
<td>Si030780m.g</td>
<td>Calvin cycle</td>
<td>BS</td>
<td>436.32</td>
<td>24891.21</td>
<td>57.05</td>
<td>5.83</td>
<td>2.01E-088</td>
</tr>
<tr>
<td>RPE</td>
<td>Si036994m.g</td>
<td>Calvin cycle</td>
<td>BS</td>
<td>1528.85</td>
<td>70559.89</td>
<td>46.15</td>
<td>5.53</td>
<td>5.86E-035</td>
</tr>
<tr>
<td>FBP</td>
<td>Si035941m.g</td>
<td>Calvin cycle</td>
<td>BS</td>
<td>3018.51</td>
<td>56039.35</td>
<td>18.57</td>
<td>4.21</td>
<td>2.46E-032</td>
</tr>
<tr>
<td>TKL</td>
<td>Si005927m.g</td>
<td>Calvin cycle</td>
<td>BS</td>
<td>24364.05</td>
<td>131028.52</td>
<td>5.38</td>
<td>2.43</td>
<td>8.46E-010</td>
</tr>
<tr>
<td>PGK</td>
<td>Si021917m.g</td>
<td>Calvin cycle</td>
<td>BS</td>
<td>59178.50</td>
<td>87873.79</td>
<td>1.48</td>
<td>0.57</td>
<td>3.05E-002</td>
</tr>
<tr>
<td>Gene</td>
<td>Accession</td>
<td>Function</td>
<td>Symbol</td>
<td>Start</td>
<td>End</td>
<td>Exon</td>
<td>Strand</td>
<td>Log10 P-value</td>
</tr>
<tr>
<td>------</td>
<td>-----------</td>
<td>-------------------</td>
<td>--------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>--------</td>
<td>---------------</td>
</tr>
<tr>
<td>GDC</td>
<td>Si000068m.g</td>
<td>Photorespiration</td>
<td>BS</td>
<td>388.71</td>
<td>74572.76</td>
<td>191.85</td>
<td>7.58</td>
<td>1.48E-103</td>
</tr>
<tr>
<td>GCH</td>
<td>Si011200m.g</td>
<td>Photorespiration</td>
<td>BS</td>
<td>65.68</td>
<td>9987.76</td>
<td>152.06</td>
<td>7.25</td>
<td>3.05E-136</td>
</tr>
<tr>
<td>SHMT</td>
<td>Si035240m.g</td>
<td>Photorespiration</td>
<td>BS</td>
<td>3133.86</td>
<td>60501.53</td>
<td>19.31</td>
<td>4.27</td>
<td>7.07E-067</td>
</tr>
<tr>
<td>GOX</td>
<td>Si040072m.g</td>
<td>Photorespiration</td>
<td>BS</td>
<td>10487.89</td>
<td>23729.01</td>
<td>2.26</td>
<td>1.18</td>
<td>4.71E-007</td>
</tr>
</tbody>
</table>
Table III: Number of genes encoding multi-protein complexes required for the light-dependent reactions of photosynthesis for which differential transcript abundance was detected in mesophyll (M) or bundle sheath (BS) cells of maize and *Setaria viridis*.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Setaria M</th>
<th>Setaria BS</th>
<th>Maize M</th>
<th>Maize BS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSII</td>
<td>23</td>
<td>9</td>
<td>34</td>
<td>4</td>
</tr>
<tr>
<td>PSI</td>
<td>3</td>
<td>10</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>NDH</td>
<td>0</td>
<td>7</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Cyclic e flow</td>
<td>2</td>
<td>12</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>
Table IV: Statistical enrichment of Mapman categories in either mesophyll (M) or bundle sheath (BS) cells of *Setaria viridis* and maize. Within each category, the number of genes that showed differential transcript abundance in each cell type is listed, along with the Benjamini-Hochberg procedure adjusted p-values.

<table>
<thead>
<tr>
<th>Category</th>
<th>Setaria M</th>
<th>BS</th>
<th>P adjusted</th>
<th>Maize M</th>
<th>BS</th>
<th>P adjusted</th>
<th>Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein synthesis</td>
<td>119</td>
<td>37</td>
<td>7.03E-8</td>
<td>443</td>
<td>41</td>
<td>2.23E-099</td>
<td>M</td>
</tr>
<tr>
<td>Protein targeting to chloroplast</td>
<td>19</td>
<td>1</td>
<td>0.002</td>
<td>33</td>
<td>5</td>
<td>7.26E-006</td>
<td>M</td>
</tr>
<tr>
<td>RNA processing</td>
<td>46</td>
<td>20</td>
<td>0.07</td>
<td>119</td>
<td>68</td>
<td>3.59E-005</td>
<td>M</td>
</tr>
<tr>
<td>Secondary metabolism isoprenoids</td>
<td>30</td>
<td>10</td>
<td>0.07</td>
<td>47</td>
<td>11</td>
<td>5.60E-007</td>
<td>M</td>
</tr>
<tr>
<td>CHO metabolism</td>
<td>47</td>
<td>77</td>
<td>0.026</td>
<td>43</td>
<td>125</td>
<td>2.68E-006</td>
<td>BS</td>
</tr>
<tr>
<td>Tricarboxylic acid cycle</td>
<td>11</td>
<td>32</td>
<td>0.01</td>
<td>16</td>
<td>37</td>
<td>0.09</td>
<td>BS</td>
</tr>
<tr>
<td>Transcription factors</td>
<td>400</td>
<td>474</td>
<td>0.004</td>
<td>521</td>
<td>736</td>
<td>0.002</td>
<td>BS</td>
</tr>
</tbody>
</table>
Table V: Enrichment of Mapman categories in either mesophyll (M) or bundle sheath (BS) cells of *Setaria viridis* and maize. These categories missed the statistical cut off, but were clearly enriched in both species. The number of genes in each category is shown.

<table>
<thead>
<tr>
<th>Category</th>
<th>Setaria</th>
<th></th>
<th>Maize</th>
<th></th>
<th>Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>BS</td>
<td>M</td>
<td>BS</td>
</tr>
<tr>
<td>Calvin-Benson cycle</td>
<td>9</td>
<td>25</td>
<td>12</td>
<td>20</td>
<td>BS</td>
</tr>
<tr>
<td>Development</td>
<td>55</td>
<td>92</td>
<td>83</td>
<td>107</td>
<td>BS</td>
</tr>
<tr>
<td>Lipid metabolism degradation</td>
<td>19</td>
<td>32</td>
<td>13</td>
<td>42</td>
<td>BS</td>
</tr>
<tr>
<td>Transport ABC</td>
<td>19</td>
<td>38</td>
<td>24</td>
<td>46</td>
<td>BS</td>
</tr>
<tr>
<td>Transport amino acids</td>
<td>23</td>
<td>32</td>
<td>14</td>
<td>36</td>
<td>BS</td>
</tr>
<tr>
<td>Lipid metabolism fatty acids</td>
<td>29</td>
<td>20</td>
<td>33</td>
<td>24</td>
<td>M</td>
</tr>
<tr>
<td>Photosystem II</td>
<td>22</td>
<td>8</td>
<td>37</td>
<td>4</td>
<td>M</td>
</tr>
<tr>
<td>Protein degradation metalloprotease</td>
<td>11</td>
<td>4</td>
<td>22</td>
<td>12</td>
<td>M</td>
</tr>
<tr>
<td>RNA binding</td>
<td>38</td>
<td>16</td>
<td>79</td>
<td>50</td>
<td>M</td>
</tr>
</tbody>
</table>
Table VI: Abundance of transcripts encoding transcriptional regulators and RNA-binding proteins implicated in chloroplast function in mesophyll (M) and bundle sheath (BS) cells of *S. viridis* and maize. Gene IDs for the homologues in *S. viridis* and maize are provided. In two cases, a gene duplication appears to have occurred in *S. viridis*. Library normalised read counts, fold enrichment in each cell type, and Benjamini-Hochberg procedure adjusted p-values are shown.

<table>
<thead>
<tr>
<th>Setaria ID</th>
<th>Maize ID</th>
<th>Role</th>
<th>Cell type</th>
<th>M</th>
<th>BS</th>
<th>Fold change BS/M</th>
<th>Log₂ Fold change BS/M</th>
<th>P Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unnamed</td>
<td>GRMZM2G177895</td>
<td>Transcription factor</td>
<td>M</td>
<td>3339.75</td>
<td>95.63</td>
<td>0.03</td>
<td>-5.13</td>
<td>8.39E-83</td>
</tr>
<tr>
<td>S011021m.g</td>
<td>GRMZM2G129261</td>
<td>Transcription factor</td>
<td>M</td>
<td>2479.71</td>
<td>155.53</td>
<td>0.06</td>
<td>-3.99</td>
<td>1.41E-58</td>
</tr>
<tr>
<td>ID5</td>
<td>GRMZM2G129261</td>
<td>Transcription factor</td>
<td>M</td>
<td>2933.07</td>
<td>115.56</td>
<td>0.04</td>
<td>-4.66</td>
<td>3.3E-74</td>
</tr>
<tr>
<td>S009633m.g</td>
<td>GRMZM2G042666</td>
<td>Transcription factor</td>
<td>M</td>
<td>3833.30</td>
<td>171.80</td>
<td>0.04</td>
<td>-4.48</td>
<td>2.95E-69</td>
</tr>
<tr>
<td>ID5</td>
<td>GRMZM2G042666</td>
<td>Transcription factor</td>
<td>M</td>
<td>2677.52</td>
<td>97.22</td>
<td>0.04</td>
<td>-4.78</td>
<td>5.33E-74</td>
</tr>
<tr>
<td>Unnamed</td>
<td>GRMZM2G026833</td>
<td>Transcription factor</td>
<td>M</td>
<td>3123.41</td>
<td>150.79</td>
<td>0.05</td>
<td>-4.37</td>
<td>1.14E-65</td>
</tr>
<tr>
<td>PTAC12</td>
<td>GRMZM5G897926</td>
<td>Transcription factor</td>
<td>M</td>
<td>368.73</td>
<td>115.97</td>
<td>0.31</td>
<td>-1.67</td>
<td>3.29E-02</td>
</tr>
<tr>
<td>SIG2</td>
<td>GRMZM2G143392</td>
<td>Transcription factor</td>
<td>M</td>
<td>2714.50</td>
<td>871.80</td>
<td>0.32</td>
<td>-1.47</td>
<td>5.89E-07</td>
</tr>
<tr>
<td>ORRM1</td>
<td>GRMZM2G090271</td>
<td>RNA binding</td>
<td>M</td>
<td>3999.80</td>
<td>840.82</td>
<td>0.21</td>
<td>-2.25</td>
<td>2.18E-13</td>
</tr>
<tr>
<td>Unnamed</td>
<td>GRMZM2G089696</td>
<td>Transcription factor</td>
<td>M</td>
<td>2311.88</td>
<td>1106.47</td>
<td>0.48</td>
<td>-1.06</td>
<td>7.18E-06</td>
</tr>
<tr>
<td>CRB</td>
<td>GRMZM2G165655</td>
<td>RNA binding</td>
<td>BS</td>
<td>768.46</td>
<td>1800.07</td>
<td>2.34</td>
<td>1.23</td>
<td>3.72E-07</td>
</tr>
</tbody>
</table>

For RNA binding proteins, only one example is provided for each category.
**Supplementary Table I:** Summary of bioanalyzer traces of RNA samples from mesophyll and bundle sheath of *Setaria viridis*.

**Supplementary Table II:** Mean fold change values for transcripts encoding C₄ proteins in *Setaria viridis* and maize. Genes are grouped according to cell type and ranked by mean fold change (high to low).

**Supplementary Table III:** Convergence in abundance of transcripts encoding components of the C₄ cycle in *Setaria viridis* and maize. Genes are grouped according to cell type and then ranked according to those that were most similar in the two C₄ lineages.

**Supplementary Table IV:** C₄ genes from maize (Chang et al., 2012; Li et al., 2010) were used to identify syntenic orthologs in *Setaria* using the grass synteny database (Schnable et al., 2012). The majority of genes recruited into the C₄ pathway in maize and *Setaria* were syntenic. For CA, RBCS, GDC, and SBP no syntenic regions were found in the database, however phylogenetic inference has identified orthologous CA genes in *Setaria* and maize (Christin et al., 2013). The number of genes in each gene family from *Setaria* and maize are listed.

**Supplementary Table V:** Differentially expressed *Setaria viridis* genes that are homologous to genes that are induced by expression of GLK1 or GLK2 in *Arabidopsis thaliana* (Waters et al., 2009). Library normalised read counts, fold enrichment in each cell type, and Benjamini-Hochberg procedure adjusted p-values are shown.

**Supplementary Table VI:** Transcriptional regulators and RNA-binding proteins implicated in chloroplast function in mesophyll (M) and bundle sheath (BS) cells of *S. viridis* and maize. Syntenic relationships are derived from Schnable et al., (2012).
**Supplementary Table VII:** Transcripts encoding chloroplast regulators are preferentially abundant in either mesophyll or bundle sheath cells of *Sorghum bicolor*. qPCR data represent mean ΔΔCt values for three biological replicates and the standard deviation is shown. Syntenic relationships are derived from Schnable et al. (2012).
Figure 1. Extraction of protein and RNA from mesophyll and bundle sheath of *Setaria viridis*.

(A) Representative image of leaf prior to rolling, with alternating bands of mesophyll (arrowhead) and bundle sheath cells. (B) After leaf rolling the bands of chloroplasts within the bundle sheath cells (circle) are visible. (C) Representative image of bundle sheath preparation after blending. (D) Immunoblotting demonstrates that CA, PEPC and NADP-MDH proteins were abundant in mesophyll samples, but not detectable in BS strands. In contrast RuBisCO LSU and NADP-ME were preferentially localised in BS strands. The molecular weight of each protein is annotated to the left of each blot. (E) Quantitative PCR for CA, PEPC, NADP-MDH, RbcS and NADP-ME indicated preferential transcript accumulation in the same cell type as each protein. Scale bars represent 200μm (A&B) and 4μm (C).
Figure 2. Abundance of transcripts encoding proteins of the C4 cycle in mesophyll and bundle sheath cells of S. viridis and maize.

(A) Summary of transcript quantification in mesophyll and bundle sheath cells, components of the Calvin-Benson cycle are shown in the lower panel. Transcripts that are more abundant in the mesophyll are coloured yellow, while those in the bundle sheath are red (scale shown in the heatmap to the right). For each component of the C4 cycle, quantifications for S. viridis and maize are shown on the left and right respectively. Log₁₀ change of transcript abundance in bundle sheath and mesophyll cells for all C₄ genes sorted by mean enrichment (high to low) in S. viridis and maize, or convergence between the two species. The top and middle panels represent transcripts that in both species were preferential to BS or M cells respectively, while the bottom panel represents transcripts that showed divergent patterns between the two species. Abbreviations are as follows: AK (adenylate kinase); ASP-AT (aspartate aminotransferase); CA (carbonic anhydrase); DCT2 (dicarboxylate transporter); DIT1 (dicarboxylate transporter); FBA (fructose-bisphosphate aldolase); FBP (fructose-1,6-bisphosphatase); GAP-DH (glyceraldehyde dehydrogenase); GCH (glycine cleavage H-protein); GDC (glycine decarboxylase); GOX (glucose oxidase); MEPS (putative protein/npyruvate synthase); NADP-ME (NADP malate dehydrogenase); OMT1 (2-exoglucuronidase/maleate transporter); PCK (phosphoenolpyruvate carboxykinase); PEPC (phosphoenolpyruvate carboxylase); PGK (phosphoglycerate kinase); PPDK (pyruvate, orthophosphate dikinase regulatory protein); PPT (phosphoenolpyruvate/phosphate translocator); PRK (phosphoribulokinase); RBCACT (ribulose-1,5-bisphosphate carboxylase oxygenase activase); RbcS (ribulose-1,5-bisphosphate carboxylase oxygenase small subunit); RPE (ribulose-phosphate 3 epimerase/mutase; ribulose-5-phosphate 3-synthase); SHMT (serine hydroxymethyltransferase); TDK (transketolase); TPI (triose phosphate isomerase); TPT (triose phosphate/phosphate antiporter).
Figure 3. Convergence in abundance of transcripts and proteins between *S. viridis* and maize.

(A) Relationship between abundance of transcripts in bundle sheath and mesophyll cells of *S. viridis* and maize. 
(B) Relationship between abundance of transcripts in bundle sheath and mesophyll cells of *S. viridis* and chloroplast proteins in maize (from Friso et al., 2010).

(A & B) Homologs were defined by best BlastP hits. All differentially expressed genes are represented in red, while C₄ genes are in black. Pearson’s correlation co-efficients (r) are shown.
Figure 4. Global convergence in transcript abundance within mesophyll and bundle sheath cells of *S. viridis* and maize.

(A) Venn diagrams show the number of homologous genes from *S. viridis* and maize that were differentially expressed (FDR = 5%) in the two cell types. (B) Functionally enriched gene categories in *Setaria viridis* and maize, defined by a Fisher exact test (FDR = 10%). (C) Venn diagrams showing the extent to which transcripts encoding homologous transcription factors accumulate in either mesophyll or bundle sheath cells of *S. viridis* and maize (FDR = 5%).

Downloaded from on July 15, 2017 - Published by www.plantphysiol.org Copyright © 2014 American Society of Plant Biologists. All rights reserved.
Figure 5. Model for the regulation of photosynthesis gene expression in mesophyll and bundle sheath cell chloroplasts of maize and *Setaria viridis*.

Genes previously implicated in the regulation of photosynthesis genes in the nucleus (GLK1 and GLK2) and chloroplast (SIG2) are depicted with solid arrows. The proposed regulation by SIG3 of PsaA/B genes is shown with a dashed black arrow. Purple and dark green ovals represent the nucleus and chloroplast respectively. Genes highlighted in bold encode components of the light harvesting complexes known to be differentially expressed between mesophyll and bundle sheath cells of maize (Li et al., 2010).