A transcription factor binding atlas for photosynthesis in cereals identifies a key role for coding sequence in the regulation of gene expression

Steven J. Burgess[*1], Ivan Reyna-Llorens*[1], Katja Jaeger[2] and Julian M. Hibberd[1]

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

[2]Sainsbury Laboratory, University of Cambridge, 47 Bateman Street, Cambridge, CB2 1LR, UK.

I-RL - suallorens@gmail.com

SJB - sjb287@cam.ac.uk

KJ - katja.jaeger@slcu.cam.ac.uk

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

[*]These authors contributed equally to this work

Keywords: Photosynthesis, gene regulation, evolution, cereals, transcription factor binding sites

Summary

The gene regulatory architecture associated with photosynthesis is poorly understood. Most plants use the ancestral C_3 pathway, but our most productive cereal crops use C_4 photosynthesis. In these C_4 cereals, large-scale alterations to gene expression allow photosynthesis to be partitioned between cell types of the leaf. Here we provide a genome-wide transcription factor binding atlas for grasses that operate either C_3 or C_4 photosynthesis. Most of the >950,000 sites bound by transcription factors are preferentially located in genic sequence rather than promoter regions, and specific families of transcription factors preferentially bind coding sequence. Cell specific patterning of gene expression in C_4 leaves is associated with combinatorial modifications to transcription factor binding despite broadly similar patterns of DNA accessibility between cell types. A small number of DNA motifs bound by transcription factors are conserved across 60 million years of grass evolution, and C_4 evolution has repeatedly co-opted at least one of these hyper-conserved *cis*-elements. The grass cistrome is highly divergent from that of the model plant *Arabidopsis thaliana*.

Introduction

Photosynthesis sets maximum crop yield, but despite millions of years of natural selection is not optimised for either current atmospheric conditions or agricultural practices (Long et al., 2015; Ort et al., 2015). The majority of photosynthetic organisms, including crops of global importance such as wheat, rice and potato use the C_3 photosynthesis pathway in which Ribulose-Bisphosphate Carboxylase Oxygenase (RuBisCO) catalyses the primary fixation of CO_2. However, carboxylation by RuBisCO is competitively inhibited by oxygen binding the active site (Bowes et al., 1971). This oxygenation reaction generates toxic waste-products that are recycled by an energy-demanding series of metabolic reactions known as photorespiration (Bauwe et al., 2010; Tolbert, 1971). The ratio of oxygenation to carboxylation increases with temperature (Jordan and Ogren, 1984; Sharwood et al., 2016) and so losses from photorespiration are particularly high in the tropics. When oxygenation is reduced through CO_2 enrichment, crops show increased photosynthetic efficiency and higher yields (Leakey et al., 2012). In addition to the inefficiency associated with oxygenation by RuBisCO, due to the rapid rise in atmospheric CO_2 concentrations from ~220 to 400ppm, the stoichiometry and kinetics of other photosynthesis enzymes are considered sub-optimal. For example, increased activity of Sedoheptulose 1,7-bisphosphatase improves photosynthesis and yield (Lefebvre et al., 2005; Miyagawa et al., 2001). Furthermore, the ability of leaves to harness light and generate chemical energy is neither optimised for current crop canopy structures (Zhu et al., 2010b) or rapid fluctuations in light (Kromdijk et al., 2016). Thus, although it is clear that improving C_3 photosynthesis would drive increased crop yields, we have an incomplete understanding of the genes that underpin this fundamental process.

The inefficiencies associated with C_3 photosynthesis in the tropics have led to multiple plant lineages evolving mechanisms that suppress oxygenation by concentrating CO_2 around RuBisCO. One such evolutionary strategy is known as C_4 photosynthesis. Species that use the C_4 pathway include maize, sorghum and sugarcane, and they represent the most productive crops on the planet (Sage and Zhu, 2011). In C_4 leaves, additional expenditure of ATP, alterations to leaf anatomy and cellular ultrastructure, as well as spatial separation of photosynthesis between compartments (Hatch, 1987) allows CO_2 concentration to be increased around tenfold compared with that in the atmosphere (Furbank, 2011). Despite the complexity of C_4 photosynthesis, it is

found in over 60 independent plant lineages and so represents one of the most remarkable examples of convergent evolution known to biology (Sage et al., 2011). In most C_4 plants the initial RuBisCO-independent fixation of CO_2 and the subsequent RuBisCO-dependent reactions take place in distinct cell-types known as mesophyll and bundle sheath cells, and so are associated with strict patterning of gene expression between these cell-types. Although the spatial patterning of gene expression is fundamental to C_4 photosynthesis, very few examples of *cis*-elements or *trans*-factors that generate cell-preferential expression required for C_4 photosynthesis have been identified (Brown et al., 2011; Gowik et al., 2004; Williams et al., 2016). In summary, in both C_3 and C_4 species, work has focussed on analysis of mechanisms controlling the expression of individual genes, and so our understanding of the overall landscape associated with photosynthesis gene expression is poor.

In yeast and animal systems, the high sensitivity of open chromatin to DNase-I (Zentner and Henikoff, 2014) has allowed comprehensive, genome-wide characterisation of transcription factor binding sites at single nucleotide resolution (Hesselberth et al., 2009; Neph et al., 2012; Thurman et al., 2012). Despite the power of this approach to define regulatory DNA and the likely transcription factors binding these sequences, this approach has not yet been used to provide insight into the regulatory architecture of photosynthetic leaves of major crops except rice (Zhang et al., 2012b). Moreover, although leaves are composed of multiple distinct cell-types, differences in transcription factor binding between cells have not yet been assessed in plants. By carrying out DNase I-SEQ on grasses that use either C_3 or C_4 photosynthesis, we provide comprehensive insight into the transcription factor binding repertoire associated with each form of photosynthesis. The data indicate that specific cell types from leaf tissue make use of a markedly distinct *cis*-regulatory code, and that transcription factor binding is more frequent within genes than promoter regions. Despite significant conservation in the transcription factors families binding DNA in grasses, it is also apparent that the binding sites they recognise are subject to high rates of mutation. Comparison of sites bound by transcription factors in both C_3 and C_4 leaves demonstrates that the repeated evolution of C_4 photosynthesis is built on combination of *de novo* gain of *cis*-elements and exaptation of highly conserved regulatory elements found in the ancestral C_3 system.

Results

A *cis*-regulatory atlas for monocotyledons

Four grass species were selected to provide insight into the regulatory architecture associated with C_3 and C_4 photosynthesis. *Brachypodium distachyon* uses the ancestral C_3 pathway (Figure 1A). *Sorghum bicolor, Zea mays* and *Setaria italica* all use C_4 photosynthesis although phylogenetic reconstructions indicate that *S. italica* represents an independent evolutionary origin of the C_4 pathway (Figure 1A). Nuclei from leaves of *S. italica* (C_4), *S. bicolor* (C_4), *Z. mays* (C_4) and *B. distachyon* (C_3) were treated with DNase I (Figure S1) and subjected to deep sequencing. A total of 799,135,794 reads could be mapped to the respective genome sequences of these species (Table S1). 159,396 DNase-hypersensitive sites (DHS) of between 150-15,060 base pairs representing broad regulatory regions accessible to transcription factor binding were identified from all four genomes (Figure 1B). Between 20,817 and 27,746 genes were annotated as containing at least one DHS (Table S2). Within these DHS, 533,409 digital genomic footprints (DGF) corresponding to individual transcription factor binding sites of between 11 and 25 base pairs were identified through differential accumulation of reads mapping to positive or negative strands around transcription factor binding sites (Figure 1B&C). At least one transcription factor footprint was identified in >75% of the broader regions defined by DHS (Table S2). In contrast to the preferential binding of transcription factors to AT-rich DNA observed in *A. thaliana* all four grasses DGF had a greater GC content compared with the genome average (Table S3).

DHS and DGF were primarily located in gene-rich regions, and depleted around centromeres (Figure 1D). Individual transcription factor binding sequences were resolved in all chromosomes from each species (Figure 1D). Many genes contained DGF that have previously been associated with specific classes of transcription factors. For example, the *SbPPC* gene (Sobic.010G160700) encoding phospho*enol*pyruvate carboxylase that catalyses the first committed step of C_4 photosynthesis, contained sixteen DGF of which six are associated with known transcription factor families (Figure 1D). On a genome-wide basis, the distribution of DGF was similar between species, with the highest proportion of such sites located in promoter, coding sequence (CDS) and intergenic regions (Figure 1E). However, when normalised to the length of such regions, the density of transcription factor recognition sites was highest in 5' untranslated regions (UTRs),

coding sequences (CDS) and 3' UTRs (Figure 1F). In all four species promoter regions contained fewer DGF than genic sequence (Figure 1F) and distribution plots showed that the density of DGF was highest in exonic sequences downstream of the annotated transcriptional start sites (Figure S2).

A distinct *cis*-regulatory lexicon for specific cells within the leaf

The above analysis provides a genome-wide overview of the *cis*-regulatory architecture associated with photosynthesis in leaves of grasses. However, as with other complex multicellular systems, leaves are composed of many specialised cell types. Because each DGF is defined by fewer sequencing reads mapping to the genome compared with the larger DHS region, depleted signals derived from low abundance cell-types cannot be detected from such tissue-level analysis (Figure 2A). Since bundle sheath strands can be separated easily (Covshoff et al., 2013) leaves of C_4 species provide a simple system to study transcription factor binding in specific cells (Figure 2B). After bundle sheath isolation from *S. bicolor, S. italica* and *Z. mays* a total of 129,137 DHS were identified (Figure 2B; Table S4) containing 452,263 DGF (Figure 2B; Table S4; FDR<0.01). Of the 452,263 DGF identified in bundle sheath strands, 170,114 were statistically enriched in the bundle sheath samples compared with whole leaves (Figure 2B; Table S4). The number of these statistically enriched DGF in bundle sheath strands of C_4 species was large and ranged from 15,880 to 85,256 in maize and *S. italica* respectively (Figure S3). Genome-wide, the number of broad regulatory regions defined by DHS in the bundle sheath that overlapped with those present in whole leaves ranged from 71 to 84% in *S. italica* and *S. bicolor* respectively (Table S5). However, only 8-23% of the narrower DGF found in the bundle sheath were also identified in whole leaves (Table S6). Taken together, these findings indicate that specific cell types of leaves share considerable similarity in the broad regions of DNA that are accessible to transcription factors, but that the short sequences actually bound by transcription factors vary dramatically.

To provide evidence that DGF predicted after analysis of separated bundle sheath strands are of functional importance, they were compared with previously validated *cis*-elements. In C_4 species, there are few such examples, but in the maize *RbcS* gene, which is preferentially expressed in bundle sheath cells, an I-box (GATAAG) is essential for light-mediated activation

(Giuliano et al., 1988) and a HOMO motif (CCTTTTTTCCTT) is important in driving bundle sheath expression (Xu et al., 2001) (Figure 2C). Both elements were detected in our pipeline. Interestingly, the HOMO motif was only bound in the bundle sheath strands (Figure 2C), and whilst the I-box was detected in both bundle sheath strands and whole leaves, the position of the DGF covering it was slightly shifted between the two samples (Figure 2C). Thus, orthogonal evidence for transcription factor binding in maize supports a functional role for DGF identified by DNaseI-SEQ in this study.

To investigate the relationship between cell specific gene expression and the position of DHS and DGF, the DNase-I data were compared with RNA-seq datasets from mesophyll and bundle sheath cells of C_4 leaves (Chang et al., 2012; Emms et al., 2016; John et al., 2014). At least three mechanisms associated with cell specific gene expression operating around individual genes were identified, and can be exemplified using three co-linear genes found on chromosome seven of *S. bicolor*. First, in the NADP-malate dehydrogenase (*MDH)* gene, which is highly expressed in mesophyll cells and encodes a protein of the core C_4 cycle (Figure S4) a broad DHS site was present in whole leaves, but not in bundle sheath strands (Figure 2D). Whilst presence of this site indicates accessibility of DNA to transcription factors that could activate expression in mesophyll cells, global analysis of all genes strongly and preferentially expressed in bundle sheath strands indicates that presence/absence of a DHS in the bundle sheath compared with the whole leaf is not sufficient to generate cell specificity (Figure S5, S6). Second, in the next contiguous gene that encodes an additional isoform of MDH that is also preferentially expressed in mesophyll cells (Figure S4) a DHS was found in both whole leaf and bundle sheath strands but DGF occupancy within this region differed between cell types (Figure 2D). Thus, despite similarity in DNA accessibility, the binding of particular transcription factors was different between cell types. However, once again, genome-wide analysis indicated that alterations to individual DGF were not sufficient to explain cell specific gene expression. For example, fewer than 30% of all enriched DGF in the bundle sheath were associated with differentially expressed genes (Table S7). Lastly, in the third gene in this region, which encodes a NAC domain transcription factor preferentially expressed in bundle sheath strands, differentially enriched DGF were associated both with regions of the gene that have similar DHS in each cell type, but also a region lacking a DHS in whole

leaves compared with bundle sheath strands (Figure 2D). These three classes of alteration to transcription factor accessibility and binding were detectable in genes encoding core components of the C_4 cycle (Figure 2E, Figure S7) implying that a complex mosaic of altered transcription factor binding mediates the cell specific expression found in the C_4 leaf.

Overall, we conclude that differences in transcription factor binding between cells is associated with both DNA accessibility defined by broad DHS, as well as fine-scale alterations to transcription factor binding defined by DGF. In addition, although bundle sheath strands possessed a distinct regulatory landscape compared with the whole leaf, we were unable to identify examples of C_4 genes in which individual transcription factor binding sites differed between bundle sheath and whole leaf samples. This finding implies that cell specific gene expression in C_4 leaves is mediated by a complex mixture of combinatorial effects mediated by alterations to gene accessibility as defined by DHS, but also changes to binding of multiple transcription factors to each C_4 gene.

Transcription factor families associated with cell specific expression

The cistrome, or set of transcription factor binding sites found in a genome, has been determined for *A. thaliana* and to date, consists of 872 experimentally verified motifs linked to 529 transcription factors (O'Malley et al., 2016). Of these 872 motifs from *A. thaliana* 525 could be identified in the *Z. mays*, *S. bicolor, S. italica* and *B. distachyon* datasets (Figure 3A). However, within individual species fewer motifs were detected and so *de novo* prediction was used to identify sequences over-represented in DGF compared with those across the whole genome. This resulted in an additional 524 novel motifs being annotated (Figure 3A). Inspection of these motifs predicted *de novo* demonstrated clear strand bias in DNase-I cuts (Figure 3B) as would expected from *bone fide* transcription factor binding. By combining known and *de novo* motifs, the percentage of DGF that could be annotated in each species increased to more than 41% (Figure 3C). The relatively high number of motifs defined by transcription factor binding sites predicted *de novo* is presumably due to the significant evolutionary time since grasses diverged from *A. thaliana*.

To define the most common motifs actually bound by transcription factors in mature leaves undertaking C_3 and C_4 photosynthesis the frequency of individual motifs was determined and ranked from most to least common in each species. The relative ranking of motifs in the four

grasses was similar (Figure 3D). Visualisation of transcription factor families associated with these DGF in word clouds showed that the most prevalent motifs are associated with the AP2-EREBP and C2H2 transcription factor families (Figure 3D). These findings indicate that across these four grasses the most commonly bound transcription factor motifs are highly conserved. There was much less conservation between transcription factor binding sites in photosynthetic leaves of these monocotyledons compared with the dicotyledon *A. thaliana* (Figure 3D). This finding combined with the large number of motifs from *A. thaliana* not detected in the grasses (Figure 3A) argue for significant divergence in the cistromes of monocotyledons and dicotyledons.

To investigate whether particular classes of transcription factor binding motifs are associated with specific genomic features, the proportion of each motif found in promoter elements, 5' UTRs, coding regions, introns and 3' UTR sequences was defined (Figure S8). In most cases, the distribution of individual motifs was similar in all genomic features, however it was noticeable that a set of motifs was particularly common in coding sequence (Figure S8). Clustering analysis indicated that a set of 96 transcription factor motifs were strongly associated with coding sequences in all four grass species (Figure S9B, S10). The clear strand-bias indicates strong protein-DNA interaction centred on these motifs within coding sequences (Figure S9C). Sequences that carry out a dual role in both coding for amino acids and in transcription factor binding have been termed duons. Thus, in grasses it appears that duons are recognised by a specific set of transcription factors.

To identify regulatory factors associated with gene expression in the C_4 bundle sheath, transcription factor motifs located in DGF enriched in either the bundle sheath or in whole leaf samples of *S. bicolor* were identified (Figure 3E). There was little difference in the ranking of the most prevalent commonly used motifs between these cell types (Figure 3E&F). For example, the AP2-EREBP and C2H2 families were dominant in both bundle sheath and whole leaf samples, indicating that cell-specificity is not associated with large-scale changes in the relative importance of transcription factor binding. However, in terms of prevalence, a small number of transcription factor binding motifs were ranked in the top 50% in whole leaves but the bottom 50% in bundle sheath strands (Figure 3F). This finding implies that quantitative modifications to the use of

particular transcription factor families are associated with the spatial patterning of gene expression that is a hallmark of C_4 photosynthesis.

Further analysis revealed that in all three C_4 species, motifs recognised by C2C2GATA, bZIP, bHLH, BZR and TCP transcription factors were enriched in whole leaf samples, whereas those bound by ARID transcription factors were enriched in the bundle sheath (Figure 3G and Table S9). Moreover, analysis of the cell-specific transcript accumulation of members of the C2C2-GATA family, revealed one orthologue which was consistently mesophyll enriched in all three C_4 species (GRMZM2G379005, Seita.1G358400, Sobic.004G337500; Figure 3H). Thus, these data implicate these transcription factor families in controlling cell-specific gene expression in C_4 leaves, and indicate that in some cases, separate C_4 lineages appear to be using orthologous transcription factors to drive cell specific expression.

Transcription factor binding sites that are conserved but mobile

As *B. distachyon, S. bicolor, Z. mays* and *S. italica* are thought to have diverged from a common ancestor around 60 million years ago (The International Brachypodium Initiative, 2010) they provide an opportunity to examine the extent to which the *cis*-regulatory code has diverged since that point. Furthermore, whilst the last common ancestor of *Z. mays* and *S. bicolor* was thought to use C_4 photosynthesis, *S. italica* belongs to a separate C_4 lineage (Zhang et al., 2012a). Thus, comparative analysis of these species should provide insight into the extent to which the *cis*-regulatory architecture is conserved in the grasses, and how it has been modified during the evolution of C_4 photosynthesis.

In pairwise comparisons of the four species, DGF fell into three categories: those for which homologous sequences were both present and bound by a transcription factor (conserved and occupied), those for which homologous sequences were present but were only bound by a transcription factor in one species (conserved but not occupied) and those for which no sequence homology could be found (not conserved) (Figure 4A). Only a small percentage of DGF were both conserved in sequence and bound by transcription factors (Figure 4B, Table S8). DGF that were conserved but unoccupied were the next most abundant group (Figure 4B) but the majority of DGF

were not conserved (Figure 4B, Table S8). These data indicate substantial turnover in the *cis*-code associated with the transcription factor binding repertoire of monocotyledons.

Consistent with the rapid turnover of DGF documented genome-wide (Figure 4B), the majority of C_4 genes did not share DGF (Table S10 and S11). However, three genes associated with the core C_4 and the Calvin-Benson-Bassham cycle that are strongly expressed in either bundle sheath or mesophyll cells contained the same *cis*-elements bound by a transcription factor in all three C_4 species. For example, in the 2-oxoglutarate/malate transporter (*OMT1*) gene, four sites defined by transcription factor binding were detected in all three C_4 species (Figure 4C). However, the position of these sites within the gene varied in each species. In the transketolase (*TKL*) gene that is preferentially expressed in bundle sheath cells, three conserved motifs defined by transcription factor binding were detected in all C_4 species, but they were also all found in the C_3 species *B. distachyon* (Figure 4D). Thus, in some cases patterning of C_4 gene expression appears linked to pre-existing regulatory architecture operating in the ancestral C_3 state, but in cases where the *cis*-regulatory code associated with C_4 gene expression is strongly conserved the position of these transcription factor binding sites within any gene is variable.

Hyper-conserved *cis*-regulators found in coding sequences of C_4 genes

To investigate the extent to which transcription factor binding sites associated with C_4 genes within a C_4 lineage are conserved, genes encoding the core C_4 cycle were compared in *S. bicolor* and *Z. mays* (Figure 5A). 28 genes associated with the C_4 and Calvin-Benson-Bassham Cycles contained a total of 531 DGF. Although many of these transcription factor footprints were conserved in sequence within orthologous genes, only twenty were both conserved and bound by a transcription factor (Figure 5A). These data therefore indicate that although many *cis*-elements found in orthologous genes of the C_4 cycle are conserved in sequence, a small proportion were bound by a transcription factor at the time of sampling.

Genome-wide, the number of DGF that were conserved in sequence and bound by a transcription factor decayed in a non-linear manner with phylogenetic distance (Figure 5B). For example, *Z. mays* and *S. bicolor* shared 9,446 DGF that were both conserved and occupied. *S. italica* shared only 1,194 DGF with *Z. mays* and *S. bicolor* (Figure 5B). Finally, comparison of these

C_4 grasses with C_3 *B. distachyon* yielded 192 DGF that have been conserved over >60Myr of evolution. 95 of these highly conserved DGF were present in whole leaf samples of the C_3 species, but in the C_4 species were restricted to the bundle sheath (Figure 5B). This set of 192 ancient and highly conserved DGF were located predominantly in 5' UTRs and coding sequence and strikingly, in bundle sheath strands, over fifty percent of these hyper-conserved DGF were in coding sequence (Figure 5B).

One such hyper-conserved DGF is found in the *NdhM* gene that encodes a subunit of the NADH complex that preferentially assembles in bundle sheath cells of C_4 plants (Figure 5C) but it is not known how this evolved. In the ancestral C_3 state a hyper-conserved DGF is found in whole leaves of *B. distachyon* (Figure 5D). However, in all three C_4 species rather than this DGF being detected in whole leaf material, it is detected in the bundle sheath. It is also noticeable that this motif has proliferated within the gene in the C_4 species compared with C_3 *B. distachyon*, and in maize and sorghum is also found in the 5' UTR as well as coding sequence. Furthermore, in whole leaf samples of these C_4 species, transcription factor binding is shifted upstream or downstream (Figure 5D). We therefore propose that preferential expression of *NdhM* in the bundle sheath is built upon a *cis*-regulator present in the C_3 state that activates expression in all photosynthetic cells of the leaf. During the evolution of C_4 photosynthesis, whilst accessibility of this ancient and highly conserved *cis*-element is maintained in the bundle sheath to allow expression of *NdhM,* in mesophyll cells an additional transcription factor(s) binds flanking sequence that blocks access to this pre-existing architecture. These findings are consistent with hyper-conserved DGF located in coding sequence playing an important role in the cell specific gene expression required in leaves of C_4 grasses.

As genome-wide analysis indicated that a specific group of DGF was associated with coding sequence (Figure S8-S10) we investigated whether motifs associated with the 192 hyper-conserved DGF found in all four grasses were over-represented in this set. Remarkably, of the 96 families of transcription factors strongly associated with binding motifs in coding sequence (Figure S10), 47 and 55 were hyper-conserved in the whole leaf and bundle sheath respectively and the ERF family was particularly common (Figure S11, S12). Overall, these data indicate that in these grasses specific families of transcription factors are particularly important in binding coding

sequences, and that the duons bound by these transcription factors are highly conserved across deep evolutionary time.

Discussion

Genome-wide transcription factor binding in grasses

The data presented here provide insight into genome-wide binding of transcription factors in photosynthetic tissue, but also maize and sorghum which represent two of the world's most productive crops. This transcription factor binding landscape shows both similarities and differences with other eukaryotic systems. For example, in contrast with *A. thaliana* in which AT-rich DNA is preferentially bound, the grasses showed preferential binding of transcription factors to GC-rich DNA. Preference for GC-rich DNA has also been observed in humans (Wang et al., 2012) and so the differences in binding likely reflect the relative proportion of nucleotides in each genome. In all these eukaryotes, individual genes are bound by a complex mosaic of transcription factors distributed across major genic feature including promoter regions, UTRs and coding sequence. However, in grasses this standard architecture exemplified by yeast, animals and *A. thaliana* appears to be have been modified such that a much higher proportion of transcription factor footprints are located in exonic and coding regions. For example, in human cells ~3% of transcription factor binding sites are exonic (Stergachis et al., 2013). In contrast, in grass leaves studied here up to 36% and 25% of transcription factor binding sites were located in exonic and coding sequence respectively. This finding is supported by the following observations. First, within individual genes the distribution of transcription factor binding sites peaked after the predicted transcriptional start site. Second, in grasses, strong and reproducible expression of transgenes is routinely achieved by fusing 5' exon and intron sequence to the promoter of interest (Cornejo et al., 1993; Jeon et al., 2000; Maas et al., 1991). Third, although the functional importance of transcription factor binding to coding sequences has been debated in animals (Xing and He, 2015), in grasses these motifs are bound by specific families of transcription factors, and so it is not the case that all transcription factors contribute to this non-random distribution. Moreover, in plants functional analysis has now indicated that duons can control the patterning of gene expression (Reyna-Llorens et al., 2016). Although it is unclear why transcription factor binding in grasses should be particularly prevalent in 5' UTR and coding sequences, these findings combined with the available literature argue for duons and the cognate transcription factors that bind them being of pervasive importance in grass genomes.

364

365 **The transcription factor landscape underpinning gene expression in specific cell types**

366 Given the central importance of cellular compartmentation to C_4 photosynthesis, there have
367 been significant efforts to identify *cis*-elements that restrict gene expression to either mesophyll or
368 bundle sheath cells of C_4 leaves (Hibberd and Covshoff, 2010; Sheen, 1999; Wang et al., 2014).
369 Along with many other systems, initial analysis focussed on regulatory elements located in
370 promoters of C_4 genes (Sheen, 1999) but, it has become increasingly apparent that the patterning
371 of gene expression between cells in the C_4 leaf can be mediated by elements in various parts of a
372 gene. This includes untranslated regions (Kajala et al., 2011; Patel et al., 2004; Viret et al., 1994;
373 Williams et al., 2016; Xu et al., 2001) and coding sequences (Brown et al., 2011; Reyna-Llorens et
374 al., 2016). The genome-wide data reported here provides an unbiased insight into where
375 transcription factors bind C_4 genes, and along with the rest of the genome, indicate that binding is
376 most dense in the 5' UTRs and coding exons.

377 Mechanistically, this DNaseI dataset also indicates that cell specific gene expression in C_4
378 leaves is not strongly correlated with changes to large-scale accessibility of DNA as defined by
379 DHS. This strongly implies that modifications to chromatin density within any one gene do not
380 impact on its expression between cell types. Rather, as only 8-24% of transcription factor binding
381 sites detected in the bundle sheath were also found in whole leaves, the data strongly implicate
382 complex modifications to patterns of transcription factor binding in controlling gene expression
383 between cell types. These findings are consistent with analogous analysis in roots where genes
384 with clear spatial patterns of expression are bound by multiple transcription factors (Sparks et al.,
385 2016) and highly combinatorial interactions between multiple activators and repressors tune the
386 output (de Lucas et al., 2016). However, it is also the case that particular classes of transcription
387 factors including the C2C2GATA, bZIP, bHLH and ARID families are implicated in patterning of
388 gene expression because they were preferentially detected as binding their cognate *cis*-elements
389 in either bundle sheath strands or whole leaves. Our findings therefore strongly imply that the
390 spatial patterning of gene expression in leaves is mediated by a quantitative switch in the
391 abundance of a group of transcription factors.

More generally, the finding that so few transcription factor binding sites were shared between different cell types in leaves of *S. bicolor, Z. mays* and *S. italica* argues strongly for the need to isolate these cells when attempting to understand the control of gene expression. Although separating bundle sheath strands from C_4 leaves is relatively trivial (Covshoff et al., 2013; Furbank et al., 1985; Leegood, 1985) this is not the case for C_3 leaves. Approaches in which nuclei from specific cell-types are labelled with an exogenous tag (Deal and Henikoff, 2011) should now allow their transcription factor landscapes to be defined. In the future, the application of DNase I-SEQ to specific cell types from both C_3 and C_4 leaves should provide insight into how the extent to which gene regulatory networks have been re-wired during the evolution of the complex C_4 trait.

Characteristics of the transcription factor repertoire facilitating evolution of the C_4 pathway

Comparison of transcription factor binding in the C_3 grass *B. distachyon* with three C_4 species provides insight into mechanisms associated with the evolution of C_4 photosynthesis. One striking finding was that in all four species, irrespective of whether they used C_3 or C_4 photosynthesis, the most abundant DNA motifs bound by transcription factors were similar. Thus, motifs recognised by the AP2EREBP, C2C2 and C2C2DOF classes of transcription factor were most commonly bound across each genome. This indicates that during the evolution of C_4 photosynthesis, there has been relatively little alteration to the most abundant classes of transcription factors that bind DNA.

The repeated evolution of the C_4 pathway has frequently been associated with convergent evolution (Sage, 2004; Sage et al., 2012). However, parallel alterations to amino acid and nucleotide sequence that allow altered kinetics of the C_4 enzymes (Christin et al., 2014, 2007) and patterning of C_4 gene expression (Brown et al., 2011) respectively have also been reported. The genome-wide analysis of transcription factor binding now indicates that parallel evolution of transcription factors has contributed to the repeated appearance of C_4 photosynthesis. This is best exemplified by the fact that in the three C_4 species that are derived from two independent C_4 lineages, motifs bound by the ARID and C2C2GATA classes of transcription factor were enriched in bundle sheath and whole leaves respectively. In the case of the C2C2GATA family, transcripts derived from one specific orthologue were more abundant in mesophyll cells of all C_4 species. Thus, within separate lineages of C_4 plant, the same classes of transcription factors have been

recruited into functioning preferentially in one cell type, and in the case of the C2C2GATA family this is associated with orthologous genes being preferentially expressed in mesophyll cells.

When orthologous genes were compared between genomes the majority of transcription factor binding sites were not conserved. Furthermore, of the DGF that were conserved, we found that their position within orthologous genes varied. This indicates that C_4 photosynthesis in grasses is tolerating both a rapid turnover of the *cis*-code, and that when motifs are conserved in sequence, their position and number within a gene can vary. It therefore appears that the cell-specific accumulation patterns of C_4 proteins can be maintained despite considerable modifications to the cistrome of C_4 leaves. This finding is analogous to the situation in yeast where the output of genetic circuits can be maintained despite rapid turnover of regulatory mechanisms underpinning them (Tsong et al., 2006). It was also the case that some conserved motifs bound by transcription factors in the C_4 species were present in *B. distachyon*, which uses the ancestral C_3 pathway. Previous work has shown that *cis*-elements used in C_4 photosynthesis can be found in orthologous genes from C_3 species (Reyna-Llorens et al., 2016; Williams et al., 2016). However, these previous studies identified *cis*-elements that were conserved in both sequence and position. As it is now clear that such conserved sites are mobile within a gene, it seems likely that many more examples of ancient *cis*-elements important in C_4 photosynthesis will be found in C_3 plants.

Although we were able to detect a small number of transcription factor binding sites that were conserved and occupied in all four species that were sampled, these ancient hyper-conserved motifs appear to have played a role in the evolution of C_4 photosynthesis. Interestingly, a large proportion of these motifs bound by transcription factors are found in coding sequence, and this bias was particularly noticeable in bundle sheath cells. Due to the amino acid code, the rate of mutation of coding sequence compared with the genome is restricted. If such regions have a longer half-life than transcription factor binding sites in other regions of the genome, then they may represent an excellent source of raw material for the repeated evolution of complex traits (Martin and Orgogozo, 2013). It remains to be determined why this characteristic is particularly noticeable in bundle sheath cells of C_4 leaves.

In summary, the data presented here provides a transcription factor binding atlas for leaves of grasses using either C_3 or C_4 photosynthesis. Surprisingly, many sequences bound by transcription

450 factors are found within genes rather than promoter regions. Indeed, particular transcription factor
451 families preferentially bind coding sequence and the motifs that they bind are highly conserved in
452 the grasses. Moreover, the canonical patterning of gene expression in C_4 leaves is underpinned by
453 complex combinatorial modifications to transcription factor binding. Lastly, consistent with the deep
454 evolutionary time associated with the divergence of the monocotyledons and dicotyledons, the
455 cistrome of grasses is highly divergent from that of the model plant *Arabidopsis thaliana*.

Figure Legends:

Figure 1: Transcription factor binding atlas for whole leaf samples of four grasses. (A) Schematic of phylogenetic relationship between species analysed. The two independent origins of C_4 photosynthesis are highlighted with black and white circles. (B) Summary of sampling and the total number of DNase I-hypersensitive sites (DHS) and Digital Genomic Footprints (DGF) identified across all four species. (C) TreeView diagrams illustrating cut density around individual digital genomic footprint (DGF). Each row represents an individual DGF, cuts are coloured according to whether they align to the positive (red) or negative (blue) strand and indicate increased cutting in a 50 bp window on either side of the DGF. The total number of DGF per sample is shown at the bottom. (D) Representation of DNase I-SEQ data from *S. bicolor*, depicting gene (grey), DHS (light blue), DGF (orange) and DNase-I cut density (dark blue) at five scales: genome wide, with chromosome number and position indicated (top), chromosomal (second level), kb genomic region (third level), individual *PPC* gene (fourth level) and individual transcription factor binding sites (fifth level). Between each level the expanded area is illustrated. (E) Pie-chart representing the distribution of DGF among genomic features. Promoters are defined as 3000 base pairs (bp) upstream the transcriptional start site while downstream and intergenic features represent regions 1000 and >1000bp downstream of the transcription termination site respectively. (F) Density of DGF per kb in each genomic feature.

Figure 2: Characterisation of the DNA binding landscape in the C_4 Bundle Sheath. (A) Schematic showing that footprints associated with low abundance cells such as the Bundle Sheath (BS) may not be detected from whole leaf (WL) samples. (B) Bundle sheath isolation for DNase I-SEQ experiments, with phylogeny (left) and workflow (right). (C) DGF identified in the maize *ZmRBCS3* gene coincide with I- and HOMO-boxes known to regulate gene expression. The gene model is annotated with whole leaf (blue) and BS (orange) DGF, and the I- and HOMO-boxes are indicated below. (D) Co-linear genes from *S. bicolor* depicting three classes in alterations to DNA accessibility and transcription factor binding to genes that are differentially expressed between whole leaf (blue) and bundle sheath (orange). The C_4 *MDH* contains a DHS and consequently DGF only in the whole leaf, the non-C_4 MDH contains the same DHS but shows variation in DGF

485 between the cell-types, and the NAC transcription factor contains DGF derived from both regions
486 that share DHS, but also one that lack a DHS in the whole leaf sample. (E) Representation of the
487 core C_4 pathway showing differentially accessible DHS, DGF and cell-specific DGF in whole leaf
488 (blue) and bundle sheath (orange) samples of *S. bicolor*. CA; Carbonic Anhydrase, PEPC;
489 Phospho*enol*pyruvate carboxylase, PPDK; Pyruvate,orthophosphate dikinase, MDH; Malate
490 dehydrogenase, NADP-ME; NADP- dependent malic enzyme, RBCS1A; Ribulose bisphosphate
491 carboxylase small subunit 1A, OAA; Oxaloacetate, Mal; Malate, PEP; Phospho*enol*pyruvate, Pyr;
492 Pyruvate, Asp; Aspartate.
493
494 **Figure 3: Distinct cistromes in monocotyledons and dicotyledons.** (A) Number of previously
495 reported motifs as well as those defined *de novo* in the grasses. (B) Density plots depicting
496 average DNase-I activity on positive (red) and negative (blue) strands centred around a *de novo*
497 motif. (C) Bar chart depicting percentage of DGF annotated with known or *de novo* motifs. (D)
498 Comparison of TF motif prevalence in Whole Leaf (WL) samples from *S. italica*, *Z. mays*, *B.*
499 *distachyon* and *A. thaliana* compared with *S. bicolor*. Word clouds depict frequency of motifs
500 associated with transcription factor families, with larger names more abundant. Scatter plots
501 compare frequency of transcription factor motifs within DGF, ranked from low (most abundant) to
502 high (least abundant). Correlation between samples is indicated as Kendell's Tau coefficient (τ).
503 (E) Comparison of transcription factor motif prevalence in BS enriched and whole leaf enriched
504 DGF from *S. bicolor*. Word clouds depict frequency of motifs associated with transcription factor
505 families. (F) Frequency plots of transcription factor motifs in whole leaf or BS samples ranked from
506 low to high. Motifs that ranked in the top 50% most prevalent in one cell type and the bottom 50%
507 in the other are depicted in red. (G) Diagram summarising motifs associated with specific
508 transcription factor families that are over-represented in BS or whole leaves from *S. italica, S.*
509 *bicolor and Z. mays*. The most common transcription factor families are represented. (H)
510 Scatterplots comparing mean transcript per million (TPM) values in bundle sheath and mesophyll
511 cells (data from Chang et al., 2012; Emms et al., 2016; John et al., 2014) for members of the
512 C2C2-GATA family *in S. bicolor, Z. mays* and *S. italica.* Orthologous genes that are consistently
513 enriched in mesophyll cells are highlighted in red.

Figure 4: *Cis*-elements show high rates of turnover and mobility in grasses. (A) Scenarios of DGF conservation between species. Reads derived from DNase-I cuts are depicted by grey, DGF that are both conserved and occupied between species by red, and DGF that are conserved but unoccupied by blue shading. (B) Bar-plot representing pairwise comparisons of DGF occupancy. (C&D) Schematic depicting the position of four transcription factor motifs that are consistently associated with oxaloacetate transporter (*OMT1*) in *S. bicolor, Z. mays* and *S. italica* (C) and three transcription factor motifs that are consistently associated with the bundle sheath enriched gene *TKL* in *S. bicolor, Z. mays*, *S. italica* and C_3 *B. distachyon* (D). The relative position of conserved motifs between orthologous genes is depicted by solid lines (blue, orange, green) and varies between species.

Figure 5: Hyper-conserved *cis*-elements in grasses recruited into C_4 photosynthesis. (A) Conservation of *cis*-regulation in C_4 and Calvin Benson Bassham cycle genes following the divergence of *Z. mays* and *S. bicolor.* The number of carbon atoms (red dots) and metabolite flow (red dashed line) between mesophyll (grey) and bundle sheath (orange) cells are illustrated along with the degree of conservation of DGF associated with BS strands. (B) Conservation of DGF occupancy in grasses across evolutionary time. Results are depicted for whole leaf (WL - blue) and bundle sheath (BS - orange) DGF. Pie-charts display the distribution of conserved and occupied DGF for whole leaf and BS strands. Promoters are defined as 3000bp upstream the transcriptional start site while downstream and intergenic features represent 1000 and >1000bp downstream of the transcription termination site respectively. (C) Evolution of *cis*-regulation in *NdhM* - the BS is enriched in the *Ndh* complex that takes part in cyclic electron flow (CEF). *NdhM* transcript abundance is higher in BS than M cells of C_4 species (data from Chang et al., 2012; Emms et al., 2016; John et al., 2014). (D) DGF in orthologous *Ndh* sequences (grey) and conserved and occupied (red). A single DGF is conserved in all four monocot species. In the ancestral C_3 state this footprint is present in whole leaf samples, but in the derived C_4 state it is occupied in the BS. An additional footprint is present upstream or downstream in whole leaf tissues which may prevent binding in mesophyll cells.

Figure S1: DNase-I digestion of nuclei for sequencing. Representative images of digested samples separated on a 2% (w/v) agarose gels by electrophoresis. (A) *S. bicolor* WL (B) *S. bicolor* BS (C) *Z. mays* WL (D) *Z. mays* BS (E) *B. distachyon* WL (F) *S. italica* WL (G) *S, italica* BS. Each gel represents a separate biological replicate, and the units of DNase-I used are illustrated above. Samples selected for sequencing are indicated in red.

Figure S2: Density plots depicting the distribution of DGF -4000 to +4000 base pairs from the transcriptional start site (TSS) of *S. bicolor, Z. mays, S. italica* and *B. distachyon* whole leaves. In each case, the distribution shows greater transcription factor binding after the transcriptional start site.

Figure S3: Bar chart depicting the number of DGF statistically enriched in bundle sheath cells of *S. bicolor, Z. mays* and *S. italica*.

Figure S4: Transcript abundance expressed as transcripts per million reads (TPM) of three co-linear genes on chromosome seven of sorghum. C_4 *MDH* (Sobic.007G166300.1), non C_4 *MDH* (non C_4, Sobic.007G166200.1), and an uncharacterised *NAC* domain protein (Sobic.007G166100.1) in BS or M cells.

Figure S5: Gene expression in mesophyll and bundle sheath cells associated DHS and DGF in *S. bicolor*. (A) Cell preferential gene expression profiles of highly abundant M and BS genes expressed as transcripts per million reads (TPM). (B) Schematic representing DHS, DGF and DE DGF present in WL (blue) and BS (orange) of *S. bicolor*.

Figure S6: Differential accessibility of broad regulatory regions in *S. bicolor* is not sufficient for cell preferential gene expression. Percentage of differentially detected DHS among BS and M specific genes (n=50) in *S. bicolor* compared against randomly generated gene samples.

Figure S7: Representation of the C_4 pathway showing differentially accessible DHS, DGF and cell specific DGF between whole leaf (blue) and bundle sheath (orange) samples in *S. italica* (A) and *Z. mays* (B). CA; Carbonic Anhydrase, PEPC; Phospho*enol*pyruvate carboxylase, PPDK; Pyruvate, orthophosphate dikinase, MDH; Malate dehydrogenase, NADP-ME; NADP-dependent malic enzyme, RBCS1A; Ribulose bisphosphate carboxylase small subnut1A, OAA; Oxaloacetate, Mal; Malate, PEP; Phospho*enol*pyruvate, Pyr; Pyruvate, Asp; Aspartate.

Figure S8: Proportion of motifs from whole leaf DGF in each genomic feature of *S. bicolor*, *Z. mays, S. italica* and *B. distachyon*. The majority of motifs are distributed across all genomic features, and so the proportion associated with any one feature is relatively low. However, a set of motifs are enriched in coding sequences.

Figure S9: A group of 96 motifs are preferentially associated with coding sequences in grasses. (A) Proportion of all annotated motifs in coding sequence for *S. bicolor* compared with *Z. mays, S. italica* and *B. distachyon*. (B) Heatmap showing the association of motifs with each genomic feature. The horizontal bar at the top represents five clusters obtained by k-means. Hierarchical clustering was used to group genomic features in *S. bicolor* (Sb)*, Z. mays* (Zm)*, S. italica* (Si) and *B. distachyon* (Bd). (C) Representative motifs associated with coding sequences in *S. bicolor* (top). TreeView diagrams illustrating cut density around representative motifs in coding sequences (middle). Each row represents an individual DGF, cuts are coloured according to strand alignment. The number of motif hits in coding sequence are listed in the left. Density plots depicting average DNase I activity on positive (red) and negative (blue) strands centred around motifs are shown at the bottom.

Figure S10: Distribution within genomic features of 96 motifs associated with DGF in coding sequences from whole leaves of *S. bicolor, Z. mays, S. italica* and *B. distachyon*.

598 **Figure S11:** Distribution within genomic features of 47 motifs associated with DGF in coding
599 sequences and found within hyper-conserved DGF (Figure 5) from whole leaves of *S. bicolor, Z.*
600 *mays, S. italica* and *B. distachyon*.

601

602 **Figure S12:** Distribution within genomic features of 55 motifs associated with DGF in coding
603 sequences and found within hyper-conserved DGF (Figure 5) from bundle sheath strands of *S.*
604 *bicolor, Z. mays, S. italica* and *B. distachyon*.

605

606 **Table S1: Summary of DNAse-SEQ Quality Metrics**. Values were calculated according to (Landt
607 et al., 2012) and include the number of reads mapped following filtering (NMAP), PCR bottleneck
608 Coefficient (PCB), Normalized Strand Cross-correlation Coefficient (NSC), Relative Strand Cross-
609 correlation Coefficient (RSC) the optimal number of peaks calculated by IDR method (PEAKS),
610 and the Signal Portion Of Tags (SPOT) for WL and BS samples from *B. distachyon*, *S. italica*, *S.*
611 *bicolor* and *Z. mays*.
612
613 **Table S2: Summary Statistics for Genomic Features Identified in Whole Leaf Samples.**
614 Including information about the number of DHS and DGF identified per sample, and the number of
615 genes that could be annotated with at least one genomic feature for *B. distachyon*, *S. italica*, *S.*
616 *bicolor* and *Z. mays*.
617
618 **Table S3: Base-pair percentages in DGFs and the genome of *S. bicolor*, *Z. mays*, *S. italica***
619 **and *B. distachyon.***
620
621 **Table S4: Summary Statistics for Genomic Features Identified in Bundle Sheath Samples.**
622 Including information about the number of DHS and DGF identified per sample, and the number of
623 genes that could be annotated with at least one genomic feature for *B. distachyon*, *S. italica*, *S.*
624 *bicolor* and *Z. mays*.
625
626 **Table S5: Summary Statistics of Overlap between DHS in Whole Leaf and Bundle Sheath**
627 **Samples.**
628
629 **Table S6: Summary Statistics of Overlap between DGF in Whole Leaf and Bundle Sheath**
630 **Samples.**
631
632 **Table S7: Summary Statistics for Differential Digital Genomic Footprint Calling.** Including the
633 total number of differential DGF (DE DGF) and the number of DE DGF in DE genes for *S. italica*,
634 *S. bicolor* and *Z. mays*.

635

636 **Table S8: Statistics for Cross Mapping of genomic features between *S. bicolor*, *S. italica*, *Z.***
637 ***mays* and *B. distachyon*.**

638

639 **Table S9: TF family motifs enriched (FDR<0.05) in differential footprints from *S.***
640 ***bicolor, Z. mays* and *S. italica* WL and BS samples.**

641

642 **Table S10: Identifiers of orthologous C_4 and CBB cycle related genes from *S. bicolor*, *S.***
643 ***italica*, *Z. mays* and *B. distachyon*.**

644

645 **Table S11: C_4 and CCB genes associated with the same known or *de novo* transcription**
646 **factor binding motifs in *Z. mays*, *S. bicolor*, *S. italica* and *B. distachyon*.**

Methods

Growth conditions and isolation of nuclei

S. bicolor, *S. italica* and *Z. mays* were grown under controlled conditions at the University of Cambridge in a chamber set to 12h/12h light/dark; 28°C light/20°C dark; 400µmol m^{-2} s^{-1} photon flux density, 60% humidity. For germination, *S. bicolor* and *Z. mays* seeds were imbibed in dH_2O for 48h, *S. italica* seeds were incubated on wet filter paper at 30°C overnight in the dark. *Z. mays*, *S. bicolor* and *S. italica* were grown on 3:1 (v/v) M3 compost to medium vermiculite mixture, with a thin covering of soil. Seedlings were hand-watered. For *B. distachyon* plants were grown under controlled conditions at the Sainsbury Laboratory Cambridge University, first under short day conditions 14h/10h, light/dark for 2 weeks and then shifted to long day 20h/4h, light/dark, for 1 week and harvested at ZT20. Temperature was set at 20°C, humidity 65% and light intensity 350µmol m^{-2} s^{-1}.

To isolate nuclei from *S. bicolor*, *Z. mays* and *S. italica* mature third and fourth leaves with a fully developed ligule were harvested 4-6 h into the light cycle on the 18th day after germination. Bundle sheath cells were mechanically isolated (Covshoff et al., 2013). At least 3 g of tissue was used for each extraction. Nuclei were isolated using a sucrose gradient adapted from (Gendrel et al., 2005) and the amount of nuclei in each preparation was quantified using a haemocytometer. For *B. distachyon* plants were flash frozen and material pulverised in a coffee grinder. 3g of plant material was added to 45 ml NIB buffer (10mM Tris-HCl, 0.2M sucrose, 0.01% (v/v) Triton X-100, pH 5.3 containing protease inhibitors (SIGMA)) and incubated at 4°C on a rotating wheel for 5 min, afterwards debris was removed by sieving through 2 layers of Miracloth (millipore) into pre-cooled flasks. Nuclei were spun down 4,000rpm, 4°C for 20 min. Plastids were lysed by adding Triton to a final concentration of 0.3% (v/v) and incubated for 15 min on ice. Nuclei were pelleted by centrifugation at 5000 rpm at 4°C for 15 min. Pellets were washed 3 times with chilled NIB buffer.

DNAse-I digestion, sequencing and library preparation

To obtain sufficient DNA each biological replicate consisted of leaves from tens of individuals and to conform to standards set by the Human Genome project at least two biological replicates

were sequenced for each sample. 2 x 10^8 of freshly extracted nuclei were re-suspended at 4°C in digestion buffer (15 mM Tris-HCl, 90 mM NaCl, 60 mM KCl, 6 mM $CaCl_2$, 0.5 mM spermidine, 1 mM EDTA and 0.5 mM EGTA, pH 8.0). DNAse-I (Fermentas) at 7.5 U was added to each tube and incubated at 37 °C for 3 min. Digestion was arrested with addition of 1:1 volume of stop buffer (50 mM Tris-HCl, 100 mM NaCl, 0.1% (w/v) SDS, 100 mM EDTA, pH 8.0, 1 mM Spermidine, 0.3mM Spermine, RNase A 40 µg/ml) and incubated at 55°C for 15 min. 50 U of Proteinase K was added and samples incubated at 55 °C for 1 h. DNA was isolated with 25:24:1 Phenol:Chloroform:Isoamyl Alcohol (Ambion) followed by ethanol precipitation. Samples were then size-selected using agarose gel electrophoresis. The extracted DNA samples were quantified fluorometrically with a Qubit 3.0 Fluorometer (Life technologies), and a total of 10 ng of digested DNA (200 pg l^{-1}) was used for library construction.

Initial sample quality control of pre-fragmented DNA was assessed using a Tapestation DNA 1000 High sensitivity Screen tape (Agilent, Cheadle UK). Sequencing ready libraries were prepared using the Hyper Prep DNA Library preparation kit (Kapa Biosystems, London UK) according to the manufacturer's instructions and indexed for pooling using NextFlex DNA barcoded adapters (Bioo Scientific, Austin TX US). Libraries were quantified using a Tapestation DNA 1000 Screen tape and by qPCR using an NGS Library Quantification Kit (KAPA Biosystems) on an AriaMx qPCR system (Agilent) and then normalised, pooled, diluted and denatured for sequencing on the NextSeq 500 (Illumina, Chesterford UK). The main library was spiked at 10% with the PhiX control library (Illumina). Sequencing was performed using Illumina NextSeq in the Departments of Biochemistry and Pathology at the University of Cambridge, UK, with 2x75 cycles of sequencing.

Data processing

Genome sequences were downloaded from Phytozome (v10) (Goodstein et al., 2012). The following genome assemblies were used: Bdistachyon_283_assembly_v2.0; Sbicolor_255_v2.0; Sitalica_164_v2; Zmays_284_AGPv3. Reads were mapped to genomes using bowtie2 (Langmead and Salzberg, 2012) with the following parameters:

--local –D 15 –R 2 –N 0 –L 20 –I S,1,0.75.

Aligned reads were then processed using samtools (Li et al., 2009) to remove those with a MAPQ score <42. DHS sites were identified using a procedure adapted from the ENCODE 3 pipeline (https://sites.google.com/site/anshulkundaje/projects/idr) (Marinov et al., 2014). Briefly DHS were called using MACS2 (Feng et al., 2012) with the following parameters to offset read locations in order to position DHS cut site in the middle of peak regions:

-p 1e-1 –nomodel –extsize 150 –shift -75 –llocal 50000

The final set of peak calls were determined using the irreproducible discovery rate (IDR (Li et al., 2011)) and calculated using the script batch_consistency_analysis.R (https://github.com/modENCODE-DCC/Galaxy/blob/master/modENCODE_DCC_tools/idr/batch-consistency-analysis.r).

Quality metrics and identification of Digital Genomic Footprints (DGF)

SPOT score (number of a subsample of mapped reads (5M) in DHS/Total number of subsampled, mapped reads (5M) (John et al., 2011)) was calculated using BEDTools (Quinlan and Hall, 2010) to determine the number of mapped reads possessing at least 1bp overlap with a DHS site. NSC and RSC scores were calculated using SPP (Kharchenko et al., 2008) and PCR bottleneck coefficient (PCB) was calculated using BEDTools and the following bash code:

```
bedtools bamtobed –bedpe –I ${FILT_BAM_FILE_n} | awk ‘BEGIN{OFS=”\t”}{print
$1,$2,$4,$6,$9,$10}’ | grep –v ‘ChrM\ChrC’| sort | uniq –c | awk ‘BEGIN{mt=0;m0=0;m1=0;m2=0}
($1==1){m1=m1+1} ($1==2){m2=m2+1} {m0=m0+1} {mt=mt+$1} END{printf
“%d\t%d\t%d\t%d\t%f\t%f\t%f\n”,mt,m0,m1,m2,m0/mt,m1/m0,m1/m2}’ > ${PBC_FILE_QC}`
```

Digital Genomic Footprints (DGF) were identified using the Wellington algorithm (Piper et al., 2013) in the pyDNase software package (http://pythonhosted.org/pyDNase/) with the following parameters:

-fdr 0.05 [regions] [reads] [output directory]

where [reads] represents a BED file of DHS locations within which footprints were called and [reads] a filtered BAM file of sequenced reads.

Differential DGF were identified using Wellington bootstrap algorithm (Piper et al., 2015) from pyDNase package with the following parameters:

-fdr 0.05 [treatment_BAM] [control_BAM] [regions] [treatment_output] [control_output]

Where [treatment_BAM] is a filtered BAM file containing sequenced reads from the sample of interest, [control_BAM] is a filtered BAM file containing mapped sequenced reads against sample for comparison; [regions] is a BED file containing DHS locations within which footprints are called. All DE DGFs with a threshold of score equal and higher than 10 were considered as differentially abundant DGFs.

Data visualisation

DHS and DGF sequences were loaded into and visualized in the Integrative Genomics Viewer (Thorvaldsdóttir et al., 2013) and figures produced in Inkscape, bar plots were generated with R package ggplot2 (Wickham, 2009), scatterplots using R function plot() and figures depicting conservation of DGF or motifs between orthologous sequences were generated using genoplotR (Guy et al., 2010). Word clouds were created with the wordcloud R package (Fellows, 2012).

TreeView images were produced in two stages. The script ‘dnase_to_javatreeview.py’ from pyDNAse was run with the following parameters to generate the input file:

[regions_BED] [reads_BAM] [OUTPUT]

Where [regions_BED] is a bed file containing locations of all DGF sites, [reads_BAM] is the BAM file containing all aligned reads, and [OUTPUT] specifies the output csv file name. To visualize files Java TreeView (Saldanha, 2004) was run with the following command:

```
java -Xmx4G -jar TreeView.jar
```

Changing the file format settings to All Files, the csv file from pyDNase was loaded into TreeView, from the dropdown menu entered Settings->Pixel Setting and checked all the Fill boxes, Contrast Value 1 and colours Red and Blue, the output was saved as .svg file.

Average cut density plots were generated using the script 'dnase_average_profile.py' from pyDNase (Piper et al., 2013, 2015) with the following parameters:

```
–w 100 –b [regions_BED] [reads_BAM] [OUTPUT]
```

Where [regions_BED] is a bed file containing locations of all DGF sites, [reads_BAM] is the BAM file containing all aligned reads, and [OUTPUT] specifies the output file name.

Genomic features were annotated and distribution calculated using the Bioconductor package ChIPpeakAnno (Zhu, 2013; Zhu et al., 2010a) interfaced with a custom R script. The required gff3 files (Goodstein et al., 2012) (Sitalica_164_v2.1.gene_exons.gff3; Sbicolor_255_v2.1.gene_exons.gff3; Zmays_284_6a.gene.exons.gff3; Bdistachyon_283_v2.1.gene_exons.gff3) downloaded from Phytozome.

In order to convert motif files into MEME format for motif scanning a multi-step procedure was necessary. Background frequency files are required when generating motifs (Thijs et al., 2001); to produce background files FASTA sequences for the regions of interest (DGF) were extracted using BEDTools suite (Quinlan and Hall, 2010) with the following command:

```
bedtools getfasta -fi [FASTA_genome] -bed [regions]-fo [FASTA_regions]
```

Background frequency files were tailored for each species for motif searching, using scripts from the meme suite (Bailey et al., 2009).

```
fasta-get-markov [FASTA_all] [background_file_MEME]
```

Motif files in FASTA format were converted to STAMP format using the online tool (http://www.benoslab.pitt.edu/stamp/) (Mahony and Benos, 2007), then RSTAT was used to convert STAMP format into TRANSFAC format (http://rsat01.biologie.ens.fr/rsa-tools/convert-matrix_form.cgi) (Medina-Rivera et al., 2015). A bug in the transfac2meme script requires that all bp frequencies are represented as floating point numbers containing two decimal places. In order to convert the TRANSFAC file to a suitable format the following code was used:

```
sed 's/0 /0.00/g' [transfac file] | sed 's/1 /1.00/g' | sed 's/2 /2.00/g' | sed 's/3 /3.00/g' | sed 's/4 /4.00/g' | sed 's/5 /5.00/g' | sed 's/6 /6.00/g' | sed 's/7 /7.00/g' | sed 's/8 /8.00/g' | sed 's/9 /9.00/g' | sed 's/0$/0.00/g' | sed 's/1$/1.00/g' | sed 's/2$/2.00/g' | sed 's/3$/3.00/g' | sed 's/4$/4.00/g' | sed 's/5$/5.00/g' | sed 's/6$/6.00/g' | sed 's/7$/7.00/g' | sed 's/8$/8.00/g' | sed 's/9$/9.00/g' | sed 's/\P0.00 /\P0/g' > [transfac_fixed]
```

MEME motif files were created from TRANSFAC files using scripts from the MEME suite (Bailey et al., 2009) with the following command:

```
transfac2meme -bg [background_file] [transfac_fixed] > [MEME_FILE]
```

where [background_file] is the background base pair distribution file and [MEME_FILE] is the motif file output.

de novo motif prediction, motif scanning and enrichment testing

_de_ novo motif prediction was performed using findMotifsGenome.pl script from the HOMER suite (Heinz et al., 2010) using digital genomic footprints (DGF) as input together with the reference genome sequence for each species with the following command:

```
findMotifsGenome.pl [INPUT_DGFs.bed] [REF_GENOME.fasta] [OUTFILE].motifs -size 200 -cpg
```

A set of 872 transcription factor binding motifs (O'Malley et al., 2016) in meme format was downloaded from

http://neomorph.salk.edu/dev/pages/shhuang/dap_web/pages/browse_table_aj.php

Motif scanning was performed using FIMO (Grant et al., 2011) with default parameters:

--bgfile [background_file] --o [OUPUT_FILE] [MOTIF_FILE] [FASTA_REGIONS]

where [background_file] is the background base pair distribution file, [OUTPUT_FILE] is the output file name, [MOTIF_FILE] is the file containing input motif(s) in MEME format and [FASTA_REGIONS] is a FASTA file containing all DGF sequences motifs are scanned against.

To determine overrepresentation of TF family motifs in samples hypergeometric tests were performed using R with the following parameters:

over<-phyper(hitInSample-1,hitInPop,failInPop,sampleSize,lower.tail=F)

where:

Population: Unique genes with an annotation in whole leaf and bundles sheath samples.

sampleSize: Number of unique genes with an annotation in whole leaf samples.

HitInPop: Total number of unique genes annotated with given transcription factor in tissue sample.

HitInSample: Number of unique genes sharing an annotation in WL and BS samples (overlap).

failInPop: Number of unique genes with annotation only in WL samples.

p-values were adjusted for the false discovery rate using the procedure of Benjamini & Hochberg (Benjamini and Hochberg, 1995).

The distribution of each motif across different genomic features was obtained for each of the 525 known annotated motifs by dividing the number of hits in a particular feature by the total number of hits in the genome. K-means clustering was then employed to group motifs by genomic feature in *Z. mays, S. italica, S. bicolor and B. distachyon*.

Whole genome alignments and pairwise cross mapping of genomic features

To cross map genomic features between species, mapping files were generated according to (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto) using tools from the

UCSC Genome Browser, including trfBig, faToNib, faSize, lavToPsl, faSplit, axtChain, chainNet (Kent et al., 2002) and LASTZ (Harris, 2007).

Genomic features where then mapped between genomes using bnMapper (Denas et al., 2015) and the following parameters:

-fBED4 –threshold 0.7 –o [outfile] [infile] [Chain file]

where [infile] is a BED file of DGF locations in the species of origin, [Chain file] is a chain file providing mapping coordinates between the species of origin and comparison.

Data

Detailed step by step methods are available for DNase I digestion are on protocols.io (dx.doi.org/10.17504/protocols.io.hdfb23n), Raw sequencing data and processed files are deposited in Gene Expression Omnibus (GSE97369).

Contributions: SJB and I-RL grew and harvested nuclei from *S. bicolor*, *S. italica* and *Z. mays*. KJ provided the nuclei from *B. distachyon*. SJB and I-RL performed DNase I digestion and data analysis. SJB, I-RL and JMH wrote the manuscript and prepared the figures.

Acknowledgements: KJ was supported by a Gatsby Career Development Fellowship, IRL was supported by CONCyT and BBSRC grant BB/L014130, whilst SJB was supported by the *3to4* grant from the EU and BB/I002243 from the BBSRC.

References

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

Bauwe, H., Hagemann, M., and Fernie, A.R. (2010). Photorespiration: players, partners and origin. Trends Plant Sci. *15*, 330–336.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B *57*, 289–300.

Bowes, G., Ogren, W.L., and Hageman, R.H. (1971). Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase. Biochem. Biophys. Res. Commun. *45*, 716–722.

Brown, N.J., Newell, C.A., Stanley, S., Chen, J.E., Perrin, A.J., Kajala, K., and Hibberd, J.M. (2011). Independent and Parallel Recruitment of Preexisting Mechanisms Underlying C_4 Photosynthesis. Science *331*, 1436–1439.

Chang, Y.M., Liu, W.Y., Shih, A.C., Shen, M.N., Lu, C.H., Lu, M.Y., Yang, H.W., Wang, T.Y., Chen, S.C., Chen, S.M., et al. (2012). Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. Plant Physiol *160*, 165–177.

Christin, P.-A., Salamin, N., Savolainen, V., Duvall, M.R., and Besnard, G. (2014). C_4 Photosynthesis Evolved in Grasses via Parallel Adaptive Genetic Changes. Curr. Biol. *17*, 1241–1247.

Christin, P.A., Salamin, N., Savolainen, V., Duvall, M.R., and Besnard, G. (2007). C_4 photosynthesis evolved in grasses via parallel adaptive genetic changes. Curr. Biol. *17*, 1241–1247.

Cornejo, M.-J., Luth, D., Blankenship, K.M., Anderson, O.D., and Blechl, A.E. (1993). Activity of a maize ubiquitin promoter in transgenic rice. Plant Mol. Biol. *23*, 567–581.

Covshoff, S., Furbank, R.T., Leegood, R.C., and Hibberd, J.M. (2013). Leaf rolling allows quantification of mRNA abundance in mesophyll cells of sorghum. J Exp Bot *64*, 807–813.

Deal, R.B., and Henikoff, S. (2011). The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana. Nat. Protoc. *6*, 56–68.

Denas, O., Sandstrom, R., Cheng, Y., Beal, K., Herrero, J., Hardison, R.C., and Taylor, J. (2015).

900 Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution.
901 BMC Genomics *16*, 87.
902 Emms, D.M., Covshoff, S., Hibberd, J.M., and Kelly, S. (2016). Independent and Parallel Evolution
903 of New Genes by Gene Duplication in Two Origins of C_4 Photosynthesis Provides New Insight into
904 the Mechanism of Phloem Loading in C_4 Species. Mol. Biol. Evol. *33*, 1796–1806.
905 Fellows, I. (2012). wordcloud: Word clouds. R Packag. Version *2*, 109.
906 Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using
907 MACS. Nat. Protoc. *7*, 1728–1740.
908 Furbank, R.T. (2011). Evolution of the C_4 photosynthetic mechanism: are there really three C_4 acid
909 decarboxylation types? J. Exp. Bot. *62*, 3103–3108.
910 Furbank, R.T., Stitt, M., and Foyer, C.H. (1985). Intercellular compartmentation of sucrose
911 synthesis in leaves of Zea mays L. Planta *164*, 172–178.
912 Gendrel, A.-V., Lippman, Z., Martienssen, R., and Colot, V. (2005). Profiling histone modification
913 patterns in plants using genomic tiling microarrays. Nat Meth *2*, 213–218.
914 Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W.,
915 Hellsten, U., Putnam, N., et al. (2012). Phytozome: a comparative platform for green plant
916 genomics. Nucleic Acids Res. *40*, D1178-86.
917 Gowik, U., Burscheidt, J., Akyildiz, M., Schlue, U., Koczor, M., Streubel, M., and Westhoff, P.
918 (2004). cis-Regulatory elements for mesophyll-specific gene expression in the C_4 plant *Flaveria*
919 *trinervia*, the promoter of the C_4 phospho*enol*pyruvate carboxylase gene. Plant Cell *16*, 1077–
920 1090.
921 Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif.
922 Bioinforma. *27*, 1017–1018.
923 Guy, L., Roat Kultima, J., and Andersson, S.G.E. (2010). genoPlotR: comparative gene and
924 genome visualization in R. Bioinformatics *26*, 2334–2335.
925 Harris, R.S. (2007). Improved pairwise alignment of genomic DNA. The Pennsylvania State
926 University.
927 Hatch, M.D. (1987). C_4 photosynthesis: a unique blend of modified biochemistry, anatomy and
928 ultrastructure. Biochim. Biophys. Acta *895*, 81–106.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., and Laslo, P. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell. *38*.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods *6*, 283–289.

Hibberd, J.M., and Covshoff, S. (2010). The regulation of gene expression required for C_4 photosynthesis. Annu Rev Plant Biol *61*, 181–207.

Jeon, J.-S., Lee, S., Jung, K.-H., Jun, S.-H., Kim, C., and An, G. (2000). Tissue-Preferential Expression of a Rice α-Tubulin Gene, OsTubA1, Mediated by the First Intron. Plant Physiol. *123*, 1005–1014.

John, C.R., Smith-Unna, R.D., Woodfield, H., Covshoff, S., and Hibberd, J.M. (2014). Evolutionary Convergence of Cell-Specific Gene Expression in Independent Lineages of C_4 Grasses. Plant Physiol. *165*, 62–75.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.-H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet *43*, 264–268.

Jordan, D.B., and Ogren, W.L. (1984). The CO_2/O_2 specificity of Ribulose 1,5-Bisphosphate Carboxylase Oxygenase - dependence on Ribulose bisphosphate concentration, pH and temperature. Planta *161*, 308–313.

Kajala, K., Williams, B.P., Brown, N.J., Taylor, L.E., and Hibberd, J.M. (2011). Multiple Arabidopsis genes primed for direct recruitment into C_4 photosynthesis. Plant J. *69*, 47–56.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, and D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

Kharchenko, P. V, Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotech *26*, 1351–1359.

Kromdijk, J., Głowacka, K., Leonelli, L., Gabilly, S.T., Iwai, M., Niyogi, K.K., and Long, S.P. (2016). Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. Science *354*, 857–861.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., and Batzoglou, S. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res *22*.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth *9*, 357–359.

Leakey, A.D.B., Bishop, K.A., and Ainsworth, E.A. (2012). A multi-biome gap in understanding of crop and ecosystem responses to elevated CO_2. Curr. Opin. Plant Biol. *15*, 228–236.

Leegood, R.C. (1985). The intercellular compartmentation of metabolites in leaves of *Zea mays* L. Planta *164*, 163–171.

Lefebvre, S., Lawson, T., Fryer, M., Zakhleniuk, O. V, Lloyd, J.C., and Raines, C.A. (2005). Increased Sedoheptulose-1,7-Bisphosphatase Activity in Transgenic Tobacco Plants Stimulates Photosynthesis and Growth from an Early Stage in Development. Plant Physiol. *138*, 451–460.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. *5*, 1752–1779.

Long, S.P., Marshall-Colon, A., and Zhu, X.-G. (2015). Meeting the Global Food Demand of the Future by Engineering Crop Photosynthesis and Yield Potential. Cell *161*, 56–66.

Maas, C., Laufs, J., Grant, S., Korfhage, C., and Werr, W. (1991). The combination of a novel stimulatory element in the first exon of the maize Shrunken-1 gene with the following intron 1 enhances reporter gene expression up to 1000-fold. Plant Mol. Biol. *16*, 199–207.

Mahony, S., and Benos, P. V (2007). STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. *35*, W253–W258.

Marinov, G.K., Kundaje, A., Park, P.J., and Wold, B.J. (2014). Large-Scale Quality Analysis of Published ChIP-seq Data. G3 *4*, 209–223.

Martin, A., and Orgogozo, V. (2013). The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution *67*, 1235–1250.

Matsuoka, M., and Numazawa, T. (1991). *Cis*-acting elements in the pyruvate, orthophosphate dikinase gene from maize. Mol. Gen. Genet. *228*, 143–152.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., et al. (2015). RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Res. 43: W50-W56.

Miyagawa, Y., Tamoi, M., and Shigeoka, S. (2001). Overexpression of a cyanobacterial fructose-1,6-/sedoheptulose-1,7-bisphosphatase in tobacco enhances photosynthesis and growth. Nat. Biotechnol. *19*, 965–969.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell *165*, 1280–1292.

Ort, D.R., Merchant, S.S., Alric, J., Barkan, A., Blankenship, R.E., Bock, R., Croce, R., Hanson, M.R., Hibberd, J.M., Long, S.P., et al. (2015). Redesigning photosynthesis to sustainably meet global food and bioenergy demand. Proc. Natl. Acad. Sci. *112*, 8529–8536.

Patel, M., Corey, A.C., Yin, L.P., Ali, S.J., Taylor, W.C., and Berry, J.O. (2004). Untranslated regions from C-4 amaranth AhRbcS1 mRNAs confer translational enhancement and preferential bundle sheath cell expression in transgenic C_4 *Flaveria bidentis*. Plant Physiol. *136*, 3550–3561.

Piper, J., Elze, M.C., Cauchy, P., Cockerill, P.N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res. *41*, e201-e201.

Piper, J., Assi, S.A., Cauchy, P., Ladroue, C., Cockerill, P.N., Bonifer, C., and Ott, S. (2015). Wellington-bootstrap: differential DNase-seq footprinting identifies cell-type determining transcription factors. BMC Genomics *16*, 1000.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. *26*, 841–842.

Reyna-Llorens, I., Burgess, S.J., Williams, B.P., Stanley, S., Boursnell, C., and Hibberd, J.M. (2016). Ancient coding sequences underpin the spatial patterning of gene expression in C_4 leaves. bioRxiv doi: https://doi.org/10.1101/085795.

Sage, R. (2004). The evolution of C_4 photosynthesis. New Phytol. *161*, 341–370.

Sage, R.F., and Zhu, X.-G. (2011). Exploiting the engine of C_4 photosynthesis. J. Exp. Bot. *62*, 2989–3000.

Sage, R.F., Christin, P.-A., and Edwards, E.J. (2011). The C_4 plant lineages of planet Earth. J. Exp. Bot. *62*, 3171–3181.

Sage, R.F., Sage, T.L., and Kocacinar, F. (2012). Photorespiration and the evolution of C_4 photosynthesis. Annu. Rev. Plant Biol. *63*, 19–47.

Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics. *20*.

Sharwood, R.E., Ghannoum, O., Kapralov, M. V, Gunn, L.H., and Whitney, S.M. (2016). Temperature responses of Rubisco from Paniceae grasses provide opportunities for improving C_3 photosynthesis. Nat. Plants *2*, 16186.

Sheen, J. (1999). C_4 gene expression. Ann. Rev Plant Physiol. Plant Mol Biol *50*, 187–217.

Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E.M., Akey, J.M., et al. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. Science *342*, 1367–1372.

The International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature *463*, 763–768.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics *17*, 1113–1122.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings Bioinforma. *14*, 178–192.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tolbert, N.E. (1971). Microbodies - peroxisomes and glyoxysomes. Annu. Rev. Plant Physiol. *22*, 45–74.

Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. (2006). Evolution of alternative transcriptional

circuits with identical logic. Nature *443*, 415–420.

Viret, J.F., Mabrouk, Y., and Bogorad, L. (1994). Transcriptional photoregulation of cell-type preferred expression of maize rbcS-m3: 3' and 5' sequences are involved. Proc. Natl. Acad. Sci. *91*, 8577–8581.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. *22*, 1798–1812.

Wang, L., Czedik-Eysenberg, A., Mertz, R.A., Si, Y., Tohge, T., Nunes-Nesi, A., Arrivault, S., Dedow, L.K., Bryant, D.W., Zhou, W., et al. (2014). Comparative analyses of C_4 and C_3 photosynthesis in developing leaves of maize and rice. Nat Biotech *32*, 1158–1165.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis (Springer New York).

Williams, B.P., Burgess, S.J., Reyna-Llorens, I., Knerova, J., Aubry, S., Stanley, S., and Hibberd, J.M. (2016). An untranslated cis-element regulates the accumulation of multiple C_4 enzymes in Gynandropsis gynandra mesophyll cells. Plant Cell *28*, 454–465.

Xing, K., and He, X. (2015). Reassessing the "Duon" Hypothesis of Protein Evolution. Mol. Biol. Evol. *32*, 1056–1062.

Xu, T., Purcell, M., Zucchi, P., Helentjaris, T., and Bogorad, L. (2001). TRM1, a YY1-like suppressor of rbcS-m3 expression in maize mesophyll cells. Proc. Natl. Acad. Sci. U. S. A. *98*, 2295–2300.

Zentner, G.E., and Henikoff, S. (2014). High-resolution digital profiling of the epigenome. Nat Rev Genet *15*, 814–827.

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z., Wang, W., et al. (2012a). Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. Nat Biotech *30*, 549–554.

Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E., and Jiang, J. (2012b). High-resolution mapping of open chromatin in the rice genome. Genome Res. *22*, 151–162.

Zhu, L.J. (2013). Integrative Analysis of ChIP-Chip and ChIP-Seq Dataset BT - Tiling Arrays: Methods and Protocols. T.-L. Lee, and A.C. Shui Luk, eds. (Totowa, NJ: Humana Press), pp. 105–124.

1074 Zhu, L.J., Gazin, C., Lawson, N.D., Pagès, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010a).

1075 ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC

1076 Bioinformatics *11*, 237.

1077 Zhu, X., Long, S., and Ort, D. (2010b). Improving Photosynthetic Efficiency for Greater Yield. Annu.

1078 Rev. Plant Biol. *61*, 235–261.

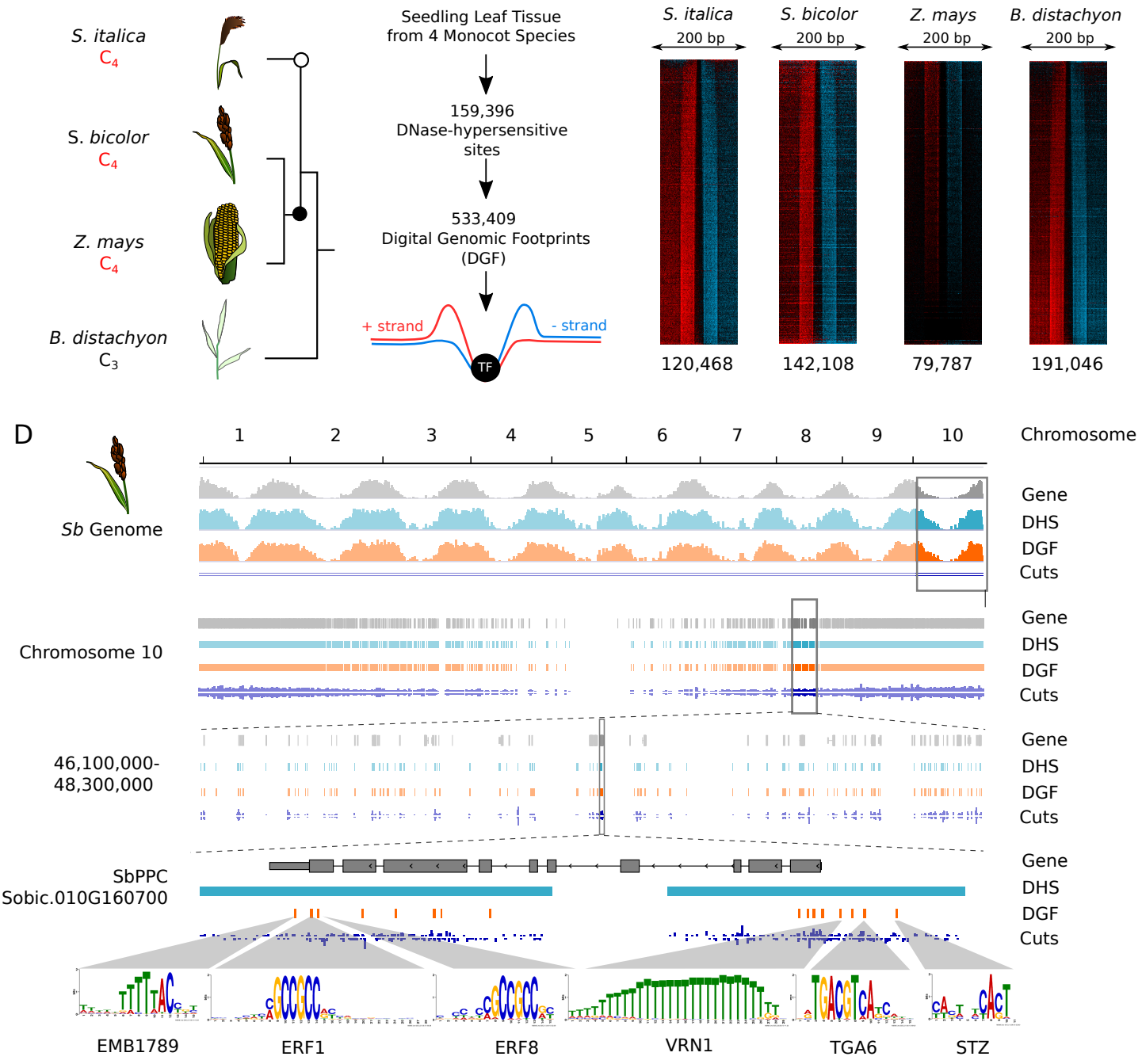
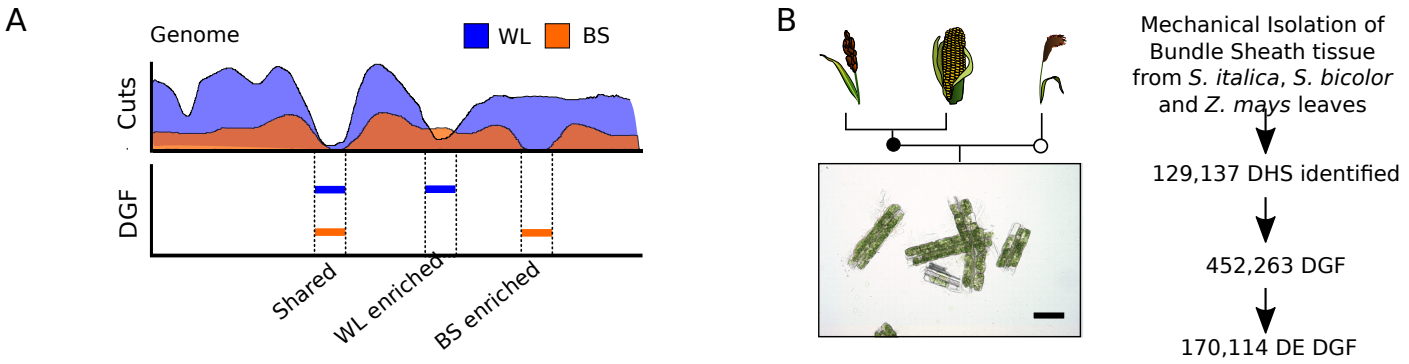
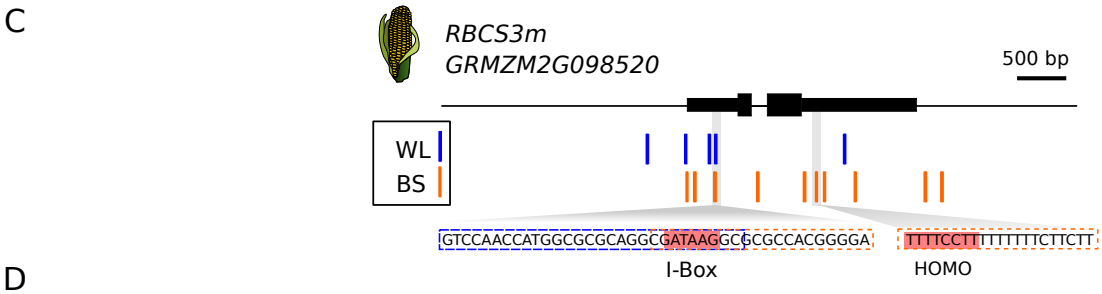
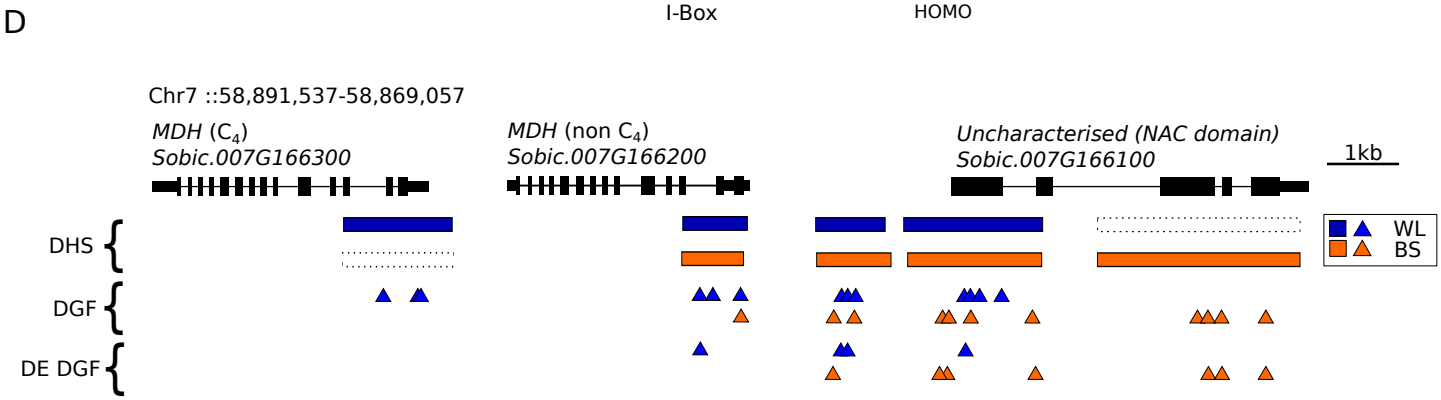


Figure 1







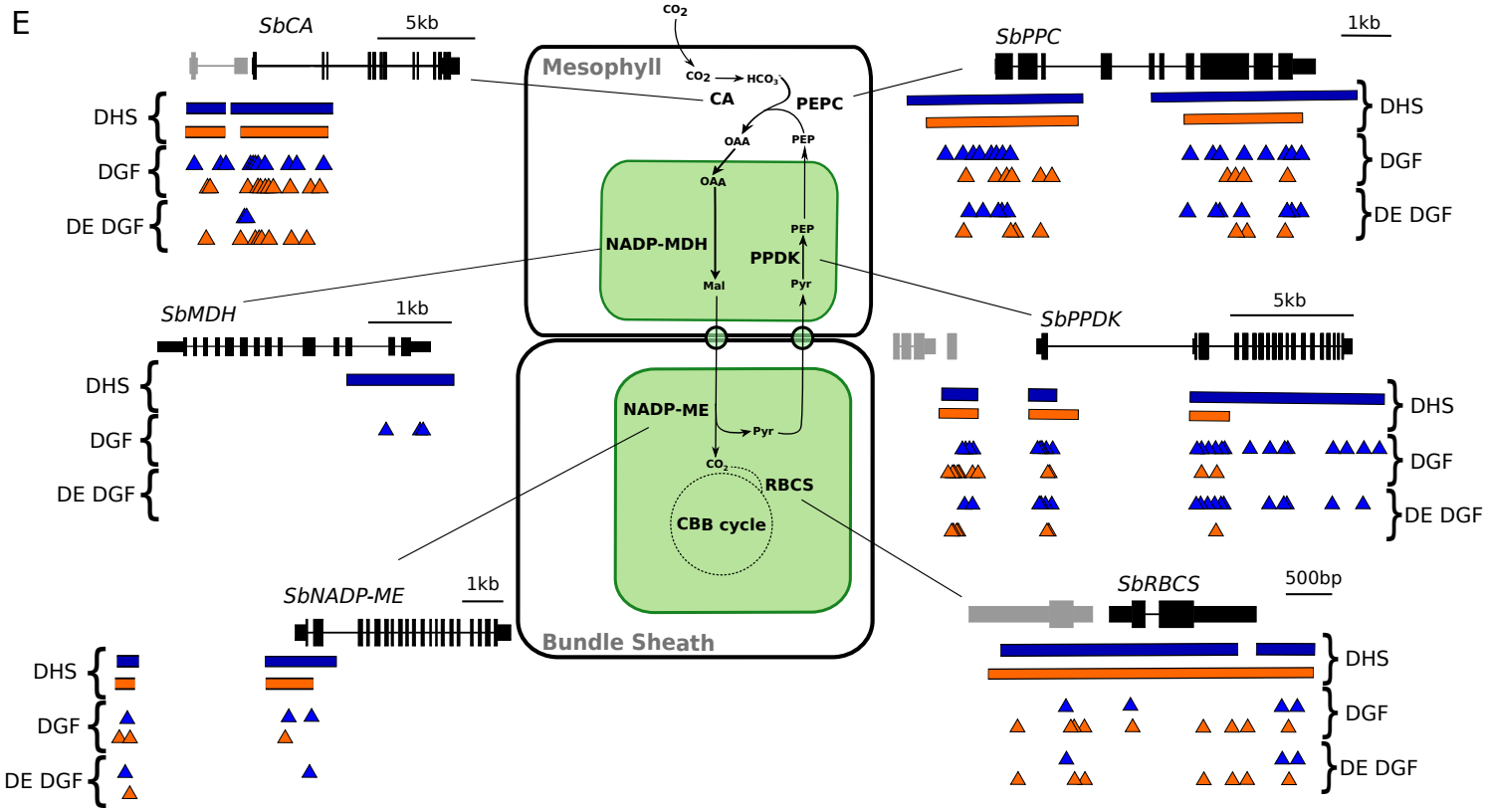


Figure 2

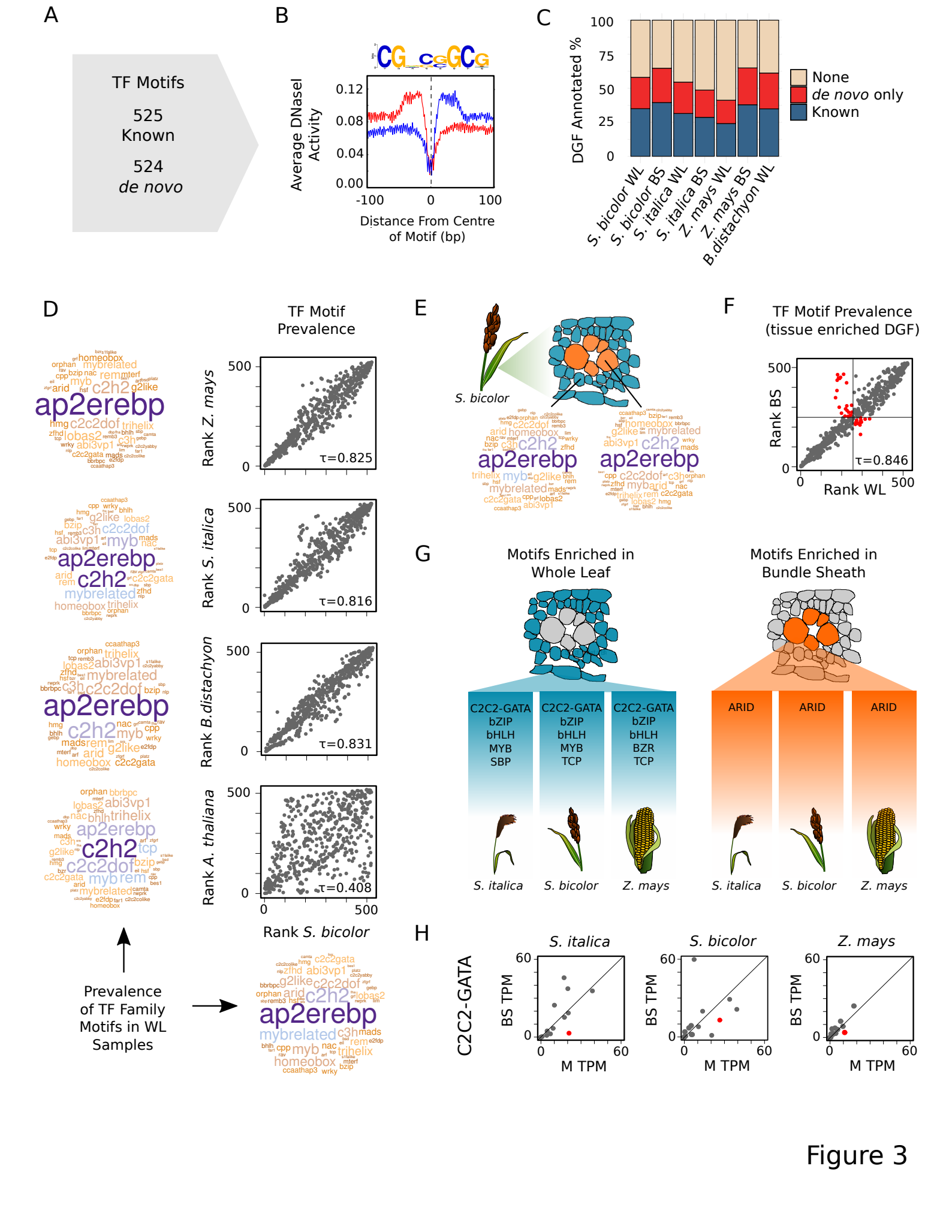


Figure 3

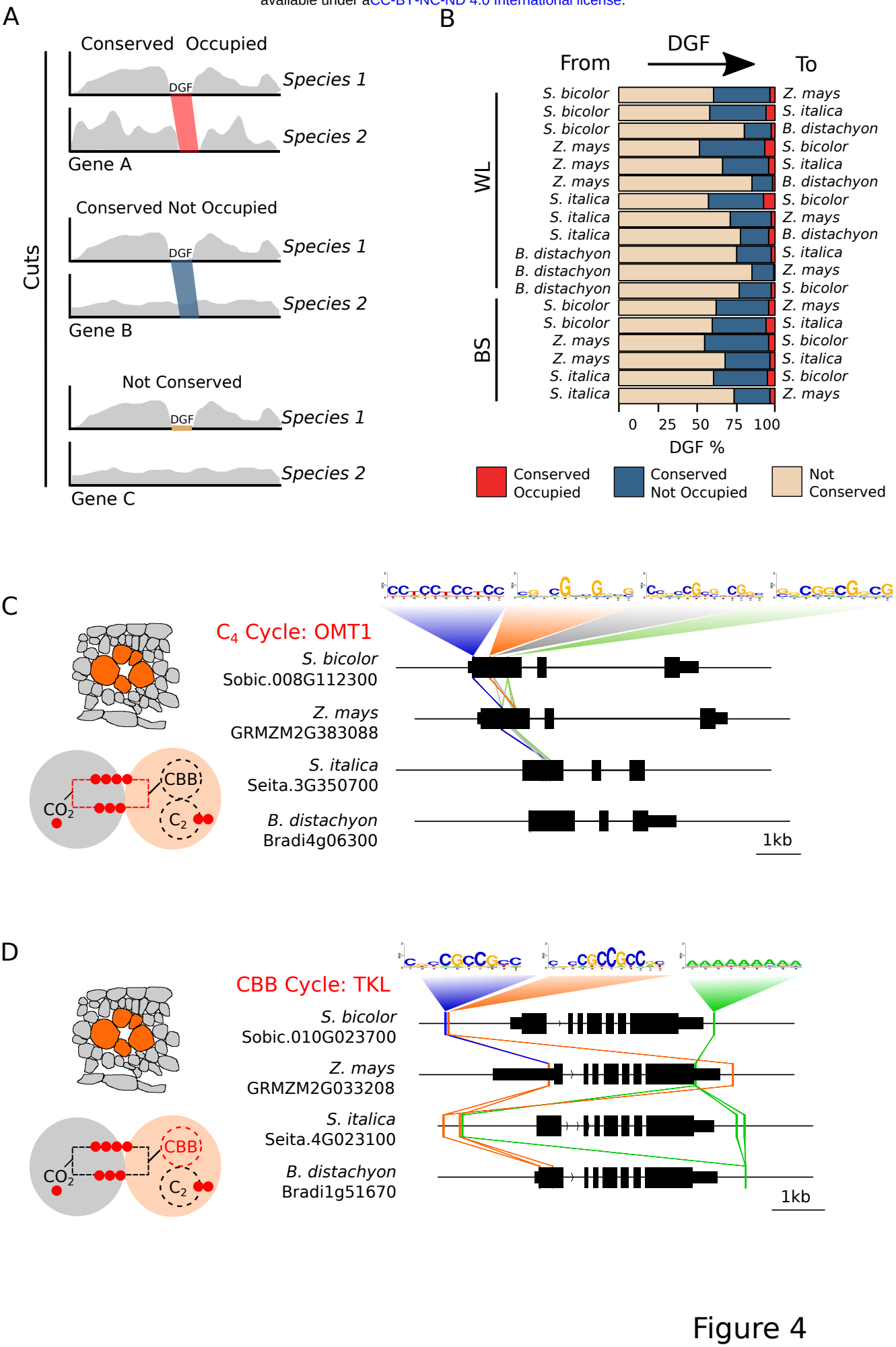
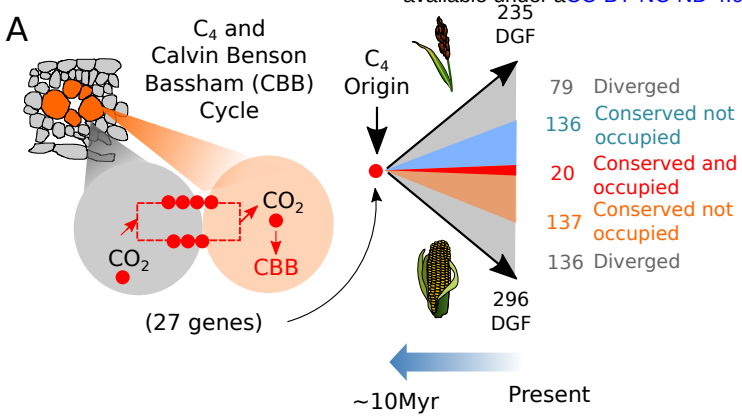
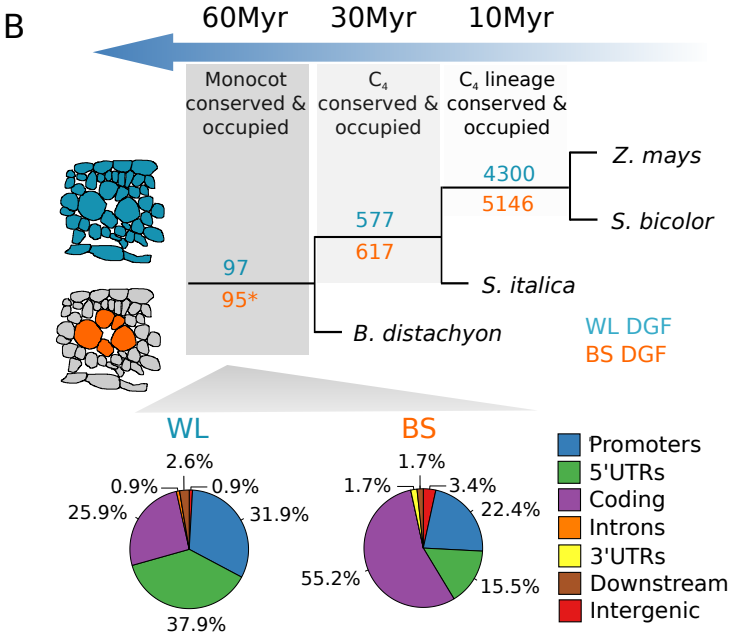
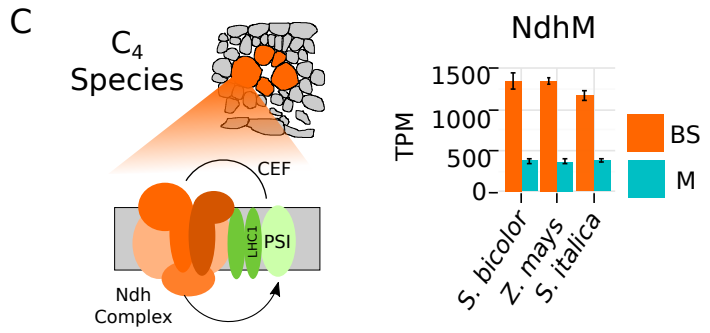


Figure 4







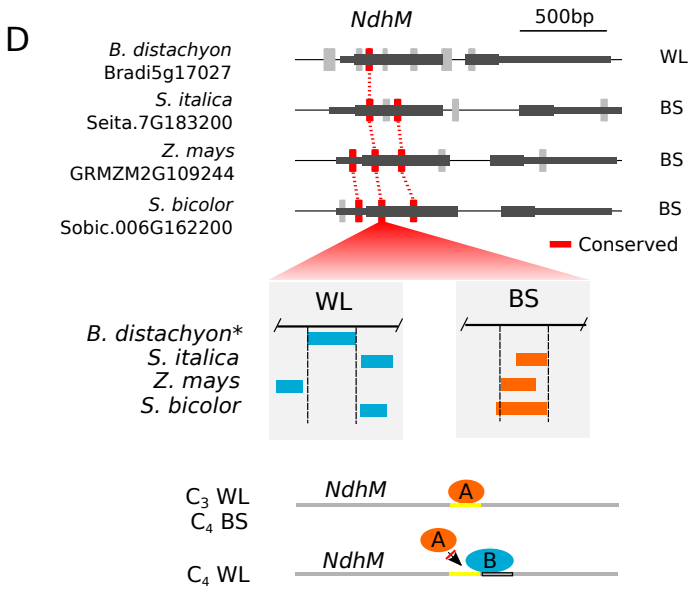


Figure 5