Genome-wide transcription factor binding in leaves from C_3 and C_4 grasses

Steven J. Burgess[1†] (0000-0003-2353-7794), Ivan Reyna-Llorens[1†] (0000-0001-7964-7306), Sean R. Stevenson[1] (0000-0001-5635-4340), Pallavi Singh[1] (0000-0003-3694-6378), Katja Jaeger[2] (0000-0002-4153-7328) and Julian M. Hibberd (0000-0003-0662-7958)[1*]

[1]Department of Plant Sciences, Downing Street, University of Cambridge, Cambridge CB2 3EA, UK.
[2]Sainsbury Laboratory, Bateman Street, University of Cambridge, Cambridge, CB2 1LR, UK.

I-RL - suallorens@gmail.com
SJB - sjb287@illinois.edu
SRS - srs62@cam.ac.uk
PS - ps753@cam.ac.uk
KJ - katja.jaeger@slcu.cam.ac.uk
JMH (corresponding) - jmh65@cam.ac.uk

[†]These authors contributed equally to this work

Short title: Transcription factor binding atlas of grass leaves

Keywords: C_4 photosynthesis, gene regulation, evolution, cereals

One sentence summary: Genome-wide patterns of transcription factor binding *in vivo* defined by DNaseI for leaves of C_3 and C_4 grasses

Abstract

The majority of plants use C_3 photosynthesis, but over sixty independent lineages of angiosperms have evolved the C_4 pathway. In most C_4 species, photosynthesis gene expression is compartmented between mesophyll and bundle sheath cells. We performed DNaseI-SEQ to identify genome-wide profiles of transcription factor binding in leaves of the C_4 grasses *Zea mays*, *Sorghum bicolor* and *Setaria italica* as well as C_3 *Brachypodium distachyon*. In C_4 species, while bundle sheath strands and whole leaves shared similarity in the broad regions of DNA accessible to transcription factors, the short sequences bound varied. Transcription factor binding was prevalent in gene bodies as well as promoters, and many of these sites could represent duons that impact gene regulation in addition to amino acid sequence. Although globally there was little correlation between any individual DNaseI footprint and cell-specific gene expression, within individual species transcription factor binding to the same motifs in multiple genes provided evidence for shared mechanisms governing C_4 photosynthesis gene expression. Furthermore, interspecific comparisons identified a small number of highly conserved transcription factor binding sites associated with leaves from species that diverged around 60 million years ago. These data therefore provide insight into the architecture associated with C_4 photosynthesis gene expression in particular and characteristics of transcription factor binding in cereal crops in general.

54 **Introduction**

55 Most photosynthetic organisms, including crops of global importance such as wheat, rice
56 and potato use the C_3 photosynthesis pathway in which Ribulose-Bisphosphate
57 Carboxylase Oxygenase (RuBisCO) catalyses the primary fixation of CO_2. However,
58 carboxylation by RuBisCO is competitively inhibited by oxygen binding the active site
59 (Bowes et al., 1971). This oxygenation reaction generates toxic waste-products that are
60 recycled by an energy-demanding series of metabolic reactions known as photorespiration
61 (Bauwe et al., 2010; Tolbert, 1971). The ratio of oxygenation to carboxylation increases with
62 temperature (Jordan and Ogren, 1984; Sharwood et al., 2016) and so losses from
63 photorespiration are particularly high in the tropics.

64 Multiple plant lineages have evolved mechanisms that suppress oxygenation by
65 concentrating CO_2 around RuBisCO. One such strategy is known as C_4 photosynthesis.
66 Species that use the C_4 pathway include maize, sorghum and sugarcane, and they
67 represent the most productive crops on the planet (Sage and Zhu, 2011). In C_4 leaves,
68 additional expenditure of ATP, alterations to leaf anatomy and cellular ultrastructure, as well
69 as spatial separation of photosynthesis between compartments (Hatch, 1987) allows CO_2
70 concentration to be increased around tenfold compared with that in the atmosphere
71 (Furbank, 2011). Despite the complexity of C_4 photosynthesis, it is found in over 60
72 independent plant lineages (Sage et al., 2011). In most C_4 plants the initial RuBisCO-
73 independent fixation of CO_2 and the subsequent RuBisCO-dependent reactions take place
74 in distinct cell-types known as mesophyll and bundle sheath cells. Although the spatial
75 patterning of gene expression that generates these metabolic specialisations is fundamental
76 to C_4 photosynthesis very few examples of *cis*-elements or *trans*-factors that restrict gene
77 expression to mesophyll or bundle sheath cells of C_4 plants have been identified (Brown et
78 al., 2011; Gowik et al., 2004; Williams et al., 2016; Reyna-Llorens et al., 2018). Moreover,
79 in grasses more generally the DNA-binding properties of relatively few transcription factors
80 have been validated (Bolduc and Hake, 2009; Yu et al., 2015; Eveland et al., 2014; Pautler
81 et al., 2015). In summary, in both C_3 and C_4 species, work has focussed on analysis of
82 mechanisms controlling the expression of individual genes, and so our understanding of the
83 overall landscape associated with photosynthesis gene expression is poor.

84 In yeast and animal systems, the high sensitivity of open chromatin to DNaseI (Zentner
85 and Henikoff, 2014) has allowed comprehensive, genome-wide characterization of
86 transcription factor binding sites at single nucleotide resolution (Hesselberth et al., 2009;
87 Neph et al., 2012; Thurman et al., 2012). In plants, DNaseI-SEQ and more recently Assay
88 for Transposase-Accessible Chromatin (ATAC-SEQ) have been employed in C_3 species

89 and provided insight into the patterns of transcription factor binding associated with
90 development (Zhang et al., 2012a; Pajoro et al., 2014; Zhang et al., 2012b, 2016), heat
91 stress (Sullivan et al., 2014) and root cell differentiation (Maher et al., 2017). By carrying out
92 DNaseI-SEQ on grass leaves that use either C_3 or C_4 photosynthesis, we aimed to provide
93 insight into the transcription factor binding repertoire associated with each form of
94 photosynthesis. Our data indicate more transcription factor binding sites are found in gene
95 bodies than promoters, and up to 25% of the footprints represent 'duons' – sequences
96 located in exons that have an impact on both gene regulation as well as the amino acid
97 sequence of the protein they encode. It is also clear that specific cell types from leaf tissue
98 make use of a markedly distinct *cis*-regulatory code and that despite significant turnover in
99 the cistrome of grasses, a small number of transcription factor motifs are conserved across
100 60 million years of evolution. Comparison of sites bound by transcription factors in both C_3
101 and C_4 leaves demonstrates that the repeated evolution of C_4 photosynthesis is built on both
102 the *de novo* gain of *cis*-elements and the exaptation of highly conserved regulatory elements
103 found in the ancestral C_3 system.

104 **Results**

105 **A *cis*-regulatory atlas for grasses**

106 To provide insight into the regulatory architecture associated with C_3 and C_4
107 photosynthesis in cereal crops, four grass were selected. *Brachypodium distachyon* uses
108 the ancestral C_3 pathway (Figure 1A). *Sorghum bicolor, Zea mays* and *Setaria italica* all use
109 C_4 photosynthesis, they were chosen as phylogenetic reconstructions indicate that *S. italica*
110 represents an independent evolutionary origin of the C_4 pathway (Figure 1A) and
111 comparison of these species can provide insight into parallel and convergent evolution of C_4
112 gene expression. Nuclei from a minimum of duplicate samples of *S. italica* (C_4), *S. bicolor*
113 (C_4), *Z. mays* (C_4) and *B. distachyon* (C_3) leaves were treated with DNaseI (Supplemental
114 Figure 1) and subjected to deep sequencing. A total of 806,663,951 reads could be uniquely
115 mapped to the respective genome sequences of these species (Supplemental Table 1).
116 From all four genomes, 159,396 DNaseI-hypersensitive sites (DHS) of between 150-15,060
117 base pairs representing broad regulatory regions accessible to transcription factor binding
118 were identified (Figure 1B). Between 20,817 and 27,746 genes were annotated as
119 containing at least one DHS (Supplemental Table 2). For subsequent analysis, only DHS
120 that were consistent between replicates as determined by the Irreproducible Discovery Rate
121 framework (Li and Dewey, 2011) were used.

122 DNaseI footprinting is a well-established technique for detecting DNA-protein interactions
123 at base pair resolution and as such has been used to generate Digital Genomic Footprints
124 (DGF) to predict transcription factor binding sites. DGF are obtained by pooling all replicates
125 to maximise the number of reads that map within each DHS, and then modelling differential
126 accumulation of reads mapping to positive or negative strands around transcription factor
127 binding sites within the DHS (Piper et al., 2013). However, the DNaseI enzyme possesses
128 some sequence bias that can affect prediction of transcription factor binding sites (He et al.,
129 2014; Yardimci et al., 2014). After performing DNaseI-SEQ on “naked DNA” that is devoid
130 of nucleosomes from each species, we identified hundreds of DGF that likely represent false
131 positives (Supplemental Figure 2A). For all species, analysis of the DGF derived from naked
132 DNA showed that treatment with DNaseI led to similar sequences being preferentially
133 digested (Supplemental Figure 2B). However, because false positive DGF predicted from
134 this approach will be influenced by the number of reads that map to each genome, and in
135 the case of maize fewer reads mapped in total, the number of false positive DGF varied
136 between species (Supplemental Figure 2A). To overcome this issue, we implemented a
137 more conservative pipeline that rather than defining false positives at specific locations
138 within the genome, calculates DNaseI cutting bias for all hexamers across each genome.

139 By employing a mixture model framework, these data are then used to generate a
140 background signal to estimate footprint likelihood scores for each putative DGF (Yardimici
141 et al., 2014; Supplemental Figure 2B). This approach removed between 15% and 30% of
142 DGF from each sample (Supplemental Figure 2C) and left a total of 430,205 DGF
143 corresponding to individual transcription factor binding sites between 11 and 25 base pairs
144 being identified (Figure 1B&C; Supplemental Table 3). At least one transcription factor
145 footprint was identified in >75% of the broader regions defined by DHS (Supplemental Table
146 2).

147 We attempted to saturate the number of predicted DGF by sequencing each species at
148 high depth (Supplemental Table 1). *In silico* sub-sampling of these data indicated that for *S.*
149 *bicolor, S. italica* and *B. distachyon*, the total number of DGF was close to saturation, but for
150 maize despite obtaining 251,955,063 reads from whole leaves this was not the case
151 (Supplementary Figure 3). Consequently, fewer DGF were predicted in maize compared
152 with the other species (Figure 1C). Since maize has a similar gene number to the other
153 species analysed, it is possible that the reduced ability to map reads to unique loci was
154 associated with the high amount of repetitive DNA in the maize genome. Another
155 contributing factor to the poor mapping rate in maize may be the low complexity found in
156 one of the libraries as reflected by the PCR bottleneck coefficient (Supplemental Table 1).
157 According to the Encyclopaedia of DNA elements (ENCODE), large number of reads from
158 low complexity libraries decreases the chances of identifying the majority of transcription
159 factor binding sites. However, despite these differences in coverage and in certain quality
160 metrics, for all four species DHS and DGF were primarily located in gene-rich regions and
161 depleted around centromeres (Figure 1D). Individual transcription factor binding sequences
162 were resolved in all chromosomes from each species (Figure 1D). On a genome-wide basis,
163 the distribution of DHS was similar between species, with the highest proportion of such
164 sites located in promoter, coding sequence and intergenic regions (Figure 1E). Notably, in
165 all four grasses, genic sequences contained more DHS than promoters (Figure 1F).

166 To provide additional evidence confirming that DGF identified in our analysis derive from
167 protein-DNA interactions, they were compared with previously identified motifs from maize.
168 Maize is the most appropriate choice for this analysis as there are more data on transcription
169 factor binding sites than in *S. bicolor* and *S. italica*. Moreover, support from previous work
170 goes some way to supporting the smaller number of DGF that we identified in this species.
171 Therefore, the literature was assessed for validated transcription factor binding sites in
172 maize. These have previously been associated with flowering (Kozaki et al., 2004; Vollbrecht
173 et al., 2005; Eveland et al., 2014), meristem development (Bolduc et al., 2012), Gibberellin

174 catabolism (Bolduc and Hake, 2009), sugar signalling (Niu et al., 2002) and leaf
175 development of maize (Yu et al., 2015), but in all cases, DGF matching these motifs were
176 found in our dataset (Figure 2A, FDR < 0.001). In addition, a larger ChIP-SEQ dataset of
177 117 transcription factors from maize leaves obtained from the pre-release maize cistrome
178 (http://www.epigenome.cuhk.edu.hk/C3C4.html, Supplemental Figure 4, Supplemental
179 Table 4) was compared with our data. Differences between specific binding sites are likely
180 because in all cases growth conditions will have varied from ours, and in some cases
181 different tissues were sampled. Despite this, 66% and 29% of the ChIP-SEQ peaks
182 overlapped with our DHS and DGF respectively. Although only 29% of DGF overlapped with
183 motifs defined by ChIP-SEQ, permutation tests performed using the "regioneR" package
184 (Gel et al., 2015) indicated a statistically greater overlap than would be expected by chance
185 (pvalue = 0.0099, 100 permutations). Moreover, when both features were systematically
186 shifted from their original position, the local z-score, which represents the strength of the
187 association at any particular position showed a sharper decrease for DGF than DHS
188 suggesting the association between ChIP-SEQ peaks and DGFs is more strongly linked to
189 the exact position of the DGF (Supplemental Figure 4B). In summary, despite detecting
190 fewer DGF in maize than in the other species, the DGF we found are supported by publicly
191 available ChIP-SEQ, EMSA and Selex datasets.
192 Consistent with the distribution of DHS (Figure 1E), annotated DGF were most common
193 in promoter, coding sequence and intergenic regions (Figure 2B) and genic sequences
194 contained more DGF than promoters (Figure 2C). Distribution plots showed that the highest
195 density of DGF was close to the annotated transcription start sites but indicated a slightly
196 skewed distribution favouring genic sequence including exons (Supplemental Figure 5). A
197 similar pattern was observed for the ChIP-SEQ signal peaks (Supplemental Figure 4C).
198 Transcription factor binding sites located in exons have been termed duons because they
199 could impact both on the regulation of transcription and amino acid sequence. Whilst in
200 general synonymous mutations not affecting amino acid sequence should be under relaxed
201 purifying selection, because of transcription factor recognition all nucleotides in duons
202 should be under purifying selection, and thus show lower mutation rates. We therefore
203 investigated the nucleotide substitution rate at Four-Fold Degenerate Sites (FFDS) using
204 variation data from 1218 maize lines (Bukowski et al., 2018) and found that it was statistically
205 significantly lower in duons than in surrounding coding sequence (Figure 2E, p= 7.04e-9).
206 This contrasts with the density of polymorphisms in non-synonymous sites (Figure 2D).
207 Although it has been proposed that GC bias of duons constrains FFDS (Xing and He, 2014)
208 we found no such bias between duon and exon sequences used in this analysis

209 (Supplemental Figure 6). Taken together, we conclude that in these cereals a significant
210 proportion of transcription factor binding likely takes place within genes.
211
212 **A distinct *cis*-regulatory lexicon for specific cells within the leaf**
213 The above analysis provides a genome-wide overview of the *cis*-regulatory architecture
214 associated with leaves of grasses. However, as with other complex multicellular systems,
215 leaves are composed of many specialised cell types. Because DGF are defined by the
216 differential DNA cleavage between protected and unprotected regions of DNA within a DHS,
217 a negative distribution compared with the larger DHS is produced (Figure 3A). Thus,
218 transcription factor binding signal from a low abundance cell-type is likely to be obscured by
219 overall signal from a tissue-level analysis (Figure 3A). Since bundle sheath strands can be
220 separated (Covshoff et al., 2013; Leegood, 1985; Furbank et al., 1985) C_4 species provide
221 a simple system to study transcription factor binding in specific cells of leaves (Figure 3B).
222 After bundle sheath isolation from *S. bicolor, S. italica* and *Z. mays*, and naked DNA
223 correction for inherent bias in DNaseI cutting, a total of 129,137 DHS were identified (Figure
224 3B; Supplemental Table 5) containing 244,554 DGF (Figure 3B; Supplemental Table 5;
225 FDR<0.01). Of these, 138,075 were statistically enriched in the bundle sheath samples
226 compared with whole leaves (Figure 3B; Supplemental Table 5). The number of these
227 statistically enriched DGF in bundle sheath strands of C_4 species was large and ranged from
228 14,250 to 73,057 in maize and *S. italica* respectively (Supplemental Table 5). The lower
229 number in maize is likely due to the reduced sequencing depth achieved. Genome-wide, the
230 number of broad regulatory regions defined by DHS in the bundle sheath that overlapped
231 with those present in whole leaves ranged from 71 to 84% in *S. italica* and *S. bicolor*
232 respectively (Supplemental Table 6). However, only 6-20% of the narrower DGF found in
233 the bundle sheath were also identified in whole leaves (Supplemental Table 7). Taken
234 together, these findings indicate that specific cell types of cereal leaves share similarity in
235 the broad regions of DNA that are accessible to transcription factors (DHS), but that the
236 short sequences actually bound by transcription factors (DGF) vary dramatically.
237 To provide evidence that DGF predicted after analysis of separated bundle sheath
238 strands are of functional importance, they were compared with previously validated
239 sequences. In C_4 grasses, to our knowledge there are no such examples in *S. bicolor* or *S.*
240 *italica*, but in the *RbcS* gene from maize, which is preferentially expressed in bundle sheath
241 cells, an I-box (GATAAG) is essential for light-mediated activation (Giuliano et al., 1988) and
242 a HOMO motif (CCTTTTTTCCTT) is important in driving bundle sheath expression (Xu et
243 al., 2001) (Figure 3C). Despite not reaching saturation in DGF prediction in maize

244 (Supplemental Figure 3) both elements were detected in our pipeline. Interestingly, a signal
245 suggesting TF binding to the HOMO motif was enriched in the bundle sheath strands (Figure
246 3C), and whilst the I-box was detected in both bundle sheath strands and whole leaves its
247 position was slightly different in each cell type (Figure 3C). These findings are therefore
248 consistent with the biochemical data implicating the I-box in control of abundance and the
249 HOMO box in control of cell-specific accumulation of *RbcS* transcripts.

250 The *ZmPEPC* gene (GRMZM2G083841) encodes the phospho*enol*pyruvate carboxylase
251 responsible for producing C_4 acids used in the C_4 pathway and is preferentially expressed
252 in mesophyll cells. Previous reports showed that a region of 600 nucleotides upstream of
253 the transcription start site carrying repeated C-rich sequences was sufficient to drive
254 expression in mesophyll cells of maize (Shaffner & Sheen, 1992; Matzuoka et al., 1994).
255 Although no DGF were detected with these C-rich sequences, they are located within a DHS
256 indicating that they are available for transcription factor binding (Figure 3D). Thus, despite
257 the fact that we had not reached saturation of DGF in maize, for both *RbcS* and *PEPC* the
258 regions of DNA accessible to transcription factor binding are consistent with previous
259 reports, and in the case of *RbcS* DGF were detected that coincide with known *cis*-elements.

260 To investigate the relationship between cell specific gene expression and the position of
261 DHS and DGF, the DNaseI data were interrogated using RNA-SEQ datasets from mesophyll
262 and bundle sheath cells of C_4 leaves (Chang et al., 2012; John et al., 2014; Emms et al.,
263 2016). At least three mechanisms associated with cell specific gene expression operating
264 around individual genes were identified and can be exemplified using three co-linear genes
265 found on chromosome seven of *S. bicolor*. First, in the *NADP-malate dehydrogenase* (*MDH*)
266 gene, which is highly expressed in mesophyll cells and encodes a protein of the core C_4
267 cycle (Figure 3E) a broad DHS site and two DGF were present in whole leaves, but not in
268 bundle sheath strands (Figure 3F). Whilst presence of this site indicates accessibility of DNA
269 to transcription factors that could activate expression in mesophyll cells, global analysis of
270 all genes strongly and preferentially expressed in bundle sheath strands versus whole
271 leaves indicates that presence/absence of a DHS in one cell type is not sufficient to generate
272 cell specificity (Supplemental Figure 7 & 8). Second, in the next contiguous gene that
273 encodes an additional isoform of MDH also preferentially expressed in mesophyll cells
274 (Figure 3E), a DHS was found in both whole leaf and bundle sheath strands but DGF within
275 this region differed between cell types (Figure 3F). Thus, despite similarity in DNA
276 accessibility, the binding of particular transcription factors varied between cell types.
277 However, once again, genome-wide analysis indicated that alterations to individual DGF
278 were not sufficient to explain cell specific gene expression. For example, only 30 to 40% of

279 all enriched DGF in the bundle sheath were associated with differentially expressed genes
280 (Supplemental Table 8). Lastly, in the third gene in this region, which encodes a NAC domain
281 transcription factor preferentially expressed in bundle sheath strands (Figure 3E),
282 differentially enriched DGF were associated both with regions of the gene that have similar
283 DHS in each cell type, but also a region lacking a DHS in whole leaves compared with
284 bundle sheath strands (Figure 3E). These three classes of alteration to transcription factor
285 accessibility and binding were detectable in genes encoding core components of the C_4
286 cycle in all three species (Supplemental Figure 9-11). Overall, we conclude that differences
287 in transcription factor binding between cells of C_4 leaves is associated with both DNA
288 accessibility defined by broad DHS, as well as fine-scale alterations to transcription factor
289 binding defined by DGF. Moreover, bundle sheath strands possessed a distinct regulatory
290 landscape compared with the whole leaf, and in genes encoding enzymes of the C_4 pathway
291 multiple transcription factor binding sites differed between bundle sheath and whole leaf
292 samples. This finding implies that cell specific gene expression in C_4 leaves is mediated by
293 combinatorial effects derived from alterations to gene accessibility as defined by DHS as
294 well as changes to binding of multiple transcription within these regions.
295
296 **DNA motifs associated with cell specific expression**
297 To provide an overview of the transcription factors most likely associated with DGF ChIP-
298 SEQ data from maize (Figure 2B) together with motifs from JASPAR plants (Khan et al.,
299 2018) and an additional 529 transcription factor motifs validated in Arabidopsis (O'Malley et
300 al., 2016) were used to annotate the DGF from *Z. mays, S. bicolor, S. italica* and *B.*
301 *distachyon* (Figure 4A). To increase the number of annotated DGF *de novo* prediction was
302 used to identify sequences over-represented in DGF compared with those across the whole
303 genome. This resulted in an additional 524 motifs being annotated (Figure 4A), but in fact
304 all of these were previously detected after *de novo* prediction from DNaseI-SEQ of rice
305 (Zhang et al., 2012b). As would be expected from *bona fide* transcription factor binding,
306 inspection of these motifs predicted *de novo* demonstrated clear strand bias in DNaseI cuts
307 (Figure 4B). By combining previously known motifs and those predicted *de novo*, the
308 percentage of DGF that could be annotated in each species increased from about 60% to
309 more than 75% (Figure 4C, Supplemental Table 9).
310 To define the most common sequences bound by transcription factors in mature leaves
311 undertaking C_3 and C_4 photosynthesis and to investigate whether C_4 photosynthesis is
312 controlled by an increase in binding of sets of transcription factors, individual motifs were
313 ranked by frequency and the Kendall rank correlation coefficient used to compare species

314 (Figure 4D). In both C_3 and C_4 species, the most prevalent transcription factor binding motifs
315 were associated with AP2-EREBP and MYB transcription factor families (p-value < 2.2^{-16};
316 Figure 4D). Next, to identify regulatory factors associated with gene expression in the C_4
317 bundle sheath, transcription factor motifs located in DGF enriched in either the bundle
318 sheath or in whole leaf samples of *S. bicolor* were identified (Figure 4E). There was little
319 difference in the ranking of the most commonly used motifs between these cell types
320 (Kendall's tau=0.815; p-value < 2.2^{-16}), indicating cell-specificity is not associated with large-
321 scale changes in the abundance of many transcription factor families (Figure 4E). After
322 performing hypergeometric tests for enrichment of individual motifs in differentially occupied
323 DGF we found 133 and 106 motifs enriched in whole leaves and bundle sheath strands
324 respectively ($p < 0.001$). Of these 239 motifs, 37 were enriched in all C_4 species with 10 and
325 27 enriched in the bundle sheath and whole leaf respectively (Figure 4F, Supplemental
326 Table 10) 66 were only enriched in bundle sheath strands and 91 to whole leaf tissue
327 (Supplemental Table 11). Some of these conserved and cell specific motifs have been
328 previously described to have a relevant role in photosynthesis. For instance, in whole leaves
329 of maize and *Setaria*, we found significant enrichment of the bHLH129 motif (Supplemental
330 Table 11) that has been proposed to act as a negative regulator of NADP-ME (Borba et al.,
331 2018).
332

333 **Multiple genes encoding enzymes of the C_4 and Calvin-Benson-Bassham cycles**
334 **share the same occupied *cis*-elements**

335 To investigate whether genes involved in the C_4 phenotype are co-regulated, we
336 compared the number of instances where the same motifs were bound in multiple C_4, Calvin-
337 Benson-Bassham and C_2 cycle genes (Supplemental Table 12). While no single *cis*-element
338 was found in all genes that are preferentially expressed in mesophyll or bundle-sheath cells,
339 the number of genes possessing the same occupied motif ranged from nine in *S. bicolor*
340 and *S. italica* to four in *S. bicolor* and *Z. mays* whole leaves respectively (Supplemental
341 Table 9 & 12). These data support a model where the combinatorial action of multiple
342 transcription factors controls groups of C_4 genes to produce the gene expression patterns
343 required for C_4 photosynthesis.

344 We next performed comparative analysis of motifs bound by transcription factors to
345 determine whether the set of *cis*-elements found in C_4 genes of each species were common,
346 or whether C_4 genes are regulated differently in each species. In pairwise comparisons,
347 DGF fell into three categories: conserved and occupied by a transcription factor, conserved
348 but only occupied in one species, and not conserved (Figure 5A). Only a small percentage

349 of DGF were both conserved in sequence and bound by transcription factors (Figure 5B,
350 Supplemental Table 13). Consistent with this, the majority of C_4 gene orthologs did not share
351 DGF. Due to the lack of DGF saturation in maize, these estimates likely set lower bounds
352 for the extent of conservation. However, in several cases, patterning of C_4 gene expression
353 correlated with a set of motifs shared across species (Figure 5C). In some cases, these
354 shared *cis*-elements were present in the ancestral C_3 state. For instance, the
355 *TRANSKETOLASE* (*TKL*) gene contains several conserved DGF that are present in the
356 bundle sheath of the C_4 species but also in whole leaves of C_3 *Brachypodium* (Figure 5).
357 This finding is consistent with the notion that C_4 photosynthesis makes use of existing
358 regulatory architecture found in C_3 plants. Nevertheless, overall, these data also indicate
359 that the majority of C_4 gene expression appears to be associated with species-specific
360 regulatory networks.
361
362 **Hyper-conserved *cis*-regulators of C_4 genes**
363 To investigate the extent to which transcription factor binding sites associated with C_4
364 genes within a C_4 lineage are conserved, genes encoding the core C_4 cycle were compared
365 in *S. bicolor* and *Z. mays* (Figure 6A, Supplemental Table 14). 27 genes associated with the
366 C_4 and Calvin-Benson-Bassham Cycles contained a total of 379 DGF. Although many of
367 these transcription factor footprints were conserved in sequence within orthologous genes,
368 only nine were both conserved and bound by a transcription factor (Figure 6A). Again, due
369 to the lack of DGF saturation in maize, these data likely represent minimum estimates of
370 conservation.
371 Genome-wide, the number of DGF that were conserved in sequence and bound by a
372 transcription factor decayed in a non-linear manner with phylogenetic distance (Figure 6B,
373 Supplemental Table 15). For example, *Z. mays* and *S. bicolor* shared 5,775 DGF that were
374 both conserved and occupied. *S. italica* shared only 670 DGF with *Z. mays* and *S. bicolor*
375 (Figure 6B). Finally, comparison of these C_4 grasses with C_3 *B. distachyon* yielded 93 DGF
376 that have been conserved over >60Myr of evolution. Because nuclei from *B. distachyon*
377 were sampled later in the photoperiod than those from the C_4 grasses, and DGF may well
378 vary over the diel cycle, it is possible that this is an underestimate of DGF conservation.
379 However, 41 of these highly conserved DGF were present in whole leaf samples of the C_3
380 species, but in the C_4 species were restricted to the bundle sheath (Figure 6B). Gene
381 Ontology analysis did not detect enrichment of any specific terms for hyper-conserved DGF
382 associated with the bundle sheath, but for whole leaves detected over-representation of "cell
383 component" categories such as membrane bound organelles and the nucleus

384 (Supplemental Table 16). In whole leaves, this set of ancient and highly conserved DGF
385 were located predominantly in 5' UTRs and coding sequences, but in bundle sheath strands
386 over fifty percent of these hyper-conserved DGF were in coding sequences (Figure 6B).
387 Overall, these data indicate that certain duons are highly conserved across deep
388 evolutionary time. The frequent use of hyper-conserved duons in the bundle sheath implies
389 that this cell type uses an ancient and highly conserved regulatory code.

390 **Discussion**

391 **Genome-wide transcription factor binding in grasses**

392 The dataset provides insight into the regulation of gene expression in cereals in general,
393 and to C_4 photosynthesis in particular. Consistent with previous analysis ranging from *A.*
394 *thaliana* (Sullivan et al., 2014) to metazoans (Natarajan et al., 2012; Stergachis et al., 2013,
395 2014), the majority of DGF detected in the four grasses were centred around annotated
396 transcription start sites. However, in these cereals it is noteworthy that transcription factor
397 binding was prevalent in genic sequence. Whilst we cannot rule out the possibility that this
398 distribution is in some way related to the methodology used in this study, there is evidence
399 that the exact distribution of transcription factor binding appears to be species specific. For
400 example, whilst in *A. thaliana* DNaseI-SEQ revealed enrichment of DHS in sequence ~400
401 base pairs upstream of transcription start sites as well as 5' UTRs (Sullivan et al., 2014) and
402 ATAC-SEQ of *A. thaliana, Medicago truncatula* and *Oryza sativa* detected most transposase
403 hypersensitive sites upstream of genes, in *Solanum lycopersicum* more were present in
404 introns and exons than upstream of annotated transcription start sites (Maher et al., 2017).

405 The prevalence of transcription factor binding to coding sequences is relevant to
406 approaches used to generate transgenic plants and test gene function and regulation. First,
407 consistent with the prevalence of DGF downstream of the annotated transcription start sites
408 that we detected, it is noteworthy that during cereal transformation, exon and intron
409 sequences are frequently used to achieve stable expression of transgenes (Maas et al.,
410 1991; Cornejo et al., 1993; Jeon et al., 2000). It is possible that this strategy is required in
411 grasses because of the high proportion of transcription factor binding downstream of
412 annotated transcription start sites. These transcription factor binding sites in coding
413 sequence also have implications for synthetic biology. Although technologies such as type-
414 IIS restriction endonuclease cloning methods allow high-throughput testing of many
415 transgenes, they rely on sequence domestication. Whilst routinely this would maintain amino
416 acid sequence, without analysis of transcription factor binding sites it could mutate motifs
417 bound by transcription factors and lead to unintended modifications to gene expression.

418

419 **The transcription factor landscape underpinning gene expression in specific tissues**

420 The finding that so few transcription factor binding sites were shared between bundle
421 sheath tissue and whole leaves of *S. bicolor, Z. mays* and *S. italica* argues for the need to
422 isolate these cells when attempting to understand the control of gene expression. Although
423 separating bundle sheath strands from C_4 leaves is relatively trivial (Covshoff et al., 2013;
424 Furbank et al., 1985; Leegood, 1985) this is not the case for C_3 leaves. Approaches in which

425 nuclei from specific cell-types are labelled with an exogenous tag (Deal and Henikoff, 2011)
426 now allow their transcription factor landscapes to be defined. The application of DNaseI-
427 SEQ to specific cell types has recently been used in roots (Maher et al., 2017) and so in the
428 future, this approach of both C_3 and C_4 leaves should provide insight into how the extent to
429 which gene regulatory networks have been re-wired during the evolution of the complex C_4
430 trait.

431 Given the central importance of cellular compartmentation to C_4 photosynthesis, there
432 have been significant efforts to identify *cis*-elements that restrict gene expression to either
433 mesophyll or bundle sheath cells of C_4 leaves (Hibberd and Covshoff, 2010; Sheen, 1999;
434 Wang et al., 2014). As previous studies of C_4 gene regulation have focused on individual
435 genes and have been performed in various species, it has not been possible to obtain a
436 coherent picture of regulation of the C_4 pathway and along with many other systems, initial
437 analysis focussed on regulatory elements located in promoters of C_4 genes (Sheen, 1999).
438 However, it has become increasingly apparent that the patterning of gene expression
439 between cells in the C_4 leaf can be mediated by elements in various parts of a gene. In
440 addition to promoter elements (Sheen, 1999; Gowik et al., 2004), this includes untranslated
441 regions (Kajala et al., 2011; Patel et al., 2004; Viret et al., 1994; Williams et al., 2016; Xu et
442 al., 2001) and coding sequences (Brown et al., 2011; Reyna-Llorens et al., 2018). By
443 providing data on *in vivo* transcription factor occupancy for the complete C_4 pathway in three
444 C_4 grasses, the data presented here allow broad comparisons and provide several insights
445 into regulatory networks controlling C_4 genes.

446 The DNaseI dataset indicates that cell specific gene expression in C_4 leaves is not
447 strongly correlated with changes to large-scale accessibility of DNA as defined by DHS. This
448 implies that modifications to transcription factor accessibility around any one gene does not
449 impact on its expression between tissues in the leaf. Rather, as only 8-24% of transcription
450 factor binding sites detected in the bundle sheath were also found in whole leaves, the data
451 strongly implicate complex modifications to patterns of transcription factor binding in
452 controlling gene expression between cell types. These findings are consistent with
453 analogous analysis in roots where genes with clear spatial patterns of expression are bound
454 by multiple transcription factors (Sparks et al., 2017) and highly combinatorial interactions
455 between multiple activators and repressors tune the output (de Lucas et al., 2016).

456 The data also provide insight into *cis*-elements that underpin the C_4 phenotype. No single
457 c*is*-element was found in all genes preferentially expressed in either mesophyll or bundle
458 sheath cells of one species. This finding is consistent with analysis of yeast where the output
459 of genetic circuits can be maintained despite rapid turnover of *cis*-regulatory mechanisms

460 underpinning them (Tsong et al., 2006). However, we did detect small numbers of C_4 genes
461 that shared common transcription factor footprints (Figure 5, Supplemental Table 14 & 15),
462 which is consistent with previous analysis that identified shared *cis*-elements in *PPDK* and
463 *CA,* or *NAD-ME1* and *NAD-ME2* in C_4 *Gynandropsis gynandra* (Williams et al., 2016; Reyna-
464 Llorens et al., 2018). Interspecific comparisons further underlined the high rate of divergence
465 in the *cis*-regulatory logic used to control C_4 genes. For example, although we detected
466 highly similar transcription factor footprints in the *OMT1* and *TKL* genes of the three C_4
467 species we assessed, this was not apparent for any other C_4 genes. As a result of the
468 apparent rapid rate of evolution in *cis*-regulatory architecture in these C_4 species, attempts
469 to engineer C_4 photosynthesis into C_3 crops to increase yield (Hibberd et al., 2008) may
470 benefit from using pre-existing regulatory mechanisms controlling mesophyll or bundle
471 sheath expression in ancestral C_3 species.
472
473 **Characteristics of the transcription factor binding in the ancestral C_3 state that have**
474 **impacted on evolution of the C_4 pathway**
475 Comparison of transcription factor binding in the C_3 grass *B. distachyon* with three C_4
476 species provides insight into mechanisms associated with the evolution of C_4
477 photosynthesis. For all four grasses, irrespective of whether they used C_3 or C_4
478 photosynthesis, the most abundant DNA motifs bound by transcription factors were similar.
479 Thus, motifs recognised by the AP2-EREBP and MYB classes of transcription factor were
480 most commonly bound across each genome. This indicates that during the evolution of C_4
481 photosynthesis, there has been relatively little alteration to the most abundant classes of
482 transcription factors that bind DNA.
483 The repeated evolution of the C_4 pathway has frequently been associated with convergent
484 evolution (Sage, 2004; Sage et al., 2012). However, parallel alterations to amino acid and
485 nucleotide sequence that allow altered kinetics of the C_4 enzymes (Christin et al., 2014,
486 2007) and patterning of C_4 gene expression (Brown et al., 2011) respectively have also been
487 reported. The genome-wide analysis of transcription factor binding reported here indicates
488 that only a small proportion of the C_4 cistrome is associated with parallel evolution. These
489 estimates regarding conservation between C_4 and C_3 species may represent
490 underestimates as whilst nuclei where all sampled in the light, those from *B. distachyon*
491 were sampled later in the photoperiod. Moreover, when orthologous genes were compared
492 between the four grasses assessed here, the majority of transcription factor binding sites
493 were not conserved, and of the DGF that were conserved, position within orthologous genes
494 varied. This indicates that C_4 photosynthesis in grasses is tolerant to a rapid turnover of the

495 *cis*-code, and that when motifs are conserved in sequence, their position and frequency
496 within a gene can vary. It therefore appears that the cell-specific accumulation patterns of
497 C_4 proteins can be maintained despite considerable modifications to the cistrome of C_4
498 leaves. It was also the case that some conserved motifs bound by transcription factors in
499 the C_4 species were present in *B. distachyon*, which uses the ancestral C_3 pathway.
500 Previous work has shown that *cis*-elements used in C_4 photosynthesis can be found in gene
501 orthologs from C_3 species (Williams et al., 2016; Reyna-Llorens et al., 2018). However,
502 these previous studies identified *cis*-elements that were conserved in both sequence and
503 position. As it is now clear that such conserved motifs are mobile within a gene, it seems
504 likely that many more examples of ancient *cis*-elements important in C_4 photosynthesis will
505 be found in C_3 plants.

506 Although we were able to detect a small number of transcription factor binding sites that
507 were conserved and occupied in all four species sampled, these ancient hyper-conserved
508 motifs appear to have played a role in the evolution of C_4 photosynthesis. Interestingly, a
509 large proportion of these motifs bound by transcription factors were found in coding
510 sequences, and this bias was particularly noticeable in bundle sheath cells. Due to the amino
511 acid code, the rate of mutation of coding sequence compared with the genome is restricted.
512 If such regions have a longer half-life than transcription factor binding sites in other regions
513 of the genome, then they may represent an excellent source of raw material for the repeated
514 evolution of complex traits (Martin and Orgogozo, 2013). Our data documenting the frequent
515 use of hyper-conserved DGF in the C_4 bundle sheath implies that this tissue may use an
516 ancient and highly conserved regulatory code. It appears that during the evolution of the C_4
517 pathway, which relies on heavy use of the bundle sheath, this ancient code has been co-
518 opted to control photosynthesis gene expression.

519 In summary, the data provide a transcription factor binding atlas for leaves of grasses
520 using either C_3 or C_4 photosynthesis. Whilst we did not achieve DGF saturation in maize,
521 commonalities between the four species were apparent. Sequences bound by transcription
522 factors were found within genes as well as promoter regions, and many of these motifs
523 represent duons. In terms of the regulation of tissue specific gene expression, whilst bundle
524 sheath strands and whole C_4 leaves shared considerable similarity in regions of DNA
525 accessible to transcription factors, the short sequences actually bound by transcription
526 factors varied dramatically. We identified a small number of transcription factor motifs that
527 were conserved in these species. The data also provide insight into the regulatory
528 architecture associated with C_4 photosynthesis more specifically. Whilst we found some
529 evidence that multiple genes important for C_4 photosynthesis share common *cis*-elements

530 bound by transcription factors, this was not widespread. This may well relate to the relatively
531 rapid turnover in the *cis*-code, and so it is possible that transcription factors interacting with
532 these motifs are more conserved. Analysis of transcription factor footprints in specific cell
533 types from leaves of C_3 grasses should in the future provide insight into the extent to which
534 gene regulatory networks have altered during the transition from C_3 to C_4 photosynthesis.

535 **Methods**

536 **Growth conditions and isolation of nuclei**

537 *S. bicolor, S. italica* and *Z. mays* were grown under controlled conditions at the Plant
538 Growth Facilities of the Department of Plant Sciences at the University of Cambridge in a
539 chamber set to 12 h/12 h light/dark; 28 ºC light/20 ºC dark; 400 µmol m^{-2} s^{-1} photon flux
540 density, 60% humidity. For germination, *S. bicolor* and *Z. mays* seeds were imbibed in H_2O
541 for 48 h, *S. italica* seeds were incubated on wet filter paper at 30 ºC overnight in the dark.
542 *Z. mays, S. bicolor* and *S. italica* were grown on 3:1 (v/v) M3 compost to medium vermiculite
543 mixture, with a thin covering of soil. Seedlings were hand-watered.

544 *B. distachyon* plants were grown in a separate growth facility under controlled conditions
545 optimised for its growth at the Sainsbury Laboratory Cambridge University, first under short
546 day conditions 14 h/10 h, light/dark for 2 weeks and then shifted to long day 20 h/4 h,
547 light/dark, for 1 week and harvested at ZT20. Temperature was set at 20 ºC, humidity 65%
548 and light intensity 350 µmol m^{-2} s^{-1}. All tissue was harvested from August to October 2015.

549 To isolate nuclei from *S. bicolor, Z. mays* and *S. italica* mature third and fourth leaves with
550 a fully developed ligule were harvested 4-6 h into the light cycle 18 days after germination.
551 Bundle sheath cells were mechanically isolated as described previously (Markelz et al.,
552 2003). At least 3 g of tissue was used for each extraction. Nuclei were isolated using a
553 sucrose gradient adapted and yield quantified using a haemocytometer. For *B. distachyon*
554 plants were flash frozen and material pulverised in a coffee grinder. 3 g of plant material was
555 added to 45 ml NIB buffer (10mM Tris-HCl, 0.2M sucrose, 0.01% (v/v) Triton X-100, pH 5.3
556 containing protease inhibitors (Sigma-Aldrich)) and incubated at 4ºC on a rotating wheel for
557 5 min, afterwards debris was removed by sieving through 2 layers of Miracloth (Millipore)
558 into pre-cooled flasks. Nuclei were spun down 4,000 rpm, 4 ºC for 20 min. Plastids were
559 lysed by adding Triton to a final concentration of 0.3% (v/v) and incubated for 15 min on ice.
560 Nuclei were pelleted by centrifugation at 5000 rpm at 4 ºC for 15 min. Pellets were washed
561 3 times with chilled NIB buffer.

562

563 **Deproteinized DNA extraction**

564 For isolation of deproteinated DNA from *S. bicolor*, *Z. mays*, *B. distachyon* and *S. italica*
565 mature third and fourth leaves with a fully developed ligule were harvested 4 h into the light
566 cycle, 18 days after germination. 100 mg of tissue was used for each extraction.
567 Deproteinated DNA was extracted using a QIAGEN DNeasy Plant Mini Kit (QIAGEN, UK)
568 according to the manufacturer's instructions.

569

570 **DNaseI digestion, sequencing and library preparation**

571 To obtain sufficient DNA each biological replicate consisted of leaves from tens of
572 individuals and to conform to standards set by the Human Genome project at least two
573 biological replicates were sequenced for each sample. 2 x 10^8 of freshly extracted nuclei
574 were re-suspended at 4 °C in digestion buffer (15 mM Tris-HCl, 90 mM NaCl, 60 mM KCl, 6
575 mM $CaCl_2$, 0.5 mM Spermidine, 1 mM EDTA and 0.5 mM EGTA, pH 8.0). DNaseI
576 (Fermentas) at 7.5 U was added to each tube and incubated at 37 °C for 3 min. Digestion
577 was arrested with addition of 1:1 volume of stop buffer (50 mM Tris-HCl, 100 mM NaCl,
578 0.1% (w/v) SDS, 100 mM EDTA, pH 8.0, 1 mM Spermidine, 0.3 mM Spermine, RNase-A 40
579 µg/ml) and incubated at 55 °C for 15 min. 50 U of Proteinase K was added and samples
580 incubated at 55 °C for 1 h. DNA was isolated with 25:24:1 Phenol:Chloroform:Isoamyl
581 Alcohol (Ambion) followed by ethanol precipitation. Fragments from 50 to 550 bp were
582 selected using agarose gel electrophoresis. The extracted DNA samples were quantified
583 fluorometrically with a Qubit 3.0 Fluorometer (Life Technologies), and a total of 10 ng of
584 digested DNA (200 pg l-1) was used for library construction.

585 Initial sample quality control of pre-fragmented DNA was assessed using a Tapestation
586 DNA 1000 High Sensitivity Screen Tape (Agilent, Cheadle UK). Sequencing ready libraries
587 were prepared using the Hyper Prep DNA Library preparation kit (Kapa Biosystems, London
588 UK) selecting fragments from 70-350 bp for optimization (see He et al., 2014) and indexed
589 for pooling using NextFlex DNA barcoded adapters (Bioo Scientific, Austin TX US). In order
590 to reduce bias due to amplification of DNA fragments by the polymerase chain reaction, as
591 recommended by the manufacturers, a low number of cycles (17 cycles) was used. Libraries
592 were quantified using a Tapestation DNA 1000 Screen Tape and by qPCR using an NGS
593 Library Quantification Kit (KAPA Biosystems) on an AriaMx qPCR system (Agilent) and then
594 normalised, pooled, diluted and denatured for sequencing on the NextSeq 500 (Illumina,
595 Chesterford UK). The main library was spiked at 10% with the PhiX control library (Illumina).
596 Sequencing was performed using Illumina NextSeq in the Departments of Biochemistry and
597 Pathology at the University of Cambridge, UK, with 2x75 cycles of sequencing. For the
598 deproteinized DNAse I seq experiments 1 µg of deproteinized DNA was resuspended in 1
599 ml of digestion buffer (15 mM Tris-HCl, 90 mM NaCl, 60 mM KCl, 6 mM $CaCl_2$, 0.5 mM
600 spermidine, 1 mM EDTA and 0.5 mM EGTA, pH 8.0). DNaseI (Fermentas) at 2.5 U was
601 added to each tube and incubated at 37 °C for 2 min. Digestion was arrested with addition
602 of 1:1 volume of stop buffer (50 mM Tris-HCl, 100 mM NaCl, 0.1% (w/v) SDS, 100 mM
603 EDTA, pH 8.0, 1 mM Spermidine, 0.3 mM Spermine, RNase A 40 µg/ml) and incubated at
604 55 °C for 15 min. 50 U of Proteinase K was added and samples incubated at 55 °C for 1 h.

605 DNA was isolated by mixing with 1 ml 25:24:1 Phenol:Chloroform:Isoamyl Alcohol (Ambion)
606 and spun for 5 min at 13,00 rpm followed by ethanol precipitation of the aqueous phase.
607 Samples were then size-selected (50-400 bp) using agarose gel electrophoresis. The
608 extracted DNA samples were quantified fluorometrically using Qubit 3.0 Fluorometer (Life
609 technologies), and a total of 1 ng of digested DNA was used for library construction.
610 Sequencing ready libraries were prepared using a KAPA Hyper Prep Kit (KAPA Biosystems,
611 London UK) according to the manufacturer's instructions. In order to reduce bias due to
612 amplification of DNA fragments by the polymerase chain reaction, as recommended by the
613 manufacturers, 17 cycles were used. Quality of the libraries were checked using a
614 Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies). Libraries were quantified by
615 Qubit 3.0 Fluorometer (Life Technologies) and qPCR using an NGS Library Quantification
616 Kit (KAPA Biosystems) and then normalised, pooled, diluted and denatured for paired end
617 sequencing using High Output 150 cycle run (2x 75 bp reads). Sequencing was performed
618 using NextSeq 500 (Illumina, Chesterford UK) in the Sainsbury Laboratory University of
619 Cambridge, UK, with 2x75 cycles of sequencing.
620
621 **DNaseI-SEQ Data processing**
622 Genome sequences were downloaded from Phytozome (v10) (Goodstein et al., 2012).
623 The following genome assemblies were used: Bdistachyon_283_assembly_v2.0;
624 Sbicolor_255_v2.0; Sitalica_164_v2; Zmays_284_AGPv3. Due to the lack of guidelines for
625 DNaseI-SEQ experiments in plants we followed the guidelines from the Encyclopaedia of
626 DNA Elements (ENCODE3). Reads were mapped to genomes using bowtie2 (Langmead
627 and Salzberg, 2012) and processed using samtools (Li et al., 2009) to remove those with a
628 MAPQ score <42. DHS were called using MACS2 (Feng et al., 2012) and the final set of
629 peak calls were determined using the irreproducible discovery rate (IDR) (Li and Dewey,
630 2011), calculated using the script batch_consistency_analysis.R
631 (https://github.com/modENCODE-
632 DCC/Galaxy/blob/master/modENCODE_DCC_tools/idr/batch-consistency-analysis.r). The
633 Irreproducible Discovery Rate framework adapted from the ENCODE 3 pipeline (Marinov et
634 al., 2014; https://sites.google.com/site/anshulkundaje/projects/idr) aims to measure the
635 reproducibility of findings by identifying the point (threshold) in which peaks are no longer
636 consistent across replicates.
637
638 **Quality metrics and identification of Digital Genomic Footprints (DGF)**

639 SPOT score (number of a subsample of mapped reads (5M) in DHS/Total number of
640 subsampled, mapped reads (5M) (John et al., 2011)) was calculated using BEDTools
641 (Quinlan and Hall, 2010) to determine the number of mapped reads possessing at least 1
642 bp overlap with a DHS site. Normalized Strand Cross-correlation coefficient (NSC) and
643 Relative Strand Cross-correlation coefficient (RSC) scores were calculated using SPP
644 (Kharchenko et al., 2008) and PCR bottleneck coefficient (PBC) was calculated using
645 BEDTools. To account for cutting bias associated with the DNaseI enzyme DNaseI-SEQ on
646 naked DNA was performed. These data were used to generate background signal profiles
647 and calculate the footprint log-likelihood ratio for each footprint using the R package
648 MixtureModel (Yardimci et al., 2014) such that those with low log likelihood ratios (FLR <0)
649 were removed. Digital Genomic Footprints (DGF) were identified using Wellington (Piper et
650 al., 2013) and differential DGF were identified using Wellington bootstrap (Piper et al., 2015).
651
652 **Data visualisation**
653 DHS and DGF sequences were loaded into and visualized in the Integrative Genomics
654 Viewer (Thorvaldsdóttir et al., 2013) and figures produced in Inkscape, plots were generated
655 with R package ggplot2 (Wickham, 2010) and figures depicting conservation of DGF or
656 motifs between orthologous sequences were generated using genoplotR (Guy et al., 2010).
657 Word clouds were created with the wordcloud R package (Fellows, 2012). TreeView images
658 were produced by processing DGF data using 'dnase_to_javatreeview.py' from pyDNAse
659 (Piper et al., 2013, 2015) and loaded into TreeView (Saldanha, 2004). Average cut density
660 plots were generated using the script 'dnase_average_profile.py' from pyDNase. Genomic
661 features were annotated and distribution calculated using PAVIS (Huang et al., 2013) and
662 plotted using ggplot2. Circular plots showing the distribution of ChIP-SEQ peaks, DHS and
663 DGF across the maize genome was generated using the R package circlize (Gu, 2014).
664
665 **DNAse cutting bias calculations and ChIP-SEQ analysis**
666 After sequencing, the number of DNA 6-mer centred at each DNase cleavage site
667 (between 3rd and 4th base) was counted and normalized by the total number of counts. Next,
668 DNA 6-mer frequencies were normalized by the frequencies of each DNA 6-mer in the
669 genome. The resulting background signal profile was used as input in the FootprintMixture.R
670 package (https://ohlerlab.mdc-berlin.de/software/FootprintMixture_109/) (Supplemental
671 Figure 2).
672 ChIP-SEQ peaks from 117 transcription factors were obtained from the pre-release maize
673 cistrome data collection (http://www.epigenome.cuhk.edu.hk/C3C4.html). Permutation tests

674 between ChIP peaks and DHS or DGF were performed using regioneR (Gel et al., 2016)
675 using 100 permutations.
676
677 ***de novo* motif prediction, motif scanning and enrichment testing**
678 *de novo* motif prediction was performed using findMotifsGenome.pl script from the
679 HOMER suite (Heinz et al., 2010) using digital genomic footprints (DGF) as input together
680 with the reference genome sequence for each species. Motif scanning was performed using
681 FIMO (Grant et al., 2011) with default parameters. To determine overrepresentation of TF
682 family motifs in samples hypergeometric tests were performed using R. The distribution of
683 each motif across different genomic features was obtained for each annotated motif by
684 dividing the number of hits in a particular feature by the total number of hits in the genome.
685
686 **Whole genome alignments, pairwise cross mapping of genomic features and variant**
687 **data processing**
688 To cross map genomic features between species, mapping files were generated
689 according to (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto)
690 using tools from the UCSC Genome Browser, including trfBig, faToNib, faSize, lavToPsl,
691 faSplit, axtChain, chainNet (Kent et al., 2002) and LASTZ (Harris, 2007). Briefly, whole
692 genome alignment was performed with LASTZ, matching alignments next to each other
693 were chained together using axtChain, sorted with axtSort, then netted together to form
694 larger blocks with chainNet. Genomic features where then mapped between genomes using
695 bnMapper (Denas et al., 2015). For the variant analysis on duons, *Z. mays* variant data
696 (Bukowski et al., 2018) was downloaded from
697 http://cbsusrv04.tc.cornell.edu/users/panzea/download.aspx?filegroupid=16 following
698 instructions. After downloading vcf files were annotated using SnpEff (Cingolani et al, 2012;
699 https://doi.org/10.4161/fly.19695) with the B73_RefGen_v4 genome assembly specifically
700 to allow identification of non-synonymous sites. A custom script was used to identify all four-
701 fold degenerate sites (FFDS) in the *Z. mays* genome. This bed file in turn was used to
702 identify which of the synonymous polymorphic sites were FFDS. Each polymorphic site had
703 its allele frequencies calculated. Putative *Z. mays* duons were identified by intersecting (with
704 bedtools intersect) the final DGF identified with exonic regions. These duons were then used
705 to extract only those exons within which a duon was found. These exons in turn had the
706 duon regions themselves subtracted to leave the exon region except the duon. This provided
707 the surrounding exonic sequences with which to compare to the duons. These two regions
708 were then intersected with the polymorphism data to identify both the number of occurrences

709 and allelic frequencies of polymorphic sites (FFDS and non-synonymous) within both the
710 duons and their surrounding exonic sequences.
711
712 **Accession numbers**
713 Methods for DNaseI digestion are on protocols.io
714 (dx.doi.org/10.17504/protocols.io.hdfb23n). Raw sequencing data and processed files are
715 deposited in Gene Expression Omnibus (GSE97369). For full methods, commands and
716 scripts, please see github (https://github.com/hibberd-lab/Burgess-Reyna_llorens-monocot-
717 DNase) and Figshare 10.6084/m9.figshare.7649450.

718 **Supplemental Legends**

719 **Supplemental Figure 1:** DNaseI digestion of nuclei for sequencing. Representative images
720 of digested samples separated on 2% (w/v) agarose gels by electrophoresis. (A) *S. bicolor*
721 whole leaf (WL); (B) *S. bicolor* Bundle Sheath (BS); (C) *Z. mays* WL; (D) *Z. mays* BS; (E) *B.*
722 *distachyon* WL; (F) *S. italica* WL; (G) *S. italica* BS. Each gel represents a separate biological
723 replicate, and the units of DNaseI used are illustrated above. Samples selected for
724 sequencing are indicated in red.

725

726 **Supplemental Figure 2:** Bias in DNaseI-SEQ cleavage. (A) TreeView diagrams illustrating
727 cut density around individual digital genomic footprint (DGF) predicted from performing
728 DNaseI-SEQ on deproteinated genomic DNA from each species. Each row represents an
729 individual DGF, cuts are coloured according to whether they align to the positive (red) or
730 negative (blue) strand and indicate increased cutting in a 100 bp window on either side of
731 the DGF. (B) Pearson correlation coefficient of DNAse I cleavage bias between *Z. mays*, *S.*
732 *bicolor*, *S. italica* and *B. distachyon*. (C) Schematic illustrating the process adopted to
733 determine DNaseI cutting bias and then normalize to allow digital genomic footprinting.

734

735 **Supplemental Figure 3:** Saturation analysis of footprints. Digital genomic footprints were
736 predicted from subsets (12.5, 25, 50, 75 and 100%) of uniquely mapped reads obtained
737 from DNAseI-SEQ of whole leaf samples in each species.

738

739 **Supplemental Figure 4:** Genome wide comparison of DGF and ChIP-SEQ peaks from 117
740 maize transcription factors. (A) Density plot of DHS, DGF, ChIP-SEQ peaks and intersecting
741 DGF/ChIP-SEQ peaks across the maize genome. The center of the plot shows a word cloud
742 representing transcription factor families in the ChIP-SEQ dataset. (B) Effect of shifting DHS
743 and DGF features from their original position on local z-scores as determined by permutation
744 tests between ChIP-SEQ peaks derived from 117 transcription factors. A total of 100
745 permutations were performed for each comparison. The sharper peak derived from shifting
746 the DGF indicates a higher sensitivity to position and therefore strong overlap with ChIP-
747 SEQ data. (C) Density plot depicting the distribution of ChIP-SEQ signals per kilobase (kb)
748 from the transcription start site (TSS) of *Z. mays.*

749

750 **Supplemental Figure 5:** Density plot depicting the distribution of DGF per kilobase (kb)
751 from the transcription start site (TSS) of *S. bicolor, Z. mays, S. italica* and *B. distachyon*
752 whole leaves.

753

754 **Supplemental Figure 6:** Nucleotide proportion of duons and surrounding exons used in the
755 substitution analysis for *Z. mays.* The frequency of each nucleotide was divided by the total
756 length to determine nucleotide proportions across duons as well as surrounding exon
757 sequences.

758

759 **Supplemental Figure 7:** Transcript abundance for genes in mesophyll and bundle sheath
760 cells associated with DHS and DGF in *S. bicolor*. (A) Cell preferential gene expression
761 profiles of highly abundant M and BS genes expressed as transcripts per million reads
762 (TPM). (B) Schematic representing DHS, DGF and DE DGF present in whole leaf (blue) and
763 BS (orange) of *S. bicolor*.

764

765 **Supplemental Figure 8:** Differential accessibility of broad regulatory regions in *S. bicolor* is
766 not sufficient for cell preferential gene expression. Percentage of differentially detected DHS
767 among BS and M specific genes in *S. bicolor* compared with randomly generated gene
768 samples (n=50).

769

770 **Supplemental Figure 9:** Representation of the C_4 pathway showing differentially accessible
771 DHS, DGF and cell specific DGF between whole leaf (blue) and bundle sheath (orange)
772 samples in *S. bicolor. CA; Carbonic Anhydrase, PEPC; Phosphoenolpyruvate carboxylase,*
773 *PPDK; Pyruvate, orthophosphate dikinase, MDH; Malate dehydrogenase, NADP-ME;*
774 *NADP-dependent malic enzyme, RBCS1A; Ribulose bisphosphate carboxylase small*
775 *subunit1A*, OAA; Oxaloacetate, Mal; Malate, PEP; Phosphoenolpyruvate, Pyr; Pyruvate,
776 Asp; Aspartate.

777

778 **Supplemental Figure 10:** Representation of the C_4 pathway showing differentially
779 accessible DHS, DGF and cell specific DGF between whole leaf (blue) and bundle sheath
780 (orange) samples in *S. italica. CA; Carbonic Anhydrase, PEPC; Phosphoenolpyruvate*
781 *carboxylase, PPDK; Pyruvate, orthophosphate dikinase, MDH; Malate dehydrogenase,*
782 *NADP-ME; NADP-dependent malic enzyme, RBCS1A; Ribulose bisphosphate carboxylase*
783 *small subunit1A*, OAA; Oxaloacetate, Mal; Malate, PEP; Phosphoenolpyruvate, Pyr;
784 Pyruvate, Asp; Aspartate.

785

786 **Supplemental Figure 11:** Representation of the C_4 pathway showing differentially
787 accessible DHS, DGF and cell specific DGF between whole leaf (blue) and bundle sheath

788 (orange) samples in *Z. mays. CA; Carbonic Anhydrase, PEPC; Phosphoenolpyruvate*
789 *carboxylase, PPDK; Pyruvate, orthophosphate dikinase, MDH; Malate dehydrogenase,*
790 *NADP-ME; NADP-dependent malic enzyme, RBCS1A; Ribulose bisphosphate carboxylase*
791 *small subunit1A*, OAA; Oxaloacetate, Mal; Malate, PEP; Phosphoenolpyruvate, Pyr;
792 Pyruvate, Asp; Aspartate.
793
794
795 **Supplemental Table 1:** Summary of DNAseI-SEQ Quality Metrics. Values include the
796 number of uniquely mapped reads with MAPQ scores >= 42 (NMAP), PCR bottleneck
797 Coefficient (PBC), Normalized Strand Cross-correlation Coefficient (NSC), Relative Strand
798 Cross-correlation Coefficient (RSC) the optimal number of peaks calculated by IDR method
799 (PEAKS), and the Signal Portion Of Tags (SPOT) for whole leaf and bundle sheath samples
800 from *B. distachyon, S. italica, S. bicolor* and *Z. mays.*
801
802 **Supplemental Table 2:** Summary Statistics for Genomic Features Identified in Whole Leaf
803 Samples. Including information about the number of DHS and DGF identified per sample,
804 and the number of genes that could be annotated with at least one genomic feature for *B.*
805 *distachyon, S. italica, S. bicolor* and *Z. mays.*
806
807 **Supplemental Table 3:** DNAseI cutting bias calculation summary for whole leaf and bundle
808 sheath data.
809
810 **Supplemental Table 4:** Transcription factors included in the ChIP-SEQ data analysis.
811
812 **Supplemental Table 5:** Summary statistics of DNAseI-SEQ analysis of Bundle Sheath
813 samples
814
815 **Supplemental Table 6:** Summary Statistics of Overlap between DHS in Whole Leaf and
816 Bundle Sheath Samples.
817
818 **Supplemental Table 7:** Summary Statistics of Overlap between DGF in Whole Leaf and
819 Bundle Sheath Samples.
820

821 **Supplemental Table 8:** Summary Statistics for Differential Digital Genomic Footprint
822 Calling. Including the total number of differential DGF (DE DGF) and the number of DE DGF
823 in DE genes for *S. italica, S. bicolor* and *Z. mays*.
824
825 **Supplemental Table 9:** Motifs mapped to genes of the C_4, CBB and C_2 cycles in *Z. mays,*
826 *S. bicolor, S. italica* for whole leaf and bundle sheath samples and in *B. distachyon* for whole
827 leaf samples
828
829 **Supplemental Table 10:** Hypergeometric tests for enrichment of individual motifs in *Z.*
830 *mays, S. bicolor, S. italica* for whole leaf and bundle sheath samples.
831
832 **Supplemental Table 11:** Hypergeometric tests for enrichment of cell specific individual
833 motifs in *Z. mays, S. bicolor, S. italica* for whole leaf and bundle sheath samples.
834
835 **Supplemental Table 12:** Number of genes in C_4, CBB and C_2 cycles annotated with a given
836 motif in *Z. mays, S. italica, S. bicolor* and *B. distachyon.*
837
838 **Supplemental Table 12:** Statistics for Cross Mapping of genomic features between *S.*
839 *bicolor, S. italica, Z. mays* and *B. distachyon.*
840
841 **Supplemental Table 13:** DGF conserved and occupied in *Z. mays, S. bicolor, S. italica* for
842 whole leaf and bundle sheath samples and in *B. distachyon* for whole leaf samples.
843
844 **Supplemental Table 14:** DGF in C_4 genes that are conserved between *Z. mays* and *S.*
845 *bicolor*.
846
847 **Supplemental Table 15:** DGF conserved in all four species.
848
849 **Supplemental Table 16:** Gene Ontology analysis on hyper-conserved DGF in whole leaf
850 samples of *S. italica, S. bicolor, Z. mays* and *B. distachyon*.
851
852
853 **Acknowledgements:** SJB was supported by the 3to4 grant from the EU and BB/I002243
854 from the BBSRC, IRL by CONACyT and BBSRC grant BB/L014130, SRS and PS by an
855 Advanced ERC grant 694733 Revolution to JMH, and KJ by a Gatsby Career Development

Fellowship. We would like to acknowledge Aslihan Karabacak for support in implementing the FootprintMixture package.

Contributions: SJB, I-RL and JMH conceptualised the experiments. 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 experiments and data analysis. PS extracted nuclei and performed DNAseI experiments on naked DNA, SRS undertook the variant analysis. SJB, I-RL, SRS and JMH wrote the manuscript and prepared the figures.

References

Bauwe, H., Hagemann, M., and Fernie, A.R. (2010). Photorespiration: players, partners and origin. Trends Plant Sci. 15: 330–6.

Bolduc, N. et al. (2012a). A Genome-Wide Regulatory Framework Identifies Maize Pericarp Color1 Controlled Genes. Plant Cell 27: 543–553.

Bolduc, N. and Hake, S. (2009). The Maize Transcription Factor KNOTTED1 Directly Regulates the Gibberellin Catabolism Gene *ga2ox1*. Plant Cell 21: 1647–1658.

Bolduc, N., Yilmaz, A., Mejia-Guerra, M.K., Morohashi, K., O'Connor, D., Grotewold, E., and Hake, S. (2012b). Unraveling the KNOTTED1 regulatory network in maize meristems. Genes Dev. 26: 1685–1690.

Borba, A. R., Serra, T. S., Górska, A., Gouveia, P., Cordeiro, A. M., Reyna-Llorens, I., Kneřová, J., Barros, P.M., Abreu, I.A., Oliveira, Hibberd, J.M., Saibo, N. (2018). Synergistic binding of bHLH transcription factors to the promoter of the maize NADP-ME gene used in C4 photosynthesis is based on an ancient code found in the ancestral C3 state. Mol. Bio and Evolution 35: 1690-1705.

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.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C. and Fan, L. (2017). Construction of the third-generation Zea mays haplotype map. GigaScience 7: gix134.

Chang YM, Liu WY, Shih ACC, Shen MN, Lu CH, Lu MYJ, Yang HW, Wang TY, Chen SC, Chen SM, 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. (2007). C_4 photosynthesis evolved in grasses via parallel adaptive genetic changes. Curr. Biol. 17: 1241–1247.

Christin, P. A., & Osborne, C. P. (2014). The evolutionary ecology of C_4 plants. New Phytologist 204: 765-781.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6: 80-92.

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). Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. BMC Genomics 16: 87.

Emms, D. M., Covshoff, S., Hibberd, J. M., & Kelly, S. (2016). Independent and parallel evolution of new genes by gene duplication in two origins of C4 photosynthesis provides new insight into the mechanism of phloem loading in C4 species. Mol Bio and Evol, *33*(7), 1796-1806.

916 Eveland, A.L., Goldshmidt, A., Pautler, M., Morohashi, K., Liseron-Monfils, C., Lewis, M.W.,
917 Kumari, S., Hiraga, S., Yang, F., Unger-Wallace, E. and Olson, A. (2014). Regulatory
918 modules controlling maize inflorescence architecture. Genome Research 24: 431-443.
919 Fellows, I. (2012). wordcloud: Word clouds. R Packag. version 2: 109.
920 Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-SEQ enrichment
921 using MACS. Nat. Protoc. 7: 1728–1740.
922 Furbank, R.T. (2011). Evolution of the C_4 photosynthetic mechanism: are there really three
923 C_4 acid decarboxylation types? J. Exp. Bot. 62: 3103–3108.
924 Furbank, R.T., Stitt, M., and Foyer, C.H. (1985). Intercellular compartmentation of sucrose
925 synthesis in leaves of *Zea mays* L. Planta 164: 172–178.
926 Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., & Malinverni, R.
927 (2015). regioneR: an R/Bioconductor package for the association analysis of genomic
928 regions based on permutation tests. Bioinformatics 32: 289-291.
929 Giuliano, G., Pichersky, E., Malik, V.S., Timko, M.P., Scolnik, P.A., and Cashmore, A.R.
930 (1988). An evolutionarily conserved protein binding sequence upstream of a plant light
931 regulated gene. Proc. Natl. Acad. Sci. 85: 7089–7093.
932 Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks,
933 W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative
934 platform for green plant genomics. Nucleic Acids Res. 40: D1178-86.
935 Gowik, U., Burscheidt, J., Akyildiz, M., Schlue, U., Koczor, M., Streubel, M., and Westhoff,
936 P. (2004). *cis*-regulatory elements for mesophyll-specific gene expression in the C_4
937 plant *Flaveria trinervia*, the promoter of the C_4 *PHOSPHOENOLPYRUVATE*
938 *CARBOXYLASE* gene. Plant Cell 16: 1077–1090.
939 Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given
940 motif. Bioinformatics 27: 1017–1018.
941 Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize implements and enhances
942 circular visualization in R. Bioinformatics 30: 2811-2812.
943 Guy, L., Roat Kultima, J., and Andersson, S.G.E. (2010). genoPlotR: comparative gene and
944 genome visualization in R. Bioinformatics 26: 2334–2335.
945 Harris, R.S. (2007). Improved pairwise alignment of genomic DNA. The Pennsylvania State
946 University. (PhD Thesis).
947 Hatch, M.D. (1987). C_4 photosynthesis: a unique elend of modified biochemistry, anatomy
948 and ultrastructure. Biochim. Biophys. Acta - Rev. Bioenerg. 895: 81–106.
949 He, H.H., Meyer, C.A., Chen, M.W., Zang, C., Liu, Y., Rao, P.K., Fei, T., Xu, H., Long, H.,
950 Liu, X.S. and Brown, M. (2014). Refined DNase-seq protocol and data analysis reveals
951 intrinsic bias in transcription factor footprint identification. Nature Methods 11: 73.
952 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., and Laslo, P. (2010). Simple
953 combinations of lineage-determining transcription factors prime cis-regulatory elements
954 required for macrophage and B cell identities. Mol. Cell. 38: 576-89
955 Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P.,
956 Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., Fields, S., and
957 Stamatoyannopoulos, J.A. (2009). Global mapping of protein-DNA interactions in vivo
958 by digital genomic footprinting. Nat. Methods 6: 283–9.
959 Hibberd, J.M. and Covshoff, S. (2010). The Regulation of Gene Expression Required for C_4
960 Photosynthesis. Annu. Rev. Plant Biol. 61: 181–207.
961 Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D., & Li, L. (2013). PAVIS: a tool for
962 Peak Annotation and Visualization. Bioinformatics 29: 3097-3099.
963 Jeon, J.-S., Lee, S., Jung, K.-H., Jun, S.-H., Kim, C., and An, G. (2000). Tissue-Preferential
964 Expression of a Rice α-Tubulin Gene, *OsTubA1*, Mediated by the First Intron. Plant
965 Physiol. 123: 1005–1014.

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.

John, C. R., Smith-Unna, R. D., Woodfield, H., Covshoff, S., & Hibberd, J. M. (2014). Evolutionary convergence of cell-specific gene expression in independent lineages of C4 grasses. Plant Phys., *165*(1), 62-75.

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.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R., Tan, G. and Baranasic, D.,. (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research 46 (D1): D260-D266.

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.

Kozaki, A., Hake, S., & Colasanti, J. (2004). The maize ID1 flowering time regulator is a zinc finger protein with novel DNA binding properties. Nucleic Acids Research, 32: 1710-1720.

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.

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

Li, B. and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

Li, C., Qiao, Z., Qi, W., Wang, Q., Yuan, Y., Yang, X., Tang, Y., Mei, B., Lv, Y., Zhao, H. and Xiao, H. (2015). Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. The Plant Cell 27: 532–545.

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.

de Lucas, M., Pu, L., Turco, G., Gaudinier, A., Morao, A.K., Harashima, H., Kim, D., Ron, M., Sugimoto, K., Roudier, F., and Brady, S.M. (2016). Transcriptional Regulation of Arabidopsis Polycomb Repressive Complex 2 Coordinates Cell-Type Proliferation and Differentiation. Plant Cell 28: 2616 -2631.

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.

Maher, K.A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D.A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M.W. and Queitsch, C. (2018). Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. The Plant Cell 30: 15-36.

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.

Markelz, N. H., Costich, D. E., & Brutnell, T. P. (2003). Photomorphogenic responses in maize seedling development. *Plant Physiology*, *133*(4), 1578-1591.

1018 Martin, A. and Orgogozo, V. (2013). The Loci of repeated evolution: a catalog of genetic
1019 hotspots of phenotypic variation. Evolution 67: 1235–50.
1020 Natarajan, A., Yardımcı, G.G., Sheffield, N.C., Crawford, G.E., and Ohler, U. (2012).
1021 Predicting cell-type–specific gene expression from regions of open chromatin. Genome
1022 Research 22: 1711–1722.
1023 Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman,
1024 R.E., John, S., Sandstrom, R., Johnson, A.K. and Maurano, M.T. (2012). An expansive
1025 human regulatory lexicon encoded in transcription factor footprints. Nature 489: 83-90.
1026 Niu, X., Helentjaris, T., & Bate, N. J. (2002). Maize ABI4 binds coupling element1 in abscisic
1027 acid and sugar response genes. The Plant Cell, 14: 2565-2575.
1028 O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M.,
1029 Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the
1030 Regulatory DNA Landscape. Cell 165: 1280–1292.
1031 Pajoro, A., Madrigal, P., Muiño, J.M., Matus, J.T., Jin, J., Mecchia, M.A., Debernardi, J.M.,
1032 Palatnik, J.F., Balazadeh, S., Arif, M. and Ó'Maoiléidigh, D.S. (2014). Dynamics of
1033 chromatin accessibility and gene regulation by MADS-domain transcription factors in
1034 flower development. Genome Biology 15: R41.
1035 Patel, M., Corey, A.C., Yin, L.P., Ali, S.J., Taylor, W.C., and Berry, J.O. (2004). Untranslated
1036 regions from C_4 Amaranth *AhRbcS1* mRNAs confer translational enhancement and
1037 preferential bundle sheath cell expression in transgenic C_4 *Flaveria bidentis*. Plant
1038 Physiol. 136: 3550–3561.
1039 Pautler, M., Eveland, A.L., LaRue, T., Yang, F., Weeks, R., Lunde, C., Je, B. Il, Meeley, R.,
1040 Komatsu, M., Vollbrecht, E., Sakai, H., and Jackson, D. (2015). *FASCIATED EAR4*
1041 Encodes a bZIP Transcription Factor That Regulates Shoot Meristem Size in Maize.
1042 Plant Cell 27: 104-120.
1043 Piper, J., Assi, S.A., Cauchy, P., Ladroue, C., Cockerill, P.N., Bonifer, C., and Ott, S. (2015).
1044 Wellington-bootstrap: differential DNase-seq footprinting identifies cell-type
1045 determining transcription factors. BMC Genomics 16: 1000.
1046 Piper, J., Elze, M.C., Cauchy, P., Cockerill, P.N., Bonifer, C., and Ott, S. (2013). Wellington:
1047 a novel method for the accurate identification of digital genomic footprints from DNase-
1048 seq data. Nucleic Acids Research 41: e201–e201.
1049 Quinlan, A.R. and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing
1050 genomic features. Bioinformatics 26: 841–842.
1051 Reyna-Llorens, I., Burgess, S.J., Reeves, G., Singh, P., Stevenson, S.R., Williams, B.P.,
1052 Stanley, S., and Hibberd, J.M. (2018). Ancient duons may underpin spatial patterning
1053 of gene expression in C_4 leaves. Proc. Natl. Acad. Sci. 115:1931-1936.
1054 Sage, R. (2004). The evolution of C_4 photosynthesis. New Phytol. 161: 341–370.
1055 Sage, R.F., Christin, P.-A., and Edwards, E.J. (2011). The C_4 plant lineages of planet Earth.
1056 J. Exp. Bot. 62: 3171–3181.
1057 Sage, R.F., Sage, T.L., and Kocacinar, F. (2012). Photorespiration and the evolution of C_4
1058 photosynthesis. Annu. Rev. Plant Biol. 63: 19–47.
1059 Sage, R.F. and Zhu, X.G. (2011). Exploiting the engine of C_4 photosynthesis. J. Exp. Bot.
1060 62: 2989–3000.
1061 Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data.
1062 Bioinformatics. 20: 3246-3248.
1063 Schäffner, A. R., & Sheen, J. (1991). Maize rbcS promoter activity depends on sequence
1064 elements not found in dicot rbcS promoters. The Plant Cell 3: 997-1012.
1065 Sharwood, R.E., Ghannoum, O., Kapralov, M. V, Gunn, L.H., and Whitney, S.M. (2016).
1066 Temperature responses of Rubisco from Paniceae grasses provide opportunities for
1067 improving C_3 photosynthesis. Nat. Plants 2: 16186.
1068 Sheen, J. (1999). C_4 gene expression. Ann. Rev Plant Physiol. Plant Mol Biol 50: 187–217.

1069 Sparks, E.E. et al. (2017). Establishment of Expression in the SHORTROOT-SCARECROW
1070 Transcriptional Cascade through Opposing Activities of Both Activators and
1071 Repressors. Dev. Cell 39: 585–596.
1072 Stergachis, A.B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A.P., Zhang, M., Byron,
1073 R., Canfield, T., Stelhing-Sun, S., Lee, K. and Thurman, R.E. (2014). Conservation of
1074 trans-acting circuitry during mammalian regulatory evolution. Nature, 515: 365-370.
1075 Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A.,
1076 Ziegler, S., LeProust, E.M., Akey, J.M., Stamatoyannopoulos, J.A. (2013). Exonic
1077 transcription factor binding directs codon choice and affects protein evolution. Science
1078 342: 1367–72.
1079 Sullivan, A.M., Arsovski, A.A., Lempe, J., Bubb, K.L., Weirauch, M.T., Sabo, P.J.,
1080 Sandstrom, R., Thurman, R.E., Neph, S., Reynolds, A.P. and Stergachis, A.B. (2014).
1081 Mapping and dynamics of regulatory DNA and transcription factor networks in A.
1082 thaliana. Cell Reports 8: 2015-2030.
1083 Taniguchi, M., Izawa, K., Ku, M.S.B., Lin, J.H., Saito, H., Ishida, Y., Ohta, S., Komari, T.,
1084 Matsuoka, M., and Sugiyama, T. (2000). Binding of cell type-specific nuclear proteins
1085 to the 5' flanking region of maize C_4 phosphoenolpyruvate carboxylase gene confers its
1086 differential transcription in mesophyll cells. Plant Mol. Biol. 44: 543–557.
1087 Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer
1088 (IGV): high-performance genomics data visualization and exploration. Briefings in
1089 Bioinformatics 14: 178–192.
1090 Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield,
1091 N.C., Stergachis, A.B., Wang, H., Vernot, B. and Garg, K. (2012). The accessible
1092 chromatin landscape of the human genome. Nature 489: 75-82.
1093 Tolbert, N.E. (1971). Microbodies - peroxisomes and glyoxysomes. Annu. Rev. Plant
1094 Physiol. 22: 45–74.
1095 Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. (2006). Evolution of alternative
1096 transcriptional circuits with identical logic. Nature 443: 415–420.
1097 Vollbrecht, E., Springer, P. S., Goh, L., Buckler IV, E. S., & Martienssen, R. (2005).
1098 Architecture of floral branch systems in maize and related grasses. Nature 436: 1119-
1099 1126.
1100 Viret, J.F., Mabrouk, Y., and Bogorad, L. (1994). Transcriptional photoregulation of cell-type
1101 preferred expression of maize *RbcS-m3*: 3' and 5' sequences are involved. Proc. Natl.
1102 Acad. Sci. 91: 8577–8581.
1103 Wickham, H. (2010). ggplot2: elegant graphics for data analysis. J. Stat. Softw. 35: 65-68.
1104 Williams, B.P., Burgess, S.J., Reyna-Llorens, I., Knerova, J., Aubry, S., Stanley, S., and
1105 Hibberd, J.M. (2016). An untranslated *cis*-element regulates the accumulation of
1106 multiple C_4 enzymes in *Gynandropsis gynandra* mesophyll cells. Plant Cell 28: 454–
1107 465.
1108 Xing, K. and He, X. (2015). Reassessing the "Duon" Hypothesis of Protein Evolution. Mol.
1109 Biol. Evol. 32: 1056–1062.
1110 Xu, T., Purcell, M., Zucchi, P., Helentjaris, T., and Bogorad, L. (2001). TRM1, a YY1-like
1111 suppressor of *RbcS-m3* expression in maize mesophyll cells. Proc. Natl. Acad. Sci. 98:
1112 2295–2300.
1113 Yardımcı, G. G., Frank, C. L., Crawford, G. E., & Ohler, U. (2014). Explicit DNase sequence
1114 bias modelling enables high-resolution transcription factor footprint detection. Nucleic
1115 acids research, 42: 11865-11878.
1116 Yu, C.P., Chen, S.C.C., Chang, Y.M., Liu, W.Y., Lin, H.H., Lin, J.J., Chen, H.J., Lu, Y.J.,
1117 Wu, Y.H., Lu, M.Y.J. and Lu, C.H. (2015b). Transcriptome dynamics of developing
1118 maize leaves and genome wide prediction of *cis*-elements and their cognate
1119 transcription factors. Proc. Natl. Acad. Sci. 112: E2477–E2486.

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

Zhang, T., Marand, A.P., and Jiang, J. (2016). PlantDHS: a database for DNase I hypersensitive sites in plants. Nucleic Acids Res. 44: D1148–D1153.

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

Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2012b). Genome-Wide Identification of Regulatory DNA Elements and Protein-Binding Footprints Using Signatures of Open Chromatin in Arabidopsis. The Plant Cell 24: 2719–2731.

1130 **Figure Legends:**

1131 **Figure 1: Transcription factor binding atlas for whole leaf samples of four grasses.**
1132 (A) Schematic of phylogenetic relationship between species analysed. The two independent
1133 origins of C_4 photosynthesis are highlighted with black and white circles (figure not drawn to
1134 scale). (B) Summary of sampling and the total number of DNaseI-hypersensitive sites (DHS)
1135 and Digital Genomic Footprints (DGF) identified across all four species. (C) TreeView
1136 diagrams illustrating cut density around individual digital genomic footprint (DGF). Each row
1137 represents an individual DGF, cuts are coloured according to whether they align to the
1138 positive (red) or negative (blue) strand and indicate increased cutting in a 50 bp window on
1139 either side of the DGF. The total number of DGF per sample is shown at the bottom. (D)
1140 Representation of DNaseI-SEQ data from *S. bicolor*, depicting gene (grey), DHS (light blue),
1141 DGF (orange) and DNaseI cut density (dark blue) at three scales: genome wide, with
1142 chromosome number and position indicated (top), chromosomal (second level) and kilobase
1143 genomic region (third level). Between each level the expanded area is denoted by dashed
1144 lines. (E) Pie-chart representing the distribution of DHS among genomic features. Promoters
1145 are defined as sequence up to 2000 base pairs (bp) upstream of the transcriptional start
1146 site, downstream represent regions 1000 downstream the transcription termination site
1147 while intergenic represent > 1000 bp downstream the transcription termination site until the
1148 next promoter region. (F) Bar chart representing number of DHS in genic and promoter
1149 regions.

1150

1151 **Figure 2: Digital genomic footprints in whole leaves of four grasses.** (A) DNA motifs
1152 from previous studies in maize ([1]Yu et al., 2015; [2]Kozaki et al., 2004; [3]Niu et al., 2002;
1153 [4]Vollbrecht et al., 2005; [5]Eveland et al., 2014; [6]Li et al., 2015; [7]Bolduc et al., 2012) were
1154 detected in whole leaves and bundle sheath strands from maize. (B) Pie-chart representing
1155 the distribution of DGF among genomic features. Promoters are defined as sequence up to
1156 2000 base pairs (bp) upstream of the transcriptional start site, downstream represent
1157 regions 1000 downstream the transcription termination site while intergenic represent >
1158 1000 bp downstream the transcription termination site until the next promoter region. (C)
1159 Bar chart representing number of DGF in genic and promoter regions. (D) Polymorphic sites
1160 per kb in duons and surrounding exons at FourFold Degenerate Sites (FFDS) and non-
1161 synonymous sites. Chi-squared tests indicate reduced rates of mutation at FFDS than
1162 expected by chance.

1163

Figure 3: Characterisation of the DNA binding landscape in the C_4 Bundle Sheath. (A) Schematic showing that due to their negative distribution below the background signal derived from reads mapping to the genome, footprints associated with low abundance cells such as the Bundle Sheath (BS) are unlikely to 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) DHS distribution across the maize *PEPC* gene in BS and WL samples. (E) Transcript abundance expressed as transcripts per million reads (TPM) of three co-linear genes on chromosome seven of sorghum - C_4 *MDH* (Sobic.007G166300), non C_4 MDH (non C_4, Sobic.007G166200) and an uncharacterised NAC domain protein (Sobic.007G166100) in bundle sheath and mesophyll cells. Schematic of these co-linear genes from *S. bicolor*, depicting three classes of alterations to DNA accessibility and transcription factor binding to genes that are differentially expressed between whole leaf and bundle sheath cells. (F) Whole leaf (blue) and bundle sheath (orange) DHS, DGF and differentially (DE) enriched DGF, as determined by the Wellington Bootstrap algorithm, are depicted. Regions where a DHS was identified in one sample but not another are indicated by dashed boxes.

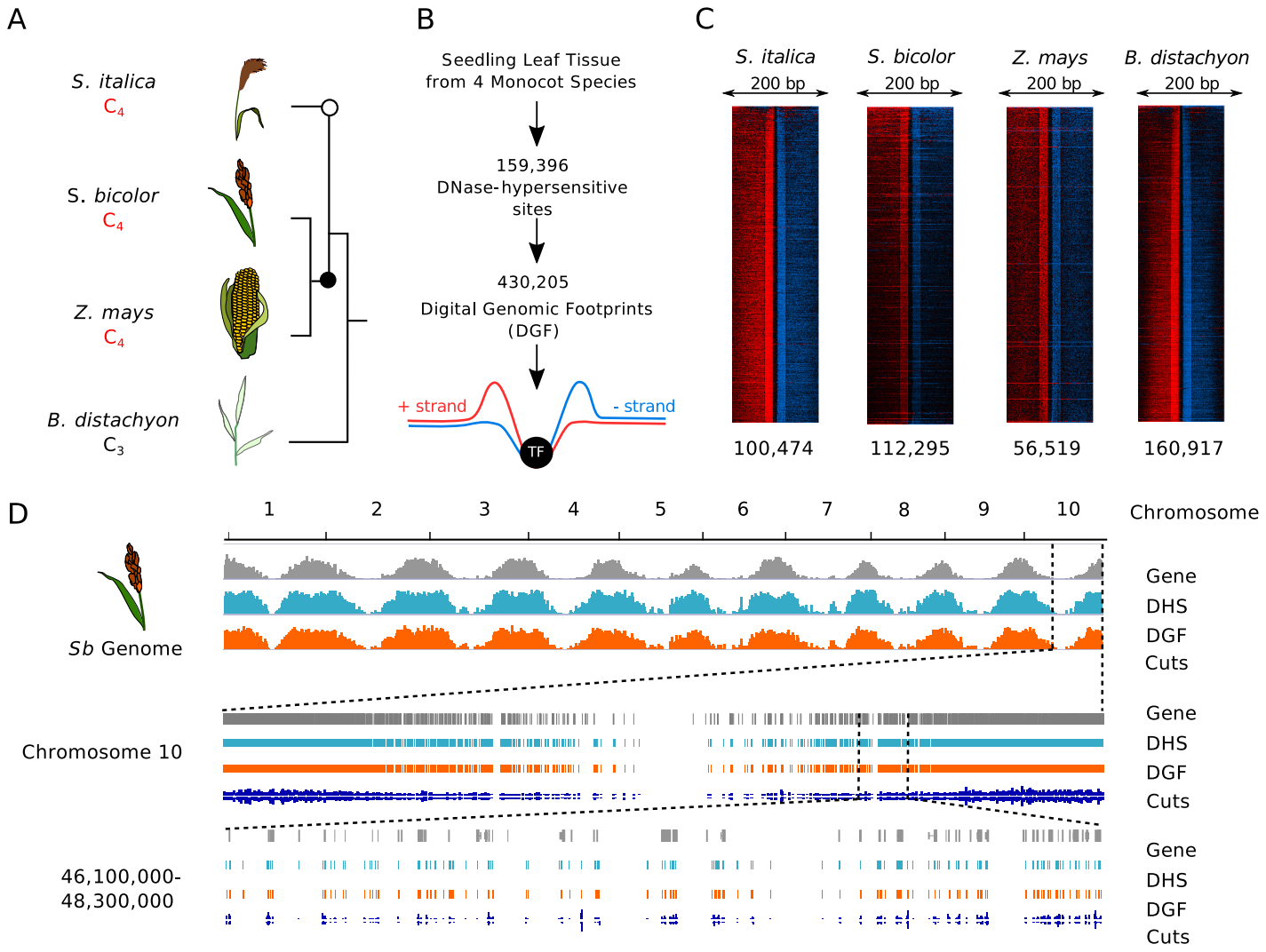
Figure 4: Cistromes associated with cell specific gene expression in C_4 grasses. (A) Number of previously reported motifs as well as those defined *de novo* in the grasses. (B) Density plots depicting average DNaseI activity on positive (red) and negative (blue) strands centred around a *de novo* motif. (C) Bar chart depicting percentage of DGF annotated with known or *de novo* motifs. (D) Comparison of transcription factor motif prevalence in Whole Leaf (WL) samples from *S. italica, Z. mays, B. distachyon* compared with *S. bicolor*. Word clouds depict frequency of motifs associated with transcription factor families, with larger names more abundant. Scatter plots compare frequency of transcription factor motifs within DGF, ranked from low (most abundant) to high (least abundant). Correlation between samples is indicated as Kendall's Tau coefficient (τ). (E) Comparison of transcription factor motif prevalence in BS enriched and WL enriched DGF from *S. bicolor*, as in (D), word clouds depict frequency of motifs associated with transcription factor families and plots compare frequency of transcription factor motifs within DGF ranked from low to high. Similarly scatter plots compare transcription factor motif prevalence in BS enriched and whole leaf enriched DGF from *S. bicolor*. (F) Venn diagram showing enriched motifs for each cell type in all three C_4 species.

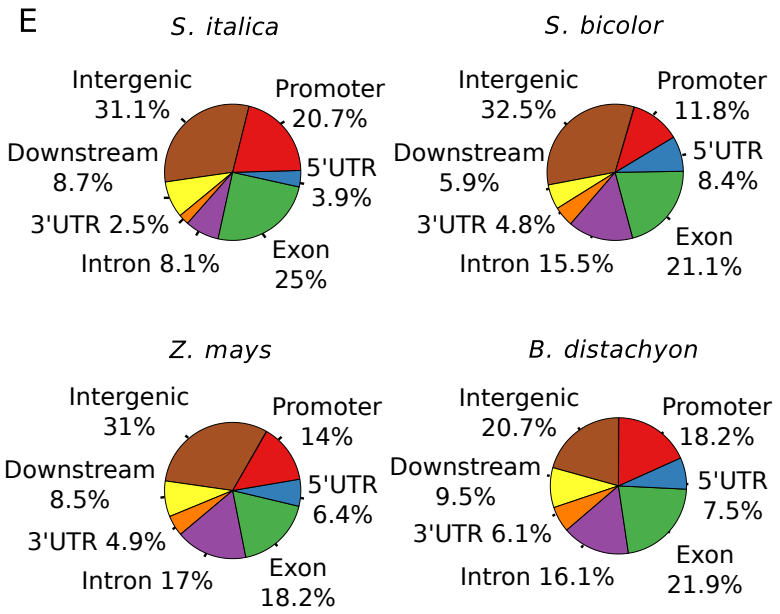
1199

1200 **Figure 5: *Cis*-elements show high rates of turnover and mobility in grasses.** (A)
1201 Scenarios for DGF conservation between species. Reads derived from DNaseI cuts are
1202 depicted in grey, DGF that are both conserved and occupied between species by red, and
1203 DGF that are conserved but unoccupied by blue shading. (B) Bar-plot representing pairwise
1204 comparisons of DGF occupancy. (C) Schematic depicting the position of a transcription
1205 factor motif consistently associated with the bundle sheath enriched *TRANSKETOLASE*
1206 (*TKL)* gene in *S. bicolor, Z. mays, S. italica* and C_3 *B. distachyon*. The position of motifs
1207 conserved between orthologous genes is depicted by solid lines and orange) and varies
1208 between species.

1209

1210 **Figure 6: Hyper-conserved *cis*-elements in grasses recruited into C_4 photosynthesis.**
1211 (A) Conservation of regulation in C_4 and Calvin Benson Bassham cycle genes following the
1212 divergence of *Z. mays* and *S. bicolor*. The number of carbon atoms (red dots) and metabolite
1213 flow (red dashed line) between mesophyll (grey) and bundle sheath (orange) cells are
1214 illustrated along with the degree of conservation of DGF associated with BS strands. (B)
1215 Conservation of DGF occupancy in grasses across evolutionary time. Results are depicted
1216 for whole leaf (WL - blue) and bundle sheath (BS - orange) DGF. The asterisk indicates 41
1217 DGF that are conserved in the BS of the C_4 species but are also found in whole leaves of *B.*
1218 *distachyon*). Pie-charts display the distribution of conserved and occupied DGF for whole
1219 leaf and BS strands. Promoters are defined as sequence up to 2000 base pairs (bp)
1220 upstream of the transcriptional start site, downstream represent regions 1000 downstream
1221 the transcription termination site while intergenic represent > 1000 bp downstream the
1222 transcription termination site until the next promoter region.


A
S. italica
C4
S. bicolor
C4
Z. mays
C4
B. distachyon
C3
B
Seedling Leaf Tissue
from 4 Monocot Species
159,396
DNase-hypersensitive
sites
430,205
Digital Genomic Footprints
(DGF)
+ strand
- strand
TF
C
S. italica
S. bicolor
Z. mays
B. distachyon
200 bp
100,474
112,295
56,519
160,917
D
1
2
3
4
5
6
7
8
9
10
Chromosome
Sb Genome
Gene
DHS
DGF
Cuts
Chromosome 10
46,100,000-
48,300,000


E
S. italica
Intergenic
31.1%
Promoter
20.7%
Downstream
8.7%
5'UTR
3.9%
3'UTR 2.5%
Intron 8.1%
Exon
25%
S. bicolor
Intergenic
32.5%
Promoter
11.8%
Downstream
5.9%
5'UTR
8.4%
3'UTR 4.8%
Intron 15.5%
Exon
21.1%
Z. mays
Intergenic
31%
Promoter
14%
Downstream
8.5%
5'UTR
6.4%
3'UTR 4.9%
Intron 17%
Exon
18.2%
B. distachyon
Intergenic
20.7%
Promoter
18.2%
Downstream
9.5%
5'UTR
7.5%
3'UTR 6.1%
Intron 16.1%
Exon
21.9%

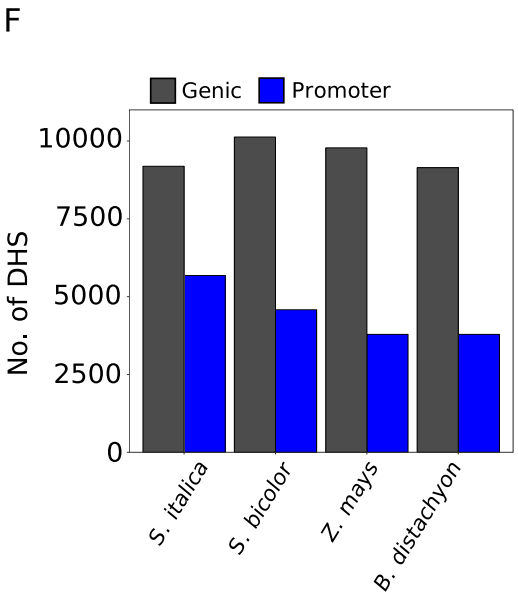

F
Genic
Promoter
No. of DHS
10000
7500
5000
2500
0
S. italica
S. bicolor
Z. mays
B. distachyon

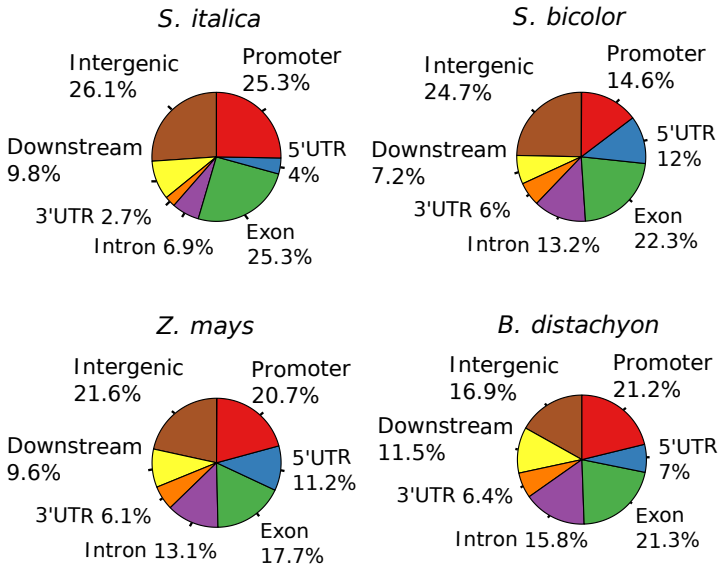
Figure 1

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 C4 photosynthesis are highlighted with black and white circles (figure not drawn to scale). (B) Summary of sampling and the total number of DNaseI-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 DNaseI-SEQ data from S. bicolor, depicting gene (grey), DHS (light blue), DGF (orange) and DNase-I cut density (dark blue) at three scales: genome wide, with chromosome number and position indicated (top), chromosomal (second level) and kb genomic region (third level). Between each level the expanded area is denoted by dashed lines. (E) Pie-chart representing the distribution of DHS among genomic features. Promoters are defined as sequence up to 2000 base pairs (bp) upstream of the transcriptional start site, downstream represent regions 1000 downstream the transcription termination site while intergenic represent > 1000 bp downstream the transcription termination site until the next promoter region. (F) Bar chart representing number of DHS in genic and promoter regions.

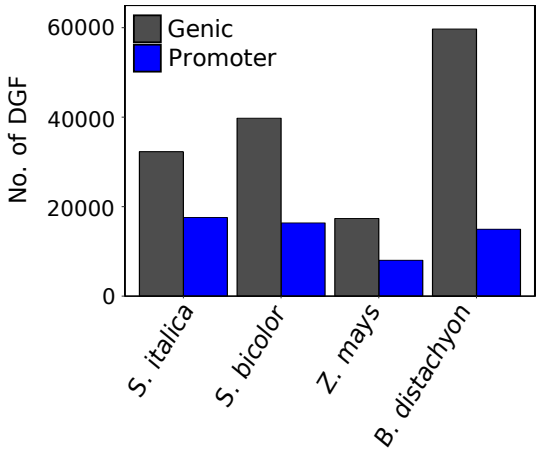
A

Motif ID	Transcription factor	Validation	Number of motif hits in DGF	
			Whole leaf	Bundle sheath
ID0008	EF2	EMSA[1]	359	1024
ID0009	ABI4/ERF11	EMSA[1]	1741	5425
ID0026	CIB1	EMSA[1]	621	1685
ID0037	WRI1	EMSA[1]	3635	13445
ID0061	HOX16	EMSA[1]	118	259
ID0062	RS2	EMSA[1]	261	581
ID0063	GBF6	EMSA[1]	1083	3476
ID0064	GLK1/GLK2	EMSA[1]	140	285
ID0439	MYB59	EMSA[1]	452	897
MA0120.1	ID1	SELEX[2]	340	802
MA0123.1	abi4	SELEX[3]	2505	9610
MA1416.1	RAMOSA1	ChIP-seq[4,5]	705	1253
MA1417.1	O2	ChIP-seq[6]	300	666
NA	KNT1	ChIP-seq[7]	2207	5327

B



C



D

Polymorphic sites per kb

	Surrounding exon	Duon	Chi Squared
Four Fold Degenerate Sites (FFDS)			
Observed	33.41	30.21	33.52, p= 7.04e-9*
Expected	33.16	33.16	
Non-synonomous			
Observed	20.96	20.27	2.48, p= 0.12
Expected	20.90	20.90	

Figure 2

Figure 2: Digital genomic footprints in whole leaves of four grasses. (A) DNA motifs from previous studies in maize (1Yu et al., 2015; 2Kozaki et al., 2004; 3Niu et al., 2002; 4Vollbrecht et al., 2005; 5Eveland et al., 2014; 6Li et al., 2015; 7Bolduc et al., 2012) were detected in whole leaves and bundle sheath strands from maize. (B) Pie-chart representing the distribution of DGF among genomic features. Promoters are defined as sequence up to 2000 base pairs (bp) upstream of the transcriptional start site, downstream represent regions 1000 downstream the transcription termination site while intergenic represent > 1000 bp downstream the transcription termination site until the next promoter region. (C) Bar chart representing number of DGF in genic and promoter regions. (D) Polymorphic sites per kb in duons and surrounding exons at FourFold Degenerate Sites (FFDS) and non-synonymous sites. Chi-squared tests indicate reduced rates of mutation at FFDS than expected by chance.

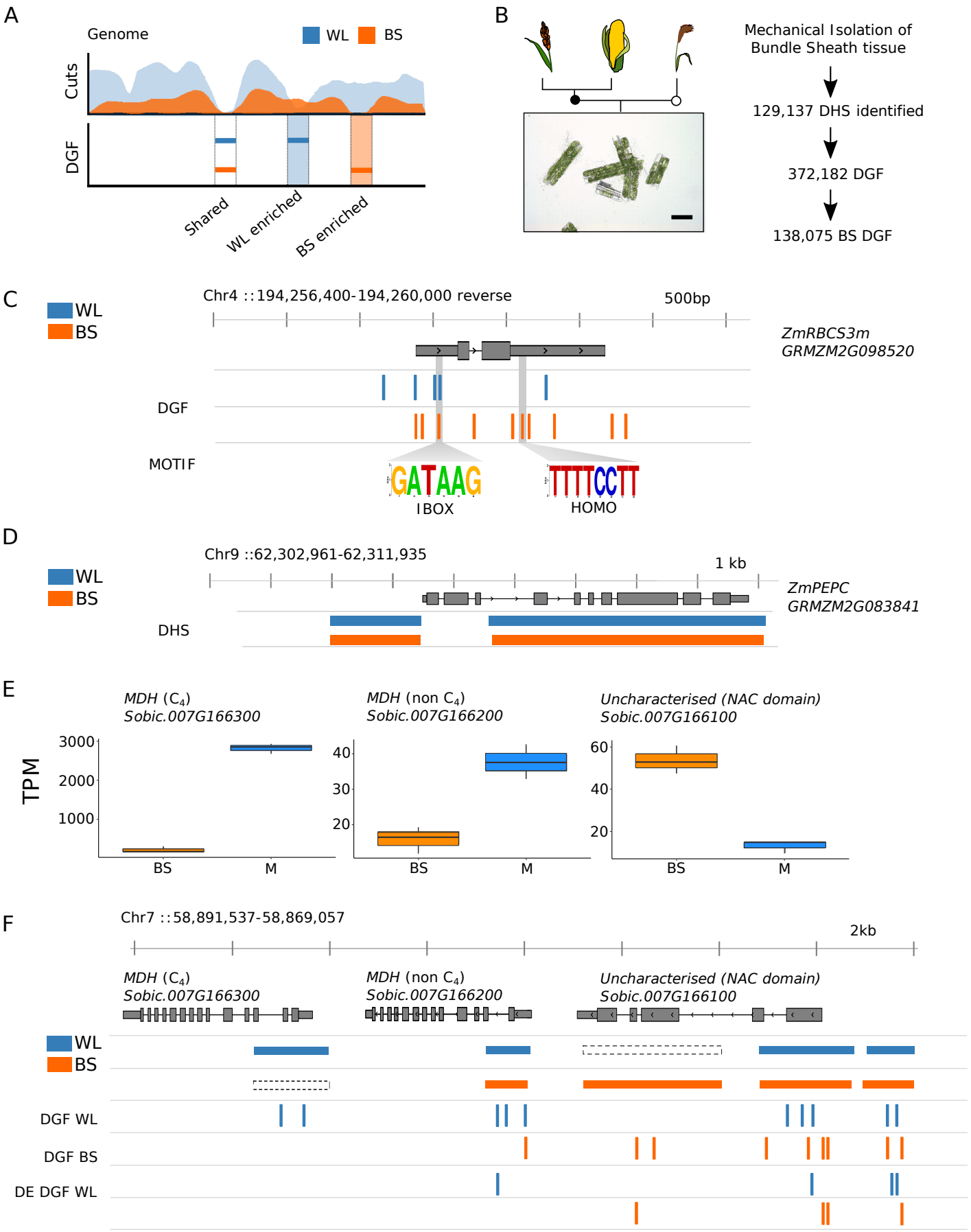

A
Genome
WL
BS
Cuts
DGF
Shared
WL enriched
BS enriched
B
Mechanical Isolation of Bundle Sheath tissue
129,137 DHS identified
372,182 DGF
138,075 BS DGF
C
WL
BS
Chr4 ::194,256,400-194,260,000 reverse
500bp
ZmRBCS3m
GRMZM2G098520
DGF
MOTIF
GATAAG
IBOX
TTTTCCTT
HOMO
D
WL
BS
Chr9 ::62,302,961-62,311,935
1 kb
ZmPEPC
GRMZM2G083841
DHS
E
MDH (C4)
Sobic.007G166300
MDH (non C4)
Sobic.007G166200
Uncharacterised (NAC domain)
Sobic.007G166100
TPM
3000
2000
1000
40
30
20
60
40
20
BS
M
F
Chr7 ::58,891,537-58,869,057
2kb
MDH (C4)
Sobic.007G166300
MDH (non C4)
Sobic.007G166200
Uncharacterised (NAC domain)
Sobic.007G166100
WL
BS
DGF WL
DGF BS
DE DGF WL

Figure 3

Figure 3: Characterisation of the DNA binding landscape in the C_4 Bundle Sheath. (A) Schematic showing that due to their negative distribution below the background signal derived from reads mapping to the genome, footprints associated with low abundance cells such as the Bundle Sheath (BS) are unlikely to be detected from whole leaf (WL) samples. (B) Bundle sheath isolation for DNaseI-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) DHS distribution across the maize PEPC gene in BS and WL samples. (E) Transcript abundance expressed as transcripts per million reads (TPM) of three co-linear genes on chromosome seven of sorghum - C_4 *MDH* (Sobic.007G166300), non C_4 *MDH* (non C_4, Sobic.007G166200) and an uncharacterised NAC domain protein (Sobic.007G166100) in bundle sheath and mesophyll cells. Schematic of these co-linear genes from *S. bicolor*, depicting three classes of alterations to DNA accessibility and transcription factor binding to genes that are differentially expressed between whole leaf and bundle sheath cells. (F) Whole leaf (blue) and bundle sheath (orange) DHS, DGF and differentially (DE) enriched DGF, as determined by the Wellington Bootstrap algorithm, are depicted. Regions where a DHS was identified in one sample but not another are indicated by dashed boxes.

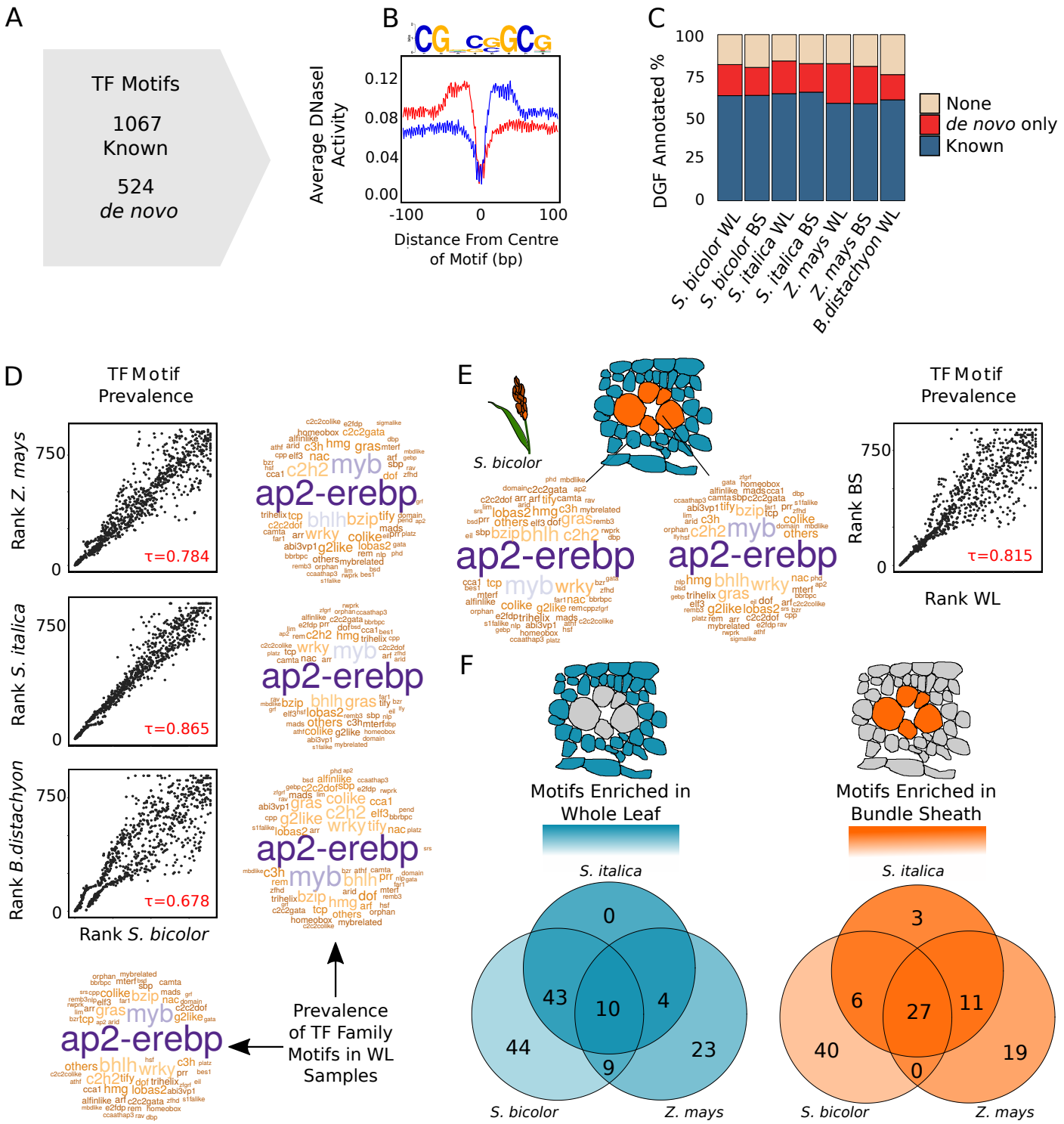

A
TF Motifs
1067
Known
524
de novo
B
Average DNaseI Activity
0.12
0.08
0.04
0.00
-100
0
100
Distance From Centre of Motif (bp)
C
DGF Annotated %
100
75
50
25
0
None
de novo only
Known
S. bicolor WL
S. bicolor BS
S. italica WL
S. italica BS
Z. mays WL
Z. mays BS
B.distachyon WL
D
TF Motif Prevalence
Rank Z. mays
τ=0.784
Rank S. italica
τ=0.865
Rank B.distachyon
τ=0.678
Rank S. bicolor
Prevalence of TF Family Motifs in WL Samples
E
S. bicolor
TF Motif Prevalence
Rank BS
τ=0.815
Rank WL
F
Motifs Enriched in Whole Leaf
Motifs Enriched in Bundle Sheath
S. italica
0
43
10
4
44
9
23
S. bicolor
Z. mays
S. italica
3
6
27
11
40
0
19
S. bicolor
Z. mays

Figure 4

Figure 4: Cistromes associated with cell specific gene expression in C_4 grasses. (A) Number of previously reported motifs as well as those defined *de novo* in the grasses. (B) Density plots depicting average DNase-I activity on positive (red) and negative (blue) strands centred around a *de novo* motif. (C) Bar chart depicting percentage of DGF annotated with known or *de novo* motifs. (D) Comparison of transcription factor motif prevalence in Whole Leaf (WL) samples from *S. italica, Z. mays, B. distachyon* compared with *S. bicolor*. Word clouds depict frequency of motifs associated with transcription factor families, with larger names more abundant. Scatter plots compare frequency of transcription factor motifs within DGF, ranked from low (most abundant) to high (least abundant). Correlation between samples is indicated as Kendall's Tau coefficient (τ). (E) Comparison of transcription factor motif prevalence in BS enriched and WL enriched DGF from *S. bicolor*, as in (D), word clouds depict frequency of motifs associated with transcription factor families and plots compare frequency of transcription factor motifs within DGF ranked from low to high. Similarly scatter plots compare transcription factor motif prevalence in BS enriched and whole leaf enriched DGF from *S. bicolor*. (F) Venn diagram showing enriched motifs for each cell type in all three C_4 species.

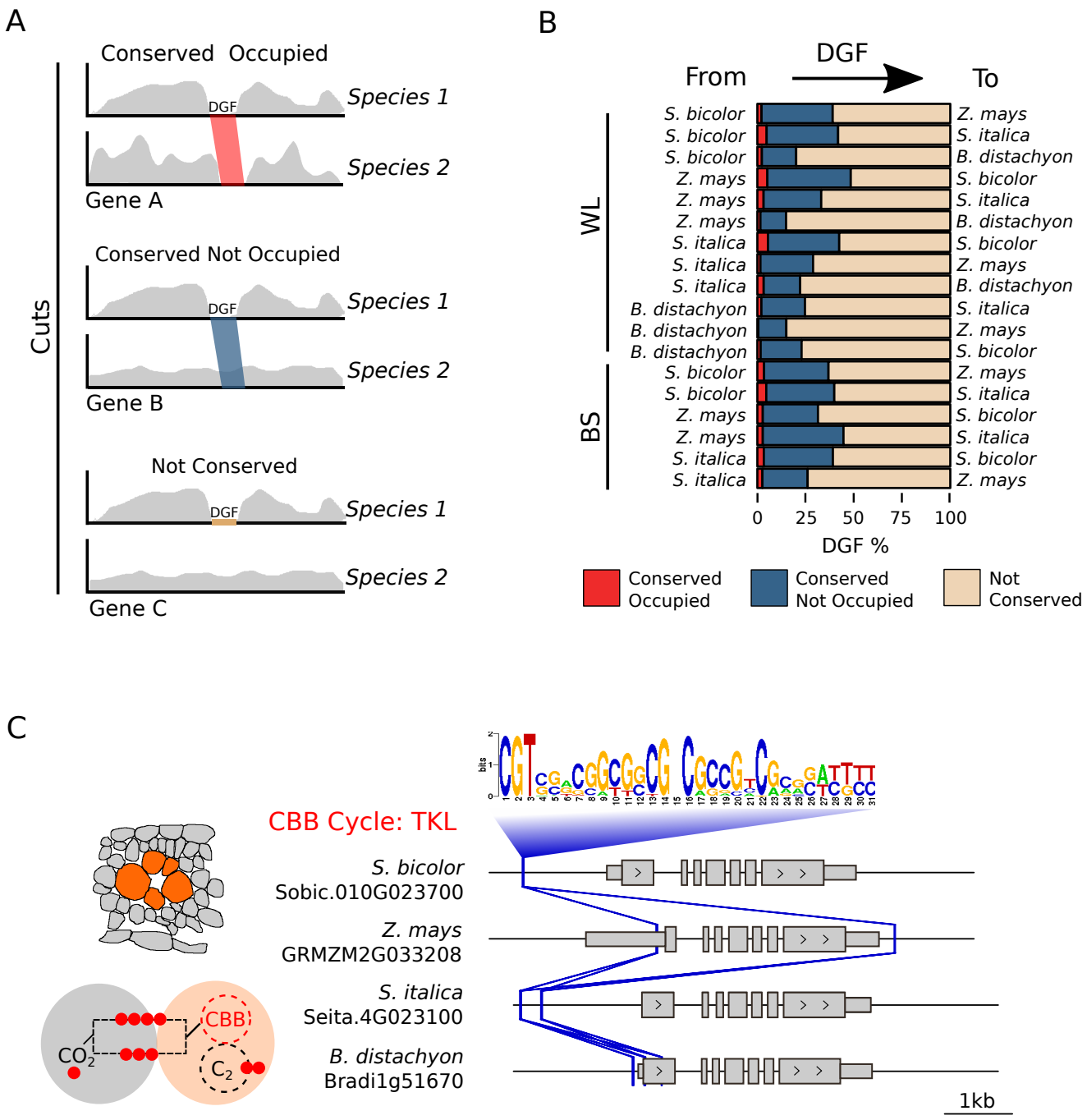

A
Conserved Occupied
Species 1
DGF
Species 2
Gene A
Conserved Not Occupied
Species 1
DGF
Species 2
Gene B
Not Conserved
Species 1
DGF
Species 2
Gene C
Cuts
B
DGF
From
To
WL
BS
S. bicolor
Z. mays
S. italica
B. distachyon
0
25
50
75
100
DGF %
Conserved Occupied
Conserved Not Occupied
Not Conserved
C
CBB Cycle: TKL
S. bicolor
Sobic.010G023700
Z. mays
GRMZM2G033208
S. italica
Seita.4G023100
B. distachyon
Bradi1g51670
CO2
CBB
C2
1kb

Figure 5

Figure 5: *Cis*-elements show high rates of turnover and mobility in grasses. (A) Scenarios for DGF conservation between species. Reads derived from DNase-I cuts are depicted in 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) Schematic depicting the position of transcription factor motif consistently associated with the bundle sheath enriched *TRANSKETOLASE* (*TKL)* gene in *S. bicolor, Z. mays, S. italica* and C_3 *B. distachyon*. The position of motifs conserved between orthologous genes is depicted by solid lines and varies between species.

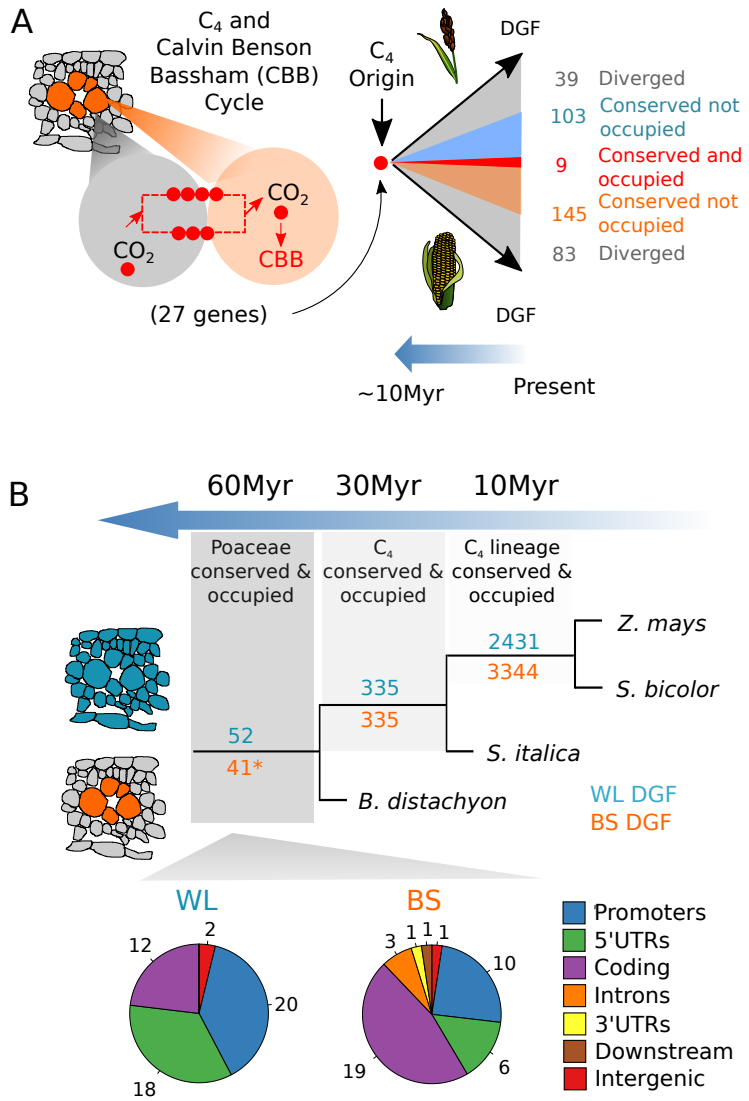


Figure 6: Hyper-conserved *cis*-elements in grasses recruited into C_4 photosynthesis. (A) Conservation of regulation in C_4 and Calvin Benson Bassham cycle genes following the divergence of *Z. mays* and *S. bicolor*. The number of carbon atoms (red dots) and metabolite flow (red dashed line) between 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. The asterisk indicates 41 DGF that are conserved in the BS of the C_4 species but are also found in whole leaves of *B. distachyon*. Pie-charts display the distribution of conserved and occupied DGF for whole leaf and BS strands. Promoters are defined as sequence up to 2000 base pairs (bp) upstream of the transcriptional start site, downstream represent regions 1000 downstream the transcription termination site while intergenic represent > 1000 bp downstream the transcription termination site until the next promoter region. (D) Bar chart representing DGF number in genic versus promoter regions.