A transcription factor ensemble orchestrating bundle sheath expression in rice

Lei Hua[1]*, Na Wang[1], Susan Stanley[1], Ruth M. Donald[1], Satish Kumar Eeda[1], Kumari Billakurthi[1], Ana Rita Borba[1] and Julian M. Hibberd[1]*

[1]Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, United Kingdom

*For correspondence

Email for contact: jmh65@cam.ac.uk, lh556@cam.ac.uk

Keywords: *Oryza sativa*; C_4 rice; Bundle Sheath; Sulfite Reductase; Promoter; cis-elements

Abstract

C_4 photosynthesis has evolved in over sixty plant lineages and improves photosynthetic efficiency by ~50%. One unifying character of C_4 plants is photosynthetic activation of a compartment such as the bundle sheath, but gene regulatory networks controlling this cell type are poorly understood. In Arabidopsis, a bipartite MYC-MYB transcription factor module restricts gene expression to these cells, but in grasses the regulatory logic allowing bundle sheath gene expression has not been defined. Using the global staple and C_3 crop rice, we find that the *SULFITE REDUCTASE* promoter is sufficient for strong bundle sheath expression. This promoter encodes an intricate *cis*-regulatory logic with multiple activators and repressors acting combinatorially. Within this landscape we identify a distal *cis*-regulatory module (CRM) activated by an ensemble of transcription factors from the WRKY, G2-like, MYB-related, DOF, IDD and bZIP families. This module is necessary and sufficient to pattern gene expression to the rice bundle sheath. Oligomerisation of the CRM and fusion to core promoters containing Y-patches allow activity to be increased 220-fold. This CRM generates bundle sheath-specific expression in *Arabidopsis* indicating deep conservation in function between monocotyledons and dicotyledons. In summary, we identify an ancient, short, and tuneable CRM patterning expression to the bundle sheath that we anticipate will be useful for engineering this cell type in various crop species.

Introduction

In plants and animals significant progress has been made in understanding transcription factor networks responsible for the specification of particular cell types. In animals, for example, homeobox transcription factors define the body plan of an embryo[1,2], and cardiac cell fate is specified by five transcription factors comprising Pnr and Doc that act as anchors for dTCF, pMad and Tin[3]. In plants the INDETERMINATE DOMAIN (IDD) transcription factors work together with SCARECROW and SHORTROOT to specify endodermal formation in the root[4,5], PHLOEM EARLY (PEAR) and VASCULAR-RELATED NAC DOMAIN (VND) transcription factors permit production of phloem and xylem vessel respectively[6,7], and basic helix–loop–helix (bHLH) transcription factors determine differentiation of guard cells[8–11]. Moreover, transcription factor networks that integrate processes as diverse as responses to external factors such as pathogens and abiotic stresses[12,13], or internal events associated with the circadian clock[14,15] and hormone signalling[16,17] have also been identified. Transcription factor activity is decoded by short *cis*-acting DNA sequences known as *cis*-regulatory elements. The binding of multiple transcription factors to a *cis*-regulatory module (CRM) thus controls transcription and the spatiotemporal patterning of gene expression by boosting or suppressing gene activity[18]. For example, the Block C enhancer interacts with the core promoter to activate expression of *FLOWERING LOCUS T* in long days[19,20], and a distant upstream enhancer controls expression of the *TEOSINTE BRANCHED1* locus in maize responsible for morphological differences compared with the wild ancestor teosinte[21,22]. In contrast to the above examples, transcription factors and cognate *cis*-elements responsible for the operation of cell types in grasses once specified have not been defined[23,24].

Given the increased specialisation of plant organs since the colonisation of land this lack of understanding of gene regulatory networks controlling cell specific gene expression is striking. For example, in the liverwort *Marchantia polymorpha* the photosynthetic thallus contains seven cell types[25], while leaves of *Oryza sativa* (rice) and *Arabidopsis thaliana* possess at least fifteen and seventeen populations of cells as defined by single-cell sequencing respectively[26]. In leaves of these angiosperms, particular cell types are specialised for photosynthesis and so activation of photosynthesis gene expression is highly responsive to light[27] One such cell type is the bundle sheath, and while these cells carry out photosynthesis they are also specialised to allow water transport from veins to mesophyll, as well as sulphur assimilation and nitrate reduction[28–30]. And, strikingly in multiple lineages, the bundle sheath has been dramatically repurposed during evolution to become fully photosynthetic and allow the complex C_4 pathway to operate[31].

Compared with the ancestral C_3 state, plants that use C_4 photosynthesis operate higher light, water and nitrogen use efficiencies[31–33]. It is estimated that introducing the C_4 pathway into C_3 rice would allow a 50% increase in yield[33,34], but it requires multiple photosynthesis genes to be expressed in the bundle sheath, including enzymes that decarboxylate C_4 acids to release CO_2 around RuBisCO, organic acid transporters, components of the Calvin-Benson-Bassham cycle, RuBisCO activase, and enzymes of starch biosynthesis[35–37]. In summary, although the bundle

sheath is found in all angiosperms and associated with multiple processes fundamental to leaf function, the molecular mechanisms responsible for directing expression to this cell type, including in global staple crops, remain undefined. We therefore studied the bundle sheath to better understand the complexity of gene regulatory networks that operate to maintain function of a cell type once it has been specified. Rice was chosen as it a global crop, and identifying how it patterns gene expression to the bundle sheath could facilitate engineering of this cell type.

We hypothesized that analysis of endogenous patterns of gene expression in the rice bundle sheath would allow us to identify a strong and early-acting promoter for this cell type. Once such a promoter was identified we also hypothesised that it could be used to initiate an understanding of the *cis*-regulatory logic that allows gene expression to be patterned to this cell type in grasses. We tested twenty-five promoters from rice genes that transcriptome sequencing indicated were highly expressed in these cells. Of these, four specified preferential expression in the bundle sheath, and one derived from the *SULFITE REDUCTASE* (*SiR*) gene (nucleotides -2571 to +42 relative to translational start site) generated strong bundle sheath expression from plastochron 3 leaves onwards. Truncation analysis showed that bundle sheath expression pattern from the *SiR* promoter is mediated by a short distal CRM and a pyrimidine patch (Y-patch) in the core promoter. This bundle sheath module is cryptic until other CRMs acting to both constitutively activate and repress expression in mesophyll cells are removed. The CRM is composed of a sextet of *cis*-elements recognised by their cognate transcription factors from the WRKY, G2-like, MYB-related, DOF, IDD and bZIP families. These transcription factors act synergistically and are sufficient to drive expression of the strong bundle sheath *SiR* promoter.

Results

The *SiR* promoter directs expression to the rice bundle sheath

To identify sequences allowing robust expression in rice bundle sheath cells, we initially used data derived from laser capture microdissection of bundle sheath strands (comprising bundle sheath, xylem and phloem) and mesophyll cells from mature leaves. To identify regulatory regions, upstream promoter sequences and where relevant DNase I hypersensitive sites that extended into coding sequence[38] from seven of the most strongly expressed genes in bundle sheath strands were cloned, fused to the β-glucoronidase (GUS) reporter and transformed into rice (**Supplementary Figure 1a, Supplementary Data 3**). Although five of these fusions (*MYELOBLASTOSIS*, *MYB*; *HOMOLOG OF E. COLI BOLA*, *bolA*; *GLUTAMINE SYNTHETASE 1*, *GS1*; *STRESS RESPONSEIVE PROTEIN*, *SRP*; *ACYL COA BINDING PROTEIN*, *ARP*) led to GUS accumulation, it was restricted to veins **(Supplementary Figure 1b,c)**. For the *SULFATE TRANSPORTER 3;1* and *3;3* (*SULT3;1* and *SULT3;3*) promoters, no staining was observed (**Supplementary Figure 1b,c**). The approach of cloning promoters from bundle sheath strands therefore appeared to be more efficient at identifying sequences capable of driving expression in veins. Using an optimised procedure to separate bundle sheath cells from veins[39], we therefore produced transcriptomes from mesophyll, bundle sheath and vascular bundles and identified clusters of genes associated with each cell type[30]. Eighteen of the genes most differentially expressed between bundle sheath and mesophyll, and associated with biological processes such as solute transport, sulfur metabolism, and nitrogen metabolism previously linked to the bundle sheath were selected (**Supplementary Figure 2a**). When the promoter from each gene was fused to GUS and transformed into rice, those from *ATP-SULFURYLASE 1B*, *ATPS1b*; *SULFITE REDUCTASE*, *SiR*; *HIGH ARSENIC CONTENT1.1*, *HAC1.1*; and *FERREDOXIN*, *Fd* were sufficient to generate expression in the bundle sheath (**Supplementary Figure 2b)**. However, *ATPS1b* and *Fd* also displayed weak activity in the mesophyll, and *HAC1.1* also led to GUS accumulation in epidermal and vascular cells. Thus, only the *SiR* promoter drove strong expression in the bundle sheath and veins with no GUS detected in mesophyll cells (**Supplementary Figure 2b,c**). An additional six promoters (*SOLUBLE INORGANIC PYROPHOSPHATASE*, *PPase*; *PLASMA MEMBRANE INTRINSIC PROTEIN1;1*, Os*PIP1;1*; *PLASMA MEMBRANE INTRINSIC PROTEIN1;3*, Os*PIP1;3*; *ACTIN-DEPOLYMERIZING FACTOR*, *ADF*; *PEPTIDE TRANSPORTER PTR2*, *PTR2*; *NITRATE REDUCTASE1*, *NIA1*) generated expression in vascular bundles, and eight promoters produced no staining (**Supplementary Figure 2b,c**). In summary, most candidate promoters failed to generate expression that was specific to bundle sheath cells, but the region upstream of the rice *SiR* gene was able to do so. We therefore selected the *SiR* promoter for further characterization.

The *SiR* promoter drives strong and early expression in bundle sheath cells

Sequence upstream of the *SiR* gene comprising nucleotides -2571 to +42 relative to the predicted translational start site was sufficient to generate expression in the rice bundle sheath. To allow faster

analysis of sequences responsible for this output, we domesticated the sequence by removing four *Bsa*I and *Bpi*I sites such that it was compatible with the modular Golden Gate cloning system. When this modified sequence was placed upstream of the GUS reporter it also generated bundle sheath preferential accumulation **(Fig. 1a)**. Fusion to a nuclear-targeted mTurquoise2 fluorescent protein confirmed that the *SiR* sequence was sufficient to direct expression to bundle sheath cells, and also revealed expression in the longer nuclei of veinal cells **(Fig. 1b)**. Expression from the domesticated and non-domesticated sequences was not different **(Fig. 1c)**. Compared with 0.58 nmol 4-MU/min/mg protein previously reported for the *Zoysia japonica PHOSPHOENOLCARBOXYKINASE (PCK)* promoter[40], activity from the *SiR* promoter was at least 36% higher. Designer Transcription Activator-Like Effector (dTALEs) and cognate Synthetic TALE-Activated Promoters (STAPs) amplify expression and allow multiple transgenes to be driven from a single promoter[41,42]. We therefore tested whether bundle sheath expression mediated by the *SiR* promoter is maintained and strengthened by the dTALE-STAP system. Stable transformants showed bundle sheath specific expression **(Supplementary Figure 3a,b)**, and GUS activity was ~18-fold higher than that from the endogenous *SiR* promoter (**Supplementary Figure 3c**). We conclude that the *SiR* promoter is compatible with the dTALE-STAP system and its activity can be strengthened.

We investigated when promoter activity was first detected during leaf development and discovered that GUS as well as fluorescence from mTurquoise2 were visible in 5-20 mm long fourth leaves at plastochron 3 (**Supplementary Figure 4**). This was not the case for the *ZjPCK* promoter even when a dTALE was used to amplify expression (**Supplementary Figure 4**). We conclude that the *SiR* promoter initiates expression in the bundle sheath before the *ZjPCK* promoter, and that it is also able to sustain higher levels of expression in this cell type.

A distal CRM necessary for expression in the bundle sheath

The *SiR* promoter contains a highly complex *cis* landscape **(Fig. 1d)** comprising at least 638 predicted motifs from 56 transcription factor families predicted using the Find Individual Motif Occurrences (FIMO) tool[43] with position weight matrix from the JASPAR database[44] (**Supplementary Data 1**). We therefore designed a 5' truncation series to investigate regions necessary for expression in the bundle sheath (**Fig. 1e**). Deleting nucleotides -2180 to -1490 and -1490 to -980 led to a statistically significant increase and then reduction in MUG activity respectively but neither truncation abolished preferential accumulation of GUS in the bundle sheath **(Fig. 1e-g).** However, when nucleotides -980 to -394 upstream of the predicted translational start site were removed GUS was no longer detectable in bundle sheath cells (**Fig. 1e-f**). Consistent with this, MUG assays showed a statistically significantly reduction in activity when these nucleotides were absent **(Fig. 1g)**. Thus, nucleotides spanning -980 to -394 of the *SiR* promoter are necessary for bundle sheath specific expression.

To test whether this region is sufficient for bundle sheath specific expression, we linked it to the minimal CaMV35S core promoter. Although weak GUS signal was detected in a few veinal cells, this

was not the case for the bundle sheath **(Fig. 1e-g)**. We conclude that sequence in two regions of the promoter (from -394 to +42 and from -980 to -394) interact to specify expression to the bundle sheath. To better understand this interaction, we next generated unbiased 5' and 3' deletions. This second deletion series further reinforced the notion that the *SiR* promoter contains a complex *cis*-regulatory landscape. For example, when nucleotides -980 to -829 were removed very weak GUS staining was observed and the MUG assay showed that activity was reduced by 73% **(Fig. 2, Supplementary Figure 5).** We conclude that nucleotides -980 to -829 from the *SiR* promoter are necessary for tuning expression in the leaf. When nucleotides -829 to -700 were removed GUS appeared in mesophyll cells **(Supplementary Figure 5)**. Truncating nucleotides -613 to -529 diminished GUS accumulation (**Supplementary Figure 5)**. The 3' deletion that removed nucleotides -251 to +42 also stopped accumulation of GUS in both bundle sheath and mesophyll cells (**Fig. 2a-c, Supplementary Figure 5)**. Notably, when the distal region required for bundle sheath expression (-980 to -829) was combined with nucleotides -251 to +42 this was sufficient for patterning to the bundle sheath (**Fig. 2a,b**).

Having identified a region in the *SiR* promoter that was necessary and sufficient for patterning to the bundle sheath, we next used phylogenetic shadowing and yeast one hybrid analysis to better understand the *cis*-elements and *trans*-factors responsible. Analysis of *cis*-elements in the *SiR* promoter that are highly conserved in grasses identified a short region located from nucleotides -588 to -539 that contained an *ETHYLENE INSENSITIVE3-LIKE 3* (*EIL3*) transcription factor binding site (**Supplementary Figure 6a,b**). Whilst deletion of this motif had no detectable effect of patterning to the bundle sheath (**Supplementary Figure 6c**), the level of expression was reduced (**Supplementary Figure 6d**). We infer that the *EIL3* motif positively regulates activity of the *SiR* promoter but is not responsible for cell specificity. These data are consistent with the promoter truncation analysis that showed nucleotides -613 to -529 containing this motif were not required for bundle sheath specific expression, but instead function as a constitutive activator (**Supplementary Figure 5**). When yeast one hybrid was used to search for transcription factors capable of binding the *SiR* promoter, sixteen were identified (**Supplementary Figure 7a,b**). For each, cognate binding sites were present. This included TCP21 and OsOBF1 that can bind to TCP motifs and Ocs/bZIP elements respectively. Consistent with the outcome of deleting the *EIL3* motif, three *EIL* transcription factors interacted with nucleotides -899 to -500 (**Supplementary Figure 7b,c**). Examination of transcript abundance in mature leaves showed that most of these transcription factors were expressed in both bundle sheath and mesophyll cells (**Supplementary Figure 7d**), implying that combinatorial interactions with cell specific factors are likely required for bundle sheath specific expression from the *SiR* promoter.

The CRM contains four subregions that simultaneously activate in bundle sheath and repress in mesophyll cells

The truncation analysis above identified two short regions comprising nucleotides -980 to -829 and -251 to +42 that were necessary and sufficient for expression in the rice bundle sheath (**Fig. 2, Supplementary Figure 8**). Sequence spanning nucleotides -251 to +42 includes both the annotated 5' untranslated region but also likely contains core promoter elements (**Supplementary Figure 9a**). Re-analysis of publicly available data identified two major transcription start sites at positions -91 (TSS1) and -41 (TSS2) (**Supplementary Figure 9a**). Although no canonical TATA-box was evident in this region, a TATA-box variant was detected at position -130 (5'-ATTAAA-3')[45] that could be responsible for transcription from TSS1. Upstream of TSS2 is a putative pyrimidine patch (Y-patch) that represents an alternate but common TC-rich core promoter motif in plant genomes)[45] (**Supplementary Figure 9a**). Scanning sequence from -251 to -1 for core promoter elements also identified MTE (Motif Ten Element), BREu (TFIIB Recognition Element upstream) and DCE-S-I (Downstream Core Element S-I) motifs associated with eukaryotic core promoters (**Supplementary Figure 9b**). We therefore assume the region upstream of TSS1 and TSS2 contains core promoter elements. When consecutive deletions to this sequence were made, statistically significant reductions in MUG activity were evident but there was no impact on accumulation of GUS in the bundle sheath (**Supplementary Figure 9c,d**). Interestingly, when the Y-patch was retained but the TATA-box like motif removed, GUS was still detected in the bundle sheath (**Supplementary Figure 9c-f**), but after deletion of the Y-patch GUS staining was no longer detectable in these cells (**Supplementary Figure 9c-f**). Consistent with the Y-patch being important for bundle sheath expression, two core promoters with only a TATA-box linked to the distal CRM did not generate detectable expression in the bundle sheath, but those from genes with one or more Y-patches did **(Fig. 3a,b)**. GUS activity was higher from the *PIP1;1* core promoter that contains three Y-patches **(Fig. 3c)**. Overall, we conclude that the TATA-box like motif is not required for expression in the bundle sheath, but the Y-patch is necessary for this patterning and in combination with a distal CRM comprising nucleotides -980 to -829, it is sufficient for expression in this cell type.

We assessed the distal CRM for transcription factor binding sites. The FIMO algorithm identified motifs associated with WRKY, G2-like, MYB-related, MADS, DOF, IDD, ARR, SNAC (Stress-responsive NAC) families. PlantPAN[46], which includes experimentally validated *cis*-elements, found an additional Dc3 Promoter Binding Factor (DPBF) binding site for group A bZIP transcription factors[47] **(Fig. 3d)**. Seven consecutive deletions spanning this CRM region and hereafter termed subregions a-g were generated **(Fig. 3d)**. Although veinal expression persisted when subregions a, b and d were absent, deletion of subregions a, b, d and f resulted in loss of GUS from bundle sheath cells (**Fig. 3e-g**). MUG analysis showed that deletion of all four regions significantly reduced promoter activity **(Fig. 3g)**. In contrast, deletions of nucleotides -938 to -923 (subregion c), -904 to -873 (subregion e), and -853 to -829 (subregion g) had no impact on the patterning **(Supplementary Figure 10)**. The subregions necessary for expression in the bundle sheath contained binding sites for WRKY, G2-like, MYB-related, DOF, IDD, SNAC and bZIP (DPBF) transcription factors. To examine the significance of these regions in the context of full-length *SiR* promoter, consecutive

deletions from subregion a to f were generated **(Supplementary Figure 11a)**. Deletion of subregion a, d or f, led to GUS accumulating primarily in mesophyll cells whereas removal of subregion b, c or e, caused GUS staining in both mesophyll cells and bundle sheath cells **(Supplementary Figure 11b)**. No significant changes in GUS activity were observed in these deletion lines **(Supplementary Figure 11c)**. We conclude that the distal CRM generates expression in the bundle sheath due to four distinct sub-regions, and that by interacting with nucleotides -829 to -251, nucleotides between -980 to -853 also function as repressors of mesophyll expression.

WRKY, G2-like, MYB-related, DOF, IDD and bZIP transcription factors activate the distal CRM

To gain deeper insight how the distal CRM operates we employed transactivation assays, co-expression analysis and site directed mutagenesis. The distal CRM contained WRKY, G2-like, MYB-related, DOF, IDD, SNAC and bZIP (DPBF) motifs **(Fig. 4a, Supplementary Figure 12)**. DOF transcription factors have recently been shown to tune expression in bundle sheath cells[27] and so using an effector assay, we tested whether the other families interacted with the distal CRM. Using publicly available data, selection was based on three criteria: first, co-expression with *SiR*; second, preferential expression in bundle sheath cells from mature rice leaves; third, rice orthologs of Arabidopsis transcription factors predicted by FIMO to have the strongest binding to each motif **(Supplementary Figure 13).** WRKY121, GLK2, MYBS1, IDD2/3/4/6/10, and bZIP3/4/9/10/11 transcription factors led to the strongest activation of expression from the bundle sheath CRM **(Fig. 4b**, **Supplementary Figure 14a-d)**, whereas the stress-responsive NAC transcription factors targeting a SNAC motif that overlaps a bZIP (DPBF) motif, activated less strongly than bZIP factors **(Supplementary Figure 14e)**. We therefore conclude that the SNAC motif is not important for activity of the bundle sheath CRM. Effector assays using pairwise combinations of transcription factors showed synergistic activation from the distal CRM when GLK2 and IDD3,4,6,10 were co-expressed **(Fig. 4c, Supplementary Figure 14f)**.

Co-expression analysis derived from a cell-specific leaf developmental gradient dataset **(Supplementary Data 5)** revealed that transcripts derived from *GLK2, MYBS1* and *IDD4,6,10* transcription factors that bind the G2-like, MYB-related and IDD motifs respectively were more abundant in mesophyll cells **(Fig. 4d)**. However, transcripts for *bZIP9, IDD2* and *WRKY121* transcription factors strongly correlated with *SiR* transcript abundance and were preferentially expressed in bundle sheath cells **(Fig. 4d)**. To test for sufficiency, we mis-expressed *bZIP9, IDD2, WRKY121* separately and in multiple combinations in mesophyll cells**.** Mesophyll expression of *bZIP9* or *WRKY121* induced GUS expression from the bundle sheath CRM in some mesophyll cells, with the effect from *WRKY121* being greater (**Fig. 4e, Supplementary Figure 15b,f**). Mesophyll expression of *IDD2* suppressed activity **(Fig. 4e, Supplementary Figure 15d)**, but when combined with *bZIP9* significant expression in the mesophyll was apparent **(Fig. 4e, Supplementary Figure 15h)**. Strikingly, expression of *bZIP9*, *IDD2* and *WRKY121* in the mesophyll strongly activated

expression in this cell type **(Fig. 4e, Supplementary Figure 15j)**. When each motif was mutated, with the exception of the WRKY, CRM activity in the bundle sheath was diminished **(Fig. 5a-c)**.

In order to test whether the distal CRM is sufficient to pattern expression to rice bundle sheath cells, we concatenated sequence containing the WRKY, G2-like, MYB-related, DOF, IDD, and bZIP sites and fused them to the core promoter of *SiR* (**Fig. 5d**). GUS staining was evident in the bundle sheath (**Fig. 5e**). Fusion to the *PIP1;1* core promoter maintained bundle sheath expression and resulted in an ~5 fold increase in activity (**Fig. 5d-f**). Oligomerisation of the CRM by repeating it three or five times increased bundle sheath specific expression 25 or 58-fold respectively when fused to *SiR* core promoter, and this effect was amplified 94 and 224-fold when fused with the *PIP1;1* core (**Fig. 5d-f**). When an oligomerised version of the CRM was linked to the *SiR* core promoter and placed in *A. thaliana*, it generated strong expression in bundle sheath cells (**Fig. 5g, Supplementary Figure 16**).

Collectively our data reveal an ensemble of transcription factors belonging to the WRKY, G2-like, MYB-related, DOF, IDD and bZIP (DPBF) families act to decode distinct *cis*-elements in a distal CRM of the *SiR* promoter, and that this transcription factor team represents an ancient and highly conserved mechanism allowing bundle sheath specific gene expression in both monocotyledons and dicotyledons.

Discussion

Expression of multiple genes in the rice bundle sheath is not associated with close upstream CRM

Gene expression is determined by interactions between elements in the core promoter allowing basal levels of transcription[48,49] with more distal *cis*-regulatory modules[50–52]. Such *cis*-regulatory modules include enhancers and silencers that act as hubs receiving input from multiple transcription factors and so allow gene expression to respond spatially and temporally to both internal and external stimuli[53,54]. After testing 25 promoters, we discovered that the majority were not capable of driving expression in the rice bundle sheath, and this included ten that generated no detectable activity of GUS in leaves. In all cases we had cloned sequence between -3191 and -960 nucleotides upstream of the predicted translational start site and so these data demonstrate that the core promoter and any CRM in these regions are not sufficient to direct expression to rice bundle sheath cells. Combined with the paucity of previously reported promoters active in this cell type[55,56], these data argue either for long range upstream enhancers[57–61] or other regulatory mechanisms being important to specify expression in the bundle sheath. Possibilities include transcription factor binding sites in introns that impact on transcription start site and strongly enhance gene expression[62,63], or in exons where because such sequences specify amino acid sequence as well as binding of *trans*-factors, they have been termed duons[64]. Functional analysis showed that duons can pattern expression to the bundle sheath of the C_4 plant *Gynandropsis gynandra*[65], and it is notable that a genome-wide analysis of transcription factor binding sites in grasses revealed genes preferentially expressed in bundle sheath cells tended to contain transcription factor binding sites in coding sequence[66]. It therefore appears possible that gene expression in the bundle sheath is commonly encoded by non-canonical architecture perhaps based on duons rather than more traditional cis-regulatory elements upstream of the core promoter.

Despite the above, we discovered four promoters capable of driving expression in the rice bundle sheath, and each was associated with a gene important in sulphur metabolism. For example, *ATPS1b*, *SiR* and *Fd* all participate in the first two steps of sulphate reductive assimilation[67–69], while *HAC1;1* encodes an arsenate reductase important in the detoxification of arsenate using glutathione that is a product of sulphur assimilation[70]. Collectively, these data support the notion that various mechanisms underpin bundle sheath specific expression in rice, but that for some genes involved in sulphur metabolism upstream regulatory regions are important.

Two distinct genetic networks governing expression in bundle sheath cells

The only other promoter for which both *cis*-elements and *trans*-factors that are necessary and sufficient to pattern bundle sheath expression have been reported is from the dicotyledonous model *A. thaliana*. In that study, a bipartite MYC-MYB module upstream of the *MYB76* gene is responsible for this output[71]. MYB76 forms part of a network governing glucosinolate biosynthesis in *A. thaliana*, and so it is notable that the gene regulatory network we report in rice is also associated with sulphur

metabolism. However, rather than the bipartite transcription factor module that activates *MYB76* in the *A. thaliana* bundle sheath, in rice we report a more complex regulatory landscape where a bundle sheath CRM is embedded in constitutive activators and an overlapping mesophyll silencer (**Fig. 6a-c**). This complexity may be associated with the fact that SiR is considered to catalyse the rate-limiting step for sulfur assimilation[68]. The CRM controlling bundle sheath *SiR* expression in rice comprises four distinct regions recognised by transcription factors belonging to the WRKY, G2-like, MYB-related, DOF, IDD and bZIP families **(Fig. 6d)**. It is of course possible that additional motifs in the CRM modulate the level of expression in bundle sheath cells. As loss of the G2-like, MYB-related, DOF, IDD and bZIP motifs all reduced expression in the bundle sheath, this implies they act co-operatively - a notion further supported by the fact that GLK2 and IDD3,4,6,10 synergistically activated promoter output in a transient assay. This cooperation between transcription factors may explain the lack of overlap between these transcription factors and ones identified by the yeast one-hybrid assay, as the latter is typically not well suited to identify transcription factor complex-DNA interactions[72]. To examine whether other bundle sheath expressed genes may be regulated by the CRM and Y-patch system we report, we assessed their prevalence in open chromatin across the genome. This demonstrated over-representation of these sequences in bundle sheath expressed genes (**Supplementary Figure 18**). Of the 283 genes containing both Y-patch and *cis*-elements found in the CRM of *SiR*, 117 were expressed in the bundle sheath, including 21 preferentially expressed in this cell type. Interestingly, this included *a Ferredoxin* annotated as provided reducing equivalents to *SiR*[67]*, a CRT-like transporter 1* (*CLT1)* allowing glutathione transport[73], a sulfurtransferase *STR22*[74], and two transcription factors *EIL2* and *EIL3* involved in ethylene signalling[75], but also considered master regulators of sulphur assimilation[76,77] (**Supplementary Data 10**).

The distal CRM in the *SiR* promoter operates in conjunction with the core promoter that contains two transcription start sites, one with an upstream TATA-box and the other a TC-rich element known as a pyrimidine (Y) patch **(Supplementary Figure 9a)**. The TATA-box is found in metazoans and plants and allows recognition by the pre-initiation complex[78], but in plants, computational analysis showed that many promoters lack a TATA-box and instead contain a Y-patch[79–81]. These genes tend to be relatively steadily expressed and associated with protein metabolism[81], and presence of a Y-patch can increase core promoter strength[82]. For *SiR,* whilst the TATA-box is not required, the Y-patch is needed for expression in the bundle sheath. Notably, both synthetically generated (**Supplementary Figure 17**) and endogenous core promoters modified to contain more Y-patches tended to drive stronger expression, indicating that in plants cell specific gene expression could be tuned by selecting different core promoters.

The distal CRM (nucleotides -980 to -829) also represses mesophyll expression if nucleotides -829 to -700 are present. This suggests these two regions interact to suppress transcription in the mesophyll (**Fig. 6a-c**). Thus, a mesophyll silencer overlaps with the bundle sheath CRM (**Fig. 6c**). Of the six families of transcription factors binding the bundle sheath CRM, WRKY121, IDD2, and

bZIP9 appear critical in controlling bundle sheath-specific expression because misexpression in the mesophyll leads to output in this cell type **(Fig. 6d)**. It is possible that transcription factors more highly expressed in the mesophyll such as GLK2, MYBS1 and IDD4/6/10 interact with additional factors binding to the -829 to -700 region and that this contributes to mesophyll-specific suppression **(Fig. 6e)**. In addition to controlling cell specificity, this complexity likely also facilitates the tuning of expression to environmental conditions. For instance, the *EIL* motif (position -588 to -539) is recognised by *ETHYLENE-INSENSITIVE LIKE* transcription factors that respond to sulphur deficiency[76,77]. As transcripts encoding *EIL* accumulate in both bundle sheath and mesophyll cells in response to sulphate deficiency, it seems likely that transcription factors repressing expression in the mesophyll respond in a dynamic manner. In addition to *EIL*, the yeast one hybrid analysis identified seven other families of transcription factor families that can bind the *SiR* promoter. Many play documented roles during abiotic or biotic stress, with for example OBF1, ERF3, NAP and FLP acting during low-temperature or drought responses[83–87], while TCP21, EREBP1, ERF3, ERF72, ERF83 are involved in both abiotic and biotic stress[88–91]. Consistent with previous *in silico* analysis[92] the presence of multiple AP2/ERF and EIL transcription factors binding sites suggests that *SiR* is likely subject to control from ethylene signalling[93] and also of transcription factors that respond to abscisic acid (ABA) and jasmonic acid (JA)[86,90,94–96]. Together this implies that multiple phytohormone signalling pathways converge on the *SiR* promoter. These data are similar to those reported for the *SHORTROOT* promoter in *A. thaliana* roots where a complex network of activating and repressing *trans*-factors also tunes expression[97]. It is also notable that the architecture we report for the bundle sheath CRM of *SiR* appears of similar complexity to the collective of five transcription factors used to specify cardiac mesoderm in *Drosophila melanogaster* and vertebrates[3]. For the five transcription factors that bind the cardiac mesoderm enhancer, the order and positioning of motifs (motif grammar) is flexible. However, this is not always the case, with for example output from the human *interferon-beta* (*INF-β*) enhancer demanding a conserved grammar[98,99]. Further work will be needed to determine if the bundle sheath CRM reported here is more similar to one of these models, or indeed, as reported for the Drosophila *eve stripe 2* enhancer, operates as a billboard in different tissues to determine patterning of expression[100].

Using the *SiR* promoter to engineer the rice bundle sheath

In addition to bundle sheath cells being important for sulphur assimilation[28–30], they have also been implicated in nitrate assimilation, the control of leaf hydraulic conductance and solute transport[30] and the systemic response to high light[101]. Moreover, in one of the most striking examples of a cell type being repurposed for a new function, bundle sheath cells have repeatedly been rewired to allow the evolution of C_4 photosynthesis[31]. To engineer these diverse processes, specific and tuneable promoters for this cell are required. However, identification of sequence capable of driving specific expression to bundle sheath strands has previously been limited to *A. thaliana* and C_4 species. For example, the *SCARECROW*[102], *SCL23*[102], *SULT2;2*[103] and *MYB76* promoters[71] are

411 derived from *A. thaliana*, whilst the *Glycine Decarboxylase P-protein (GLDP*) promoter is from the
412 C_4 dicotyledon *Flaveria trinervia*[104,105]*.* In rice, only the C_4 *Zoysia japonica PCK* and the C_4 *Flaveria*
413 *trinervia GLDP* promoters are known to pattern expression to the bundle sheath[55,56]. Both are
414 capable of conditioning expression in this cell type, but are weak, turn on late during leaf
415 development and the molecular basis underpinning their ability to restrict expression to the bundle
416 sheath has not been defined. It has therefore not been possible to rationally design or tune
417 expression to this important cell type in rice. The architecture of the *SiR* promoter we report here
418 now provides an opportunity to engineer the bundle sheath.

419 In summary, from analysis of the ~2600 nucleotide *SiR* promoter we identify an CRM comprising
420 81 nucleotides that with the Y-patch is sufficient to drive expression to bundle sheath cells. Moreover,
421 we show that output from the sequence can be tuned via two approaches. First, oligomerising the
422 distal CRM can drastically increase expression. Second, combining it with different core promoters
423 achieved the same output, and correlated with copy number of the Y-patch present in natural and
424 synthetic promoters. Our identification of a minimal promoter that drives expression in bundle sheath
425 cells of rice now provides a tool to allow this important cell type to be manipulated. Cell specific
426 manipulation of gene expression has many perceived advantages. For example, when constitutive
427 promoters have been used to drive gene expression gene silencing and reduction of plant fitness
428 due to metabolic penalties[106,107]. In contrast, tissue specific promoters allow targeted gene
429 expression either spatially or at particular developmental stages and so allow increased precision in
430 trait engineering[108]. The *SiR* promoter and the bundle sheath *cis*-regulatory module that we identify
431 thus provide insights into mechanisms governing cell specific expression in plants, and may also
432 contribute to our ability to engineer and improve cereal crops.

Methods

Plant material and growth conditions

Kitaake (*O. sativa* ssp. *japonica*) was transformed using *Agrobacterium tumefaciens* as described previously[109] with the following modifications. Mature seeds were sterilized with 2.5% (v/v) sodium hypochlorite for 15 mins, and calli were induced on NB medium with 2 mg/L 2,4-D at 30 ºC in darkness for 3-4 weeks. Actively growing calli were then co-incubated with *A. tumefaciens* strain *LBA4404* in darkness at 25ºC for 3 days, selected on NB medium supplied with 35 mg/L hygromycin B for 4 weeks, and those that proliferated, placed on NB medium with 10 mg/L hygromycin B for 4 weeks at 28 ºC under continuous light. Plants resistant to hygromycin were planted in 1:1 mixture of topsoil and sand and placed in a greenhouse at the Botanic Garden, University of Cambridge under natural light conditions but supplemented with a minimum light intensity of 390 µmol m^{-2} s^{-1}, a humidity of 60%, temperatures of 28ºC and 20ºC during the day and night respectively, and a photoperiod of 12 h light, 12 h dark. Subsequent generations were grown in a growth cabinet in 12 h light/12 h dark, at 28 ºC, a relative humidity of 65%, and a photon flux density of 400 µmol m^{-2} s^{-1}.

Laser capture microdissection RNA-seq and data analysis

For the first Laser Capture Microdissection (LCM) RNA-seq analysis, the middle 1-cm of the fourth fully expanded IR64 (*Oryza sativa* ssp. *indica*) leaves were sampled 4 hours after dawn. Leaf tissue was fixed with acetone and embedded into Steedman's wax[39]. Paradermal sections of 7µm were prepared with a microtome and mounted on PEN membrane slides (Applied Biosystems, LCM0522). Bundle sheath strands (bundle sheath as well as phloem and xylem cells) and mesophyll cells were isolated using CapSure Macro Caps (Applied Biosystems, LCM0211). To prepare cDNA libraries, 20-40 ng RNA was depleted for ribosomal RNA using Ribozero kit (Illumina), and then the 1st strand cDNA was generated and amplified using Ovation v2 RNA-seq system (Nugen) according to the user manual. Single stranded cDNA was digested using S1 Nuclease (Promega, M5761) then 100 ng of amplified cDNA sheared using a Covaris E220 focused ultra-sonicator. cDNA libraries were prepared using Truseq nano DNA library preparation kit (Illumina). Paired-ended 75-bp sequencing were performed using Nextseq 550 sequencer. Poly A/T/N and low-quality reads were removed using AfterQC and Trimmomatic, and gene expression in transcripts per million (TPM) quantified using Salmon v1.1.0[110]. Differential gene expression analysis was carried out using DESeq2[111], with differentially expressed genes being defined with an adjusted $P < 0.05$ and log2 fold change > 1.

For subsequent analysis, the middle 1-cm of the fourth fully expanded leaves from Kitaake (*Oryza sativa* ssp. *japonica*) were fixed and embedded. Bundle sheath, mesophyll and veinal cells were separated, library preparation and RNA-seq data analysis were conducted according to Hua et al., 2021. To identify bundle sheath specific promoters, candidate genes from bundle sheath and vein or bundle sheath alone clusters were filtered based on pairwise comparisons among the three tissue types using DESeq2 and EdgeR[112](*Log2FC(BS/M)>2 & adjusted P (BS/M) < 0.01 & FDR (BS/M) <*

0.01 & Log2FC(BS/V) > -0.5) to identify 210 genes specifically expressed in bundle sheath and vein or bundle sheath alone (**Supplementary Data 4**).

A leaf developmental gradient was also used for LCM RNA-seq. Here, Kitaake leaf tissue was also fixed and embedded according to Hua et al., 2021[30]. Material from the shoot apical meristem and leaf four primordia (referred to as Stage 1) were isolated from 3-day-old seedlings. Subsequent samples were taken from the leaf blade 5-mm above the leaf ligule at day 6 (Stage 2), 8 (Stage 3), 9 (Stage 4), 10 (Stage 5), 13 (Stage 6) and 17 (Stage 7) after sowing, with the midrib removed before embedding into Steedman's wax. Prior to LCM, sections of Stage 2, 3, and 4 were treated with 1% iodine (w/v) in acetone for 1 min. 1.4 - 3ng of RNA from each cell type (three to four biological replicates for each stage) were reverse transcribed, and the resultant cDNA amplified using the SMART-Seq v4 Ultra Low Input RNA Kit (Clontech) according to the manufacturer's instructions. 11 PCR cycles were used for all samples. 1 ng of amplified cDNA was used as input for library preparation using the Nextera XT DNA Library Preparation Kit (Illumina). Libraries were then sequenced using Illumina's NovaSeq 6000 sequencing platform to generate 31 - 50 million paired-ended 150-bp reads for each sample. Transcript abundance was determined after reads were quantified using Salmon v1.5.1[110] against *Oryza sativa* v7.0 transcripts from Phytozome V13[113] with selective alignment ("--validateMappings") enabled. Gene-level abundance (transcripts per million, TPM) was summarized using tximport 1.18.0[114], and genes with TPM > 1 in at least three samples of at least one developmental stage were defined as expressed genes. Co-expression network analysis was performed using Log2 transformed TPM and Weighted Correlation Network Analysis (WGCNA (v.1.63))[115] with soft threshold of 12, minimal module size of 100 genes, MEDissThres cut-off of 0.1.

Construct preparation

To test promoters for bundle sheath specific expression, a minimal of 1.5-kb upstream sequences from translational start site or entire intergenic region if it is shorter than 1.5-kb were amplified from genomic DNA with primers listed in **Supplementary Data 7**; in some cases, this was extended both upstream and downstream of the TSS to capture regulatory elements identified through DNase hypersensitive sites (DHS) at the Plant DHSs Database (https://plantdhs.org/)[38]. Gel-purified PCR products were cloned into a Gateway pENTR vector using directional D-TOPO Cloning Kit (Invitrogen, K240020), and the promoters recombined into the *pGWB3* expression vector and fused with *GUS* gene using LR reaction. The resultant vector was transformed into *A. tumefaciens* strain *LBA4404* and then Kitaake.

To engineer the *SiRpro* such that it is compatible with the Golden gate system, four *Bsa*I or *Bpi*I restriction enzyme recognition sites at -214, -298, -1468, and -2309 nucleotide relative to translational start site were mutated from T to A as described[116] using PCR primers listed in **Supplementary Data 8**, PCR fragments were assembled into the *pAGM9121* vector[117] using Golden Gate reactions, and used to drive *kzGUS* (intronless *GUS*)[55]and *H2B-mTurquoise2* reporter genes

using *Tnos* as a terminator. A 5' prime deletion series was generated using the domesticated *SiRpro* as template and prepared as level 0 PU modules. The 3' prime deletion series was prepared as level 0 P modules with the minimal *CaMV35S* promoter as the U module and linked with *kzGUS* and terminated with *Tnos*. To test *SiRpro* with the dTALE/STAP system[41], the 42-bp coding region was excluded and the 2571-bp resultant fragment placed into a level 0 PU module EC14330 and used to drive *dTALE*. Two reporters were used. For the GUS reporter *kzGUS* was linked with *STAP62* and terminated with *Tnos*. In the fluorescent reporter construct, a chloroplast targeting peptide fused to *mTurquoise 2* was linked with *STAP4* and terminated with *Tact2*. In both constructs, *OsAct1pro* driving *HYG* (the hygromycin resistant gene) was terminated with *Tnos* and used as the selection marker during rice transformation.

Consecutive deletions and site directed mutagenesis within nucleotides -980 to -829 were cloned as level 0 PU modules. Core promoter sequences from *SiR*, *PIP1;1*, *PIP1;3, NRT1.1A* and *ATPS1b* were amplified using primers listed in **Supplementary Data 8** and cloned as level 0 U modules. For oligomerisation, the bundle sheath CRMs was cloned as level 0 P modules, and then assembled into level 1 modules with respective U modules as well as *kzGUS* and the *Tnos* terminator. To mis-express *bZIP9, IDD2*, and *WRKY121* individually or in combination in mesophyll cells, the *ZmPEPC* promoter[118,119] was used to drive *dTALE*, expression of *bZIP9, IDD2*, and *WRKY121* controlled by different STAPs: *bZIP9* under STAP62, *IDD2* under STAP45, and *WRKY121* under STAP56. When *IDD2* was expressed alone its was driven by the *ZmUBI* promoter. A reporter construct containing nucleotides -980 to -829 and the endogenous core promoter (-250 to +42) driving kzGUS was used to assess the CRM activity in the same level 2 constructs used to mis-express each transcription factor.

Motif analysis

The Find Individual Motif Occurrences (FIMO) tool[43] from the Multiple Em for Motif Elucidation (MEME) suite v.5.4.0[120] was used to search for individual motifs within promoter sequences using default parameters with "--thresh" of "1e-3". Position weight matrix of 656 non-redundant plant motifs and 13 RNA polymerase II (POLII) core promoter motifs were obtained from JASPAR (https://jaspar.elixir.no/downloads/)[121], the *DPBF* binding sites and the Y-patch was included as previously described[47,122]. To cluster the transcription factor binding motifs, the RSAT matrix-clustering tool[123] was run on all 656 non-redundant plant motifs using the default parameters. This yielded 51 motif clusters, that were classified based on transcription factor families (**Supplementary Data 2**).

To assess the co-occurrence of the Y-patch motif and the sextet of *cis*-elements present in the distal CRM of *SiR* across all genes in the rice genome, we first examined the presence of the Y-patch motif in the core promoter. Specifically, we extracted DNA regions (5' to 3') spanning from 200 bp upstream of the transcriptional start site (TSS) to the translational start site and scanned this for the Y-patch motif using FIMO. The Y-patch motif was defined based on a *p*-value threshold of

<0.0004 and was required to be in the same orientation as the gene. To investigate the co-occurrence of the sextet of motifs, we extracted accessible chromatin as defined by *DNase* I Hypersensitive Sites[38] within 2000 bp upstream of the TSS to the end of the 3' UTR. Motif scanning was performed using FIMO with default parameters and a significance threshold ("--thresh") of 1e-3. The presence of the WRKY and G2-like motif families was determined using a *p*-value cutoff of <1e-3, whereas MYBR_B, DOF, IDD, and DPBF motif families were identified using a *p*-value cutoff of <1e-4 within a 300-bp window. Finally, 283 genes containing both the Y-patch motif in the core promoter and the distal CRM across the promoter and gene body were identified (**Supplementary Data 10**).

Analysis of GUS and fluorescent reporters

In all cases, to account for position effects associated with transformation via *A. tumefaciens*, multiple T_0 lines were assessed for each construct. GUS staining was performed as described previously[124] with the following minor modifications. Leaf tissue was fixed in 90% (v/v) acetone overnight at 4 ºC, after washing with 100 mM phosphate buffer (pH 7.0), leaf samples were transferred into 1 mg/ml 5-bromo-4-chloro-3-indolyl glucuronide (X-GlcA, Melford, B72200) GUS staining solution, subjected to 2 mins vacuum infiltration 5 times, and then incubated at 37 ºC for between 0.5 and 168 hours. Chlorophyll was cleared further using 90% (v/v) ethanol overnight at room temperature. Cross sections were prepared manually using a razor blade and images were taken using an Olympus BX41 light microscope. Quantification of GUS activity was performed using a fluorometric MUG assay[124]. ~200 mg mature leaves from transgenic plants were frozen in liquid nitrogen and ground into fine powder with a Tissuelyser (Qiagen). Soluble protein was extracted in 1 mL of 50 mM phosphate buffer (pH 7.0) supplemented with 0.1% [v/v] Triton X-100 and cOmplete™ Protease Inhibitor Cocktail (Roche, 11873580001, one tablet per 100 mL). Protein concentration then determined using a Qubit protein assay kit (Invitrogen, Q33212). The MUG fluorescent assay was performed in duplicates with 20 µl protein extract in MUG assay buffer (50 mM phosphate buffer (pH 7.0), 10 mM EDTA-Na_2, 0.1% [v/v] Triton X-100 (Sigma, X100), 0.1% [w/v] N-lauroylsarcosine sodium (Fluka analytical, 61747), 10 mM DTT (Melford, D11000), 2 mM 4-methylumbelliferyl-β-D-glucuronide (4-MUG, Sigma, M9130)) in a 200 µl total volume. The reaction was conducted at 37 ◦C in GREINER 96 F-BOTTOM microtiter plate using a CLARIOstar plate reader. 4-Methylumbelliferone (4-MU) fluorescence was recorded every 2 minutes for 20 cycles with excitation at 360 nm and emission detected at 450 nm. 4-MU concentration was determined based on a standard curve of ten 4-MU standards placed in the same plate. GUS enzymatic rates were calculated by averaging the slope of MU production from each of the duplicate reactions.

In order to visualize mTurquoise2 signal, mature leaves were dissected into 2 cm sections, leaf epidermal cells were removed by scraping the leaf surface with a razor blade and then mounted with deionized water. Imaging was then performed using a Leica TCS SP8 confocal laser-scanning microscope using a 20x air objective. mTurquoise2 fluorescence was excited at 442 nm with

emission at 471-481 nm, chlorophyll autofluorescence was excited at 488 nm with emission at 672-692 nm.

Yeast one hybrid, protoplast isolation and transactivation assay

The yeast one hybridisation assay was performed by Hybrigenics (https://www.hybrigenics-services.com/). DNA fragments were synthesized and used as bait, rice leaf and root cDNA libraries were used as prey. The number of clones screened and concentration of 3-AT were as follows: fragment 1, 70.2 million clones screened with 0 mM 3-AT; fragment 2, 61.5 million clones screened with 0 mM 3-AT; fragment 3, 68.4 million clones screened with 20 mM 3-AT; fragment 4, 57.4 million clones screened with 100 mM 3-AT; fragment 5, 94.2 million clones screened with 200 mM 3-AT.

For transactivation effector assays, transcription factors were cloned using primers list in Supplementary Figure 13 based on a number of selection criteria: first, transcription factors from families that were preferentially expressed in the bundle sheath of mature leaves (belonging to BS or BSV clusters in Hua et al., 2021[30]) including transcription factors that were preferentially expressed in BSS during leaf maturation (co-expression modules 15&17 in the leaf developmental gradient dataset (**Supplementary Data 5**)), and transcription factors that were co-expressed with *SiR* (module 0047) were identified from a publicly available co-expression network RiceGGM2021[125]. Additionally, other members of the IDD and group-A bZIP families were cloned (Supplementary Figure 13). Coding sequences of these candidate transcription factors were domesticated as described[116] and amplified using primers listed in **Supplementary Data 9** and prepared as level 0 SC module in the background of *pAGM9121*. They were then assembled into a level 1 module with a *ZmUBIpro* promoter and *Tnos* terminator as effector plasmids. Nucleotides -980 to -829 with the endogenous core promoter (nucleotide -250 to +42) were fused with the *LUCIFERASE* (*LUC*) reporter. All golden gate level 1 modules for protoplast transfection were extracted using ZymoPURE™ II Plasmid Midiprep Kit (Zymo research, D4201).

Rice leaf protoplasts and PEG-mediated transformation were performed as described previously[126]. In each transformation, 2 µg of transformation control plasmids (*ZmUBIpro::GUS-Tnos*), 5 µg of reporter plasmids, and 5 µg of effector plasmids per transcription factor were combined and mixed with 170 µl protoplasts. After incubation on the benchtop for overnight, protein was extracted using passive lysis buffer (Promega, E1941), GUS activity was determined with 20 µl of protein sample and MUG fluorescent assay as described above, LUC activity was measured with 20 µl of protein sample and 100 µl of LUC assay reagent (Promega, E1483) using Clariostar plate reader. Transcriptional activity from the promoter was calculated as LUC luminescence / rate of MUG accumulation per second.

Data availability

RNA sequencing data for bundle sheath strands and mesophyll from mature rice leaf (IR64) have been deposited in the Sequence Read Archive (SRA) under accession PRJNA1205909 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1205909]. RNA sequencing data of bundle sheath strands and mesophyll during leaf developmental gradient are available under accession PRJNA1205924 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1205924]. Sequences of *ZmPEPC* promoter (PQ873046 [https://www.ncbi.nlm.nih.gov/nuccore/PQ873046.1/]), *SiR* promoter (PQ873047 [https://www.ncbi.nlm.nih.gov/nuccore/PQ873047]), the bundle sheath CRM (PQ873048 [https://www.ncbi.nlm.nih.gov/nuccore/PQ873048]), three copies of the bundle sheath CRM (PQ873049 [https://www.ncbi.nlm.nih.gov/nuccore/PQ873049]), five copies of the bundle sheath CRM (PQ873050 [https://www.ncbi.nlm.nih.gov/nuccore/PQ873050]), *SiR* core promoter (PQ873051 [https://www.ncbi.nlm.nih.gov/nuccore/PQ873051]) and *PIP1.1* core promoter (PQ873052 [https://www.ncbi.nlm.nih.gov/nuccore/PQ873052]) are deposited in GenBank. Source data are provided with this paper.

Code availability

Scripts used for searching Y-patch and the motif sextet can be accessed at https://github.com/hibberd-lab/rice-bundle-sheath-cis-regulatory-module (http://doi.org/10.5281/zenodo.15681256).

References

1. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**, 565–570 (1978).
2. Krumlauf, R. Hox genes in vertebrate development. *Cell* **78**, 191–201 (1994).
3. Junion, G. *et al.* A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**, 473–486 (2012).
4. Drapek, C., Sparks, E. E. & Benfey, P. N. Uncovering gene regulatory networks controlling plant cell differentiation. *Trends in Genetics* **33**, 529–539 (2017).
5. Moreno-Risueno, M. A. *et al.* Transcriptional control of tissue formation throughout root development. *Science (1979)* **350**, 426–430 (2015).
6. Kubo, M. *et al.* Transcription switches for protoxylem and metaxylem vessel formation. *Genes Dev* **19**, 1855–1860 (2005).
7. Miyashima, S. *et al.* Mobile PEAR transcription factors integrate positional cues to prime cambial growth. *Nature* **565**, 490–494 (2019).
8. MacAlister, C. A., Ohashi-Ito, K. & Bergmann, D. C. Transcription factor control of asymmetric cell divisions that establish the stomatal lineage. *Nature* **445**, 537–540 (2006).
9. Ohashi-Ito, K. & Bergmann, D. C. *Arabidopsis* FAMA controls the final proliferation/differentiation switch during stomatal development. *Plant Cell* **18**, 2493–2505 (2006).
10. Pillitteri, L. J., Sloan, D. B., Bogenschutz, N. L. & Torii, K. U. Termination of asymmetric cell division and differentiation of stomata. *Nature* **445**, 501–505 (2006).
11. Kanaoka, M. M. *et al. SCREAM/ICE1* and *SCREAM2* Specify Three Cell-State Transitional Steps Leading to *Arabidopsis* Stomatal Differentiation. *Plant Cell* **20**, 1775–1785 (2008).
12. Nakashima, K., Ito, Y. & Yamaguchi-Shinozaki, K. Transcriptional regulatory networks in response to abiotic stresses in Arabidopsis and grasses. *Plant Physiol* **149**, 88 (2009).
13. Tsuda, K. & Somssich, I. E. Transcriptional networks in plant immunity. *New Phytol* **206**, 932–947 (2015).
14. McClung, C. R. Plant circadian rhythms. *Plant Cell* **18**, 792–803 (2006).
15. Nagel, D. H. & Kay, S. A. Complexity in the wiring and regulation of plant circadian networks. *Curr Biol* **22**, (2012).
16. Depuydt, S. & Hardtke, C. S. Hormone signalling crosstalk in plant growth regulation. *Curr Biol* **21**, (2011).
17. Verma, V., Ravindran, P. & Kumar, P. P. Plant hormone-mediated regulation of stress responses. *BMC Plant Biol* **16**, 1–10 (2016).
18. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Elsevier, 2006). doi:10.1016/B978-0-12-088563-3.X5018-4.
19. Adrian, J. *et al.* Cis-regulatory elements and chromatin state coordinately control temporal and spatial expression of *FLOWERING LOCUS T* in Arabidopsis. *Plant Cell* **22**, 1425–1440 (2010).
20. Liu, L. *et al.* Induced and natural variation of promoter length modulates the photoperiodic response of *FLOWERING LOCUS T*. *Nat Commun* **5**, (2014).

21. Clark, R. M., Wagler, T. N., Quijada, P. & Doebley, J. A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. *Nat Genet* **38**, 594–597 (2006).

22. Stam, M., Belele, C., Dorweiler, J. E. & Chandler, V. L. Differential chromatin structure within a tandem array 100 kb upstream of the maize *b1* locus is associated with paramutation. *Genes Dev* **16**, 1906–1918 (2002).

23. Schmitz, R. J., Grotewold, E. & Stam, M. Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *Plant Cell* **34**, 718–741 (2022).

24. Weber, B., Zicola, J., Oka, R. & Stam, M. Plant enhancers: a call for discovery. *Trends Plant Sci* **21**, 974–987 (2016).

25. Wang, L. *et al.* The maturation and aging trajectory of *Marchantia polymorpha* at single-cell resolution. *Dev Cell* **58**, 1429-1444.e6 (2023).

26. Wang, Y., Huan, Q., Li, K. & Qian, W. Single-cell transcriptome atlas of the leaf and root of rice seedlings. *Journal of Genetics and Genomics* **48**, 881–898 (2021).

27. Swift, J. *et al.* Exaptation of ancestral cell-identity networks enables C_4 photosynthesis. *Nature* **636**, 143–150 (2024).

28. Leegood, R. C. Roles of the bundle sheath cells in leaves of C_3 plants. *J Exp Bot* **59**, 1663–1673 (2008).

29. Aubry, S., Smith-Unna, R. D., Boursnell, C. M., Kopriva, S. & Hibberd, J. M. Transcript residency on ribosomes reveals a key role for the *Arabidopsis thaliana* bundle sheath in sulfur and glucosinolate metabolism. *Plant Journal* **78**, 659–673 (2014).

30. Hua, L. *et al.* The bundle sheath of rice is conditioned to play an active role in water transport as well as sulfur assimilation and jasmonic acid synthesis. *The Plant Journal* **107**, 268–286 (2021).

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

32. Makino, A., Sakuma, H., Sudo, E. & Mae, T. Differences between maize and rice in N-use efficiency for photosynthesis and protein allocation. *Plant Cell Physiol* **44**, 952–956 (2003).

33. Mitchell, P. L. & Sheehy, J. E. Supercharging rice photosynthesis to increase yield. *New Phytol* **171**, 688–693 (2006).

34. Hibberd, J. M., Sheehy, J. E. & Langdale, J. A. Using C_4 photosynthesis to increase the yield of rice—rationale and feasibility. *Curr Opin Plant Biol* **11**, 228–231 (2008).

35. Kajala, K. *et al.* Strategies for engineering a two-celled C_4 photosynthetic pathway into rice. *J Exp Bot* **62**, 3001–3010 (2011).

36. Aubry, S., Kelly, S., Kümpers, B. M. C. C., Smith-Unna, R. D. & Hibberd, J. M. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C_4 photosynthesis. *PLoS Genet* **10**, e1004365 (2014).

37. Ermakova, M. *et al.* Installation of C_4 photosynthetic pathway enzymes in rice using a single construct. *Plant Biotechnol J* **19**, 575–588 (2021).

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

39. Hua, L. & Hibberd, J. M. An optimized protocol for isolation of high-quality RNA through laser capture microdissection of leaf material. *Plant Direct* **3**, e00156 (2019).

40. Emmerling, J. Studies into the regulation of C_4 photosynthesis – towards factors controlling bundle sheath expression and Kranz anatomy development. *Doctoral thesis, Heinrich Heine University, Düsseldorf* (2018).

41. Brückner, K. *et al.* A library of synthetic transcription activator-like effector-activated promoters for coordinated orthogonal gene expression in plants. *The Plant Journal* **82**, 707–716 (2015).

42. Danila, F. *et al.* A single promoter-TALE system for tissue-specific and tuneable expression of multiple genes in rice. *Plant Biotechnol J* **20**, 1786–1806 (2022).

43. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

44. Fornes, O. *et al.* JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87–D92 (2020).

45. Civáň, P. & Švec, M. Genome-wide analysis of rice (*Oryza sativa* L. subsp. *japonica*) TATA box and Y Patch promoter elements. *Genome* **52**, 294–297 (2009).

46. Chow, C.-N. *et al.* PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Res* **47**, D1155–D1163 (2018).

47. Kim, S. Y., Ma, J., Perret, P., Li, Z. & Thomas, T. L. Arabidopsis ABI5 subfamily members have distinct DNA-binding and transcriptional activities. *Plant Physiol* **130**, 688 (2002).

48. Juven-Gershon, T. & Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* **339**, 225–229 (2009).

49. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19**, 621 (2018).

50. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).

51. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272–286 (2014).

52. Ray-Jones, H. & Spivakov, M. Transcriptional enhancers and their communication with gene promoters. *Cellular and Molecular Life Sciences* **78**, 6453 (2021).

53. Li, L., Zhu, Q., He, X., Sinha, S. & Halfon, M. S. Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. *Genome Biol* **8**, R101 (2007).

54. Buecker, C. & Wysocka, J. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet* **28**, 276 (2012).

55. Lee, D. Y. *et al.* Engineering chloroplast development in rice through cell-specific control of endogenous genetic circuits. *Plant Biotechnol J* **19**, 2291–2303 (2021).

56. Nomura, M. *et al.* Differential expression pattern of C_4 bundle sheath expression genes in rice, a C_3 plant. *Plant Cell Physiol* **46**, 754–761 (2005).

57. Li, E. *et al.* Long-range interactions between proximal and distal regulatory regions in maize. *Nat Commun* **10**, 2633 (2019).

58. Zhao, H. *et al.* Identification and functional validation of super-enhancers in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* **119**, e2215328119 (2022).

59. Liu, L. *et al. KRN4* controls quantitative variation in maize kernel row number. *PLoS Genet* **11**, e1005670 (2015).

60. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat Genet* **43**, 1160 (2011).

61. Yan, W. *et al.* Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. *Nat Commun* **10**, 1705 (2019).

62. Gallegos, J. E. & Rose, A. B. The enduring mystery of intron-mediated enhancement. *Plant Science* **237**, 8–15 (2015).

63. Rose, A. B. & Beliakoff, J. A. Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing. *Plant Physiol* **122**, 535–542 (2000).

64. Stergachis, A. B. *et al.* Exonic transcription factor binding directs codon choice and impacts protein evolution. *Science* **342**, 1367 (2013).

65. Reyna-Llorens, I. *et al.* Ancient duons may underpin spatial patterning of gene expression in C4 leaves. *Proc Natl Acad Sci U S A* **115**, 1931–1936 (2018).

66. Burgess, S. J. *et al.* Genome-wide transcription factor binding in leaves from C_3 and C_4 grasses. *Plant Cell* **31**, 2297–2314 (2019).

67. Yonekura-Sakakibara, K. *et al.* Analysis of reductant supply systems for ferredoxin-dependent sulfite reductase in photosynthetic and nonphotosynthetic organs of maize. *Plant Physiol* **122**, 887–894 (2000).

68. Khan, M. S. *et al.* Sulfite reductase defines a newly discovered bottleneck for assimilatory sulfate reduction and is essential for growth and development in *Arabidopsis thaliana*. *Plant Cell* **22**, 1216–1231 (2010).

69. Takahashi, H., Kopriva, S., Giordano, M., Saito, K. & Hell, R. Sulfur assimilation in photosynthetic organisms: Molecular functions and regulations of transporters and assimilatory enzymes. *Annu Rev Plant Biol* **62**, 157–184 (2011).

70. Shi, S. *et al.* OsHAC1;1 and OsHAC1;2 function as arsenate reductases and regulate arsenic accumulation. *Plant Physiol* **172**, 1708–1719 (2016).

71. Dickinson, P. J. *et al.* A bipartite transcription factor module controlling expression in the bundle sheath of *Arabidopsis thaliana*. *Nat Plants* **6**, 1468–1479 (2020).

72. Sewell, J. A. & Fuxman Bass, J. I. Options and considerations when using a Yeast One-Hybrid system. *Methods Mol Biol* **1794**, 119–130 (2018).

73. Yang, J. *et al. OsCLT1*, a CRT-like transporter 1, is required for glutathione homeostasis and arsenic tolerance in rice. *New Phytologist* **211**, 658–670 (2016).

74. Selles, B., Moseler, A., Rouhier, N. & Couturier, J. Rhodanese domain-containing sulfurtransferases: multifaceted proteins involved in sulfur trafficking in plants. *J Exp Bot* **70**, 4139–4154 (2019).

75. Yang, C., Lu, X., Ma, B., Chen, S. Y. & Zhang, J. S. Ethylene signaling in rice and Arabidopsis: conserved and diverged aspects. *Mol Plant* **8**, 495–505 (2015).

76. Dietzen, C. *et al.* The transcription factor EIL1 participates in the regulation of sulfur-deficiency response. *Plant Physiol* **184**, 2120–2136 (2020).

77. Maruyama-Nakashita, A., Nakamura, Y., Tohge, T., Saito, K. & Takahashi, H. *Arabidopsis* SLIM1 is a central transcriptional regulator of plant sulfur response and metabolism. *Plant Cell* **18**, 3235–3251 (2006).

78. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**, 449–479 (2003).

79. Yamamoto, Y. *et al.* Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Res* **35**, 6219 (2007).

80. Yamamoto, Y. Y. *et al.* Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* **8**, 1–23 (2007).

81. Bernard, V., Brunaud, V. & Lecharny, A. TC-motifs at the TATA-box expected position in plant genes: A novel class of motifs involved in the transcription regulation. *BMC Genomics* **11**, 1–15 (2010).

82. Jores, T. *et al.* Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat Plants* **7**, 842–855 (2021).

83. Shimizu, H. *et al.* LIP19, a basic region leucine zipper protein, is a Fos-like molecular switch in the cold signaling of rice plants. *Plant Cell Physiol* **46**, 1623–1634 (2005).

84. Zhang, C. *et al.* Transcriptomic analysis reveals the role of FOUR LIPS in response to salt stress in rice. *Plant Mol Biol* **110**, 37–52 (2022).

85. Zhang, H. *et al.* EAR motif mutation of rice OsERF3 alters the regulation of ethylene biosynthesis and drought tolerance. *Planta* **237**, 1443–1451 (2013).

86. Chen, X. *et al.* The NAC family transcription factor OsNAP confers abiotic stress response through the ABA pathway. *Plant Cell Physiol* **55**, 604–619 (2014).

87. Qu, X. *et al.* A rice R2R3-Type MYB transcription factor OsFLP positively regulates drought stress response via OsNAC. *Int J Mol Sci* **23**, 5873 (2022).

88. Lin, R., Zhao, W., Meng, X. & Peng, Y. L. Molecular cloning and characterization of a rice gene encoding AP2/EREBP-type transcription factor and its expression in response to infection with blast fungus and abiotic stresses. *Physiol Mol Plant Pathol* **70**, 60–68 (2007).

89. Jung, S. E. *et al.* Overexpression of *OsERF83*, a vascular tissue-specific transcription factor gene, confers drought tolerance in rice. *Int J Mol Sci* **22**, 7656 (2021).

90. Jisha, V. *et al.* Overexpression of an AP2/ERF Type transcription factor *OsEREBP1* confers biotic and abiotic stress tolerance in rice. *PLoS One* **10**, e0127831 (2015).

91. Tezuka, D. *et al.* The rice ethylene response factor OsERF83 positively regulates disease resistance to *Magnaporthe oryzae*. *Plant Physiol Biochem* **135**, 263–271 (2019).

92. Kurt, F., Filiz, E. & Aydın, A. Sulfite Reductase (SiR) gene in rice (*Oryza sativa*): bioinformatics and expression analyses under salt and drought Stresses. *J Plant Growth Regul* **41**, 2246–2260 (2022).

93. Binder, B. M. Ethylene signaling in plants. *J Biol Chem* **295**, 7710 (2020).

94. Yaish, M. W. *et al.* The APETALA-2-like transcription factor OsAP2-39 controls key interactions between abscisic acid and gibberellin in rice. *PLoS Genet* **6**, (2010).

95. Zhang, C. *et al.* Suppression of Jasmonic Acid-mediated defense by viral-inducible microRNA319 facilitates virus infection in rice. *Mol Plant* **9**, 1302–1314 (2016).

96. Zhang, X. *et al. Magnaporthe oryzae* induces the expression of a microRNA to suppress the immune response in rice. *Plant Physiol* **177**, 352 (2018).

97. Sparks, E. E. *et al.* Establishment of expression in the *SHORTROOT-SCARECROW* transcriptional cascade through opposing activities of both activators and repressors. *Dev Cell* **39**, 585–596 (2016).

98. Thanos, D. & Maniatis, T. Virus induction of human *IFNβ* gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).

99. Panne, D. The enhanceosome. *Curr Opin Struct Biol* **18**, 236–242 (2008).

100. Kulkarni, M. M. & Arnosti, D. N. Information display by transcriptional enhancers. *Development* **130**, 6569–6575 (2003).

101. Xiong, H. *et al.* Photosynthesis-independent production of reactive oxygen species in the rice bundle sheath during high light is mediated by NADPH oxidase. *Proc Natl Acad Sci U S A* **118**, e2022702118 (2021).

102. Cui, H., Kong, D., Liu, X. & Hao, Y. SCARECROW, SCR-LIKE 23 and SHORT-ROOT control bundle sheath cell fate and function in *Arabidopsis thaliana*. *Plant Journal* **78**, 319–327 (2014).

103. Kirschner, S. *et al.* Expression of *SULTR2;2*, encoding a low-affinity sulphur transporter, in the Arabidopsis bundle sheath and vein cells is mediated by a positive regulator. *J Exp Bot* **69**, 4897–4906 (2018).

104. Engelmann, S. *et al.* The gene for the P-subunit of glycine decarboxylase from the C_4 species *Flaveria trinervia*: analysis of transcriptional control in transgenic *Flaveria bidentis* (C_4) and *Arabidopsis* (C_3). *Plant Physiol* **146**, 1773–1785 (2008).

105. Wiludda, C. *et al.* Regulation of the photorespiratory *GLDPA* gene in C_4 flaveria: an intricate interplay of transcriptional and posttranscriptional processes. *Plant Cell* **24**, 137–151 (2012).

106. Glick, B. R. Metabolic load and heterologous gene expression. *Biotechnol Adv* **13**, 247–261 (1995).

107. Que, Q., Wang, H. Y., English, J. J. & Jorgensen, R. A. The frequency and degree of cosuppression by sense Chalcone Synthase transgenes are dependent on transgene promoter strength and are reduced by premature nonsense codons in the transgene coding sequence. *Plant Cell* **9**, 1357 (1997).

108. Kummari, D. *et al.* An update and perspectives on the use of promoters in plant genetic engineering. *J Biosci* **45**, 1–24 (2020).

109. Hiei, Y. & Komari, T. *Agrobacterium*-mediated transformation of rice using immature embryos or calli induced from mature seed. *Nat Protoc* **3**, 824–834 (2008).

110. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).

111. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

112. Robinson, M. D., Mccarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *BIOINFORMATICS APPLICATIONS NOTE* **26**, 139–140 (2010).

113. Goodstein, D. M. *et al.* Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res* **40**, D1178 (2012).

114. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2016).

115. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 1–13 (2008).

116. Marillonnet, S. & Grützner, R. Synthetic DNA assembly using Golden Gate cloning and the hierarchical modular cloning pipeline. *Curr Protoc Mol Biol* **130**, e115 (2020).

117. Engler, C. *et al.* A golden gate modular cloning toolbox for plants. *ACS Synth Biol* **3**, 839–843 (2014).

118. Gupta, S. Das *et al.* The *C4Ppc* promoters of many C_4 grass species share a common regulatory mechanism for gene expression in the mesophyll cell. *The Plant Journal* **101**, 204–216 (2020).

119. Matsuoka, M., Kyozuka, J., Shimamoto, K. & Kano-Murakami, Y. The promoters of two carboxylases in a C_4 plant (maize) direct cell-specific, light-regulated expression in a C_3 plant (rice). *The Plant Journal* **6**, 311–319 (1994).

120. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res* **43**, 39–49 (2015).

121. Fornes, O. *et al.* JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87–D92 (2020).

122. Molina, C. & Grotewold, E. Genome wide analysis of Arabidopsis core promoters. *BMC Genomics* **6**, 1–12 (2005).

123. Castro-Mondragon, J. A., Jaeger, S., Thieffry, D., Thomas-Chollier, M. & Van Helden, J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res* **45**, e119–e119 (2017).

124. Jefferson, R. A., Kavanagh, T. A. & Bevan, M. W. GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J* **6**, 3901–7 (1987).

125. Zhang, Y. *et al.* Rice co-expression network analysis identifies gene modules associated with agronomic traits. *Plant Physiol* **190**, 1526–1542 (2022).

126. Page, M. T., Parry, M. A. J. & Carmo-Silva, E. A high-throughput transient expression system for rice. *Plant Cell Environ* **42**, 2057–2064 (2019).

Acknowledgements

This work was funded by the Bill and Melinda Gates Foundation C_4 Rice grant awarded to the University of Oxford (2015-2019 OPP1129902 and 2019-2024 INV-002970) L.H., N.W., S.S., R.D., K.B., S.K.E., A.R.B., and J.M.H.. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Author contributions Statement

L.H. and J.M.H. conceived the work. J.M.H. guided execution of experiments and oversaw the project. L.H., N.W., S.S., R.D., K.B., S.K.E and A.R.B. did the experiments and analysed the data. L.H. and J.M.H. wrote the manuscript with input from all authors.

Competing Interests Statement

The authors declare no competing interests.

Figure legends

Fig. 1: Nucleotides -980 to -394 of the *SiR* promoter are necessary for bundle sheath expression.

a, Domesticated *SiR* promoter generates strong GUS staining in bundle sheath. **b**, mTurquoise2 signal driven by the domesticated *SiR* promoter in nuclei of bundle sheath cells (marked by yellow dashed lines) and vein cells in mature leaves, red indicates chlorophyll autofluorescence. **c**, The fluorometric 4-methylumbelliferyl-β-D-glucuronide (MUG) assay shows no statistically significant difference between the endogenous and domesticated *SiR* promoter activity, data subjected to two-sided Wilcoxon rank-sum test, n indicates the number of biological replicates (independent T_0 transgenic plants), median catalytic rate of GUS indicated with red lines. Source data are provided as a Source Data file. **d**, Landscape of transcription factor binding sites in the *SiR* promoter using the Find Individual Motif Occurrences (FIMO) program. The likelihood of match to 656 plant known transcription factor motifs in the *SiR* promoter is shown by transcription factor families (**Supplementary Data 1**), *P* values calculated from the log-likelihood score by the FIMO program. **e**, Schematics showing 5' truncations. **f**, Representative images of leaf cross sections from transgenic lines after GUS staining. Zoomed-in images of lateral veins shown in right panels, the staining duration is displayed in the bottom-left corner, bundle sheath cells highlighted with dashed red line, scale bars = 50 μm. **g**, Promoter activity determined by the fluorometric 4-methylumbelliferyl-β-D-glucuronide (MUG) assay. Data subjected to pairwise two-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction for multiple comparison. Lines with differences in activity that were statistically significant (adjusted $P<0.05$) were labelled with different letters. The median represents the median GUS activity value and is shown as red line, n indicates the number of T_0 transgenic plants analyzed. Source data are provided as a Source Data file.

Fig. 2: A distal cis-regulatory module (CRM) and the core promoter that are necessary and sufficient for bundle sheath expression.

a, Schematics showing deletions of nucleotides -980 to -829 and -251 to +42. **b**, Representative image of leaf cross sections of transgenic lines after GUS staining. Zoomed-in images of lateral veins shown in right panels, the staining duration is displayed in the bottom-left corner, bundle sheath cells highlighted with dashed red line, scale bars = 50 μm. **c**, Promoter activity determined by the fluorometric 4-methylumbelliferyl-β-D-glucuronide (MUG) assay. Data subjected to pairwise two-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction for multiple comparison. Lines with differences in activity that were statistically significant (adjusted $P < 0.05$) were labelled with different letters. The median represents the median GUS activity value and is shown as red line in the plot, n indicates the number of T_0 transgenic plants analyzed. Source data are provided as a Source Data file.

Fig. 3: The Y-patch and four distinct regions in the distal cis-regulatory module (CRM) are required for bundle sheath specific expression.

a-**c**, Nucleotides -980 and -829 from the *SiR* promoter pattern expression to the bundle sheath when linked with the *PIP1;1* and *NRT1.1A* core promoters containing Y-patches. **a**, Schematics showing Y-patch and TATA-box in core promoters of *ATPSb, PIP1;3, NRT1.1A* and *PIP1;1* which have been used to initiate the transcription of nucleotides -980 and -829. **b**, Representative cross sections of transgenic rice leaves after GUS staining, zoomed-in image of lateral veins shown in the right panel, bundle sheath cells highlighted with red dashed lines, the staining duration is displayed in the bottom-left corner, scale bars = 50 μm. **c**, Promoter activity determined by the fluorometric 4-methylumbelliferyl-β-D-glucuronide (MUG) assay. **d**, Schematics showing transcription factor binding sites between nucleotides -980 and -829. **e**, Schematics showing consecutive deletions between nucleotides -980 and -829 fused to the GUS reporter. **f**, Representative images of cross sections from transgenic lines after GUS staining, zoomed-in images of lateral veins shown in right panels, the staining duration is displayed in the bottom-left corner, bundle sheath cells highlighted with red dashed lines, scale bars = 50 μm. **g**, Promoter activity determined by the fluorometric 4-methylumbelliferyl-β-D-glucuronide (MUG) assay. In **c**&**g**, data were subjected to pairwise two-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction. Lines with differences in activity that were statistically significant (adjusted $P < 0.05$) were labelled with different letters. Median catalytic rate of GUS indicated with red lines, n indicates total number of T_0 transgenic plants assessed. Source data are provided as a Source Data file.

Fig. 4: WRKY, G2-like, MYB-related, IDD and bZIP transcription factors interact and activate with the distal cis-regulatory module (CRM).

a, Schematics showing transcription factor binding sites between nucleotides -980 and -829 which are likely required for bundle sheath specific expression. **b**, Effector assays showing that each transcription factor activates expression from the distal CRM. **c**, Effector assays showing synergistic activation from the distal CRM when GLK2 and IDD3,4,6,10 were co-expressed. In **b** and **c**, LUC/GUS ratio from four biological replicates were $\log_2$ transformed and analyzed using pairwise t-tests with Benjamini-Hochberg correction for multiple comparisons. Statistically significant differences (adjusted $P < 0.05$) are indicated by different letters. Box plots show the 25th, median, and 75th percentiles; whiskers extend to the most extreme values within 1.5× the interquartile range. The assay was independently repeated three times with similar results. Source data are provided as a Source Data file. **d**, Transcript abundance of transcription factors in bundle sheath strands (BSS) and mesophyll (M) cells during maturation. Leaf developmental stage S2 to S7 represent base of the 4th leaf at the 6th, 8th, 9th, 10th, 13th and 17th day after sowing. **e**, Representative images of transgenic lines mis-expressing WRKY121, IDD2 and bZIP9 in mesophyll cells, staining duration is displayed in the bottom-left corner, zoom-in of mesophyll shown in right panel, red arrows indicate GUS expressing mesophyll cells. The bundle sheath CRM activity is determined by the fluorometric 4-methylumbelliferyl-β-D-glucuronide (MUG) assay, data subjected to pairwise two-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction. Lines showing statistically significant differences in activity (adjusted $P < 0.05$) were labeled with different letters. The median represents the median GUS activity value and is shown as red lines in the plot, n indicates the number of T_0 transgenic plants analyzed. Source data are provided as a Source Data file.

Fig. 5: Oligomerisation of bundle sheath cis-regulatory module (CRM) increases bundle sheath expression. a&**d**, Schematics showing site-directed mutagenesis of WRKY, G2-like, MYBR, IDD and bZIP motifs, mutated nucleotides highlighted in red (**a**), and constructs to test the impact of oligomerization of the CRM (**d**). **b**&**e**, Representative images of cross sections from transgenic lines after GUS staining, zoomed-in images of lateral veins shown in right panel, the staining duration is displayed in the bottom-left corner, bundle sheath cells highlighted with red dashed lines, scale bars = 50 µm. **c**&**f**, Promoter activity determined by the fluorometric 4-methylumbelliferyl-β-D-glucuronide (MUG) assay. Data subjected to pairwise two-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction. Lines with differences in activity that were statistically significant (adjusted $P < 0.05$) were labelled with different letters. Median catalytic rate of GUS indicated with red lines, n indicates total number of T_0 transgenic plants assessed. Source data are provided as a Source Data file. **g**, Paradermal view of Arabidopsis leaf expressing GUS under the control of 3x BS CRM combined with *OsSiR* core promoter, the staining duration is displayed in the bottom-left corner. M indicate mesophyll, BS for bundle sheath, V for vein. Zoomed in images for the red box shown on right, bundle sheath highlighted with red dash lines.

Fig. 6: Model of mechanism underpinning bundle sheath expression from *SiR* promoter.

a, Schematic with location of the bundle sheath (BS) cis-regulatory module (CRM), constitutive activators and mesophyll silencer. **b**, Bundle sheath expression is a result of the CRM, constitutive activators and mesophyll silencer acting in concert. Schematic indicating how the CRM operates within a broader *cis*-regulatory landscape. **c**-**e**, Model depicting transcription factors and cognate *cis*-elements responsible for bundle sheath expression (**d**) and mesophyll repression (**e**).

Figure 1

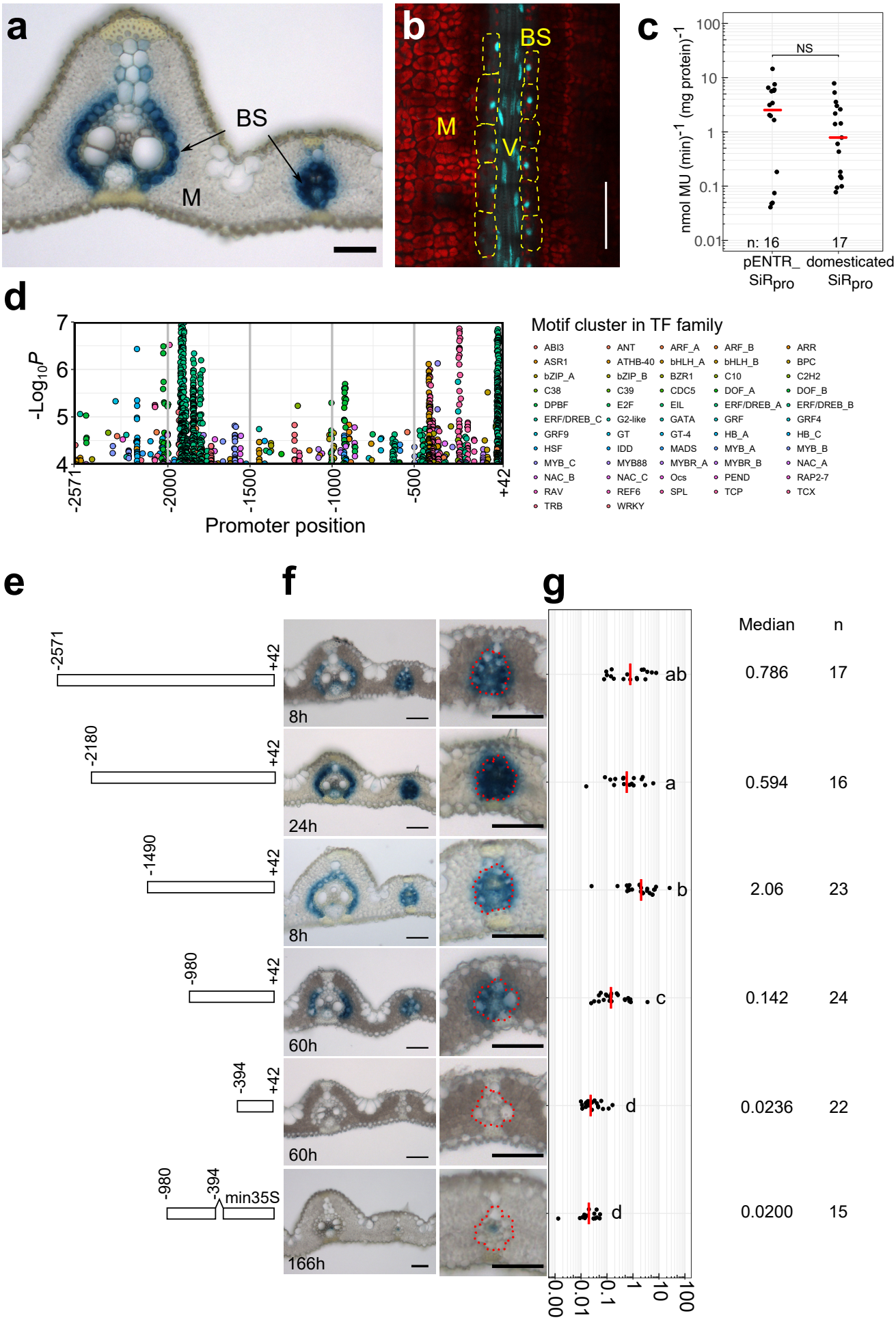

a
BS
M
b
BS
M
V
c
NS
nmol MU (min)-1 (mg protein)-1
100
10
1
0.1
0.01
n: 16
17
pENTR_SiRpro
domesticated SiRpro
d
-Log10P
7
6
5
4
-2571
-2000
-1500
-1000
-500
+42
Promoter position
Motif cluster in TF family
ABI3
ANT
ARF_A
ARF_B
ARR
ASR1
ATHB-40
bHLH_A
bHLH_B
BPC
bZIP_A
bZIP_B
BZR1
C10
C2H2
C38
C39
CDC5
DOF_A
DOF_B
DPBF
E2F
EIL
ERF/DREB_A
ERF/DREB_B
ERF/DREB_C
G2-like
GATA
GRF
GRF4
GRF9
GT
GT-4
HB_A
HB_C
HSF
IDD
MADS
MYB_A
MYB_B
MYB_C
MYB88
MYBR_A
MYBR_B
NAC_A
NAC_B
NAC_C
Ocs
PEND
RAP2-7
RAV
REF6
SPL
TCP
TCX
TRB
WRKY
e
-2571
+42
-2180
+42
-1490
+42
-980
+42
-394
+42
-980
-394
min35S
f
8h
24h
8h
60h
60h
166h
g
Median
n
ab
0.786
17
a
0.594
16
b
2.06
23
c
0.142
24
d
0.0236
22
d
0.0200
15
0.001
0.01
0.1
1
10
100
nmol MU min-1(mg protein)-1

Figure 2

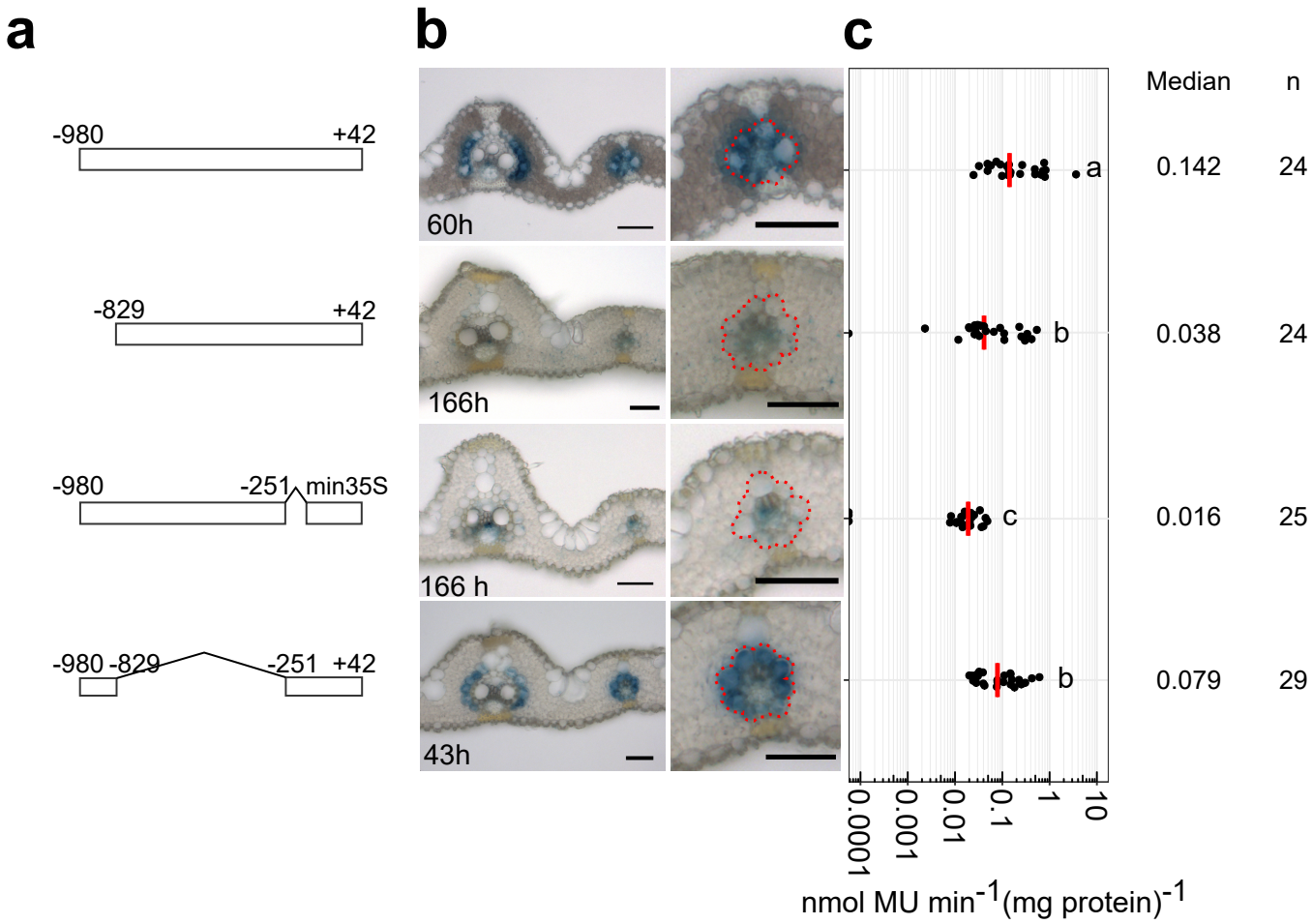

a
b
c
-980
+42
-829
+42
-980
-251
min35S
-980
-829
-251
+42
60h
166h
166 h
43h
Median
n
a
0.142
24
b
0.038
24
c
0.016
25
b
0.079
29
0.0001
0.001
0.01
0.1
1
10
nmol MU min-1(mg protein)-1

Figure 3

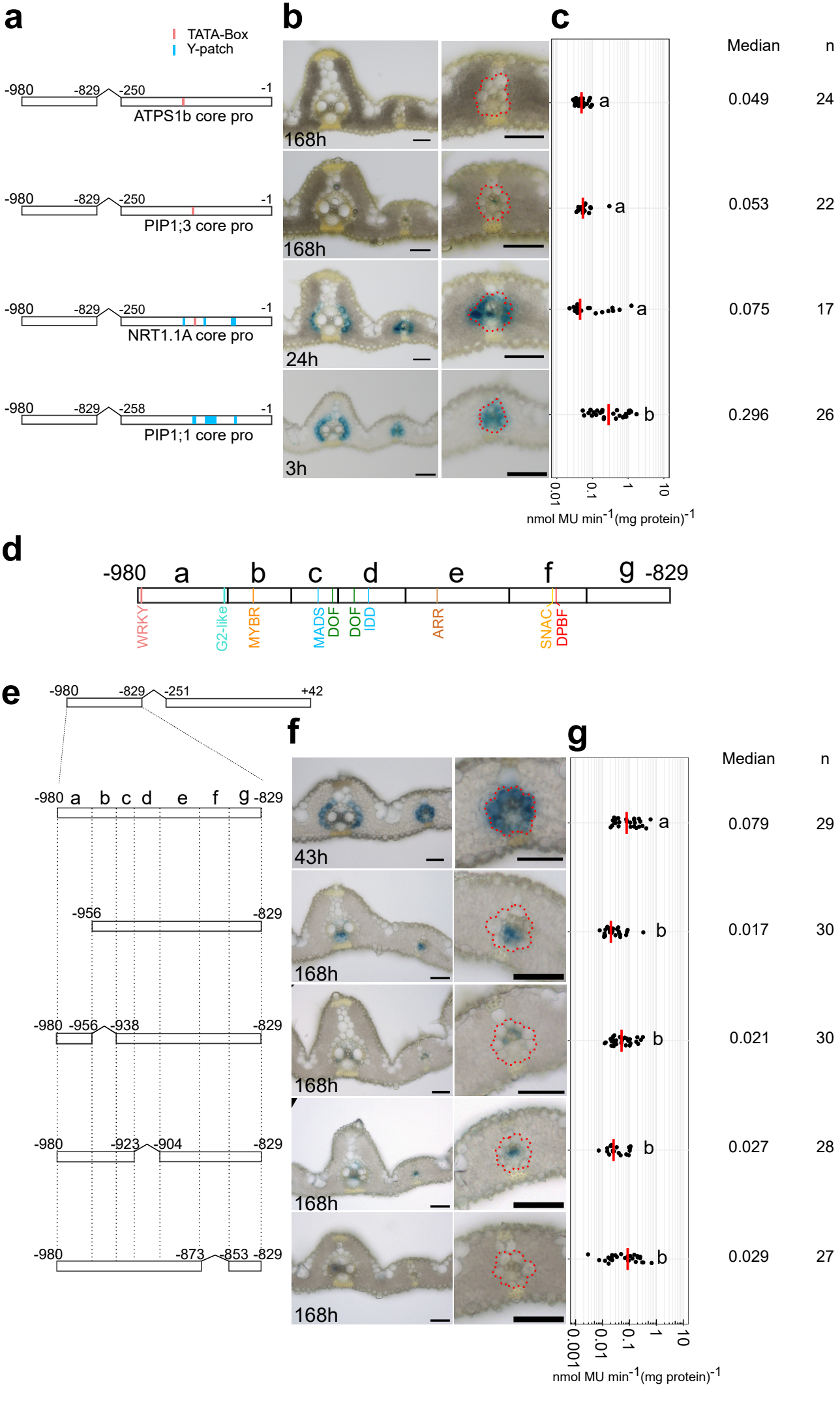

a
TATA-Box
Y-patch
-980
-829
-250
-1
ATPS1b core pro
PIP1;3 core pro
NRT1.1A core pro
-258
PIP1;1 core pro
b
168h
24h
3h
c
Median
n
0.049
24
0.053
22
0.075
17
0.296
26
0.01
0.1
1
10
nmol MU min-1(mg protein)-1
d
-980
a
b
c
d
e
f
g
-829
WRKY
G2-like
MYBR
MADS
DOF
IDD
ARR
SNAC
DPBF
e
-980
-829
-251
+42
-956
-938
-923
-904
-873
-853
f
43h
g
0.079
29
0.017
30
0.021
30
0.027
28
0.029
27
0.001
nmol MU min-1(mg protein)-1

Figure 4

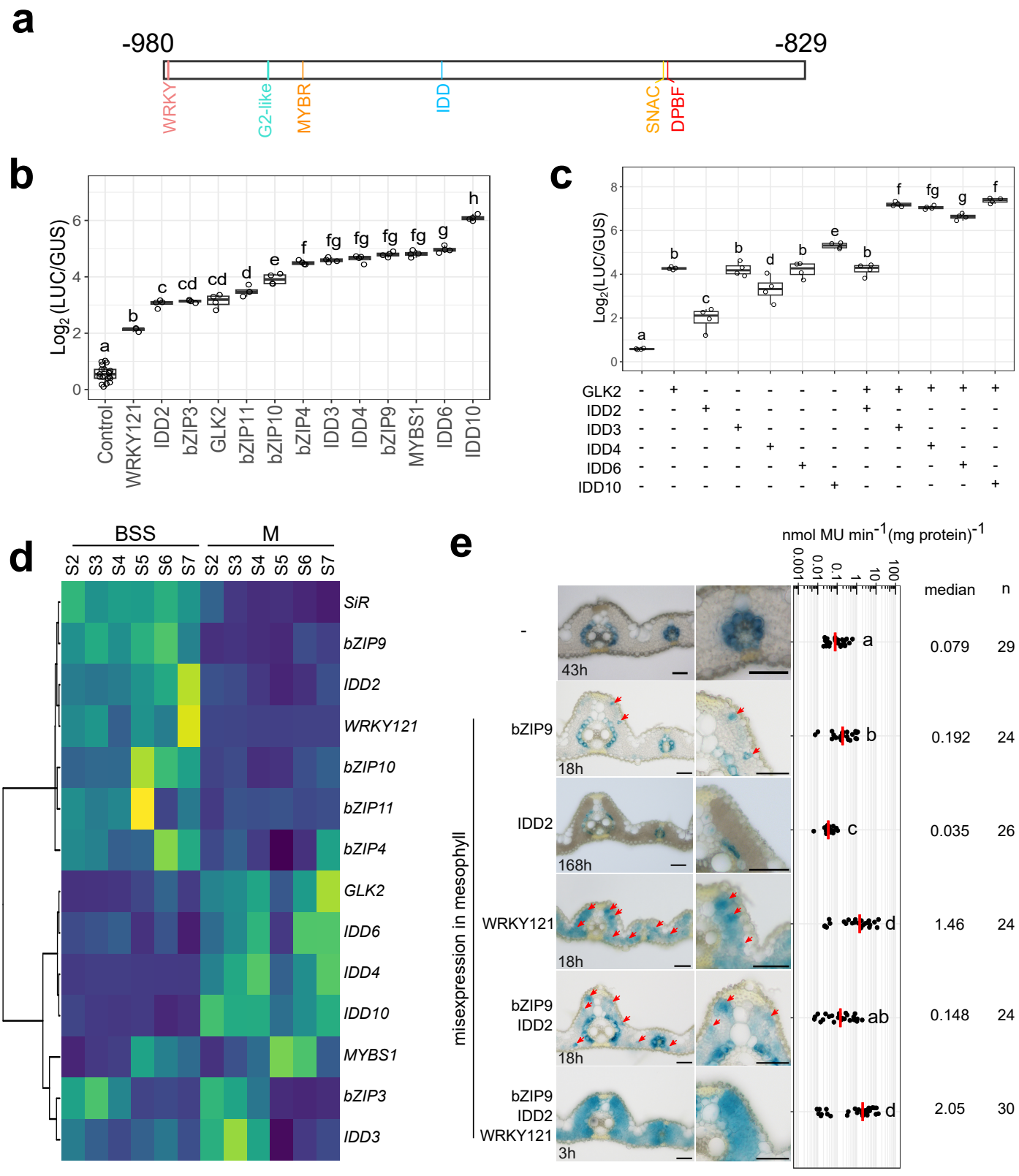
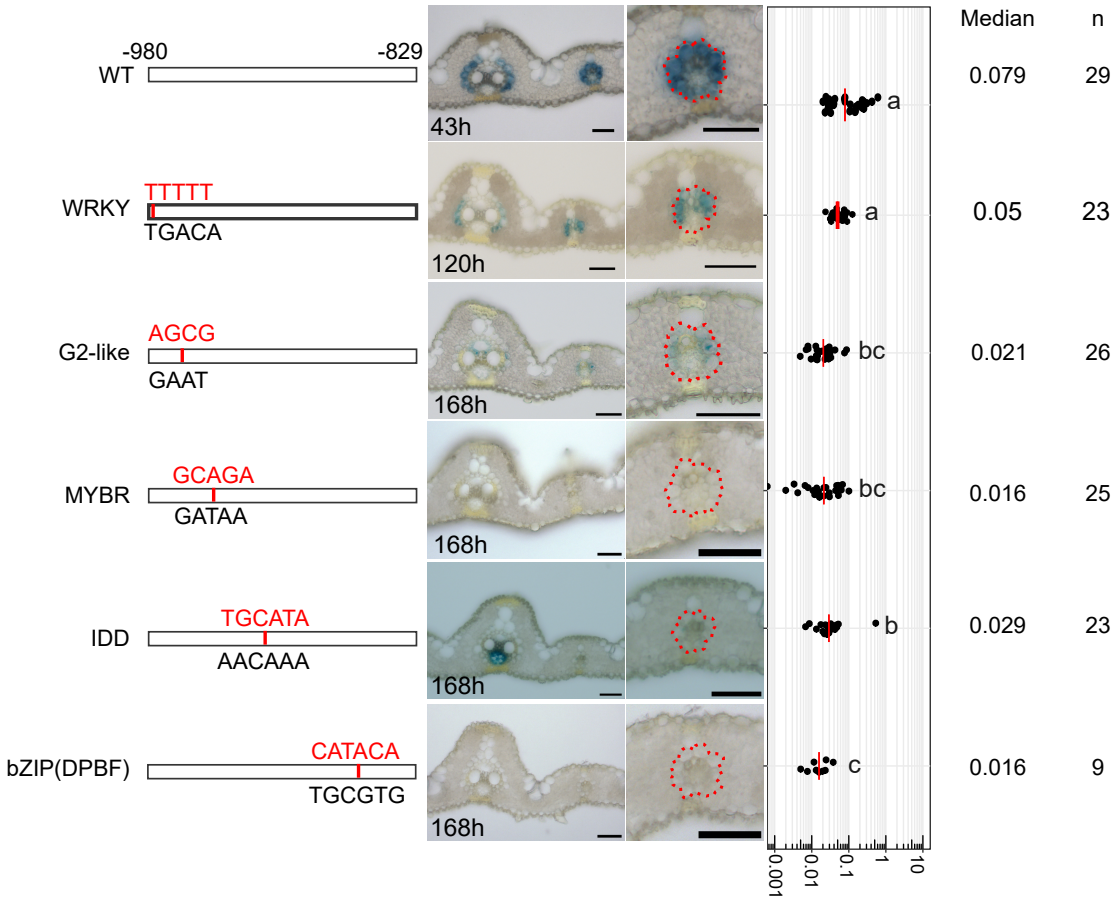
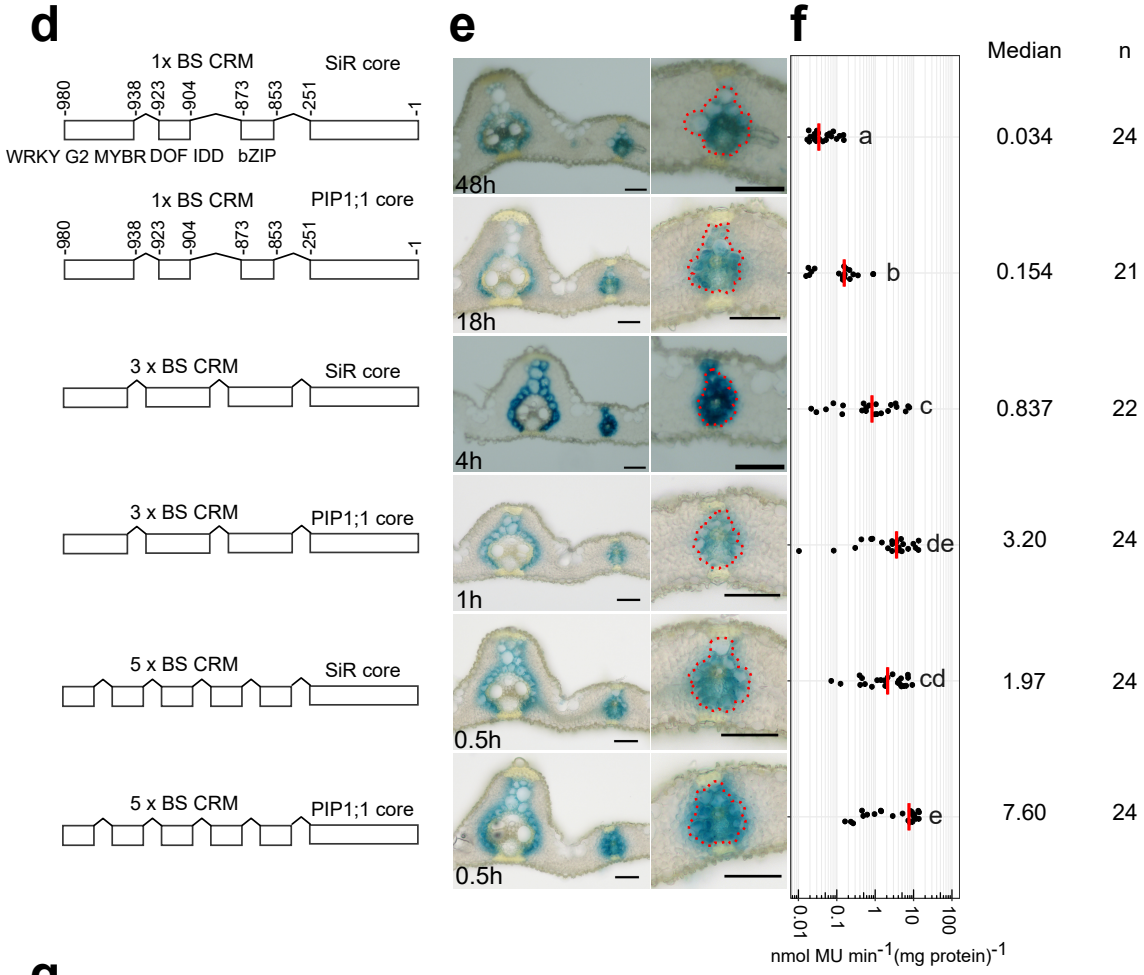

a
-980
-829
WRKY
G2-like
MYBR
IDD
SNAC
DPBF
b
Log2 (LUC/GUS)
Control
WRKY121
IDD2
bZIP3
GLK2
bZIP11
bZIP10
bZIP4
IDD3
IDD4
bZIP9
MYBS1
IDD6
IDD10
c
Log2(LUC/GUS)
GLK2
IDD2
IDD3
IDD4
IDD6
IDD10
d
BSS
M
S2 S3 S4 S5 S6 S7 S2 S3 S4 S5 S6 S7
SiR
bZIP9
IDD2
WRKY121
bZIP10
bZIP11
bZIP4
GLK2
IDD6
IDD4
IDD10
MYBS1
bZIP3
IDD3
e
nmol MU min-1(mg protein)-1
0.001 0.01 0.1 1 10 100
median
n
misexpression in mesophyll
-
43h
bZIP9
18h
IDD2
168h
WRKY121
18h
bZIP9
IDD2
18h
bZIP9
IDD2
WRKY121
3h
a 0.079 29
b 0.192 24
c 0.035 26
d 1.46 24
ab 0.148 24
d 2.05 30

Figure 5


a
b
c
-980
-829
WT
WRKY
TTTTT
TGACA
G2-like
AGCG
GAAT
MYBR
GCAGA
GATAA
IDD
TGCATA
AACAAA
bZIP(DPBF)
CATACA
TGCGTG
43h
120h
168h
168h
168h
168h
Median
n
0.079
29
0.05
23
0.021
26
0.016
25
0.029
23
0.016
9
a
a
bc
bc
b
c
0.001
0.01
0.1
1
10


d
e
f
1x BS CRM
SiR core
-980
-938
-923
-904
-873
-853
-251
-1
WRKY G2 MYBR DOF IDD bZIP
1x BS CRM
PIP1;1 core
3 x BS CRM
SiR core
3 x BS CRM
PIP1;1 core
5 x BS CRM
SiR core
5 x BS CRM
PIP1;1 core
48h
18h
4h
1h
0.5h
0.5h
Median
n
0.034
24
0.154
21
0.837
22
3.20
24
1.97
24
7.60
24
a
b
c
de
cd
e
0.01
0.1
1
10
100
nmol MU min-1(mg protein)-1

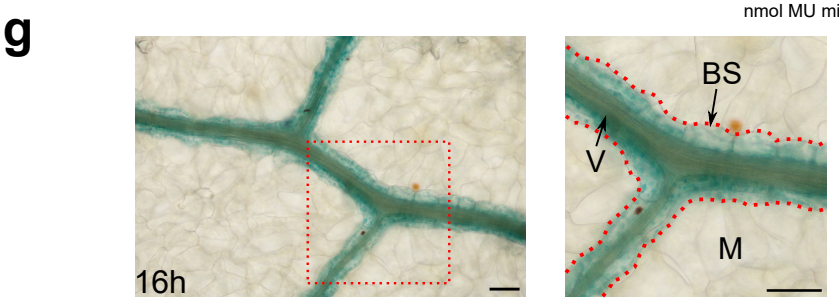
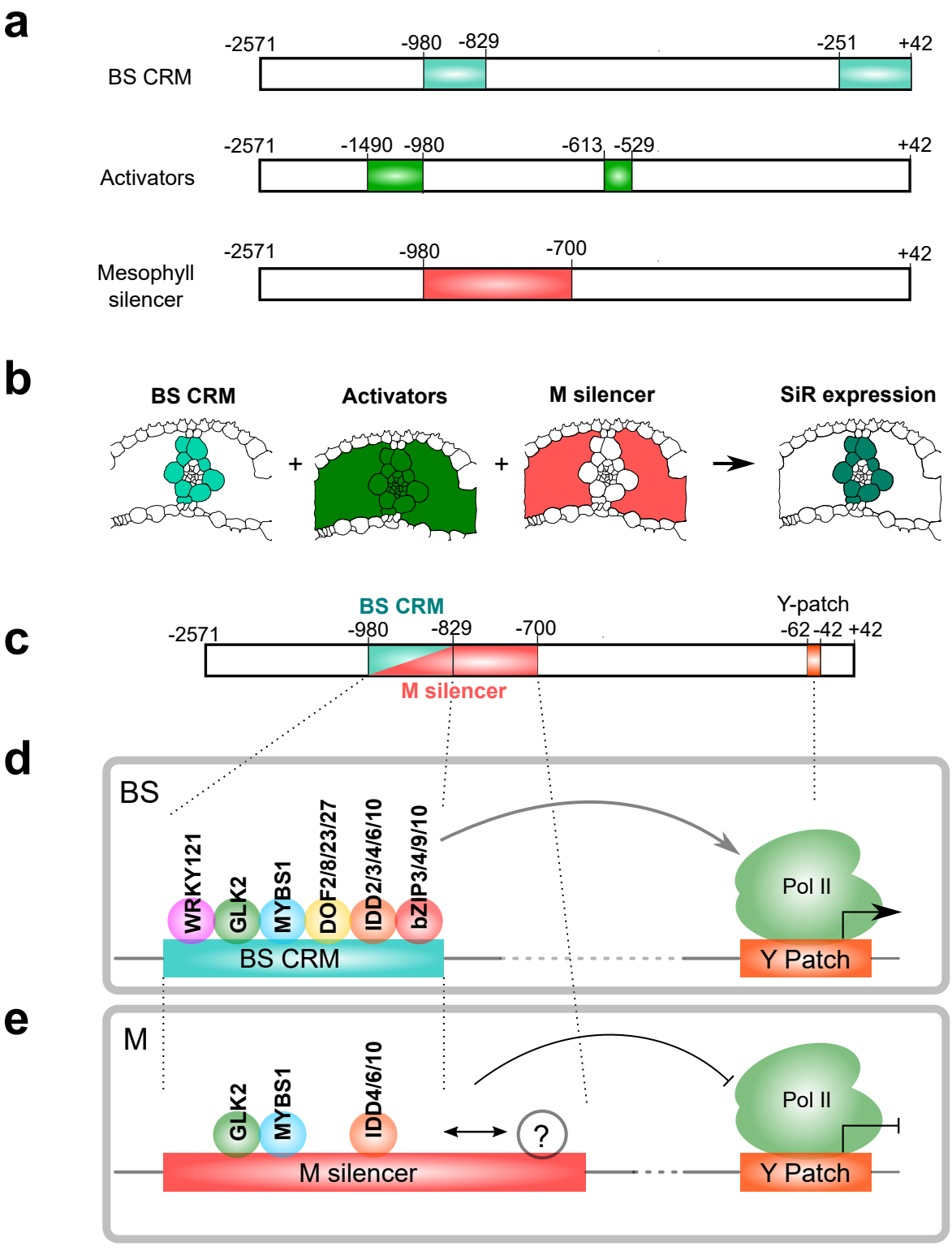

g
16h
BS
V
M

Figure 6


a
BS CRM
-2571
-980
-829
-251
+42
Activators
-1490
-613
-529
Mesophyll silencer
-700
b
BS CRM
Activators
M silencer
SiR expression
c
Y-patch
-62
-42
M silencer
d
BS
WRKY121
GLK2
MYBS1
DOF2/8/23/27
IDD2/3/4/6/10
bZIP3/4/9/10
Pol II
Y Patch
e
M
IDD4/6/10
?