Studies into the Regulation of C_4 Photosynthesis – Towards Factors Controlling Bundle Sheath Expression and Kranz Anatomy Development

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Jan Emmerling
aus Aachen

Düsseldorf, September 2018

aus dem Institut für Entwicklungs- und Molekularbiologie der Pflanzen
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Peter Westhoff

2. Prof. Dr. Maria von Korff Schmising

Tag der mündlichen Prüfung: 17.12.2018

Eidesstattliche Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der “Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist. Die Dissertation habe ich in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 21.09.2018

(Jan Emmerling)

Content

Abbreviations

2PG	2-Phospoglycolate
3PG	3-Phosphoglycerate
ABA	Abscisic acid
AtHB8	*Arabidopsis thaliana* homeobox gene 8
AtML1	*Arabidopsis thaliana* MERISTEM LAYER 1
ATP	Adenosine triphosphate
BASS2	Bile acid/sodium symporter 2
bp	Base pairs
BSA	Bovine serum albumin
BSC	Bundle sheath cell
BSM1	Bundle sheath motif 1
CA	Carbonic anhydrase
CaMV	Cauliflower Mosaic Virus
CAT	Catalase
CBB-cycle	Calvin-Bassham-Benson cycle
CCM	Carbon concentrating mechanism
cDNA	Complementary DNA
CDS	Coding sequence
CRE	*Cis*-regulatory element
cTP	Chloroplast transit peptide
DiT1	Dicarboxylate transporter 1
DiT2	Dicarboxylate transporter 1
DNA	Deoxyribonucleic acid
DOF	DNA-BINDING WITH ONE ZINC-FINGER
Fd-GOGAT	Glutamine:oxoglutarate aminotransferase
G2	GOLDEN 2
GDC	Glycine decarboxylase
GGAT	glyoxylate:glutamate aminotransferase
GLDH	Glycine decarboxylase subunit H
GLDL	Glycine decarboxylase subunit L
GLDP	Glycine decarboxylase subunit P
GLDT	Glycine decarboxylase subunit T
GLK	G2-like
GLYK	Glycerate kinase
GOX	Glycolate oxidase
GS	Glutamine synthetase
GUS	ß-glucuronidase
H2B	Histone 2B
HPR	Hydroxypyruvate reductase
hpRNA	hairpin RNA
IVD	Interveinal distance
kb	Kilo base pairs
kDa	Kilo Dalton
LCA	Last common ancestor
Mal	Malate

MC	Mesophyll cell
MDH	Malate dehydrogenase
MEM1	Mesophyll expression module 1
MITE	Miniature inverted-repeat transposable element
MP	MONOPTEROS
mRNA	messenger RNA
MU	Methylumbelliferone
NAD	Nicotinamide adenine dinucleotide
NAD-ME	NAD-dependent malic enzyme
NADP	Nicotinamide adenine dinucleotide phosphate
NADP-ME	NADP-dependent malic enzyme
OAA	Oxaloacetic acid
P1 to P5	Plastochrons 1 to 5
PCK	Phospoenolpyruvate carboxykinase
PCR	Polymerase chain reaction
PEPC	Phosphoenolpyruvate carboxylase
PGLP	Phosphoglycolate phosphatase
PIN1	PIN-FORMED 1
PLGG1	Plastidial glycerate/glycolate antiporter
PPDK	Pyruvate orthophosphate dikinase
PPT	Phosphate/PEP translocator
PTGS	Post-transcriptional gene silencing
PWM	Positional weight matrix
Pyr	Pyruvate
qRT-PCR	Quantitative real time PCR
RACE	Rapid amplification of cDNA-ends
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
Rubisco	RubP carboxylase/oxidase
RubP	Ribulose-1,5-bisphosphate
SCR	SCARECROW
SGAT	serine:glyoxylate aminotransferase
SHMT	Serine hydroxymethyltransferase
SHR	SHORTROOT
SNP	Single nucleotide polymorphism
T-DNA	Transfer DNA
TE	Transposable element
TF	Transcription Factor
THF	Tetrahydrofolate
TIR	Terminal inverted repeat
TSD	Target site duplication
TSS	Transcriptional start site
UTR	Untranslated region
YFP	Yellow fluorescent protein
ZmUbi1	*Zea mays* UBIQUITIN 1

I. Introduction

1. What is C_4 photosynthesis?

1.1 A common theme

C_4 photosynthesis is a series of anatomic and metabolic adaptions that evolved to evade high rates of photorespiration. These are caused by the central enzyme Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) under CO_2 delimiting conditions, like high radiation and temperature, low water availability and open habitats (Osborne and Sack, 2012; Sage *et al.*, 2012). It arose over 60 times independently in several plant genera within the last ~30 million years (Christin *et al.*, 2014; Sage *et al.*, 2012). The recurrent evolution of C_4 is reflected by a high diversity in its metabolic pathway (Wang *et al.*, 2014b) and anatomy (Edwards and Voznesenskaya, 2011), all converging in one common theme: the prefixation of CO_2 by phosphoenolpyruvate carboxylase (PEPC) into a C_4 compound in an outer compartment and its subsequent release in an inner, O_2-depleted compartment, which exclusively houses Rubisco. By this theme, CO_2 is increased manifold in the internal compartment, allowing Rubisco to operate more efficiently (von Caemmerer and Furbank, 1999). Due to the variability in manifestations, C_4 photosynthesis is often rather described as a syndrome than as a single trait and recognized as an outstanding example of convergent evolution (Sage, 2004). Since its discovery more than 50 years ago by Hatch and Slack (1966), over 8100 C_4 species in more than 400 genera have been identified (Sage, 2016), initially by typical characteristics like low CO_2 compensation point, low carbon isotope discrimination and its distinct anatomic features and subsequently by interpolation from phylogenetic relationship (Sage, 2016). Soon after its discovery further molecular analysis identified the central biochemical pathway, which is known as C_4-cycle (Hatch, 1987).

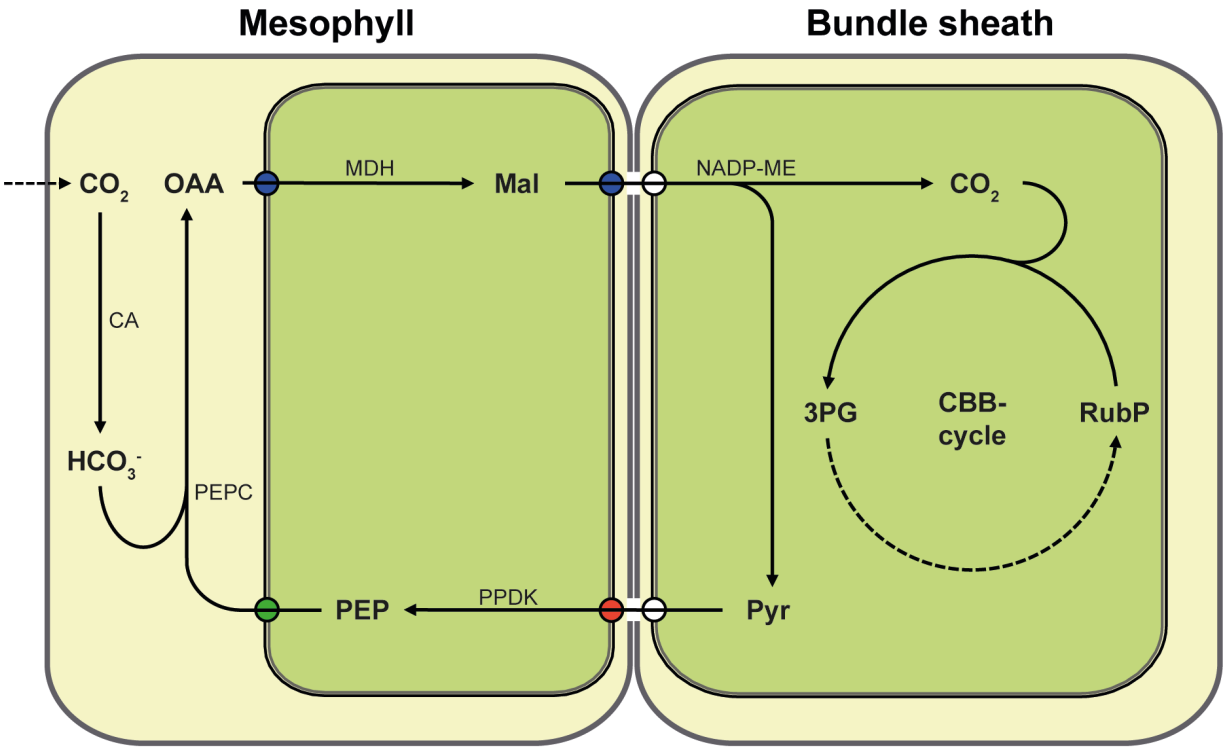


Figure 1 Core C4-cycle of the NADP-ME subtype. 3PG, 3-Phosphoglycerate; CA, Carbonic anhydrase; Mal, Malate; MDH, Malate dehydrogenase; OAA, Oxaloacetic acid; PEP, Phosphoenolpyruvate; PEPC, PEP carboxylase; PPDK, Pyruvate orthophosphate dikinase; Pyr, Pyruvate; RubP, Ribulose-1,5-bisphosphate. Coloured circles represent known transporters: Red, Bile acid/sodium symporter 2 (BASS2; Furumoto *et al.*, 2011); Blue, Dicarboxylate transporter 1 (DiT1; Renné *et al.*, 2003); Green, Phosphate/PEP translocator (PPT; Bräutigam *et al.*, 2008).

1.2 The C_4-cycle biochemistry

Approximately 3 % of known land plants (Sage, 2004) exhibit a supplementary metabolic pathway, which concentrates CO_2 around Rubisco – the C_4-cycle. Most plants that conduct C_4 photosynthesis exhibit a characteristic leaf anatomy, consisting of close vein spacing and large mestome sheath or bundle sheath cells (BSC), surrounded by often only one layer of mesophyll cells (MC), forming a wreath-like structure, the so-called Kranz anatomy. This specialised anatomy is highly linked to the C_4-cycle (Fig. 1), which relies on the spatial separation between CO_2 initial fixation and final assimilation. In contrast to C_3 photosynthesis, where the first product of CO_2 fixation is the C_3 compound 3-phosphoglycerate (3PG), generated by Rubisco, the C_4-cycle uses a MC specific PEPC to fix HCO_3^-, provided by a cytosolic carbonic anhydrase (CA), to generate the C_4 compound oxaloacetic acid (OAA) – hence, the designation as C_3 and C_4 photosynthesis. In the NADP-ME C_4 subtype, OAA is reduced to malate by malate dehydrogenase (MDH) and diffuses into the

BSC where it is decarboxylated by an NADP-dependent malic enzyme (NADP-ME). The generated pyruvate is transferred back to the MC where it is phosphorylated by pyruvate orthophosphate dikinase (PPDK) to regenerate the carboxy-acceptor. Two alternative canonical C_4 subtypes are distinguished, which use aspartate as transport metabolite and the eponymous decarboxylases NAD-dependent malic enzyme (NAD-ME) or phosphoenolpyruvate carboxykinase (PCK). However, for several C_4 species, enzyme activity and abundance indicate a rather composite cycle of either NADP-ME and PCK or NAD-ME and PCK (Furbank, 2011; Meister *et al.*, 1996; Muhaidat and McKown, 2013; Wang *et al.*, 2014b; Wingler *et al.*, 1999), which can exhibit developmental or age dependent plasticity in its composition (Pick *et al.*, 2011; Sommer *et al.*, 2012) and may even be temperature dependent, as observed in *Alloteropsis semialata* (Ueno and Sentoku, 2006). This might be particularly apparent in the grass subfamily of Chloridoideae. Most chloridoid species are classified as either NAD-ME or PCK subtype but do not point back to monophyletic origins, suggesting several transitions between those subtypes (Christin *et al.*, 2009a), which could be related to a redundant presence of both subtypes and a species specific shift of preference for one or the other.

2. Why did C_4 photosynthesis evolve?

2.1 Rubisco – The bottleneck

The initial metabolic step of carbon assimilation in most photosynthetic organisms is the conversion of CO_2 and ribulose-1,5-bisphosphate (RubP) to two molecules of 3PG by Rubisco in a process termed Calvin-Bassham-Benson cycle (CBB-cycle). With an average turnover rate of 5 conversions per second and composed of eight large and small subunits each – summing up to 560 kDa – Rubisco is one of the slowest and largest proteins known (Spreitzer and Salvucci, 2002; Tabita *et al.*, 2007). Owing to its size and slow kinetics, plants have to invest up to 30 % of total leaf nitrogen into Rubisco biogenesis, rendering it the probably most abundant protein in the world (Ellis, 1979; Evans, 1989; Raven, 2013). But Rubisco is not only slow and *oversized*, it is also promiscuous. The most fatal side-reaction of Rubisco is – due to its affinity for O_2 – the oxygenation of RubP, resulting in one molecule of

3PG and one molecule of 2-phosphoglycolate (2PG). The latter is a potent inhibitor of triosephosphate isomerase and phosphofructo-kinase (Anderson, 1971; Kelly and Latzko, 1977). When Rubisco evolved, about 3 billion years ago (Nisbet *et al.*, 2007), atmospheric CO_2 levels were likely to be 100 fold higher than today (Kasting and Howard, 2006), while O_2 levels were 10^{14} times lower (Catling and Zahnle, 2002). Up until 400 million years ago atmospheric $[CO_2]/[O_2]$ did not promote significant oxygenation activity (Sage, 2004), explaining why Rubisco was able to evolve in this way at all. At present atmospheric concentrations the oxygenation of RubP accounts for approximately 25 % of Rubisco activity (Sharkey, 1988), even increasing in hot and arid environments, due to i) the temperature dependent differential solubility of CO_2 and O_2, ii) the temperature dependent decrease in CO_2 affinity of Rubisco and iii) reduced gas exchange in plants under arid conditions, to minimize evaporation (Brooks and Farquhar, 1985; Jordan and Ogren, 1984; Schulze and Hall, 1982).

2.2 The photorespiratory cycle

Since 2PG can neither be fed into the CBB-cycle nor be accumulated (due to its toxicity), it has to be recycled in a complex process termed photorespiration – or more appropriately, C_2-cycle – regenerating one molecule of 3PG from two molecules of 2PG, consuming ATP and NADPH and releasing previously fixed CO_2. This results in an approximately 50 % extra energy demand at moderate temperatures (Peterhänsel *et al.*, 2010).

The C_2-cycle involves eight core enzymes and several (partially unknown) transporters (Hagemann and Bauwe, 2016) and takes place in at least three different organelles: chloroplast, mitochondrion and peroxisome (Fig. 2).

Starting with RubP oxygenation, 2PG is dephosphorylated to glycolate by phosphoglycolate phosphatase (PGLP), which is exported by a glycerate/glycolate antiporter (Pick *et al.*, 2013) and diffuses to the peroxisome, where it is oxidised by glycolate oxidase (GOX) to glyoxylate, generating hydrogen peroxide. Subsequently, glyoxylate is aminated to glycine by glutamate:glyoxylate aminotransferase (GGAT)

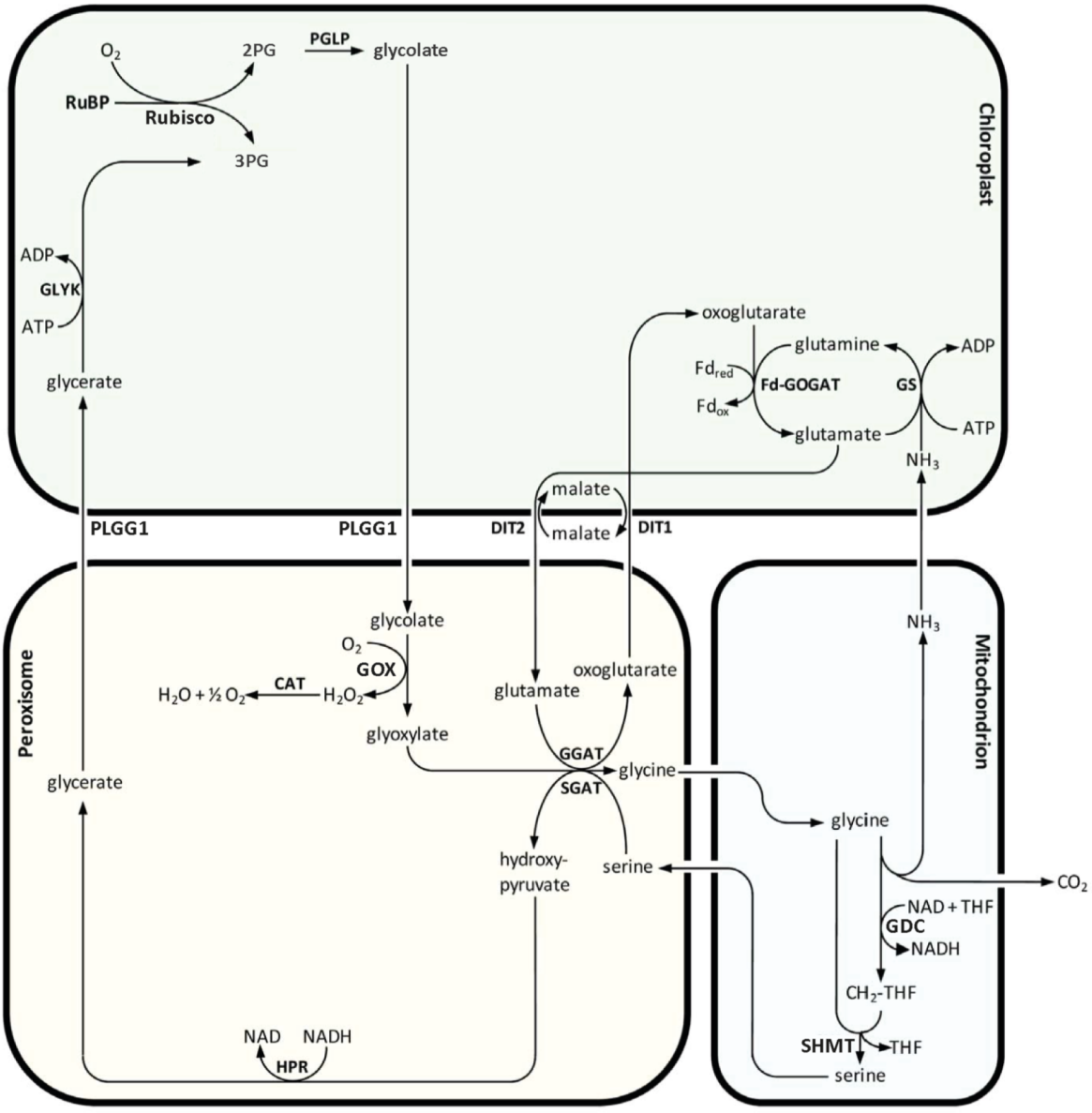


Figure 2 Central reactions of the photorespiratory cycle. 2PG, 2-Phosphoglycolate; 3PG, 3-Phosphoglycerate; CAT, Catalase; DiT1 and 2, Dicarboxylate transporter 1 and 2; Fd-GOGAT, Glutamine:oxoglutarate aminotransferase; GDC, Glycine decarboxylase; GGAT Glutamate:glyoxylate aminotransferase; GLYK Glycerate kinase; GOX Glycolate oxidase; GS, Glutamine synthetase; HPR Hydroxypyruvate reductase; PGLP Phosphoglycolate phosphatase; PLGG1, Plastidial glycerate/glycolate antiporter; Rubisco, RubP carboxylase/oxidase; RubP, Ribulose-1,5-bisphosphate; SGAT, Serine:glyoxylate aminotransferase; SHMT, Serine hydroxymethyltransferase; THF, Tetrahydrofolate. Adapted from Peterhänsel *et al.* (2010)

and serine:glyoxylate aminotransferase (SGAT). The required glutamate is imported from the chloroplast by dicarboxylate translocators (Renné *et al.*, 2003; Somerville and Somerville, 1985) in exchange for malate, while glycine is shuttled to the mitochondrion. Here, two molecules of glycine are converted to one molecule of serine by the catalytic interplay of glycine decarboxylase (GDC) and serine hydroxymethyltransferase (SHMT), releasing ammonia and the aforementioned CO_2. Serine is exported from the mitochondrion back into the peroxisome, where it is

deaminated to hydroxypyruvate by SGAT, donating the amino group for the transamination of glyoxylate. Subsequently, hydroxypyruvate is reduced by hydroxypyruvate reductase (HPR) to glycerate, which finally returns to the chloroplast where it is phosphorylated to 3PG by glycerate-3 kinase (GLYK) and re-enters the CBB-cycle. The ammonia released by GDC in the mitochondrion can be re-assimilated in the chloroplast by glutamine synthetase (GS) and ferredoxin-dependent glutamine:oxoglutarate aminotransferase (Fd-GOGAT), providing the glutamate for the GGAT reaction.

The photorespiratory C_2-cycle does not only recycle 2PG, but was also reported to protect from photoinhibition (Kozaki and Takeba, 1996; Takahashi *et al.*, 2007), provide C_1 units to several essential biosynthesis pathways as nucleotide or amino acid synthesis (Hanson and Roje, 2001) and glycine to the glutathione synthesis during stress (Noctor *et al.*, 1999).

3. How did C_4 photosynthesis evolve?

Mapping the approximately 8000 known C_4 species to the taxonomy of plants currently suggests that C_4 photosynthesis evolved more than 60 times independently (Sage, 2016), rendering it one of the most remarkable examples for convergent evolution of a complex trait and suggesting that evolution of C_4 depends only on a small number of key factors. In fact, all of the proteins known to be involved in C_4 photosynthesis are already present in C_3 plants, where most of them carry out housekeeping functions. Consequently, a detailed model of C_4 evolution has been elaborated (Bräutigam and Gowik, 2016; Heckmann *et al.*, 2013; Monson, 1999; Monson *et al.*, 1984; Rawsthorne, 1992; Sage, 2004), largely based on the analysis of C_3-C_4 intermediate species. Although those species do not necessarily pose evolutionary intermediates, they do exhibit discrete anatomical and physiological characteristics between C_3 and C_4 photosynthesis, thus allowing inference of common features that correlate with their degree of C_4-ness and lead to a more or less stepwise model (Fig. 3).

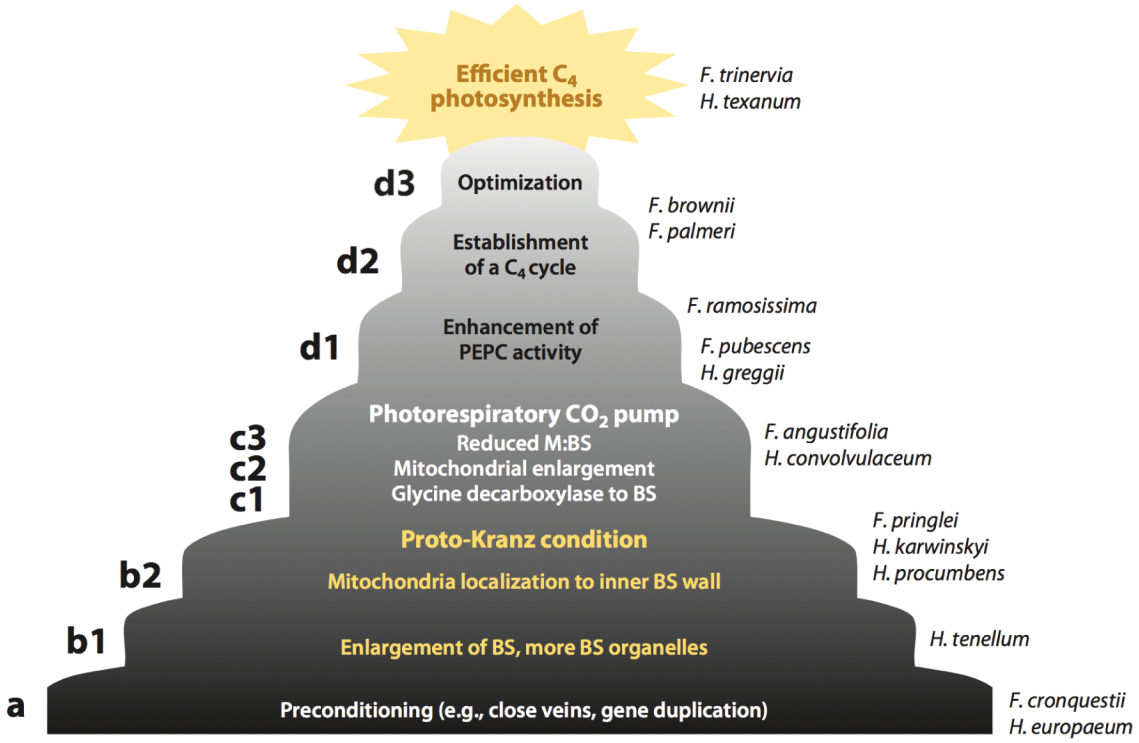


Figure 3 Conceptual linear model of C4 evolution. Numbers within each phase indicate important steps. *Flaveria* and *Heliotropium* species corresponding to each phase are shown on the right. Stages: (a) preconditioning, (b) evolution of proto-Kranz anatomy, (c) evolution of the photorespiratory CO2 pump, (d) establishment of a C4 cycle. BS, bundle sheath; M, mesophyll; PEPC, phosphoenolpyruvate carboxylase. Adapted from Sage *et al.* (2012)

3.1 Preconditioning and evolution of proto-Kranz anatomy

It is obvious that not all plants evolved C_4 photosynthesis, not even in hot and arid environments, although, under these conditions, C_4 photosynthesis unequivocally does pose a fitness gain. This suggests that certain traits need to be present before establishment of a C_4-cycle can be beneficial. Supporting this assumption is the observation that C_4 origins are not evenly distributed across the plant phylogeny (Sage *et al.*, 2011). For instance the grass PACMAD clade comprises 22-24 independent C_4 origins while its evenly-sized sister clade (BEP) does not contain any C_4 species (Grass Phylogeny Working Group II, 2012).

A yet quite undefined group of those preconditions might be associated to the composition of the genome. Since all genes involved in a C_4-cycle already fulfil – likely essential – functions in C_3 plants, they need to be duplicated in order to adapt to a new function (Monson, 2003). This suggests that whole or partial genome duplications might contribute to the establishment of a complex trait from existing

genes. Consistently, most of the genes associated to the evolution of C_4 arose from multi-gene families, e.g. PEPC, PCK and NADP-ME (Christin *et al.*, 2009a; Christin *et al.*, 2007; Christin *et al.*, 2009b; Westhoff and Gowik, 2004).

Active transposable elements (TEs) might also affect C_4-evolvability. Considering the necessary changes in anatomy, regulatory networks and metabolism and the large amount of differential expressed genes observed in closely related C_3 and C_4 species (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011), it is presumable that C_4 ancestors possessed a higher genomic plasticity. This might be owed to the content of TEs and their ability to alter gene expression (Feschotte, 2008; Rebollo *et al.*, 2012) and increase recombination (McClintock, 1984). Suggesting that highly active TEs, induced by environmental stress like water, temperature and radiation (Capy *et al.*, 2000; McClintock, 1984; Wessler, 1996), may have facilitated the establishment of C_4 photosynthesis.

Finally, a short generation time, frequent sexual reproduction and the population size are factors, which – in general – greatly influence the frequency and fixation of novel mutations and thus might also affect evolvability of C_4 photosynthesis (Monson, 2003).

Besides genomic plasticity and *life style* of a species, some anatomic preconditions have to be present before a C_4 cycle can be established. Once selection for metabolic steps towards C_4 sets off, these, in turn, also exert selective pressure on anatomical features, e.g. an increased demand for CBB-cycle activity, at some point, will also select for chloroplast size, since larger chloroplasts can harbour more CBB-cycle enzymes. But other anatomical features cannot, or cannot yet, be explained by selective pressure through metabolic steps of C_4 evolution, suggesting that they had to be present beforehand. As such, close vein spacing is assumed to be an early step towards C_4, since it increases the BSC:MC ratio and allows faster diffusion between those cells (Ehleringer *et al.*, 1997). Also, vein density is greatly affected by similar selective pressures as C_4 evolution, like high light, temperature and open and arid environments. This is because greater hydraulic capacity increases the rate of photosynthesis, when water is abundant and decreases the demand for closed stomata and the risk of hydraulic failure under low water availability (Osborne and Sack, 2012).

Further, an increase in BSC size is also assumed to precede C_4 evolution, which was supported by statistical analyses of PACMAD and BEP grasses. Although both clades exhibit comparably high vein density, only the C_3 species of the C_4-rich PACMAD clade show an increased proportion of bundle sheath tissue compared to the C_4-absent BEP clade (Christin *et al.*, 2013; Griffiths *et al.*, 2013). This suggests that an increase of the bundle sheath area promoted C_4 evolvability, but could not explain why it evolved. Griffiths *et al.* (2013) suggested that larger BSC could act in cavitation repair and maintaining hydraulic conductance. Another explanation could be increased water storage in the leaf as an adaption to saline or arid environments (Sage and Coleman, 2001).

With an increase in vein density also comes a decrease in photosynthetically active mesophyll, which in turn might select for photosynthetic capacity of the BSC, increasing chloroplast number and size and subsequently, due to photorespiration, the number of mitochondria (Bräutigam and Gowik, 2016; Sage *et al.*, 2012). These often concentrate at the centripetal cell-wall, potentially to increase scavenging of photorespiratory CO_2 (Muhaidat *et al.*, 2011).

Together, those traits form the so-called proto-Kranz anatomy found in several C_3 species closely related to C_4 origins (Marshall *et al.*, 2007; Muhaidat *et al.*, 2011; Sage *et al.*, 2013). However, the molecular mechanisms that direct vein density, bundle sheath size and its photosynthetic capacity remain largely unknown.

3.2 The photorespiratory CO_2 pump – evolutionary link between C_3 and C_4

Once proto-Kranz anatomy is established, investment into a first two-celled molecular carbon concentrating mechanism (CCM) becomes beneficial. The characteristic feature of all C_3-C_4 intermediate species – the photorespiratory CO_2 pump. This step is assumed to be initialised by the loss of GDC activity in MC. Hence, photorespiratory glycine has to move to the BSC to be decarboxylated, leading to a local increase of CO_2 and thus reducing Rubisco oxygenation activity (Monson *et al.*, 1984). As Rubisco now can work more efficiently in BSC than MC, further selective pressure for close vein distance, BSC size and photosynthetic capacity increases, as can be observed in C_3-C_4 intermediate species. It was shown that, in the C_4 model species *Flaveria*, the loss of GDC in MC did not occur abruptly but gradually (Schulze *et al.*, 2013), corresponding to an increase in Kranz anatomy features (Holaday *et al.*,

1984; Ku *et al.*, 1983; Monson and Moore, 1989; Sage *et al.*, 2013). Radioactive carbon labelling and metabolic modelling suggested that this CCM leads to a three-fold increase in BSC CO_2 levels (Keerberg *et al.*, 2014). Although highly linked to the evolution of C_4, this CO_2 pump is also found in plant lineages not directly related to C_4 origins (Sage *et al.*, 2012), suggesting a stable evolutionary trait on its own. Since it utilises the C_2 compound glycine to shuttle CO_2 and to emphasise its independency to C_4 photosynthesis, this CCM is also termed C_2 photosynthesis (Sage *et al.*, 2012; Vogan *et al.*, 2007).

As indicated above, C_4 photosynthesis is not one distinctive metabolic cycle but a series of anatomical and metabolic adaptions that lead to a common phenotype – the spatial separation of initial carbon fixation by PEPC and final assimilation by Rubisco, while the intermediate evolutionary and metabolic steps (Furbank, 2011; Heckmann *et al.*, 2013; Williams *et al.*, 2013), as well as the anatomic implementation (Edwards and Voznesenskaya, 2011; Williams *et al.*, 2013), are versatile and flexible. Since several C_4 origins lack intermediate species, which would allow inference of evolutionary steps, the *canonical* model of C_4 evolution mainly is based on a few genera, rich in intermediate species, like *Flaveria* or *Heliotropium*. This might generate a biased and probably too linear view on the individual steps or their sequence.

Nonetheless, the intersecting step in the evolution of C_4 appears to be the establishment of C_2 photosynthesis, since a restriction of GDC to the bundle sheath can be observed even in C_4 families that lack closely related intermediate species, such as maize (Chang *et al.*, 2012), *Sorghum* (Döring *et al.*, 2016) *Setaria* (John *et al.*, 2014) or *Chloris* (Ohnishi and Kanai, 1983). Emphasising the central role of the glycine decarboxylase.

The glycine decarboxylase is key component to the photorespiratory cycle and connective link between C_3 and C_4 photosynthesis. It consists of four proteins (GLDH, GLDL, GLDP and GLDT) that do not form a heteromeric complex, but are loosely assembled together in the mitochondrial matrix (why it is also termed *glycine cleavage system*), where it can account for more than 30 % of soluble protein (Oliver *et al.*, 1990). All subunits are nuclear encoded, usually by several gene copies, with the exception of GLDT (Bauwe, 2011), which often consists of a single gene (Phytozome v12.1.15).

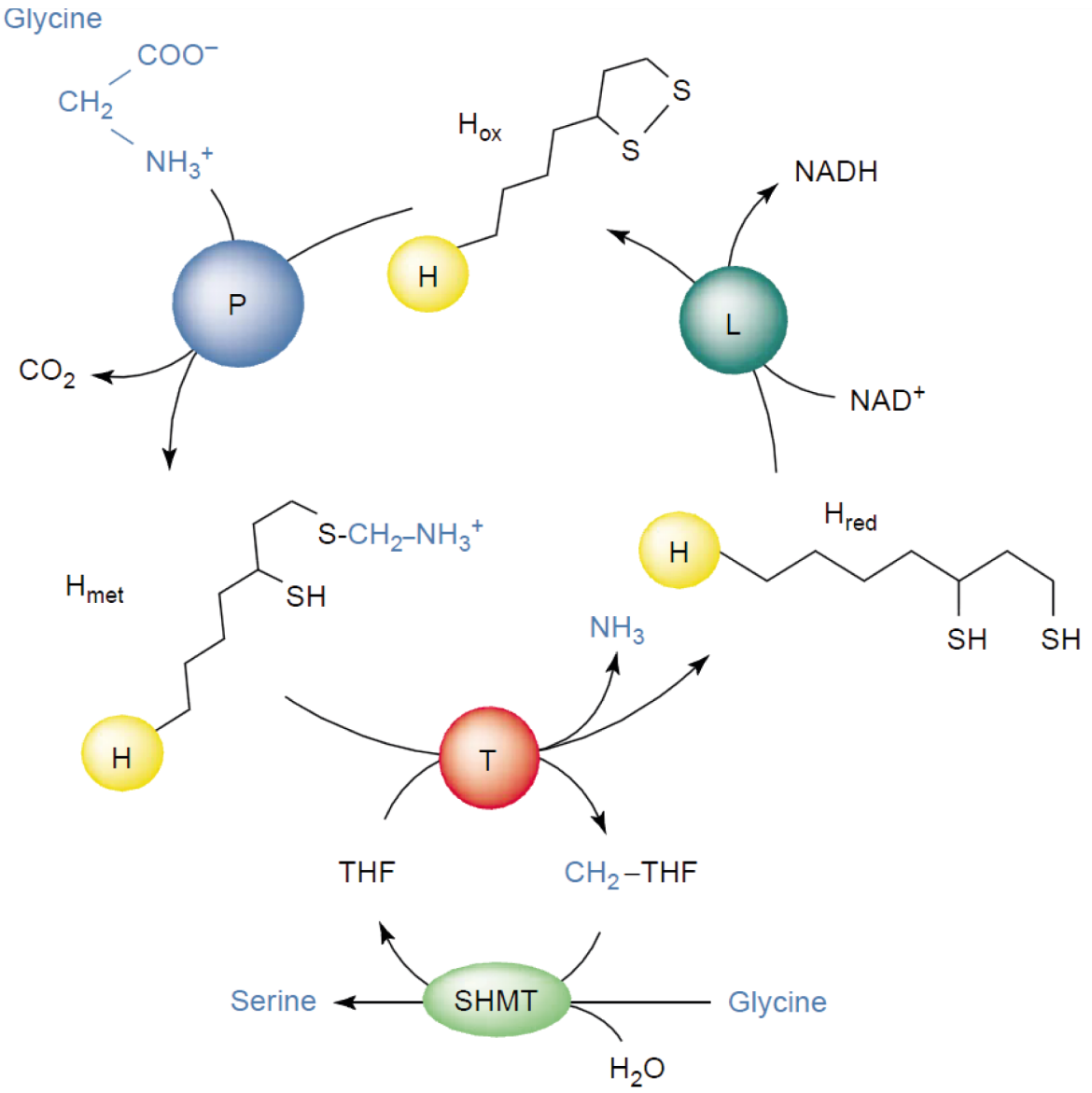


Figure 4 Schematic outline of glycine to serine conversion. Glycine decarboxylase is composed of the four subunits H, L, P and T that are loosely assembled in the mitochondrial matrix, where they, in cooperation with serine hydroxymethyltransferase (SHMT), catalyse the tetrahydrofolate (THF) dependent conversion of glycine to serine. Taken from Douce *et al.* (2001).

The decarboxylation of glycine in plant mitochondria occurs in tight interaction with SHMT, together converting two molecules of glycine to one molecule of serine, reducing NAD^+ and releasing CO_2 and NH_3 (Fig. 4). The initial step and actual decarboxylation of glycine is conducted by the homodimeric GLDP, which subsequently transfers the residual methylamine group onto a lipoamide arm of GLDH. GLDT then catalyses the cleavage of the methylamine group, releasing NH_3 and methylating the cofactor tetrahydrofolate (THF). While GLDH is regenerated by the homodimeric GLDL, the methylated THF serves as substrate for the methylation of another molecule of glycine by SHMT to form serine. In contrast to the other

subunits, GLDL is also part of other enzyme complexes, such as pyruvate dehydrogenase (Luethy *et al.*, 1996).

In C_3 plants all four subunits are highly expressed throughout the photosynthetic active tissue and often exhibit light dependent regulation (Srinivasan and Oliver, 1995; Vauclare *et al.*, 1998; Walker and Oliver, 1986). Knockout of GLDP or GLDT is lethal, even under elevated CO_2, clearly showing that GDC is not only essential to photorespiration (Engel *et al.*, 2007; Timm *et al.*, 2017) and emphasising the central role of GDC in providing C_1 units (Hanson and Roje, 2001). In turn, overexpression of GLDH or GLDL leads to an increase in biomass production and decrease in accumulation of transitory starch. Both not visible under elevated CO_2, suggesting an, at least, indirect function of GDC in balancing starch and sugar biosynthesis (Timm *et al.*, 2012; Timm *et al.*, 2015).

Since GLDP is the actual decarboxylase of GDC, its delocalisation to the BSC is most likely the initial step – or at least the most efficient one – in establishing C_2 photosynthesis. This can be observed in the C_2 species *Moricandia arvensis*, where GLDP is confined to the BSC, while the other subunits and SHMT are also highly expressed in the mesophyll *(*Morgan *et al.*, 1993; Rawsthorne *et al.,* 1988*)*. For the intermediate-rich genus of *Flaveria* it was shown that all species contain two GLDP genes, one ubiquitously expressed in photosynthetic tissue and one confined to the BSC. A shift of GDC activity to the BSC was facilitated by a gradual loss of the ubiquitous isoform from C_3 to C_4 species (Schulze *et al.*, 2013). Consequently, although not comprehensively analysed yet, at least in *Panicum* and *Flaveria* intermediate species, the other GDC subunits appear to be confined to the BSC as well (Morgan *et al*., 1993; this study: Chapter 1), supporting that excess expression of the other subunits in MC is futile after delocalisation of GLDP. In this context, it might be interesting to analyse why – within ~6 million years (Apel *et al.*, 1997; Perfectti *et al.*, 2017) – the other subunits never were reallocated in *Moricandia arvensis.* Potentially giving insight on the constraints of C_4 evolution.

3.3 The slippery slope towards C_4

Comparative RNA-seq experiments of closely related C_3 and C_4 species indicate, that several hundred to thousand genes exhibit differential expression patterns (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011). Modelling approaches suggest that,

once preconditions are met, the trajectory towards establishment of a C_4-cycle is very smooth and flexible (Heckmann *et al.*, 2013; Williams *et al.*, 2013). All in all, indicating that the amount of changes necessary to generate C_4 might not be as low as previously thought, but rather that, after acquiring certain key factors, the onset to evolve C_4 becomes inevitable, as long as selective pressure is maintained – depicting a slippery slope towards C_4 (Bräutigam and Gowik, 2016).

Observable in intermediate species and supported by modelling, C_2 photosynthesis is followed by an increase in C_4 enzyme activity (Edwards and Ku, 1987; Heckmann *et al.*, 2013; Williams *et al.*, 2013). This might likely be due to the concomitant release of ammonia in the BSC by GDC, generating a drastic nitrogen imbalance under photorespiratory conditions (Rawsthorne *et al.*, 1988). Flux balance and RNA-seq analysis indicate that low C_4 cycle activity was likely established to solve this ammonia imbalance, which in turn transports CO_2 to the BSC (Mallmann *et al.*, 2014). From this point any optimisation of the involved enzymes, in terms of increased expression, shifted expression or kinetic adaption, as well as further concomitant adaption of anatomy, might directly translate into a gain of carbon assimilation.

Finally, a loss of MC Rubisco followed by a decrease of the – now largely unnecessary – C_2-cycle completes the C_4 evolution. Although this last step likely fixes the trait and renders C_4 to C_3 reversions unlikely (Bräutigam and Gowik, 2016), one has to consider the flexibility observed in some C_4 species, like *Salsola* (Lauterbach *et al.*, 2017; Li *et al.*, 2015; Pyankov *et al.*, 2001) or *Eleocharis* (Ueno *et al.*, 1988), that can conduct C_4 or C_3 photosynthesis, depending on the type of leaf. Suggesting, that as long as a functional C_3 cycle is available, it might also be re-establishable (Kadereit *et al.*, 2012).

4. Introducing C_4 photosynthesis into C_3 crops – Why and how?

4.1 Benefits of C_4 photosynthesis

Although only 2 % of the 390 000 known plants species conduct C_4 photosynthesis, they account for ~23 % of the world's primary biomass production (Sage, 2016; Still *et al.*, 2003; Willis, 2016). Despite its high efficiency, only ten out of 150 listed crops

worldwide conduct C_4 photosynthesis. While the three most important C_4 species – maize, sugar cane and *Sorghum* – are mainly grown for livestock feed, sugar and bioethanol production, most of the plants that contribute to food production – like rice, wheat and soybean – conduct C_3 photosynthesis (Sage, 2016). With an estimated population of 10 billion people by 2050 and stagnating productivity improvement for the major food crops (Cassman, 1999; Sheehy, 2001; Zhu *et al.*, 2010), the world is facing a looming food crisis, further exacerbated by the increasing competition to biofuel production (Cassman and Liska, 2007) and limited capacities to increase agricultural space. Thus, a second Green Revolution (Surridge, 2002) is required to improve photosynthetic efficiency (Long *et al.*, 2006; Zhu *et al.*, 2010).

C_4 photosynthesis can increase intercellular CO_2 levels by an order of magnitude (von Caemmerer and Furbank, 1999), nearly saturating Rubisco. Comparison of rice and maize, as well as estimation approaches, suggest that a C_4-cycle increases the radiation use efficiency by ~50 % (Kiniry *et al.*, 1989; Sheehy, 2000; Zhu *et al.*, 2008). This also results in a better water use efficiency, compared to C_3 plants, due to the decreased demand for stomata opening to exchange gasses. Consequently, C_4 plants exhibit very low expression of C_2-cycle genes (Bräutigam *et al.*, 2011; Bräutigam *et al.*, 2014; Gowik *et al.*, 2011; Mallmann *et al.*, 2014) and over 60% less Rubisco (Ghannoum *et al.*, 2011), resulting in a higher nitrogen use efficiency.

For those very reasons, the introduction of C_4 photosynthesis into C_3 crops seems as desirable as it is ambitious to solve the projected future demand on food supply (Covshoff and Hibberd, 2012; Hibberd *et al.*, 2008). A basic concept to achieve this is to study how nature did. C_4 photosynthesis evolved several times independently, suggesting a common theme that can be traced and ultimately imitated (Westhoff and Gowik, 2010).

4.2 Engineering C_4

Although other, and by far simpler, approaches to optimise photosynthetic efficiency have been proposed (reviewed in Evans, 2013; Ort *et al.*, 2014; Zhu *et al.*, 2010), engineering C_4 photosynthesis is probably the most profitable and tempting. On the one hand, C_4 photosynthesis has dramatically improved radiation, water and nitrogen use efficiencies and on the other, its incredibly convergent evolution promises easy acquisition. However, the master switch is yet to be found, but accumulating

evidence suggests that C_4 photosynthesis rather evolved so many times because its trajectory is smooth and flexible (Bräutigam and Gowik, 2016; Heckmann *et al.*, 2013; Williams *et al.*, 2013), i.e. there are no fitness valleys between evolutionary steps, while its sequence is widely arbitrary and likely attainable by several ways, as long as genomic plasticity and selective pressure are maintained. Nonetheless, efforts have not ceased and suggestions on how to proceed are manifold (Covshoff and Hibberd, 2012; Fouracre *et al.*, 2014; Hibberd *et al.*, 2008; Kajala *et al.*, 2011; Leegood, 2013; Mitchell and Sheehy, 2006; Sage and Zhu, 2011; Schuler *et al.*, 2016), but convergent in the opinion that knowledge about two central aspects is still lacking – the regulation of Kranz anatomy and spatio-temporal gene expression.

4.2.1 Regulation of leaf anatomy

Due to the tight connection of the C_4-cycle and Kranz anatomy, changing the morphology is paramount to engineering a two-celled C_4-cycle. Characteristics of Kranz anatomy are i) increased venation, ii) increased BSC:MC area, iii) increased number and size of chloroplasts and mitochondria in BSC, iv) high number of plasmodesmatal connections between MC and BSC and v) dimorphic chloroplasts, dependent on the cell- and C_4 subtype (Edwards and Voznesenskaya, 2011).
Comparative analysis of leaf development from closely related C_3 and C_4 species indicate that high vein density is conveyed by accelerated higher order vein formation, accompanied by delayed cell differentiation (Külahoglu *et al.*, 2014; McKown and Dengler, 2009). Correspondingly, genes that were already known to be related to auxin signalling, transport and biosynthesis, as well as cell cycle regulation, are differentially expressed across leaf gradients of closely related C_3 and C_4 species (Huang *et al.*, 2017; Külahoglu *et al.*, 2014; Kümpers *et al.*, 2017). Polar auxin efflux determines the site of procambial cell formation, from which new vascular tissue arises and excess auxin or reduced transport of it lead to surplus vein formation (reviewed in Scarpella and Helariutta, 2010). However, the developmental program initiated by auxin, and particularly the genes that control the development of BSC, are not well understood. In Arabidopsis the formation of procambial cells is preceded by expression of the auxin efflux transporter PIN-FORMED 1 (PIN1) and the auxin response factor MONOPTEROS (MP). The latter directly regulates expression of the homeobox gene ATHB8, a regulator of procambium development (Donner *et al.*,

2009). ATHB8 activity is accompanied by expression of DOF genes (DNA-BINDING WITH ONE ZINC-FINGER) and the GRAS family transcription factor (TF) SHORTROOT (SHR) (Gardiner *et al.*, 2011; Gardiner *et al.*, 2010). Although, whether or how these factors interact is not known, increasing evidence indicates that the SHR pathway is involved in the subsequent specification of BSC and MC. Maize mutants of SHR and its adjacently expressed interaction partner SCARECROW (SCR) exhibit an impaired venation pattern and ectopic formation of BSC (Slewinski *et al.*, 2014; Slewinski *et al.*, 2012). Perturbed BSC development was also observed in Arabidopsis mutants of SCR and SHR homologues (Cui *et al.*, 2014). Accordingly, differential expression of these genes was observed in several systems biology approaches (Aubry *et al.*, 2014a; Külahoglu *et al.*, 2014; Li *et al.*, 2010; Wang *et al.*, 2013). The SHR pathway is key to the specification of the endodermal and cortical layers in roots (reviewed in Benfey, 2016) and similarities in leaves, stems and roots in mutants of the SHR pathway suggest that the development of root endodermis, stem starch sheath and leaf bundle sheath is governed by the same regulatory mechanism (Slewinski *et al.*, 2012), with organ specific adaptions (Fouracre *et al.*, 2014).

Auxin signalling evidently controls vein formation and subsequent activity of the SHR pathway seems to regulate cell differentiation and thus impact vein density, either by stalling cell differentiation, which allows formation of higher order veins from undifferentiated ground tissue (Huang *et al.*, 2017; Külahoglu *et al.*, 2014) or by promoting early MC differentiation and thus ceasing further proliferation (McKown and Dengler, 2010). While both positively affect BSC:MC ratio, increased BSC size is also a commonly observed characteristic of Kranz anatomy and Külahoglu *et al.* (2014) attributed this to endoreduplication of BSC chromosomes, which was accompanied by prolonged expression of cell-cycle marker genes during C_4 leaf development. Ploidy and cell size are well known to be correlated (Sugimoto-Shirasu and Roberts, 2003). Although the mechanism is not fully understood, it seems to be tightly regulated by the homeobox TF MERISTEM LAYER1 (ATML1; Meyer *et al.*, 2017). Thus, it would be intriguing to analyse the expression of ATML1 orthologues across a C_4 leaf developmental gradient, particularly because the dosage effect might already generate a general bias in gene expression and thus could impact the early developmental program of organelle and plasmodesmata development and subsequently, the photosynthetic capacity of BSC. This could explain the observed

correlation between ploidy and chloroplast number (Butterfass, 1988; Ho and Rayburn, 1991). The dosage effect could also impact GOLDEN2-like genes (GLK), homologues of the GARP transcription factor GOLDEN2 (G2) that regulates chloroplast biogenesis (Hall *et al.*, 1998; Rossini *et al.*, 2001). Overexpression of a rice GLK gene led to increased chloroplast development in the vascular bundles of rice seedlings (Nakamura *et al.*, 2009) and ectopic expression of maize G2 and its paralogue GLK1 led to a sustained phenotype, accompanied by increased mitochondria and plasmodesmata number (Wang *et al.*, 2017). In maize both genes are expressed differentially between BSC and MC (Rossini *et al.*, 2001), but act redundantly (Wang *et al.*, 2017). This supports that a dosage effect of G2 could impact BSC organelle development, but also suggests that some C_4 species, like maize might have acquired Kranz anatomy rather by differential regulation of individual paralogues. Aubry *et al.* (2014b) found massive endoreduplication in whole-leave nuclei extracts of the C_3 species *A. thaliana*, including nuclei from BSCs. This led them to assume that the endoreduplication observed in BSC of the closely related C_4 species *G. gynandropsis* (Aubry *et al.*, 2014b; Külahoglu *et al.*, 2014) could not be related to BSC development. However, Aubry *et al.* (2014b) could not unequivocally disprove endoreduplication, since neither the effect of differential endoreduplication in MC and BSC was considered nor the impact of vascular nuclei on quantitative analysis, which were co-isolated by using FtGLDPA promoter. This promoter is active in BSC and vascular tissue (Engelmann *et al.*, 2008).

Finally, dimorphic chloroplasts – that is, apart from size, the presence or absence of thylakoid stacking, photosystem II and starch accumulation – might not be a necessity of C_4, but rather a consequence of differentially expressed photosynthetic genes. Targeted knockdown of GLDH in rice MC leads to a decrease in MC chloroplast area (Lin *et al.*, 2016) and knockdown of Rubisco small subunit directly affects starch accumulation (Fichtner *et al.*, 1993). More strikingly, overexpression of the C_4 decarboxylase NADP-ME in rice causes chloroplasts to be depleted in photosystem II and thylakoid stacking (Takeuchi *et al.*, 2000). This also displays the tight interplay of photosynthesis and anatomy and emphasises the importance of proper control over gene expression.

4.2.2 Regulation of gene expression

A two-celled C_4 cycle heavily relies on differential expression of photosynthetic genes in BSC and MC. Most of the key components are identified and cell specific promoters for BSC and MC expression in rice are available (Matsuoka *et al.*, 1994; Nomura *et al.*, 2005). This suggests that the introduction of a rudimentary C_4-cycle into the C_3 crop rice might be feasible soon.

However, the large number of genes necessary to solely express a C_4-cycle already exceeds current transformation capacities. While stacking of three to four small genes was already successfully conducted (Halpin, 2005), transformation of the expected 14 metabolic core genes will only be possible by repeated transformation. Since each stack of genes has to include a selection marker, introduction of C_4 needs at least five transformation events. Subsequently, those stacks will segregate independently, requiring a roughly 32-fold increased breeding population to maintain the trait (Claire, 2005) and genes for anatomic adaptions, correct enzyme functionality and auxiliary fluxes are not even included. Furthermore, the recurrent use of the same two, comparatively large, promoters will not only consume transformation capacity but also inevitably lead to homology-based gene silencing (Meyer and Saedler, 1996). Thus, generation of short promoters or ultimately isolation of the necessary *cis*-regulatory elements (CREs) will be vital to C_4 engineering. Particularly since, on a long term, several genes will have to be responsive to multiple inductive or suppressive signals for proper regulation of Kranz anatomy and C_4 cycle. Consequently, several system wide approaches tried to identify CREs or whole networks that control C_4 photosynthesis (Burgess *et al.*, 2017; Cao *et al.*, 2016; John *et al.*, 2014; Wang *et al.*, 2014a; Xu *et al.*, 2016), but still lack experimental validation. However, system wide prediction of CREs is difficult, since affinity of TFs is not only determined by motif sequence. In fact, some TFs exhibit very low sequence specificity *in vitro*, but are dependent on dimerisation with other TFs (Escalante *et al.*, 2002; Isakova *et al.*, 2016; Ptashne *et al.*, 1980) or collaborative DNA binding (Mirny, 2010). CREs are also influenced by order, orientation and spacing (Farley *et al.*, 2016) or DNA shape, determined by the motif's flanking sequences (Mathelier *et al.*, 2016).

Concurrently to system wide prediction, several individual regulatory mechanisms driving C_4 gene expression have been detected by conventional approaches (reviewed in Hibberd and Covshoff, 2010; Reeves *et al.*, 2017), but only a few were characterised in detail. As such, the mesophyll expression module 1 (MEM1) of

Flaveria and the bundle sheath motif 1 (BSM1) of *Gynandropsis* are probably the best described CREs.

MEM1 is composed of two submodules, which together drive MC expression of PEPC in C_4 *Flaveria* species, but exhibit sequence alterations in their C_3 cognates (Gowik *et al.*, 2004). Identified by dissection of the *PEPC* upstream flanking sequence from *F. trinervia*, MEM1 was first narrowed down to a 41 bp fragment, approximately 2 kb upstream of the translational start, but is also present as two submodules in other *Flaveria* species, seperated by ~100 bp (Gowik *et al.*, 2004). Akyildiz *et al.* (2007) later showed that the intervening distance had no impact, but that both submodules are necessary and dependent on two small polymorphisms that distinguish MC specific expression in C_4 and ubiquituous expression in C_3 *Flaveria* species. A MEM1-like sequence was also found in the upstream flanking sequence of a CA gene from *F. bidentis* and shown to be necessary for MC specificity (Gowik *et al.*, 2017).

BSM1 was discovered in the coding sequence of both NAD-ME genes from *G. gynandra* (Brown *et al.*, 2011). The authors found that fusion of a 240 bp fragment of the coding sequence to a reporter gene was sufficient to direct expression to the BSC, although it was under control of a ubiquituously active *Cauliflower Mosaic Virus* 35S promoter. Moreover, they showed that a similar fragments from maize and rice NADP-ME genes had the same effect. Reyna-Llorens *et al.* (2018) later dissected this region and found that two short sequence motifs are necessary for preferential expression in the BSC of *G. gynandra*. Intriguingly, although the BSM1 sequence was also conserved in NAD-ME genes of the closely related C_3 species Arabidopsis, the expression pattern was not, suggesting that spatial expression of the corresponding TF was altered during the transition from C_3 to C_4.

However, although both motifs have been shown to be functional in other genes (Gowik *et al.*, 2017; Reyna-Llorens *et al.*, 2018), they lack cell specific expression, when transferred to the C3 model species Arabidopsis (Akyildiz *et al.*, 2007; Brown *et al.*, 2011). This suggests evolutionary adaption to the controlling TFs, which remain to be identified and exemplifies the current lack of knowledge about these processes, which are fundamental to C4 engineering.

5. References

Akyildiz M, Gowik U, Engelmann S, Koczor M, Streubel M, Westhoff P. 2007. Evolution and function of a cis-regulatory module for mesophyll-specific gene expression in the C4 dicot Flaveria trinervia. *Plant Cell* **19**, 3391-3402.
Anderson LE. 1971. Chloroplast and cytoplasmic enzymes II. Pea leaf triose phosphate isomerases. *Biochimica et Biophysica Acta (BBA) - Enzymology* **235**, 237-244.
Apel P, Horstmann C, Pfeffer M. 1997. The Moricandia syndrome in species of the Brassicaceae - evolutionary aspects. *Photosynthetica* **33**, 205-215.
Aubry S, Kelly S, Kumpers BM, Smith-Unna RD, Hibberd JM. 2014a. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis. *PLoS Genet* **10**, e1004365.
Aubry S, Kneřová J, Hibberd JM. 2014b. Endoreduplication is not involved in bundle-sheath formation in the C4 species Cleome gynandra. *Journal of Experimental Botany* **65**, 3557-3566.
Bauwe H. 2011. Chapter 6 Photorespiration: The Bridge to C_4 Photosynthesis. In: Raghavendra AS, Sage RF, eds. *C4 Photosynthesis and Related CO2 Concentrating Mechanisms*, Vol. 32: Springer Netherlands, 81-108.
Benfey PN. 2016. Chapter Three - Defining the Path from Stem Cells to Differentiated Tissue. In: Wassarman PM, ed. *Current Topics in Developmental Biology*, Vol. 116: Academic Press, 35-43.
Bräutigam A, Gowik U. 2016. Photorespiration connects C3 and C4 photosynthesis. *Journal of Experimental Botany*.
Bräutigam A, Hoffmann-Benning S, Weber APM. 2008. Comparative Proteomics of Chloroplast Envelopes from C3 and C4 Plants Reveals Specific Adaptations of the Plastid Envelope to C4 Photosynthesis and Candidate Proteins Required for Maintaining C4 Metabolite Fluxes. *Plant Physiology* **148**, 568-579.
Bräutigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Mass J, Lercher MJ, Westhoff P, Hibberd JM, Weber AP. 2011. An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiol* **155**, 142-156.
Bräutigam A, Schliesky S, Külahoglu C, Osborne CP, Weber APM. 2014. Towards an integrative model of C(4) photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C(4) species. *Journal of Experimental Botany* **65**, 3579-3593.
Brooks A, Farquhar GD. 1985. Effect of temperature on the CO2/O2 specificity of ribulose-1,5-bisphosphate carboxylase/oxygenase and the rate of respiration in the light. *Planta* **165**, 397-406.
Brown NJ, Newell CA, Stanley S, Chen JE, Perrin AJ, Kajala K, Hibberd JM. 2011. Independent and Parallel Recruitment of Preexisting Mechanisms Underlying C4 Photosynthesis. *Science* **331**, 1436-1439.
Burgess SJ, Reyna-Llorens I, Jaeger K, Hibberd JM. 2017. A transcription factor binding atlas for photosynthesis in cereals identifies a key role for coding sequence in the regulation of gene expression. *bioRxiv* **165787**, https://doi.org/10.1101/165787.
Butterfass T. 1988. Nuclear control of plastid division. In: Boffey S, Lloyd D, eds. *Division and segregation of organelles*, Vol. 35. Cambridge: Cambridge University Press, 21-38.
Cao C, Xu J, Zheng G, Zhu X-G. 2016. Evidence for the role of transposons in the recruitment of cis-regulatory motifs during the evolution of C4 photosynthesis. *BMC Genomics* **17**, 1-11.

Capy P, Gasperi G, Biémont C, Bazin C. 2000. Stress and transposable elements: co-evolution or useful parasites? *Heredity* **85**, 101.
Cassman KG. 1999. Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. *Proceedings of the National Academy of Sciences* **96**, 5952-5959.
Cassman KG, Liska AJ. 2007. Food and fuel for all: realistic or foolish? *Biofuels, Bioproducts and Biorefining* **1**, 18-23.
Catling DC, Zahnle K. 2002. Evolution of Atmospheric Oxygen. *Encyclopedia of Atmospheric Sciences*: Academic Press Inc., 754-761.
Chang Y-M, Liu W-Y, Shih AC-C, Shen M-N, Lu C-H, Lu M-YJ, Yang H-W, Wang T-Y, Chen SC-C, Chen SM, Li W-H, Ku MSB. 2012. Characterizing Regulatory and Functional Differentiation between Maize Mesophyll and Bundle Sheath Cells by Transcriptomic Analysis. *Plant Physiology* **160**, 165-177.
Christin P-A, Petitpierre B, Salamin N, Büchi L, Besnard G. 2009a. Evolution of C4 Phosphoenolpyruvate Carboxykinase in Grasses, from Genotype to Phenotype. *Mol Biol Evol* **26**, 357-365.
Christin P-A, Spriggs E, Osborne CP, Strömberg CAE, Salamin N, Edwards EJ. 2014. Molecular Dating, Evolutionary Rates, and the Age of the Grasses. *Syst Biol* **63**, 153-165.
Christin PA, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ. 2013. Anatomical enablers and the evolution of C4 photosynthesis in grasses. *Proc Natl Acad Sci U S A* **110**, 1381-1386.
Christin PA, Salamin N, Savolainen V, Duvall MR, Besnard G. 2007. C4 Photosynthesis evolved in grasses via parallel adaptive genetic changes. *Curr Biol* **17**, 1241-1247.
Christin PA, Samaritani E, Petitpierre B, Salamin N, Besnard G. 2009b. Evolutionary insights on C4 photosynthetic subtypes in grasses from genomics and phylogenetics. *Genome Biol Evol* **1**, 221-230.
Claire H. 2005. Gene stacking in transgenic plants – the challenge for 21st century plant biotechnology. *Plant Biotechnol J* **3**, 141-155.
Covshoff S, Hibberd JM. 2012. Integrating C4 photosynthesis into C3 crops to increase yield potential. *Curr Opin Biotechnol* **23**.
Cui H, Kong D, Liu X, Hao Y. 2014. SCARECROW, SCR‐LIKE 23 and SHORT‐ROOT control bundle sheath cell fate and function in Arabidopsis thaliana. *The Plant Journal* **78**, 319-327.
Donner TJ, Sherr I, Scarpella E. 2009. Regulation of preprocambial cell state acquisition by auxin signaling in Arabidopsis leaves. *Development* **136**, 3235-3246.
Döring F, Streubel M, Bräutigam A, Gowik U. 2016. Most photorespiratory genes are preferentially expressed in the bundle sheath cells of the C4 grass Sorghum bicolor. *Journal of Experimental Botany* **67**, 3053-3064.
Douce R, Bourguignon J, Neuburger M, Rebeille F. 2001. The glycine decarboxylase system: a fascinating complex. *Trends Plant Sci* **6**, 167-176.
Edwards GE, Ku MSB. 1987. Biochemistry of C3–C4 Intermediates. *Photosynthesis*: Academic Press, 275-325.
Edwards GE, Voznesenskaya EV. 2011. Chapter 4 C4 Photosynthesis: Kranz Forms and Single-Cell C4 in Terrestrial Plants. In: Raghavendra AS, Sage RF, eds. *C4 Photosynthesis and Related CO2 Concentrating Mechanisms*. Dordrecht: Springer Netherlands, 29-61.
Ehleringer JR, Cerling TE, Helliker BR. 1997. C4 photosynthesis, atmospheric CO2, and climate. *Oecologia* **112**, 285-299.
Ellis RJ. 1979. The most abundant protein in the world. *Trends in Biochemical Sciences* **4**, 241-244.

Engel N, van den Daele K, Kolukisaoglu Ü, Morgenthal K, Weckwerth W, Pärnik T, Keerberg O, Bauwe H. 2007. Deletion of Glycine Decarboxylase in Arabidopsis Is Lethal under Nonphotorespiratory Conditions. *Plant Physiology* **144**, 1328-1335.
Engelmann S, Wiludda C, Burscheidt J, Gowik U, Schlue U, Koczor M, Streubel M, Cossu R, Bauwe H, Westhoff P. 2008. The gene for the P-subunit of glycine decarboxylase from the C4 species Flaveria trinervia: analysis of transcriptional control in transgenic Flaveria bidentis (C4) and Arabidopsis (C3). *Plant Physiol* **146**, 1773-1785.
Escalante CR, Brass AL, Pongubala JMR, Shatova E, Shen L, Singh H, Aggarwal AK. 2002. Crystal Structure of PU.1/IRF-4/DNA Ternary Complex. *Molecular Cell* **10**, 1097-1105.
Evans JR. 1989. Photosynthesis and nitrogen relationships in leaves of C3 plants. *Oecologia* **78**, 9-19.
Evans JR. 2013. Improving Photosynthesis. *Plant Physiology* **162**, 1780-1793.
Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS. 2016. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc Natl Acad Sci U S A* **113**, 6508-6513.
Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397-405.
Fichtner K, Quick WP, Schulze E-D, Mooney HA, Rodermel SR, Bogorad L, Stitt M. 1993. Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic tobacco transformed with "antisense" rbcS. *Planta* **190**, 1-9.
Fouracre JP, Ando S, Langdale JA. 2014. Cracking the Kranz enigma with systems biology. *J Exp Bot* **65**, 3327-3339.
Furbank RT. 2011. Evolution of the C4 photosynthetic mechanism: are there really three C4 acid decarboxylation types? *Journal of Experimental Botany* **62**, 3103-3108.
Furumoto T, Yamaguchi T, Ohshima-Ichie Y, Nakamura M, Tsuchida-Iwata Y, Shimamura M, Ohnishi J, Hata S, Gowik U, Westhoff P, Bräutigam A, Weber APM, Izui K. 2011. A plastidial sodium-dependent pyruvate transporter. *Nature* **476**, 472.
Gardiner J, Donner TJ, Scarpella E. 2011. Simultaneous activation of SHR and ATHB8 expression defines switch to preprocambial cell state in Arabidopsis leaf development. *Developmental Dynamics* **240**, 261-270.
Gardiner J, Sherr I, Scarpella E. 2010. Expression of DOF genes identifies early stages of vascular development in Arabidopsis leaves. *Int J Dev Biol* **54**, 1389-1396.
Ghannoum O, Evans JR, von Caemmerer S. 2011. Chapter 8 Nitrogen and Water Use Efficiency of C4 Plants. In: Raghavendra AS, Sage RF, eds. *C4 Photosynthesis and Related CO2 Concentrating Mechanisms*. Dordrecht: Springer Netherlands, 129-146.
Gowik U, Bräutigam A, Weber KL, Weber AP, Westhoff P. 2011. Evolution of c4 photosynthesis in the genus flaveria: how many and which genes does it take to make c4? *Plant Cell* **23**, 2087-2105.
Gowik U, Burscheidt J, Akyildiz M, Schlue U, Koczor M, Streubel M, Westhoff P. 2004. cis-Regulatory elements for mesophyll-specific gene expression in the C4 plant Flaveria trinervia, the promoter of the C4 phosphoenolpyruvate carboxylase gene. *Plant Cell* **16**, 1077-1090.
Gowik U, Schulze S, Saladié M, Rolland V, Tanz SK, Westhoff P, Ludwig M. 2017. A MEM1-like motif directs mesophyll cell-specific expression of the gene encoding the C4 carbonic anhydrase in Flaveria. *Journal of Experimental Botany* **68**, 311-320.
Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *New Phytol* **193**, 304-312.
Griffiths H, Weller G, Toy LF, Dennis RJ. 2013. You're so vein: bundle sheath physiology, phylogeny and evolution in C3 and C4 plants. *Plant Cell Environ* **36**, 249-261.

Hagemann M, Bauwe H. 2016. Photorespiration and the potential to improve photosynthesis. *Current Opinion in Chemical Biology* **35**, 109-116.
Hall LN, Rossini L, Cribb L, Langdale JA. 1998. GOLDEN 2: A Novel Transcriptional Regulator of Cellular Differentiation in the Maize Leaf. *Plant Cell* **10**, 925-936.
Halpin C. 2005. Gene stacking in transgenic plants--the challenge for 21st century plant biotechnology. *Plant Biotechnol J* **3**, 141-155.
Hanson AD, Roje S. 2001. One-Carbon Metabolism in Higher Plants. *Annu Rev Plant Physiol Plant Mol Biol* **52**, 119-137.
Hatch MD. 1987. C4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics* **895**, 81-106.
Hatch MD, Slack CR. 1966. Photosynthesis by sugar-cane leaves: A new carboxylation reaction and the pathway of sugar formation. *Biochemical Journal* **101**, 103-111.
Heckmann D, Schulze S, Denton A, Gowik U, Westhoff P, Weber AP, Lercher MJ. 2013. Predicting C4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell* **153**, 1579-1588.
Hibberd JM, Covshoff S. 2010. The regulation of gene expression required for C4 photosynthesis. *Annu Rev Plant Biol* **61**, 181-207.
Hibberd JM, Sheehy JE, Langdale JA. 2008. Using C4 photosynthesis to increase the yield of rice—rationale and feasibility. *Current Opinion in Plant Biology* **11**, 228-231.
Ho I, Rayburn AL. 1991. The relationship between chloroplast number and genome size in Zea mays ssp. mays. *Plant Science* **74**, 255-260.
Holaday AS, Lee KW, Chollet R. 1984. C3–C4 Intermediate species in the genus Flaveria: leaf anatomy, ultrastructure, and the effect of O2 on the CO2 compensation concentration. *Planta* **160**, 25-32.
Huang C-F, Yu C-P, Wu Y-H, Lu M-YJ, Tu S-L, Wu S-H, Shiu S-H, Ku MSB, Li W-H. 2017. Elevated auxin biosynthesis and transport underlie high vein density in C4 leaves. *Proceedings of the National Academy of Sciences* **114**, E6884-E6891.
Isakova A, Berset Y, Hatzimanikatis V, Deplancke B. 2016. Quantification of Cooperativity in Heterodimer-DNA Binding Improves the Accuracy of Binding Specificity Models. *Journal of Biological Chemistry* **291**, 10293-10306.
John CR, Smith-Unna RD, Woodfield H, Covshoff S, Hibberd JM. 2014. Evolutionary Convergence of Cell-Specific Gene Expression in Independent Lineages of C(4) Grasses. *Plant Physiology* **165**, 62-75.
Jordan DB, Ogren WL. 1984. The CO2/O2 specificity of ribulose 1,5-bisphosphate carboxylase/oxygenase. *Planta* **161**, 308-313.
Kadereit G, Ackerly D, Pirie MD. 2012. A broader model for C(4) photosynthesis evolution in plants inferred from the goosefoot family (Chenopodiaceae s.s.). *Proceedings of the Royal Society B: Biological Sciences* **279**, 3304-3311.
Kajala K, Covshoff S, Karki S, Woodfield H, Tolley BJ, Dionora MJA, Mogul RT, Mabilangan AE, Danila FR, Hibberd JM, Quick WP. 2011. Strategies for engineering a two-celled C4 photosynthetic pathway into rice. *Journal of Experimental Botany* **62**, 3001-3010.
Kasting JF, Howard MT. 2006. Atmospheric composition and climate on the early Earth. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 1733-1742.
Keerberg O, Parnik T, Ivanova H, Bassuner B, Bauwe H. 2014. C2 photosynthesis generates about 3-fold elevated leaf CO2 levels in the C3-C4 intermediate species Flaveria pubescens. *J Exp Bot* **65**, 3649-3656.
Kelly GJ, Latzko E. 1977. Chloroplast Phosphofructokinase: II. Partial Purification, Kinetic and Regulatory Properties. *Plant Physiol* **60**, 295-299.

Kiniry JR, Jones CA, O'Toole JC, Blanchet R, Cabelguenne M, Spanel DA. 1989. Radiation-use efficiency in biomass accumulation prior to grain-filling for five grain-crop species. *Field Crops Research* **20**, 51-64.
Kozaki A, Takeba G. 1996. Photorespiration protects C3 plants from photooxidation. *Nature* **384**, 557.
Ku MSB, Monson RK, Littlejohn RO, Nakamoto H, Fisher DB, Edwards GE. 1983. Photosynthetic Characteristics of C(3)-C(4) Intermediate Flaveria Species : I. Leaf Anatomy, Photosynthetic Responses to O(2) and CO(2), and Activities of Key Enzymes in the C(3) and C(4) Pathways. *Plant Physiology* **71**, 944-948.
Külahoglu C, Denton AK, Sommer M, Maß J, Schliesky S, Wrobel TJ, Berckmans B, Gongora-Castillo E, Buell CR, Simon R, De Veylder L, Bräutigam A, Weber APM. 2014. Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C(3) and C(4) Plant Species. *Plant Cell* **26**, 3243-3260.
Kümpers BMC, Burgess SJ, Reyna-Llorens I, Smith-Unna R, Boursnell C, Hibberd JM. 2017. Shared characteristics underpinning C4 leaf maturation derived from analysis of multiple C3 and C4 species of Flaveria. *Journal of Experimental Botany* **68**, 177-189.
Lauterbach M, Billakurthi K, Kadereit G, Ludwig M, Westhoff P, Gowik U. 2017. C3 cotyledons are followed by C4 leaves: intra-individual transcriptome analysis of Salsola soda (Chenopodiaceae). *Journal of Experimental Botany* **68**, 161-176.
Leegood RC. 2013. Strategies for engineering C4 photosynthesis. *Journal of Plant Physiology* **170**, 378-388.
Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR. 2010. The developmental dynamics of the maize leaf transcriptome. *Nat Genet* **42**.
Li Y, Ma X, Zhao J, Xu J, Shi J, Zhu X-G, Zhao Y, Zhang H. 2015. Developmental Genetic Mechanisms of C(4) Syndrome Based on Transcriptome Analysis of C(3) Cotyledons and C(4) Assimilating Shoots in Haloxylon ammodendron. *PLoS ONE* **10**, e0117175.
Lin H, Karki S, Coe RA, Bagha S, Khoshravesh R, Balahadia CP, Ver Sagun J, Tapia R, Israel WK, Montecillo F, de Luna A, Danila FR, Lazaro A, Realubit CM, Acoba MG, Sage TL, von Caemmerer S, Furbank RT, Cousins AB, Hibberd JM, Quick WP, Covshoff S. 2016. Targeted Knockdown of GDCH in Rice Leads to a Photorespiratory Deficient Phenotype Useful as a Building Block for C4 Rice. *Plant and Cell Physiology*.
Long SP, Zhu X-G, Naidu SL, Ort DR. 2006. Can improvement in photosynthesis increase crop yields? *Plant, Cell and Environment* **29**, 315-330.
Luethy MH, Miernyk JA, David NR, Randall DD. 1996. Plant pyruvate dehydrogenase complexes. In: Patel MS, Roche TE, Harris RA, eds. *Alpha-Keto Acid Dehydrogenase Complexes*. Basel: Birkhäuser Basel, 71-92.
Mallmann J, Heckmann D, Bräutigam A, Lercher MJ, Weber AP, Westhoff P, Gowik U. 2014. The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. *eLife* **3**, e02478.
Marshall DM, Muhaidat R, Brown NJ, Liu Z, Stanley S, Griffiths H, Sage RF, Hibberd JM. 2007. Cleome, a genus closely related to Arabidopsis, contains species spanning a developmental progression from C3 to C4 photosynthesis. *The Plant Journal* **51**, 886-896.
Mathelier A, Xin B, Chiu T-P, Yang L, Rohs R, Wasserman Wyeth W. 2016. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Systems* **3**, 278-286.e274.
Matsuoka M, Kyozuka J, Shimamoto K, Kano-Murakami Y. 1994. The promoters of two carboxylases in a C4 plant (maize) direct cell-specific, light-regulated expression in a C3 plant (rice). *Plant J* **6**, 311-319.

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* **226**, 792-801.
McKown AD, Dengler NG. 2009. Shifts in leaf vein density through accelerated vein formation in C4 Flaveria (Asteraceae). *Ann Bot* **104**, 1085-1098.
McKown AD, Dengler NG. 2010. Vein patterning and evolution in c4 plants. *Botany* **88**, 775-786.
Meister M, Agostino A, Hatch MD. 1996. The roles of malate and aspartate in C4 photosynthetic metabolism of Flaveria bidentis (L.). *Planta* **199**, 262-269.
Meyer HM, Teles J, Formosa-Jordan P, Refahi Y, San-Bento R, Ingram G, Jönsson H, Locke JCW, Roeder AHK. 2017. Fluctuations of the transcription factor ATML1 generate the pattern of giant cells in the Arabidopsis sepal. *eLife* **6**, e19131.
Meyer P, Saedler H. 1996. HOMOLOGY-DEPENDENT GENE SILENCING IN PLANTS. *Annual Review of Plant Physiology and Plant Molecular Biology* **47**, 23-48.
Mirny LA. 2010. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences* **107**, 22534-22539.
Mitchell PL, Sheehy JE. 2006. Supercharging rice photosynthesis to increase yield. *New Phytologist* **171**, 688-693.
Monson RK. 1999. 11 - The Origins of C4 Genes and Evolutionary Pattern in the C4 Metabolic Phenotype. In: Sage RF, Monson RK, eds. *C4 plant biology*. San Diego: Academic Press, 377-410.
Monson RK. 2003. Gene Duplication, Neofunctionalization, and the Evolution of C4Photosynthesis. *International Journal of Plant Sciences* **164**, S43-S54.
Monson RK, Edwards GE, Ku MSB. 1984. C3-C4 Intermediate Photosynthesis in Plants. *BioScience* **34**, 563-574.
Monson RK, Moore Bd. 1989. On the significance of C3-C4 intermediate photosynthesis to the evolution of C4 photosynthesis. *Plant, Cell and Environment* **12**, 689-699.
Morgan CL, Turner SR, Rawsthorne S. 1993. Coordination of the cell-specific distribution of the four subunits of glycine decarboxylase and of serine hydroxymethyltransferase in leaves of C3-C4 intermediate species from different genera. *Planta* **190**, 468-473.
Muhaidat R, McKown AD. 2013. Significant involvement of PEP-CK in carbon assimilation of C4 eudicots. *Ann Bot* **111**, 577-589.
Muhaidat R, Sage TL, Frohlich MW, Dengler NG, Sage RF. 2011. Characterization of C(3)--C(4) intermediate species in the genus Heliotropium L. (Boraginaceae): anatomy, ultrastructure and enzyme activity. *Plant Cell Environ* **34**, 1723-1736.
Nakamura H, Muramatsu M, Hakata M, Ueno O, Nagamura Y, Hirochika H, Takano M, Ichikawa H. 2009. Ectopic Overexpression of The Transcription Factor OsGLK1 Induces Chloroplast Development in Non-Green Rice Cells. *Plant and Cell Physiology* **50**, 1933-1949.
Nisbet EG, Grassineau NV, Howe CJ, Abell PI, Regelous M, Nisbet RER. 2007. The age of Rubisco: the evolution of oxygenic photosynthesis. *Geobiology* **5**, 311-335.
Noctor G, Arisi A-CM, Jouanin L, Foyer CH. 1999. Photorespiratory glycine enhances glutathione accumulation in both the chloroplastic and cytosolic compartments. *Journal of Experimental Botany* **50**, 1157-1167.
Nomura M, Higuchi T, Ishida Y, Ohta S, Komari T, Imaizumi N, Miyao-Tokutomi M, Matsuoka M, Tajima S. 2005. Differential expression pattern of C4 bundle sheath expression genes in rice, a C3 plant. *Plant Cell Physiol* **46**, 754-761.
Ohnishi J-i, Kanai R. 1983. Differentiation of Photorespiratory Activity between Mesophyll and Bundle Sheath Cells of C4 Plants I. Glycine Oxidation by Mitochondria. *Plant and Cell Physiology* **24**, 1411-1420.

Oliver DJ, Neuburger M, Bourguignon J, Douce R. 1990. Interaction between the Component Enzymes of the Glycine Decarboxylase Multienzyme Complex. *Plant Physiology* **94**, 833-839.
Ort DR, Merchant SS, Alric J, Barkan A, Blankenship RE, Bock R, Croce R, Hanson MR, Hibberd JM, Long SP, Moore TA, Moroney J, Niyogi KK, Parry MAJ, Peralta-Yahya PP, Prince RC, Redding KE, Spalding MH, van Wijk KJ, Vermaas WFJ, von Caemmerer S, Weber APM, Yeates TO, Yuan JS, Zhu XG. 2015. Redesigning photosynthesis to sustainably meet global food and bioenergy demand. *Proc Natl Acad Sci U S A* **112**, 8529-8536.
Osborne CP, Sack L. 2012. Evolution of C4 plants: a new hypothesis for an interaction of CO2 and water relations mediated by plant hydraulics. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 583-600.
Perfectti F, Gómez JM, González-Megías A, Abdelaziz M, Lorite J. 2017. Molecular phylogeny and evolutionary history of Moricandia DC (Brassicaceae). *PeerJ* **5**, e3964.
Peterhänsel C, Horst I, Niessen M, Blume C, Kebeish R, Kurkcuoglu S, Kreuzaler F. 2010. Photorespiration. *Arabidopsis Book* **8**, e0130.
Pick TR, Bräutigam A, Schluter U, Denton AK, Colmsee C, Scholz U, Fahnenstich H, Pieruschka R, Rascher U, Sonnewald U, Weber AP. 2011. Systems analysis of a maize leaf developmental gradient redefines the current C4 model and provides candidates for regulation. *Plant Cell* **23**, 4208-4220.
Pick TR, Bräutigam A, Schulz MA, Obata T, Fernie AR, Weber APM. 2013. PLGG1, a plastidic glycolate glycerate transporter, is required for photorespiration and defines a unique class of metabolite transporters. *Proceedings of the National Academy of Sciences* **110**, 3185-3190.
Ptashne M, Jeffrey A, Johnson AD, Maurer R, Meyer BJ, Pabo CO, Roberts TM, Sauer RT. 1980. How the lambda repressor and cro work. *Cell* **19**, 1-11.
Pyankov V, Ziegler H, Kuz'min A, Edwards G. 2001. Origin and evolution of C4 photosynthesis in the tribe Salsoleae (Chenopodiaceae) based on anatomical and biochemical types in leaves and cotyledons. *Plant Systematics and Evolution* **230**, 43-74.
Raven JA. 2013. Rubisco: still the most abundant protein of Earth? *New Phytologist* **198**, 1-3.
Rawsthorne S. 1992. C3–C4 intermediate photosynthesis: linking physiology to gene expression. *The Plant Journal* **2**, 267-274.
Rawsthorne S, Hylton CM, Smith AM, Woolhouse HW. 1988. Distribution of photorespiratory enzymes between bundle-sheath and mesophyll cells in leaves of the C3–C4 intermediate species Moricandia arvensis (L.) DC. *Planta* **176**, 527-532.
Rebollo R, Romanish MT, Mager DL. 2012. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annu Rev Genet* **46**, 21-42.
Reeves G, Grangé-Guermente MJ, Hibberd JM. 2017. Regulatory gateways for cell-specific gene expression in C4 leaves with Kranz anatomy. *Journal of Experimental Botany* **68**, 107-116.
Renné P, Dreßen U, Hebbeker U, Hille D, Flügge UI, Westhoff P, Weber APM. 2003. The Arabidopsis mutant dct is deficient in the plastidic glutamate/malate translocator DiT2. *The Plant Journal* **35**, 316-331.
Reyna-Llorens I, Burgess SJ, Reeves G, Singh P, Stevenson SR, Williams BP, Stanley S, Hibberd JM. 2018. Ancient duons may underpin spatial patterning of gene expression in C4 leaves. *Proceedings of the National Academy of Sciences*.
Rossini L, Cribb L, Martin DJ, Langdale JA. 2001. The Maize Golden2 Gene Defines a Novel Class of Transcriptional Regulators in Plants. *Plant Cell* **13**, 1231-1244.
Sage RF. 2004. The evolution of C4 photosynthesis. *New Phytologist* **161**, 341-370.

Sage RF. 2016. A portrait of the C4 photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineages, and Hall of Fame. *J Exp Bot* **67**, 4039-4056.
Sage RF, Christin PA, Edwards EJ. 2011. The C(4) plant lineages of planet Earth. *J Exp Bot* **62**, 3155-3169.
Sage RF, Coleman JR. 2001. Effects of low atmospheric CO2 on plants: more than a thing of the past. *Trends Plant Sci* **6**, 18-24.
Sage RF, Sage TL, Kocacinar F. 2012. Photorespiration and the evolution of C4 photosynthesis. *Annu Rev Plant Biol* **63**, 19-47.
Sage RF, Zhu X-G. 2011. Exploiting the engine of C4 photosynthesis. *Journal of Experimental Botany* **62**, 2989-3000.
Sage TL, Busch FA, Johnson DC, Friesen PC, Stinson CR, Stata M, Sultmanis S, Rahman BA, Rawsthorne S, Sage RF. 2013. Initial Events during the Evolution of C4 Photosynthesis in C3 Species of Flaveria. *Plant Physiology* **163**, 1266-1276.
Scarpella E, Helariutta Y. 2010. Vascular pattern formation in plants. *Current Topics in Developmental Biology*, Vol. 91, 221-265.
Schuler ML, Mantegazza O, Weber APM. 2016. Engineering C4 photosynthesis into C3 chassis in the synthetic biology age. *The Plant Journal* **87**, 51-65.
Schulze E-D, Hall AE. 1982. Stomatal Responses, Water Loss and CO2 Assimilation Rates of Plants in Contrasting Environments. In: Lange OL, Nobel PS, Osmond CB, Ziegler H, eds. *Physiological Plant Ecology II: Water Relations and Carbon Assimilation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 181-230.
Schulze S, Mallmann J, Burscheidt J, Koczor M, Streubel M, Bauwe H, Gowik U, Westhoff P. 2013. Evolution of C4 photosynthesis in the genus flaveria: establishment of a photorespiratory CO2 pump. *Plant Cell* **25**, 2522-2535.
Sharkey TD. 1988. Estimating the rate of photorespiration in leaves. *Physiologia Plantarum* **73**, 147-152.
Sheehy J. 2001. Will yield barriers limit future rice production. *Crop Science and Prospect, CABI Publishing*, 281-306.
Sheehy JE. 2000. Limits to yield for C3 and C4 rice: an agronomist's view. In: Sheehy JE, Mitchell PL, Hardy B, eds. *Redesigning rice photosynthesis to increase yield*: Elsevier, 39-52.
Slewinski TL, Anderson AA, Price S, Withee JR, Gallagher K, Turgeon R. 2014. Short-Root1 Plays a Role in the Development of Vascular Tissue and Kranz Anatomy in Maize Leaves. *Molecular Plant* **7**, 1388-1392.
Slewinski TL, Anderson AA, Zhang C, Turgeon R. 2012. Scarecrow Plays a Role in Establishing Kranz Anatomy in Maize Leaves. *Plant and Cell Physiology* **53**, 2030-2037.
Somerville SC, Somerville CR. 1985. A mutant of Arabidopsis deficient in chloroplast dicarboxylate transport is missing an envelope protein. *Plant Science Letters* **37**, 217-220.
Sommer M, Bräutigam A, Weber APM. 2012. The dicotyledonous NAD malic enzyme C4 plant Cleome gynandra displays age‐dependent plasticity of C4 decarboxylation biochemistry. *Plant Biology* **14**, 621-629.
Spreitzer RJ, Salvucci ME. 2002. RUBISCO: Structure, Regulatory Interactions, and Possibilities for a Better Enzyme. *Annual Review of Plant Biology* **53**, 449-475.
Srinivasan R, Oliver DJ. 1995. Light-Dependent and Tissue-Specific Expression of the H-Protein of the Glycine Decarboxylase Complex. *Plant Physiology* **109**, 161-168.
Still CJ, Berry JA, Collatz GJ, DeFries RS. 2003. Global distribution of C3 and C4 vegetation: Carbon cycle implications. *Global Biogeochemical Cycles* **17**, 6-1-6-14.

Sugimoto-Shirasu K, Roberts K. 2003. “Big it up”: endoreduplication and cell-size control in plants. *Current Opinion in Plant Biology* **6**, 544-553.
Surridge C. 2002. The rice squad. *Nature* **416**, 576.
Tabita FR, Hanson TE, Li H, Satagopan S, Singh J, Chan S. 2007. Function, Structure, and Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs. *Microbiology and Molecular Biology Reviews* **71**, 576-599.
Takahashi S, Bauwe H, Badger M. 2007. Impairment of the photorespiratory pathway accelerates photoinhibition of photosystem II by suppression of repair but not acceleration of damage processes in Arabidopsis. *Plant Physiol* **144**, 487-494.
Takeuchi Y, Akagi H, Kamasawa N, Osumi M, Honda H. 2000. Aberrant chloroplasts in transgenic rice plants expressing a high level of maize NADP-dependent malic enzyme. *Planta* **211**, 265-274.
Timm S, Florian A, Arrivault S, Stitt M, Fernie AR, Bauwe H. 2012. Glycine decarboxylase controls photosynthesis and plant growth. *FEBS Lett* **586**, 3692-3697.
Timm S, Giese J, Engel N, Wittmiß M, Florian A, Fernie AR, Bauwe H. 2017. T-protein is present in large excess over the other proteins of the glycine cleavage system in leaves of Arabidopsis. *Planta*.
Timm S, Wittmiß M, Gamlien S, Ewald R, Florian A, Frank M, Wirtz M, Hell R, Fernie AR, Bauwe H. 2015. Mitochondrial Dihydrolipoyl Dehydrogenase Activity Shapes Photosynthesis and Photorespiration of Arabidopsis thaliana. *Plant Cell* **27**, 1968-1984.
Ueno O, Samejima M, Muto S, Miyachi S. 1988. Photosynthetic characteristics of an amphibious plant, Eleocharis vivipara: Expression of C4 and C3 modes in contrasting environments. *Proceedings of the National Academy of Sciences* **85**, 6733-6737.
Ueno O, Sentoku N. 2006. Comparison of leaf structure and photosynthetic characteristics of C3 and C4Alloteropsis semialata subspecies. *Plant, Cell & Environment* **29**, 257-268.
Vauclare P, Macherel D, Douce R, Bourguignon J. 1998. The gene encoding T protein of the glycine decarboxylase complex involved in the mitochondrial step of the photorespiratory pathway in plants exhibits features of light-induced genes. *Plant Mol Biol* **37**, 309-318.
Vogan PJ, Frohlich MW, Sage RF. 2007. The functional significance of C3–C4 intermediate traits in Heliotropium L. (Boraginaceae): gas exchange perspectives. *Plant, Cell & Environment* **30**, 1337-1345.
von Caemmerer S, Furbank RT. 1999. 6 - Modeling C4 Photosynthesis. In: Sage RF, Monson RK, eds. *C4 plant biology*. San Diego: Academic Press, 173-211.
Walker JL, Oliver DJ. 1986. Light-induced increases in the glycine decarboxylase multienzyme complex from pea leaf mitochondria. *Archives of Biochemistry and Biophysics* **248**, 626-638.
Wang L, Czedik-Eysenberg A, Mertz RA, Si Y, Tohge T, Nunes-Nesi A, Arrivault S, Dedow LK, Bryant DW, Zhou W, Xu J, Weissmann S, Studer A, Li P, Zhang C, LaRue T, Shao Y, Ding Z, Sun Q, Patel RV, Turgeon R, Zhu X, Provart NJ, Mockler TC, Fernie AR, Stitt M, Liu P, Brutnell TP. 2014a. Comparative analyses of C4 and C3 photosynthesis in developing leaves of maize and rice. *Nature Biotechnology* **32**, 1158.
Wang P, Kelly S, Fouracre JP, Langdale JA. 2013. Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz anatomy. *Plant J* **75**, 656-670.
Wang P, Khoshravesh R, Karki S, Tapia R, Balahadia CP, Bandyopadhyay A, Quick WP, Furbank R, Sage TL, Langdale JA. 2017. Re-creation of a Key Step in the Evolutionary Switch from C3 to C4 Leaf Anatomy. *Current Biology* **27**, 3278-3287.e3276.

Wang Y, Bräutigam A, Weber APM, Zhu X-G. 2014b. Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis. *Journal of Experimental Botany* **65**, 3567-3578.
Wessler SR. 1996. Plant retrotransposons: Turned on by stress. *Current Biology* **6**, 959-961.
Westhoff P, Gowik U. 2004. Evolution of c4 phosphoenolpyruvate carboxylase. Genes and proteins: a case study with the genus Flaveria. *Ann Bot* **93**, 13-23.
Westhoff P, Gowik U. 2010. Evolution of C4 photosynthesis--looking for the master switch. *Plant Physiol* **154**, 598-601.
Williams BP, Johnston IG, Covshoff S, Hibberd JM. 2013. Phenotypic landscape inference reveals multiple evolutionary paths to C4 photosynthesis. *eLife* **2**, e00961.
Willis K, ed. 2016. *The state of the world's plants 2017*. Richmond, Surrey, UK: Royal Botanic Gardens, Kew. ISBN: 978-1-84246-647-6
Wingler A, Walker RP, Chen Z-H, Leegood RC. 1999. Phosphoenolpyruvate Carboxykinase Is Involved in the Decarboxylation of Aspartate in the Bundle Sheath of Maize. *Plant Physiology* **120**, 539-546.
Xu J, Bräutigam A, Weber APM, Zhu X-G. 2016. Systems analysis of cis-regulatory motifs in C(4) photosynthesis genes using maize and rice leaf transcriptomic data during a process of de-etiolation. *Journal of Experimental Botany* **67**, 5105-5117.
Zhu X-G, Long SP, Ort DR. 2008. What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? *Current Opinion in Biotechnology* **19**, 153-159.
Zhu X-G, Long SP, Ort DR. 2010. Improving Photosynthetic Efficiency for Greater Yield. *Annual Review of Plant Biology* **61**, 235-261.

II. Scientific Aims

With an estimated population of 10 billion people by 2050 and stagnating crop productivity improvement by conventional breeding, the world is facing a looming food crisis. Engineering C_4 photosynthesis could improve productivity of C_3 crops by up to 50 %. C_4 photosynthesis is a metabolic and anatomical syndrome, which allows plants to concentrate CO_2 around Rubisco, reducing futile photorespiration and at the same time improving water and nitrogen usage efficiencies. For this, C_4 photosynthesis relies on spatially confining carbon fixation to an outer compartment and its assimilation to an inner compartment of the leaf, requiring a specialised leaf morphology, called Kranz anatomy. However, neither is known how to express genes efficiently in the respective compartment nor how to alter leaf anatomy accordingly.

(1) The glycine decarboxylase (GDC) is a central enzyme of photorespiration and composed of the four subunits GLDP, GLDT, GLDH and GLDL. On the evolutionary trajectory from C_3 to C_4, GDC expression becomes confined to an inner compartment – the bundle sheath cells (BSC) – creating an early CO_2 concentrating mechanism. The genus *Flaveria*, as one of the youngest C_4 origins, encompasses several closely related species with varying degree of "C_4-ness", reflecting intermediate states of C_4 evolution. In this study, promoter:reporter fusion constructs were generated from several *Flaveria* species and transformed into C_3 and C_4 plants, to elucidate the evolution and regulation of GLDT expression. Analyses of promoter truncations and fusions were coupled with discriminative sequence analysis to identify conserved polymorphisms that associate with differential expression of GLDT.

(2) Introduction of C_4 photosynthesis into C_3 crops requires artificial promoters that harbour only necessary and functional CRE, to cope with limited transformation bandwidth and to efficiently regulate spatio-temporal C_4 gene expression. However, so far, neither precise nor sufficient reporter gene expression in BSC of the C_3 crop rice was reported. This study addressed the phosphoenolpyruvate carboxykinase (PCK) promoter from the C_4 grass *Zoysia japonica*, which was previously shown to be active in rice vascular bundles. Promoter:reporter fusion constructs were generated and transformed into rice. Analyses of promoter truncation and fusion

constructs were used to delimit a region of interest. Subsequent CRE prediction and comparison with RNA-seq studies was conducted to identify relevant CRE and corresponding transcription factors.

(3) Kranz anatomy is, amongst others, characterised by close vein spacing and large BSC. In maize, both traits are established early in leaf development and differ dramatically between its foliar and husk leaves. Comparative RNA-seq on leaf developmental gradients from both leaf types identified a set of transcription factors that are upregulated in husk leaves during the stage of vein patterning, suggesting a negative regulatory effect on Kranz anatomy traits. In this study, orthologues of these potential negative regulators were knocked down by hairpin-RNA induced post-transcriptional gene silencing. Quantitative real time PCR was conducted, to confirm knock down of target genes. Kranz anatomy related parameters were determined in cross-sections of T0 plants, from knock down and reference constructs, to assess the impact of each transcription factor on leaf morphology.

III. Summary

Synthetic C_4 photosynthesis could be the answer to ensure food security in face of an exponentially growing world population. Thorough understanding of the molecular factors that control C_4 photosynthesis is pivotal to its engineering. Since C_4 photosynthesis requires precise spatial gene expression and a specialised leaf anatomy, this study strived to elucidate genetic factors that regulate these traits.

Upstream flanking sequences of the GLDT gene from C_3, C_3-C_4 and C_4 *Flaveria* species were fused to reporter genes and analysed in C_3 and C_4 background. The results revealed that GLDT is differentially localised in C_3 and C_4 *Flaveria* species. It was shown that the underlying regulatory mechanism is also functional in distantly related C_3 species and acts on transcriptional level. Sequence analysis indicated that spatially confined expression correlated with the presence of large conserved regions upstream of the GLDT gene. The insertions are exclusive to the only *Flaveria* clade that contains fully evolved C_4 species. Deletion and substitution of these conserved regions showed that one of these regions is necessary for spatially confined expression. Consensus reconstruction from RNA-seq data suggested that this region corresponds to an ancient insertion of a small transposable element – a MITE – which seemed to be highly abundant in *Flaveria* species. Subsequent dissection narrowed down the region of interest to 59 bp. Additional swapping of conserved regions confirmed the presence of a second *cis*-regulatory element and sequence analysis identified 11 polymorphisms, which associate with differential expression of GLDT.

In a similar approach, the upstream flanking sequence of a PCK gene from the C_4 grass *Zoysia japonica* was isolated and analysed in the C_3 background of rice. Reporter gene localisation showed expression exclusively in bundle sheath cells. Subsequent truncation and fusion constructs revealed that deletion of a 300 bp fragment altered the expression pattern from bundle sheath to mesophyll specific. Prediction of CREs indicated that only three known binding sites were exclusive to this region. Two of these corresponded to GOLDEN 2-like transcription factors (TFs) – known regulators of C4 photosynthesis and morphology.

A comprehensive study previously identified 18 candidate-TFs that negatively correlated with development of C_4 leaf anatomy. Here, orthologues of these negative regulator candidates were knocked down in rice, by hairpin-RNA-induced post-transcriptional gene silencing. The knock down was confirmed by quantitative real time PCR. Its impact on leaf anatomy was quantitatively assessed by measurement of three to four relevant anatomical parameters in the T0 generation. Although, some candidates induced strong phenotypes, none of the relevant parameters were severely affected. Similar was observed in a corresponding study, where 60 potential positive regulators were ectopically expressed in rice, but were unable to affect relevant parameters. These results indicated that either the early leaf developmental program is highly buffered against changes on transcriptional level or that the *ZmUbi1* promoter – used by default in monocot species and utilised in both studies – is not as ubiquitously expressed as generally assumed.

IV. Chapters

1. **Insight into the evolution of GLDT expression in the asterid genus Flaveria.**

2. **Towards mapping of *cis*-regulatory elements in the upstream flanking sequence of *GLDT* from the genus *Flaveria.***

3. **Dissection of the phosphoenolpyruvate carboxykinase upstream flanking sequence from the C_4 grass *Zoysia japonica.***

4. **Knockdown of potential negative Kranz anatomy regulators in rice.**

1. Insight into the evolution of GLDT expression in the asterid genus *Flaveria*

Introduction

Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) is an ambivalent enzyme, being able to fix either CO_2 or O_2, depending on their ratio. Under high CO_2, it prevalently catalyses the reaction of ribulose-1,5-bisphosphate to two molecules of 3-phosphoglycerate (3PGA) as part of the Calvin cycle. With decreasing CO_2 levels, though, Rubisco's affinity for O_2 becomes problematic, since it generates not only one molecule of 3PGA but also one molecule 2-phosphoglycolate (2PG), which is useless to the Calvin cycle and even toxic when accumulated (Anderson, 1971; Kelly and Latzko, 1977). The 2PG has to be recycled in a process termed photorespiration or C_2 cycle, consuming ATP and releasing previously fixed NH_3 and CO_2.

Glycine decarboxylase T-protein (GLDT) is one of four subunits (P, H, L and T) of the glycine decarboxylase (GDC), also termed glycine cleavage system. GDC is loosely assembled in the mitochondrial matrix, where it catalyses a central reaction of the C_2 cycle. Together with serine hydroxymethyltransferase (SHMT) it converts two molecules of glycine to one molecule of serine, releasing the aforementioned NH_3 and CO_2. Reduction in GDC activity by knockdown of P, H, or T subunit leads to accumulation of glycine and a chlorotic phenotype, which recovers under elevated CO_2 (Engel *et al.*, 2007; Timm *et al.*, 2017; Zhou *et al.*, 2013). However complete knockout of either P or T subunit in *Arabidopsis thaliana* is lethal even under elevated CO_2 (Engel *et al.*, 2007; Timm *et al.*, 2017), supporting the hypothesis that low amounts of GDC are essential to C_1 metabolism (Bauwe, 2011; Hanson and Roje, 2001).

To reduce photorespiration, mechanisms evolved that concentrate CO_2 around Rubisco. One of those carbon concentrating mechanisms (CCM), the economically most important one, is C_4 photosynthesis.

C_4 photosynthesis relies on uncoupling the initial CO_2 fixation and Rubisco by prefixing it through phosphoenolpyruvate carboxylase (PEPC), which has no affinity for O_2. Very briefly, the prefixed CO_2 diffuses as malate or aspartate to the site of Rubisco where it is released, while the carboxy donor diffuses back and is regenerated. In most higher plants this is facilitated by localising PEPC to mesophyll cells (MC) and Rubisco and the Calvin cycle to bundle sheath cell (BSC). This shift in metabolic activity is supported by a change in leaf anatomy, resulting in closely spaced veins, wreath-like surrounded by large, organelle rich BSC and one layer of

MC. This so called Kranz anatomy increases MC:BSC ratio and interface and it isolates Rubisco from the site of CO_2 entry into the leaf chlorenchyma. Consequently, photorespiration rates are low in C_4 species and water and nitrogen use efficiencies are increased, allowing them to grow under hot and arid conditions (Ehleringer and Monson, 1993).

Despite this brief description, C_4 photosynthesis is a rather complex trait and its establishment likely encompasses differential regulation of several hundred genes (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011). Nonetheless, it evolved more than 60 times independently (Sage, 2016), not only raising interest in how C_4 photosynthesis evolved but also why it was able to evolve in such a repeatable fashion (Sage *et al.*, 2012).

One excellent model to study the evolution of C_4 is the genus *Flaveria*, which is not only composed of C_4 and C_3 species but also of several intermediate species that exhibit varying degrees of C_4-ness (Holaday *et al.*, 1984; Ku *et al.*, 1983; Ku *et al.*, 1991). Those C_3-C_4 intermediates are, amongst other features, characterised by photorespiratory rates between C_3 and C_4 photosynthesis and an intermediate leaf anatomy (Edwards and Ku, 1987). But the decreased photorespiratory rates are not necessarily associated with the presence of a functional C_4 cycle (Ku *et al.*, 1991; Rawsthorne *et al.*, 1988b), instead intermediate species rely on a simpler CCM, termed now C_2 photosynthesis (Sage *et al.*, 2012; Vogan *et al.*, 2007).

C_2 photosynthesis is facilitated by an early form of Kranz-anatomy and localisation of GDC to the BSC (Sage *et al.*, 2014). This way, photorespiratory glycine has to move to BSC, where its decarboxylation by GDC increases CO_2 concentration up to 3-fold (Keerberg *et al.*, 2014). This, in return, might have favoured the evolution of increased photosynthetic and photorespiratory capacity of BSC and reduction of MC:BSC ratio (Bauwe, 2011; Sage *et al.*, 2012). Restriction of GDC to the BSC also leads to massive imbalance of NH_3 between MC and BSC, which needs to be fixed (Rawsthorne *et al.*, 1988a). Computer modelling suggests that this can be facilitated by increased C_4 enzyme activity (Mallmann *et al.*, 2014). However, the initial step of GDC is the decarboxylation of glycine by GLDP. Thus, efficient CO_2 enrichment can only occur when GLDP is restricted to the BSC. Indeed, for the C_3-C_4 species *Moricandia arvensis* it was shown that only the P subunit is localised to BSC, while in *Flaveria* C_3-C_4 species also other subunits are absent from MC (Morgan *et al.*, 1993; this study). Additionally, it was shown that restriction of GLDP, the actual

decarboxylase of GDC, to the BSC occurred gradually (Schulze *et al.*, 2013), resulting in increased selective pressure for the other subunits to trail GLDP. Here we present evidence that the localisation of GLDT in the genus *Flaveria* might have occurred in at least two steps, presumably in response to declining GLDP levels in MC.

Results

Promoter activities of the GLDT 5' upstream sequences of F. trinervia (C_4), F. ramosissima (C_3-C_4) and F. robusta (C_3) in transgenic F, bidentis (C_4) and Arabidopsis thaliana (C_3)

So far, we have a quite reasonable understanding of how evolution shaped expression of *GLDP* in *Flaveria* (Engelmann *et al.*, 2008; Schulze *et al.*, 2013; Schulze *et al.*, 2016; Wiludda *et al.*, 2012). This involved at least two gene copies, establishment of a cryptic promoter and nonsense mediated decay. In contrast, it was long known that the GLDT protein of *F. trinervia* localises to BS mitochondria (Morgan *et al.*, 1993), but nothing was known about the mechanism of regulation or how it was established. To address this matter we isolated upstream flanking sequences of *GLDT* coding regions from representative C_3, C_3-C_4 and C_4 *Flaveria* species either by genome walking (*F. robusta* (C_3) and *F. ramosissima* (C_3-C_4)) or based on the publicly available sequence (*F. trinervia* (C_4), accession no. Z99769).
Rapid amplification of cDNA ends (RACE) of all three species mainly correlated to the annotated transcriptional start site (TSS) of *F. trinervia* and confirmed the position of the first intron inside the 5’ untranslated region (5’-UTR). Notably, mapping of short reads from RNA-seq experiments (Mallmann *et al.*, 2014) also indicated low transcription from an alternative TSS inside the first intron for *F. ramosissima* and *F. trinervia* (data not shown).
In contrast to *GLDP*, neither genome walking, 5’-RACE nor RNA-seq *de novo* assemblies indicated the presence of a second copy of *GLDT*. Coinciding with current genomic assemblies of all sequenced Asterid species, including the closest sequenced relative *Helianthus annuus* (Badouin *et al.*, 2017).

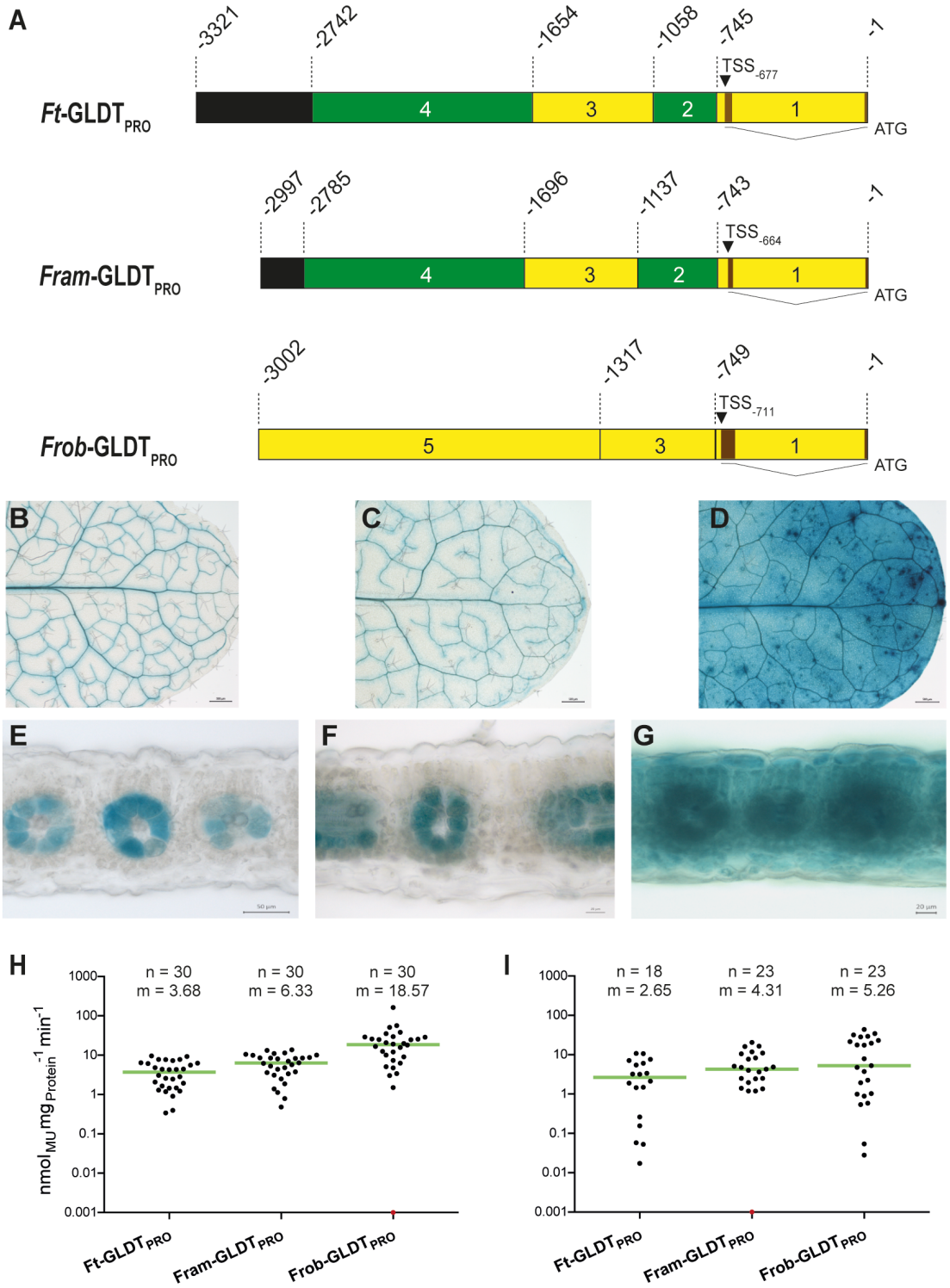


Figure 1 GUS expression analysis of Flaveria *GLDT* upstream flanking sequence. 5'-flanking sequence of different Flaveria species were isolated by genome walking, fused to a GUS reporter gene and analysed in transgenic *A. thaliana* and *F. bidentis*. **(A)** Schematic representation of *GLDT* upstream flanking sequence from *F. trinervia*, *F. ramosissima* and *F. robusta* drawn to scale. Regions conserved between flanking sequences of different species were assigned corresponding numbers. Regions that are only conserved in *F. trinveria* and *F. ramosissima* are highlighted in green. Black regions indicate no similarity to other sequences. Transcriptional start sites (TSS) were determined by 5'-RACE. The 5'-UTR (brown) contains an intron of app. 630 bp (indicated by open triangle). **(B – G)** Histochemical GUS localisation of *Ft*-GLDT$_{PRO}$, *Fram*-GLDT$_{PRO}$ and *Frob*-GLDT$_{PRO}$, respectively, in whole leaves of *A. thaliana* (B – D) and leaf cross sections of *F. bidentis* (E – F). Incubation times were 6h, 2h, 2h (B – D) and 2h, 15h, 3h (E – F). **(H – I)** Fluorometric quantification of GUS activity in transgenic *A. thaliana* (H) and *F. bidentis* (I). Median values are indicated by green line and are stated above (m). Red dots indicate measurements below detection limit.

Comparison of all three isolated upstream flanking sequences revealed two conserved regions exclusive to *F. trinervia* and *F. ramosissima* (regions 2 and 4, Fig. 1A), while the structure of the upstream flanking sequence of *F. robusta* mainly corresponded to the publicly available sequence of *F. anomala* (a C_3-C_4 species, not part of this study; accession no. Z71184).

The isolated sequences were fused to a GUS reporter gene and transformed into the C_4 species *F. bidentis* and the C_3 species *A. thaliana*. In leaves of *A. thaliana* the upstream flanking sequences of *F. trinervia* and *F. ramosissima* mediated GUS expression confined to vascular bundles (Fig. 1B, C), while in leaves of *F. bidentis* GUS was expressed exclusively in the BSC (Fig. 1E, F). In contrast, the upstream flanking sequence of *F. robusta* conveyed ubiquitous expression in leaves of both *A. thaliana* and *F. bidentis* (Fig. 1D, G). This indicated that BSC localisation of GLDT is i) conveyed on transcriptional level, ii) associated to one or both conserved regions exclusive to *F. ramosissima* and *F. trinervia* and iii) mediated by a mechanism that is conserved, at least, between Rosids and Asterids, which putatively split about 120 million years ago (Magallón and Castillo, 2009). Further, we found that the restriction of GUS expression in leaf vascular bundles of *A. thaliana* appeared less confined from *F. ramosissima* upstream flanking sequence compared to *F. trinervia* (cmp. Fig. 1B and C), while localisation was comparable in transgenic *F. bidentis* (Fig. 1E, F).

The upstream flanking sequence of F. trinervia GLDT gene

As indicated in Fig 1, the mechanism that confines spatial expression from the *GLDT* upstream flanking sequence of *F. trinervia* appeared to be conserved between the Asteracean C_4 species *F. bidentis* and the Brassicacean C_3 species *A. thaliana.*

However, while restriction of GUS activity to BSC was highly consistent in hand cross sections of transgenic *F. bidentis*, cellular localisation in *A. thaliana* was more cumbersome, due to the small size of the bundle sheath and vascular cells. In cross sections of resin-embedded leaves, GUS staining was observed in the whole vascular bundle, that is BSC *plus* vascular tissue, and occasionally even showed concentric staining in surrounding mesophyll cells (Fig. 2A). Therefore, we re-investigated the promoter activity of the GLDT upstream sequence of *F. trinervia* by fusing it to a fluorescent reporter gene encoding a nuclear-localised, non-diffusible

yellow fluorescent protein (H2B:YFP). Fig 2B illustrates that the H2B:YFP protein was only visible in BSC and vascular tissue of *A. thaliana*. This suggested that the observed GUS staining in MC was rather due to diffusion of the dye, than activity of the reporter gene. This also showed that the underlying mechanism does not act completely redundant in *F. bidentis* and *A. thaliana*, since additional expression was observed in the vascular tissue of *A. thaliana* (Fig. 2) but not in that of *F. bidentis* (Fig. 1E).

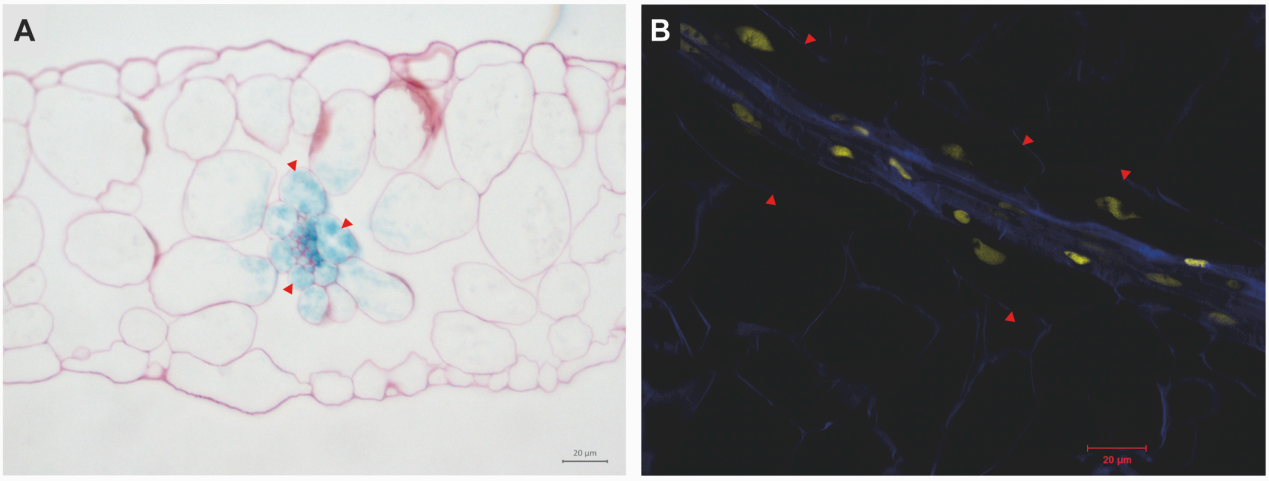


Figure 2 Cellular localisation of *F. trinervia GLDT* upstream flanking sequence activity in Arabidopsis. (A) Histochemical GUS localisation in transverse leaf sections of *A. thaliana* transformed with *Ft*-GLDT$_{PRO}$. Whole leaves were stained (5h), cleared and embedded in LR white (see Methods section). Cell walls were stained with Safranine O. **(B)** Expression of nuclear localised H2B:YFP under control of *F. trinervia GLDT* upstream flanking sequence in *A. thaliana* leaves. Optical parallel section by confocal microscopy. Cell walls were stained with calcofluor white. Red arrows mark bundle sheath cells.

Consecutive deletions and substitutions reveal pivotal roles for regions 3 and 2

Despite additional activity in the vascular tissue, spatial expression of *Ft*-GLDT$_{PRO}$ was comparable in F. bidentis and A. thaliana. For this reason we used the latter as host for dissecting the upstream flanking sequence of GLDT, allowing faster and easier transformation.

Alignment of the isolated *GLDT* upstream flanking sequences identified two conserved regions that were unique to *F. ramosissima* and *F. trinervia* (regions 2 and 4). Since both sequences generated spatially confined expression (see Fig. 1), it seemed plausible that spatial expression was associated with those conserved regions. Therefore, we generated three consecutive deletion constructs of *Ft*-

GLDT$_{PRO}$ (*Ft*-GLDT$_{PRO}$321, *Ft*-GLDT$_{PRO}$21 and *Ft*-GLDT$_{PRO}$1; see Fig. 3A) and analysed them in *A. thaliana*.

The GUS localisation (Fig. 3A) showed that regions 3, 2 and 1 (construct *Ft*-GLDT$_{PRO}$321) were able of providing spatial expression specificity, i.e. in bundle sheath and vascular tissue, indistinguishable from that of construct *Ft*-GLDT$_{PRO}$ (Fig. 2). This indicates that region 4 of the *GLDT* upstream flanking sequence (Fig. 2) is not needed for promoter activity in the bundle sheath and the vasculature. When region 3 was deleted (construct *Ft*-GLDT$_{PRO}$21) a drastic drop in promoter activity was observed (Fig. 3B), but not in the spatial expression pattern (Fig. 3D). Only when region 2 was also removed (construct *Ft*-GLDT$_{PRO}$1; Fig. 3) the promoter activity was completely lost (Fig. 3E). Similar was observed in analogous experiments with the isolated upstream flanking sequence of *F. robusta*, where region 1 alone did not convey any detectable GUS expression (data not shown).

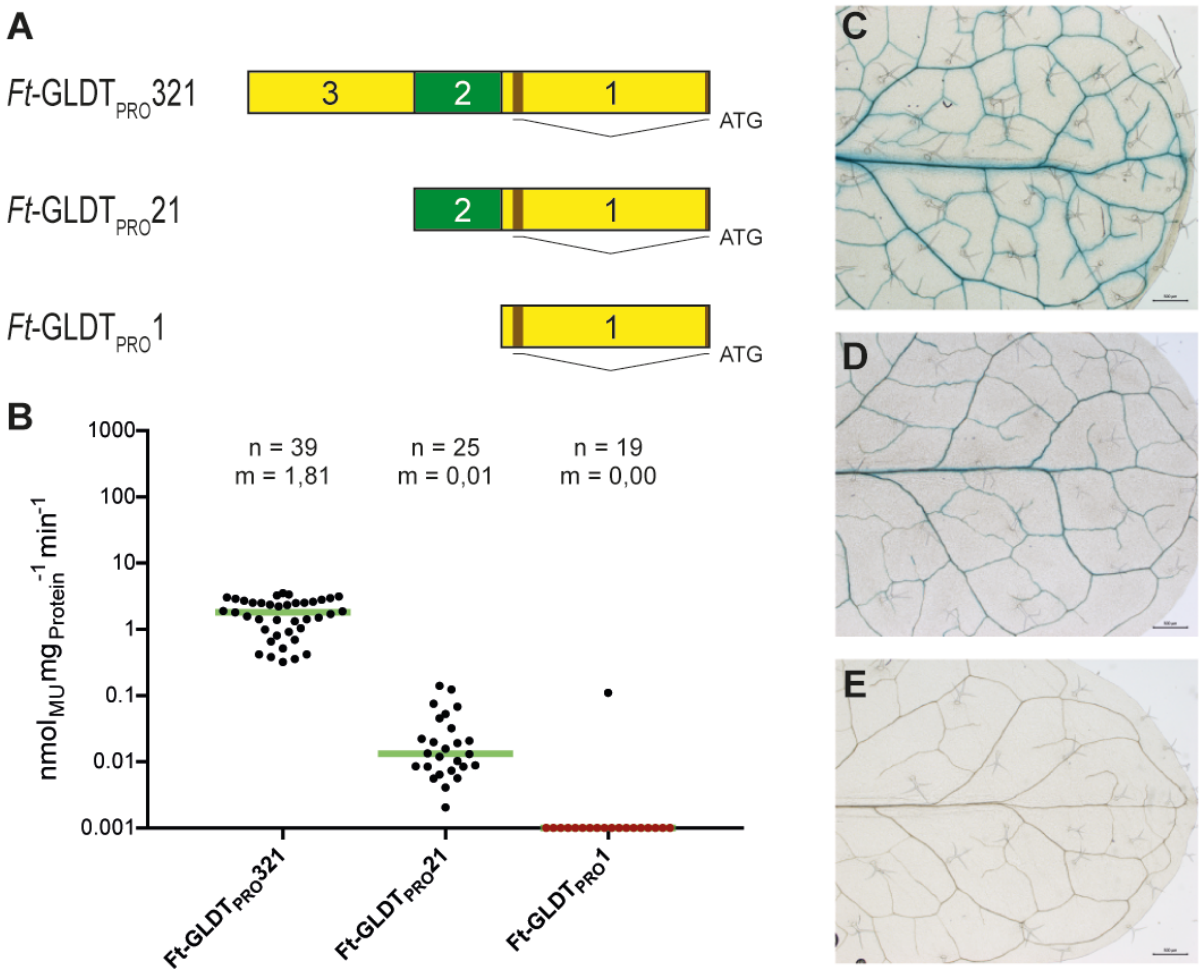


Figure 3 GUS expression analyses of *Ft*-GLDT$_{PRO}$ consecutive deletion constructs in *A. thaliana*. (A) Schematic representation of *Ft*-GLDT$_{PRO}$ deletion constructs. **(B)** Fluorometric quantification of GUS activity in transgenic A. thaliana. Median values are indicated by green lines and are stated above (m). Red dots indicate measurements below detection limit. **(C - E)** Histochemical GUS localisation in A. thaliana transformed with *Ft*-GLDT$_{PRO}$321, *Ft*-GLDT$_{PRO}$21 and *Ft*-GLDT$_{PRO}$1, respectively. Incubations times for staining were 10h, 24h and 6 days.

To confirm that construct *Ft*-GLDT$_{PRO}$21 was also able to convey confined expression in Flaveria this construct was additionally transformed into *F. bidentis*. Like the full construct *Ft*-GLDT$_{PRO}$ (Fig. 1E) *Ft*-GLDT$_{PRO}$21 mediated expression exclusively in BSC of transgenic *F. bidentis* (Supplemental Fig. S1).

These results demonstrated that indeed one of the regions that are unique to *F. trinervia* and *F. ramosissima*, i.e. region 2, positively affects spatial expression of *GLDT* and thus might be essential. However, region 3 could also harbour relevant *cis*-regulatory elements, as the drastic quantitative effect on GUS expression indicated.

We generated two additional constructs (*Ft*-GLDT$_{PRO}$32ID1 and *Ft*-GLDT$_{PRO}$32-60; see Fig. 4A) to address the question, whether promoter activity in the bundle sheath

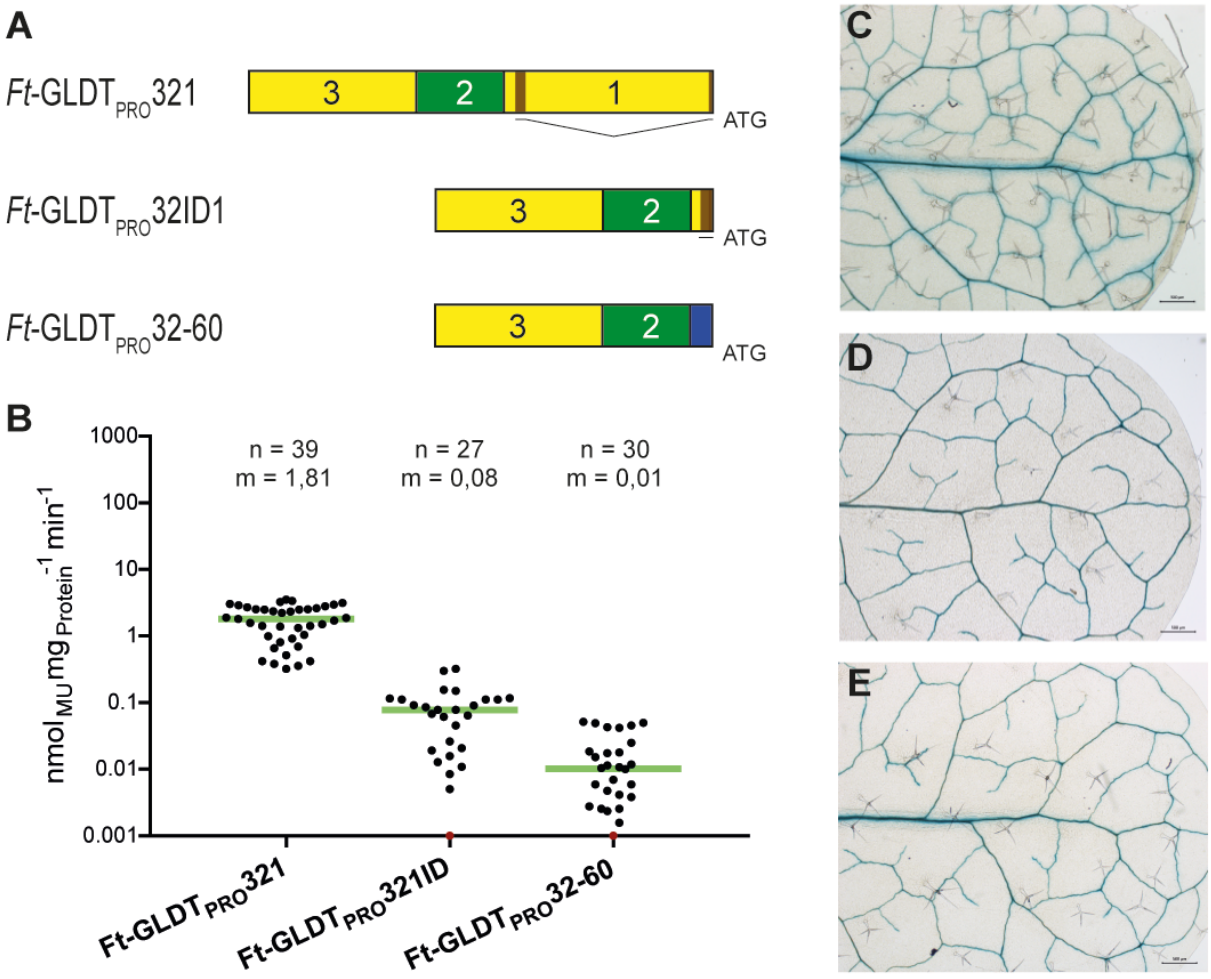


Figure 4 GUS expression analyses of region 1 truncation and substitution constructs. (A) Schematic representation of constructs transformed into Arabidopsis. Blue region represents *Cauliflower Mosaic Virus* (CaMV) minimal promoter (-60 to -1). **(B)** Fluorometric quantification of GUS activity in Arabidopsis. Median values are indicated by green lines and are stated above (m). **(C - E)** Histochemical GUS localisation in Arabidopsis transformed with *Ft*-GLDT$_{PRO}$321, *Ft*-GLDT$_{PRO}$321ID and *Ft*-GLDT$_{PRO}$32-60, respectively. Incubation times for staining were 10h, 7h and 2 days. Data from *Ft*-GLDT$_{PRO}$321 was included for comparison and is the same as in Fig. 3.

and the vasculature originates from region 3 and 2 or whether they only enhance expression in a general fashion, while the localisation is mediated by elements in the 5'-UTR or the leader intron, as it has been shown for other plant genes (Kim *et al.*, 2004; Patel *et al.*, 2004; Patel *et al.*, 2006). Deletion of the leader intron in region 1 (*Ft*-GLDT$_{PRO}$32ID1) led to a 20-fold decrease in GUS activity (Fig. 4B), but promoter activity in the bundle sheath and the vasculature could still be detected (Fig. 4D). Similarly, when region 1 was completely replaced by the *CaMV* 35S minimal promoter (*Ft*-GLDT$_{PRO}$32-60), the promoter activity was reduced 180-fold (Fig. 4B), but the spatial expression pattern was not affected (Fig. 4E).
These results indicated that, despite a strong quantitative effect, spatial expression of *GLDT* is not mediated by the leader intron and 5'-UTR or at least does not depend on it.

Excision and substitution of region 2 indicate a spacer function

Based on the finding that region 2 and 1 convey tissue specific expression, while region 1 alone does not, we hypothesized that region 2 might harbour *cis*-regulatory elements, causing the observed expression pattern. Hence, we generated an excision construct, where region 3 and 1 are fused in tandem (*Ft*-GLDT$_{PRO}$31; Fig. 5A). This construct is in line with the topology of the upstream flanking sequence of *F. robusta* in which a region 2 equivalent is missing (see Fig. 1). GUS staining of leaves of *A. thaliana* transformed with this construct revealed that promoter activity was not confined anymore to bundle sheath and vasculature, as background activity was clearly visible in mesophyll tissue (Fig. 5C). The GUS expression levels were not as high as those of the *GLDT* upstream flanking sequence of *F. robusta* (*Frob*-GLDT$_{PRO}$; Fig. 1H), neither appeared expression in the mesophyll as strong and ubiquitous as with the upstream flanking sequence of *F. robusta* (cmp. Fig. 1D and 5C). This clearly showed that region 2 is necessary to confine Ft-GLDT promoter activity to the leaf vascular bundles, i.e. bundle sheath and vasculature.
To elucidate the functional role of region 2 for *Ft*-GLDT promoter specificity, two different constructs were designed and tested in *A. thaliana*. Firstly, region 2 was replaced by its reverse complement (*Ft*-GLDT$_{PRO}$32rc1) to test whether this region might contain enhancer-like *cis*-regulatory modules. Secondly, an independent

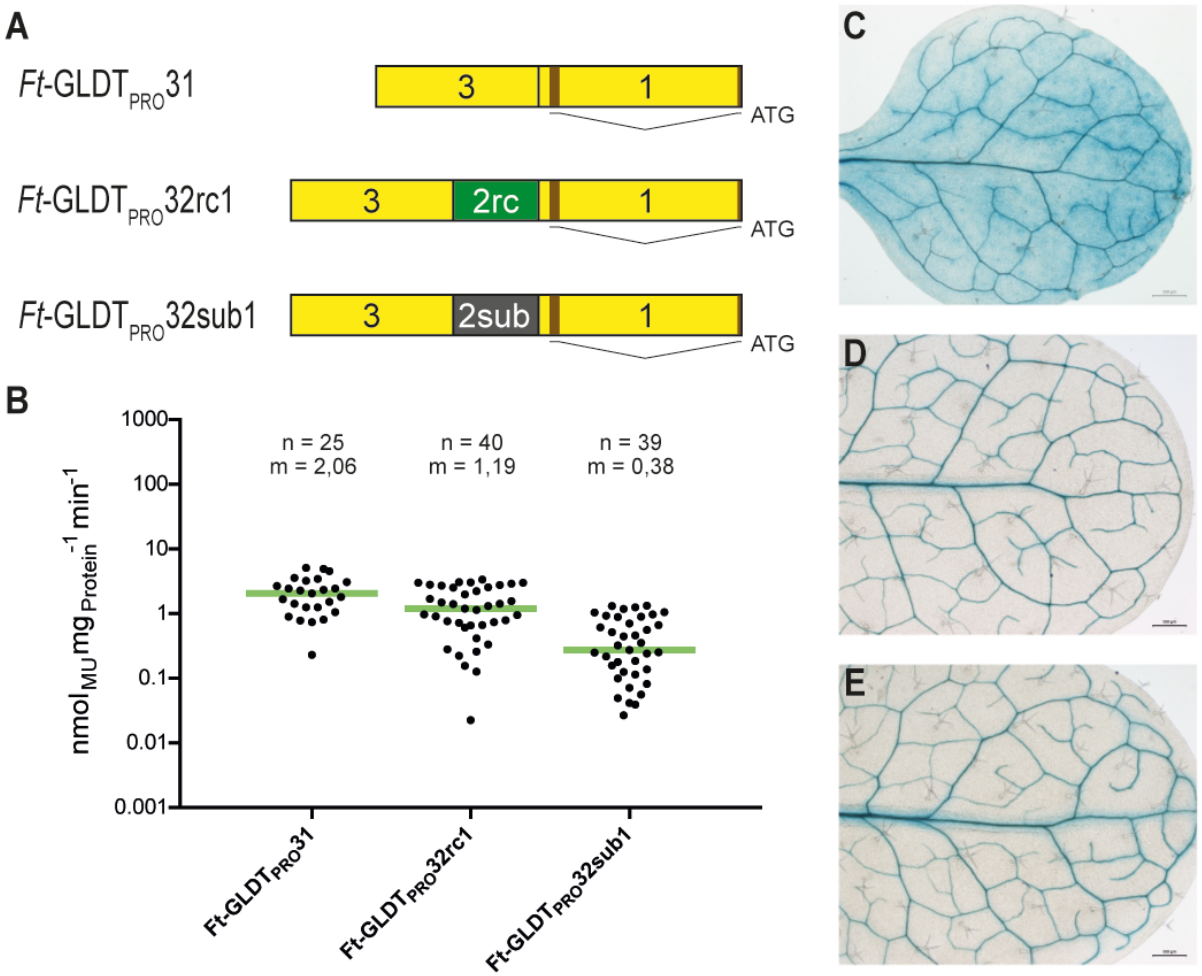


Figure 5 GUS analysis of region 2 excision and substitution constructs. (A) Schematic representation of *Ft*-GLDT$_{PRO}$321 deletion and substitution construct transformed into Arabidopsis. In constructs *Ft*-GLDT$_{PRO}$32rc1 and *Ft*-GLDT$_{PRO}$32sub1 region 2 was substituted with its reverse complement and a part of YFP coding region, respectively. (B) Fluorometric quantification of GUS activity in A. thaliana. Median values are indicated by green line and stated above (m). (C - E) Histochemical GUS localisation in A. thaliana transformed with *Ft*-GLDT$_{PRO}$31, *Ft*-GLDT$_{PRO}$32rc1 and *Ft*-GLDT$_{PRO}$32sub1, respectively. Incubation times for staining were 6h, 6h and 7h.

spacer sequence of the same size, an arbitrary fragment of the YFP coding region, was inserted between regions 1 and 3 resulting in *Ft*-GLDT$_{PRO}$32sub1 to investigate whether region 2 might rather function as a type of spacer segment. Fig. 5D and E show that the GUS expression pattern of both constructs was undistinguishable from the expression pattern of the native *GLDT* upstream sequence of *F. trinervia* (cmp. Fig. 3C), albeit with reduced activity. More precisely, substitution of region 2 with the YFP spacer sequence led to an about 5-fold decrease in activity, while substitution with its reverse complement decreased activity, non-significantly, by 1.5-fold median activity (Fig. 5B). Taken together, these results confirm the notion that the presence of region 2 is crucial to maintain a "C_4-like" expression pattern, but, although it does

mediate confined expression, its particular sequence is apparently not. We conclude, therefore, that region 2 functions primarily as a spacer to maintain the distance between region 3, which may contain an upstream enhancer, and the core promoter located in region 1.

Region 2 exhibits characteristics of transposable elements

Region 2 is not found in the upstream flanking sequence of the *GLDT* gene of F. robusta (C_3), but in those of the C_3-C_4 intermediate F. ramosissima and the C_4 species F. trinervia. Since this region is mandatory to maintain a "C_4-like" expression pattern, but also may contain enhancing activity on its own, the question towards the evolutionary origin of this region arose. Since transposable elements are known to be drivers of evolutionary changes and may confer novel expression characteristics on nearby genes (Feschotte, 2008; Kejnovsky *et al.*, 2012) we searched region 2 for hints of transposable activity.

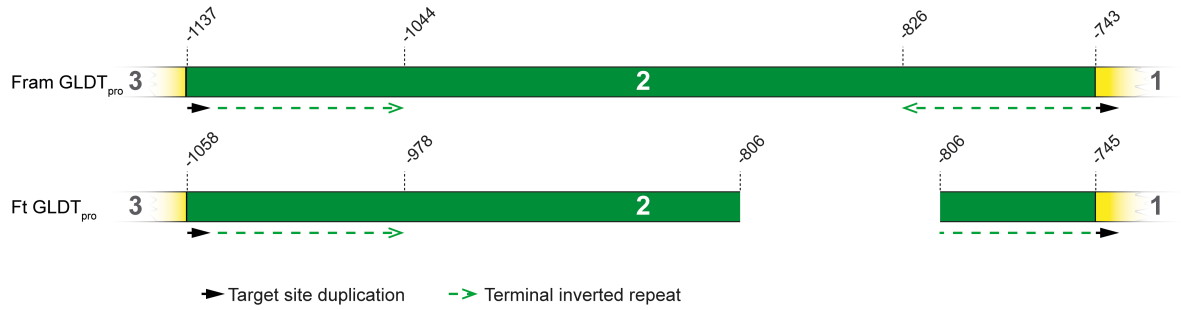


Figure 6 Structural analysis and comparison of region 2. Region 2 *of F. ramosissima* and *F. trinervia GLDT* upstream flanking sequence exhibits characteristics of a DNA transposon. (A) Schematic representation of region 2 from *F. ramosissima* and *F. trinervia*. Black arrows indicate target site and its duplication. Dashed green lines indicate putative terminal inverted repeats.

Sequence dot plotting indicated potential terminal inverted repeat (TIR) of approximately 80 bp, albeit only with a sequence similarity about 65% in the GLDT upstream flanking sequences of *F. ramosissima* and *F. trinervia*. (Fig. 6) Upstream, the potential TIRs are followed by an almost perfect 10 bp direct repeat of the 5'-end of region 1, representing a putative target site duplication (TSD) (Fig. 6).

Additional analysis of available RNA-seq data indicated that region 2-like elements are also present several times in transcribed genes of *F. robusta* (Supplemental data S1). The size, structure and lack of open reading frames are characteristic for miniature inverted-repeat transposable elements (MITEs) (Feschotte *et al.*, 2002).

Dissection of region 2 reveals a 59 bp fragment with enhancing activity

To further delimit potential *cis*-regulatory elements in region 2, it was divided into four subregions overlapping by 20 bp and termed 2.1 to 2.4 in 3'->5' direction (Fig. 7A). Since construct *Ft*-GLDT$_{PRO}$21 exhibited expression close to the visual and measurable detection limit (Figure 3B), dissection of region 2 was conducted in the context of construct *Ft*-GLDT$_{PRO}$32ID1 (Figure 4), i.e. in presence of regions 3 and 1 but absence of the leader intron. This allowed a more sensitive range for measuring GUS activity than construct *Ft*-GLDT$_{PRO}$321, where potential deactivation of region 2 by replacing it with its reverse complement lowered GUS expression, but not significantly (Figure 5B).

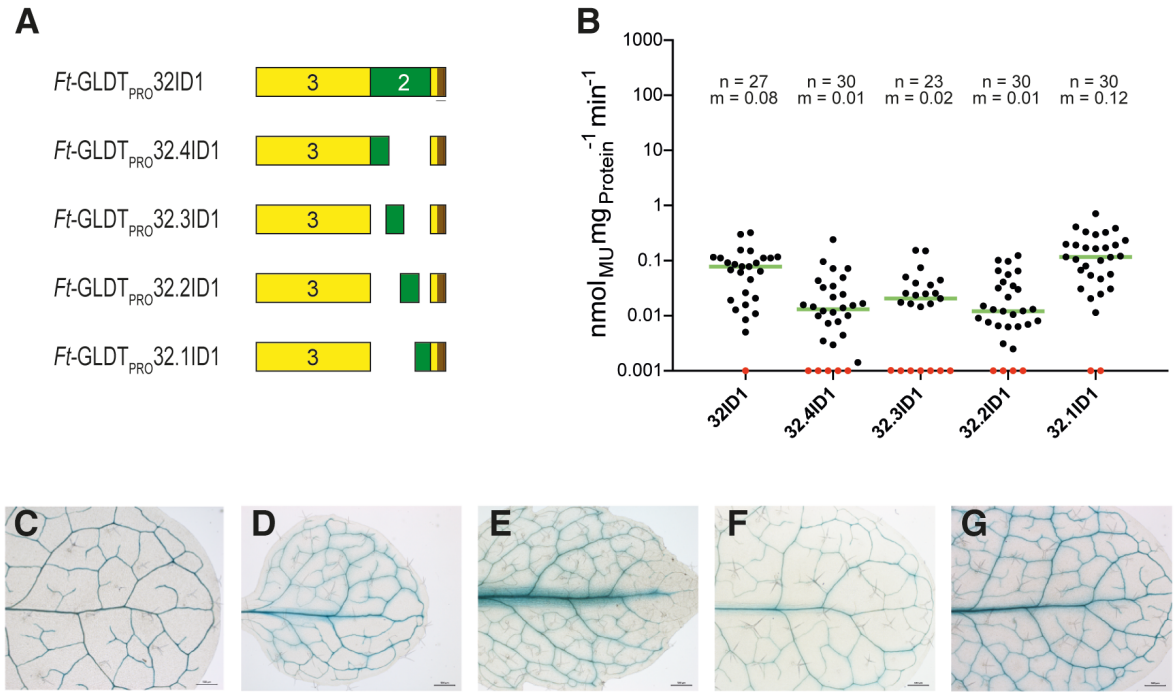


Figure 7 Dissection of region 2 from *F. trinervia*. (A) Schematic representation of reference (*Ft*-GLDT$_{PRO}$32ID1) and region 2 dissection constructs. Region 2 was divided into 4 subregions, overlapping by 20 bp and termed 2.1 to 2.4 in 3'-> 5' direction. **(B)** Fluorometric quantification of GUS activity. Green bars correspond to median values. Red dots indicate measurements below detection limit. Quantity (n) and median (m) are depicted above the corresponding scatter plot. **(C –G)** GUS expression in leaves of transgenic *A. thaliana* of constructs *Ft*-GLDT$_{PRO}$32ID1, *Ft*-GLDT$_{PRO}$32.4ID1, *Ft*-GLDT$_{PRO}$32.3ID1, *Ft*-GLDT$_{PRO}$32.2ID1 and *Ft*-GLDT$_{PRO}$32.1ID1, respectively. Note that the data of construct *Ft*-GLDT$_{PRO}$32ID1 are the same as in Fig. 4 and are included for easier comparison.

None of the subregions mediated a change in GUS localisation (Fig. 7C – G). Again confirming the presence of redundant CREs, most likely in region 3. However, while subregions 2.4 to 2.2 mediated significantly lower GUS activity than the reference construct, it significantly increased when only subregion 2.1 was present (Fig. 7B). This did not only show the existence of enhancing elements in this subregion, but also excluded the other subregions. Further, this supported the hypothesis that the distance between region 3 and 1 affects transcriptional strength, since construct *Ft*-$GLDT_{PRO}$32.1ID1 exhibited 50 % higher GUS activity than *Ft*-$GLDT_{PRO}$32ID1. The non-overlapping part of subregion 2.1 consists of 59 bp and ranges from -804 bp to -745 bp, in respect to the translational start site.

Discussion

GDC is a central player in photorespiration (Bauwe, 2011; Bauwe and Kolukisaoglu, 2003; Sage, 2004). Its restriction to the bundle sheath is considered an early and important step in evolution of C_4 photosynthesis (Bauwe, 2011; Rawsthorne, 1992; Sage *et al.*, 2012). But how this restriction evolved has only been highlighted for the P subunit (GLDP) of *Flaveria* (Engelmann *et al.*, 2008; Schulze *et al.*, 2013; Wiludda *et al.*, 2012). Here we wanted to shed first light on how restriction of GLDT to BSC was accomplished.

In contrast to all other subunits, GLDT of *Flaveria* is comprised of a single gene locus, as *de novo* RNA-seq assemblies indicate (Mallmann *et al.*, 2014), coinciding with all Asterid genome sequences published yet (Phytozome v12.1). Fusion of the upstream flanking sequence of *F. robusta* (C_3), *F. ramosissima* (C_3-C_4) and *F. trinervia* (C_4) to a GUS reporter gene showed accumulation of GUS in BSC for the C_3-C_4 and C_4 species, while for the C_3 species, GUS accumulated in BSC and M of stably transformed *F. bidentis* (C_4; Fig. 1E – G). Comparable localisation was found when transformed into the C_3 plant *A. thaliana* (Fig. 1B – D). This indicates that the regulatory mechanism locating GLDT expression to BSC of *Flaveria* is already present and fulfils at least a comparable function in C_3 plants. However, while hand cross-sections of transformed *F. bidentis* showed a strict GUS localisation to the BSC cells (Fig. 1E – F), microtome sections and confocal microscopy of transformed *A. thaliana* additionally exhibited high activity in vascular cells (Fig. 2). We therefore

assume that either the expression profile of a corresponding transcription factor is not completely conserved in both species or that additional factors mediating vascular suppression or post-transcriptional regulation are present in the *GLDT* upstream flanking sequence of *F. trinervia* but are not functional in *A. thaliana*.

It is assumed that restriction of GLDP to BSC was sufficient to generate a first carbon concentrating mechanism (CCM) (Bauwe, 2011; Morgan *et al.*, 1993), while restriction of the other subunits likely constituted subsequent optimisation steps. Recently it was shown, that *GLDPA* of the basal C_3 species *F. pringlei* also localises to BSC, indicating that all *Flaveria* species harbour one BSC restricted GLDP, while a second gene (*GLDPB*) is also expressed in MC, but was gradually inactivated from C_3 over C_3-C_4 to C_4 (Schulze *et al.*, 2013), coinciding with immunolocalisation studies in other C_3-C_4 and C_3 *Flaveria* species (Sage *et al.*, 2013). In contrast, GLDT of the C_3-C_4 species *F. ramosissima* appears to be highly restricted to BSC, at least in transgenic *F. bidentis* (Fig. 1G). Hypothetically, leaving *F. ramosissima* with still relatively high amounts of GLDPB in MC, while GLDT expression in MC is below visual GUS detection level. One possible explanation could be that additional changes in *trans* occurred during transition from C_3-C_4 to C_4 and the actual expression of GLDT in *F. ramosissima* might be better represented by GUS expression in the C_3 background of *A. thaliana,* where confinement appeared less stringent than in C_4 *F. bidentis* (cmp. Fig 1C and 1F). On the other hand, although GLDP catalyses the initial step and actual decarboxylation of glycine, restriction of any other subunit to BSC would also lead to a moderate accumulation of glycine in MC (Bauwe, 2011; Engel *et al.*, 2007; Timm *et al.*, 2017; Zhou *et al.*, 2013) that would have to diffuse to the BSC for decarboxylation. This suggests that, at least in *F. ramosissima,* GLDT could have preceded GLDP localisation to BSC and thus be the evolutionary driving factor for a CO_2 pump.

The genus *Flaveria* is comprised of C_3, C_3-C_4, C_4-like and C_4 species. All phylogenetic reconstructions place C_3 species, such as *F. robusta*, to the basis, followed by first C_3-C_4 species, which split up into two distinct clades A and B. Clade A harbours all C_4 species including *F. trinervia*, C_4-like and C_3-C_4 species as *F. ramosissima* (Lyu *et al.*, 2015; McKown *et al.*, 2005). Sequence comparison of *GLDT* upstream flanking sequence indicate, that region 2 is exclusive to clade A (*F. ramosissima, F. trinervia*), since basal C_3 species (*F. robusta*) and C_3-C_4 species of

clade B (*F. anomala*) lack region 2. While the upstream flanking sequence of *F. robusta* mediates transcriptional activity throughout the leave, the activity of *F. anomala* has not been analysed, but we assume it might show a preference for leaf vascular bundles corresponding to decreased MC GLDPB levels (Schulze *et al.*, 2013) and maybe similar to the expression pattern of *F. trinervia* fusion construct of regions 3 and 1 (Fig. 5C). In addition, we provide evidence that region 2 is an ancient copy of a MITE, since it is present several times in *F. robusta* transcriptome, showing characteristic structural features (Fig. 6) and even similarity to an interspersed element in *H. annuus* (Supplemental Data S1 and S2). Given the vast amount of differentially expressed genes between closely related C_3 and C_4 species (Gowik *et al.*, 2011) and the potential of transposable elements (TEs) to convey such (Feschotte, 2008; Rebollo *et al.*, 2012), it is very likely that TEs contributed significantly to the evolution of C_4, particularly, when their potential to be activated by environmental stress, i.e. selective pressure, is considered (Capy *et al.*, 2000; McClintock, 1984; Wessler, 1996). In this context, we assume, that future genome sequencing projects of young and closely related C_3 and C_4 species, like *Flaveria*, will significantly illuminate, how and to which extent TEs contributed to the evolution of C_4.

However, despite its apparent effect on *GLDT* expression, a molecular model for its action is yet hard to infer. Dissection of region 2 showed that an enhancing CRE resides in the 59 bp of the 3'-end (Fig. 7) and insertion of region 2 occurred closely to the TSS, 15 bp upstream of a putative TATA-box. This suggests that insertion might have disrupted the core promoter, supported by the result that region 1 alone does not convey any detectable expression (Fig. 3E). Regions 2 and 1 together mediate weak but confined expression (Fig. 3D), suggesting that insertion of region 2 either restored a core promoter that already drove expression confined to the vascular bundle including its sheath (in this case putatively directed by elements in 5'-UTR or leader intron) or added *cis*-elements mediating the observed expression pattern. On the other hand, in the presence of region 3, either region 1 or 2 were substitutable (Fig. 4 and 5, respectively), indicating that region 3 harbours *cis*-elements that are functionally redundant to regions 2 and/or 1, in terms of spatial confinement. Speculatively, this element could have already been present in the last common ancestor (LCA) of clade A and B, allowing intermediate species of the latter to

express *GLDT* preferentially in BSC, coinciding with decreased glycine decarboxylation capacity in MC (Schulze *et al.*, 2013; Schulze *et al.*, 2016).
Nevertheless, region 3 of *F. trinervia* also conveyed expression in other leaf tissues, when fused directly to region 1 (Fig. 5C; which is reconstitution of the hypothetical LCA state, albeit separated by a restriction site), indicating that region 3 also harbours a general enhancer mediating expression throughout, at least, the chlorenchyma tissues of the leaf. A possible explanation for conservation of a general enhancer, although its effect on constitutive activity in the native promoter was below visual detectability, might be the low but ubiquitous demand for GLDT in C_1 metabolism (Hanson and Roje, 2001; Timm *et al.*, 2017). Since the effect of the general enhancer was not visual in presence of region 2 or likely any sequence of comparable length separating regions 3 and 1 (Fig. 5E), we assume region 2 also acts in a sequence independent manner, by spatially separating regions 3 and 1, thus hampering interaction of *trans*-factors bound in region 3 and transcription initiation around the TSS. Intriguingly, despite apparent expression in other leaf tissues, quantitative GUS analysis of this construct did not show significantly increased activity (cmp. Fig. 3B and 5B). We can only speculate, that GUS expression in leaf vascular bundles was decreased by a comparable amount as it was enhanced in other leaf tissues.
Nonetheless, these results also exemplified the importance of considering positional effects in promoter studies, as a gain of expression upon sequence deletion is never a sufficient proof for the concomitant identification of negative *cis*-regulatory elements.

Conclusion

Here we presented first insights on how *GLDT* expression was shaped on the evolutionary trajectory from C_3 to C_4 in the genus *Flaveria*. We showed that the underlying mechanism for GLDT confinement to leaf vascular bundles acts on transcriptional level and is also functional in distantly related C_3 species. In case of *GLDT*, it relies on at least two independent *cis*-elements, of which one is exclusive to the *Flaveria* clade harbouring the only C_4 species of this genus and was possibly acquired by insertion of a MITE. We gave evidence that this insertion has also a

sequence independent function and proposed that it acts by spatially separating upstream enhancers and transcriptional start site. Nonetheless, dissection of this region indicated that a transcriptional enhancer resides at its 3'-end and effectively narrowed down its position to a 59 bp area. Future experiments will be conducted to fine-map this enhancer and to analyse if it is conserved within the MITE family or was acquired after its insertion into the *GLDT* upstream flanking sequence of *Flaveria*.

Material and Methods

Plant transformation

Flaveria bidentis was transformed as described by Chitty *et al.* (1994). *Arabidopsis thaliana* was transformer by floral dip according to Clough and Bent (1998) and Logemann *et al.* (2006). Both were transformed using the *Agrobacterium tumefaciens* strain AGL1 (Lazo *et al.*, 1991). Insertion of T-DNA into the host genome was confirmed by PCR.

Genome walking and 5'-RACE

Genomic DNA of *Flaveria robusta*, *Flaveria ramosissima* was isolated using DNeasy Plant Mini Kit (QIAGEN). Genome walking was carried out by using Universal Genome Walker Kit (Clontech). Two gene specific primers of *GLDT* coding region were designed, for each species, based on *de novo* assembled contigs from 454 sequencing (Bräutigam *et al.*, 2011). Nested PCR was carried out using adaptor primers and gene specific primers FraGSP1, FraGSP2 and FroGSP1, FroGSP2 for *F. ramosissima* and *F. robusta,* respectively (Table S1). For both species two subsequent genome walking steps were carried out using adapter primers and gene specific primers designed to hybridize with the obtained fragments (GSP3, GSP4 and GSP5, GSP6 for the corresponding species; Table S1). Subsequently obtained fragments were 887, 1703 and 983 bp for *F. ramosissima* and 909, 2154 and 885 bp for *F. robusta.*

For amplification of cDNA ends, total RNA of *F. robusta, F. ramosissima* and *F. trineriva* was isolated using RNeasy Plant Mini Kit (QIAGEN). Preparation of cDNA libraries and subsequent 5'-RACE was conducted by using SMARTer RACE cDNA Amplification Kit (Clontech) with gene specific primers FraGSP1, FroGSP1 and

FtGSP1-FW (Table S1) for *F. ramosissima, F. robusta* and *F.trinervia* cDNA, respectively.

Cloning of reporter gene constructs

All PCR products were purified by gel extraction (QIAquick Gel Extraction Kit, QIAGEN), cloned into cloning vector pJET1.2 (CloneJET PCR Cloning Kit, Fermentas/Thermo Fisher Scientific) and confirmed by sequencing prior to digestion and ligation with the expression vector. All oligonucleotide sequences mentioned in the following are listed in supplementary table S1.

F. robusta and *F. ramosissima* upstream flanking sequence were isolated from genomic DNA by nested PCR using primers pairs FroGSP1/FroGSP7 and FroGLDT-FW:SgsI/FroGLDT-RV:XmaI for *F. robusta* and FraGSP1/FraGSP7 and FraGLDT-FW:SfaAI/FraGLDT-RV:SgsI for *F. ramosissima* upstream flanking sequence. Primers were designed on basis of fragments isolated from genome walking. Both upstream flanking sequences were cloned, using the PCR attached restriction sites, into an in-house version of pBI121 expression vector (Jefferson *et al.*, 1987), harbouring the *Hind*III/*Xma*I cloned custom multiple cloning site GCGATCGCGGCCGGCCGGCGCGCC instead of the *Cauliflower Mosaic Virus* (CaMV) 35S promoter. The resulting constructs were termed *Fram*-GLDT$_{PRO}$ and *Frob*-GLDT$_{PRO}$.

F. trinervia upstream flanking sequence was isolated from genomic DNA by nested PCR using primer pairs FtGSP-FW/FtGSP-RV and FtGLDT-FW:SalI/FtGLDT-RV:XmaI on basis of the publicly available sequence (accession no. Z99769). The isolated fragment was cloned by *Sal*I/*Xma*I digestion into construct *GLDPA*-FtΔ6 described in Engelmann *et al.* (2008) and termed *Ft*-GLDT$_{PRO}$. H2B:YFP construct was generated by PCR amplification of H2B:YFP (Boisnard-Lorig *et al.*, 2001) from construct GLDPA-Ft::H2B:YFP described in Engelmann *et al.* (2008), adding 5'-*Xma*I and *Sac*I-3' with primers H2B:YFP-FW:XmaI and H2B:YFP-RV:SacI. The product was cloned by *Xma*I/*Sac*I digestion into construct *Ft*-GLDT$_{PRO}$, replacing the *uidA* reporter gene.

Consecutive deletion constructs of *F. trinervia* upstream flanking sequence were generated by PCR amplification of the corresponding fragment from *Ft*-GLDT$_{PRO}$, attaching an *SfaA*I restriction site to the 5'-end. Primer combinations were FtGLDTR3-FW:SfaAI/FtGLDT-RV-XmaI, FtGLDTR2-FW:SfaAI/FtGLDT-RV-XmaI

and FtGLDTR1-FW:SfaAI/FtGLDT-RV-XmaI for constructs *Ft*-GLDTP$_{RO}$321, *Ft*-GLDT$_{PRO}$21 and *Ft*-GLDT$_{PRO}$1, respectively. Each fragment was cloned by *SfaA*I/*Xma*I digestion into the modified pBI121.

Construct *Ft*-GLDT$_{PRO}$31 was generated by PCR amplification of region 3 and 1 with primer pairs FtGLDTR3-FW:SfaAI/FtGLDTR3-RV:XhoI and FtGLDTR1-FW:XhoI/ FtGLDT-RV:XmaI, respectively. Both fragments were digested with corresponding restriction enzymes and assembled in *SfaA*I/*Xma*I digested modified pBi121.

Constructs *Ft*-GLDT$_{PRO}$32rc1 and *Ft*-GLDT$_{PRO}$32sub1 were generated by PCR amplification of region 2 and a size corresponding fragment of H2B:YFP with primer pairs FtGLDTR2-FW:XhoI/ FtGLDTR2-RV:XhoI and YFP313-FW:XhoI/ YFP313-RV:XhoI, respectively. Fragments were digested by *Xho*I and ligated with *Xho*I digested construct *Ft*-GLDT$_{PRO}$31. Resulting clones were selected by PCR for reverse orientation of the corresponding fragment.

Construct *Ft*-GLDT$_{PRO}$32ID1 was generated by PCR amplification of the fragment from region 3 to the 5'-splice site of the first intron (-1654 to -664 bp) from construct *Ft*-GLDT$_{pro}$321 with primers FtGLDTR3-FW:SfaAI and FtGLDTR1ID-RV:XmaI, attaching residual nucleotides of the 5'-UTR (-11 to -1 bp) and an *Xma*I restriction site. The resulting fragment was cloned into *Ft*-GLDT$_{PRO}$321 under digestion with *SfaA*I and *Xma*I. For construction of *Ft*-GLDT$_{PRO}$32-60, regions 3 and 2 were amplified by PCR in one Fragment using primers FtGLDTR3-FW and FtGLDTR2-RV from construct *Ft*-GLDT$_{PRO}$321. The 35S minimal promoter (-60 to -1 bp) of CaMV was amplified from pBi121 binary vector (Jefferson *et al.*, 1987) using primers 35S-60-FW and 35S-60-RV. The PCR product was phosphorylated and ligated with regions 3 and 2 amplicon. A subsequent PCR on the ligation product was conducted using primers FtGLDTR3-FW:SfaAI and 35S-60-RV:XmaI. The resulting product was cloned under *SfaA*I/*Xma*I digestion into *Ft*-GLDT$_{PRO}$321.

Constructs *Ft*-GLDT$_{PRO}$32.4ID1 to *Ft*-GLDT$_{PRO}$32.1ID1 were synthesised and assembled in pUC57 cloning vector with flanking 5'-*SfaA*I and 3'-*Xma*I restrictions sites by GenScript (USA) and cloned accordingly into the modified pBI121 expression vector.

In situ detection of GUS activity and fluorometric measurement

For transgenic *F. bidentis*, hand cross-sections of the fifth leaf from the top of an approximately 40 cm tall plant were prepared and incubated as described in

(Engelmann *et al.*, 2008). For transgenic *A. thaliana*, whole leaves were incubated approximately 3 weeks after germination. Chlorophyll was removed from samples of *F. bidentis* and *A. thaliana* by treatment with 3:1 ethanol/acetic acid. Samples were cleared and mounted in 10:1:3 chloral hydrate/glycerol/water.

Transversal sections of *A. thaliana* were obtain by embedding the incubated leaves, after removal of chlorophyll, in LR White resin (agar scientific) according to Khoshravesh *et al.* (2017). Samples were cut into 5-10 µm thick sections, stained with Safranine O (Sigma-Aldrich) and mounted in Entellan (Sigma-Aldrich).

Quantification of β-glucuronidase (GUS) activity was conducted according to Jefferson *et al.* (1987) and Kosugi *et al.* (1990) by continuous measurement (Fior *et al.*, 2009) using a Synergy HTX Multi-Mode Microplate Reader (BioTek Instruments). Samples were incubated at 37° C and excited at 360 nm in 5 minute intervals. Emission was measured at 460 nm. Fluorescent emission per unit 4-methylumbelliferone (MU) was calculated from a standard row ranging from 0.1 – 10 µM MU. Obtained results were normalised to minutes and mg protein, as determined by Bradford Protein Assay (Bio-Rad) using a bovine serum albumin (BSA) standard. Statistical confidence was calculated using Mann-Whitney test.

Confocal microscopy

Samples for confocal microscopy were prepared according to Kurihara *et al.* (2015). Cell walls were stained with calcofluor white M2R (Sigma-Aldrich). Calcofluor white and H2B:YFP were excited at 405 and 514 nm and their emission was detected at 410 – 490 nm and 517 – 579 nm, respectively.

Consensus reconstruction

To analyse the presence of region-2-*like* elements in transcriptomic space, short read archives from *F. robusta* root and shoot RNA-seq experiments (SRA Accessions SRX794075 and SRX794076, respectively) described in Mallmann *et al.* (2014) were used. Reads were trimmed to a PHRED score >= 10 using Trimmomatic v0.36 (Bolger *et al.*, 2014). Trimmed reads were searched for any similarities to region 2 of *F. ramosissima GLDT* upstream flanking sequence using BLASTN v2.6.0 (Altschul *et al.*, 1990) with parameter [word_size 4, reward 1, penalty 1, gapopen 1, gapextend 2, dust no]. Scoring reads were *de novo* assembled using CAP3 v10.2011 (Huang and Madan, 1999) as provided by Bioconda repository (https://bioconda.github.io) with

parameters [-p 100, -o 35, -n -15]. Assembled contig sequences were mapped on region 2 of *F. ramosissima* upstream flanking sequence using BWA-MEM v0.7.17 (Li, 2013) with parameters [-k 4, -B 2, -O 2, -T 20]. A majority consensus was generated from mapped contigs and used as reference for BWA-MEM mapping, to refine the consensus sequence. For conservative estimation of the abundance of region-2-*like* insertions in transcribed genes, contigs exceeding 5' and 3'-ends of the consensus sequence were clustered with CD-HIT-EST v4.7 (Li and Godzik, 2006) using parameters [-g 1, -c 0.9, -aS 0.9]. Consensus sequences of resulting clusters were generated and mapped again on refined region-2-*like* consensus sequence.

Accession numbers

F. robusta and *F. ramosissima GLDT* upstream flanking sequence were deposited at NCBI GenBank under the accession numbers MG977011 and MG977012.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**, 403-410.

Anderson LE. 1971. Chloroplast and cytoplasmic enzymes II. Pea leaf triose phosphate isomerases. *Biochimica et Biophysica Acta (BBA) - Enzymology* **235**, 237-244.

Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Brière C, Owens GL, Carrère S, Mayjonade B, Legrand L, Gill N, Kane NC, Bowers JE, Hubner S, Bellec A, Bérard A, Bergès H, Blanchet N, Boniface M-C, Brunel D, Catrice O, Chaidir N, Claudel C, Donnadieu C, Faraut T, Fievet G, Helmstetter N, King M, Knapp SJ, Lai Z, Le Paslier M-C, Lippi Y, Lorenzon L, Mandel JR, Marage G, Marchand G, Marquand E, Bret-Mestries E, Morien E, Nambeesan S, Nguyen T, Pegot-Espagnet P, Pouilly N, Raftis F, Sallet E, Schiex T, Thomas J, Vandecasteele C, Varès D, Vear F, Vautrin S, Crespi M, Mangin B, Burke JM, Salse J, Muños S, Vincourt P, Rieseberg LH, Langlade NB. 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148-152.

Bauwe H. 2011. Chapter 6 Photorespiration: The Bridge to C_4 Photosynthesis. In: Raghavendra AS, Sage RF, eds. *C4 Photosynthesis and Related CO2 Concentrating Mechanisms*, Vol. 32: Springer Netherlands, 81-108.

Bauwe H, Kolukisaoglu U. 2003. Genetic manipulation of glycine decarboxylation. *J Exp Bot* **54**, 1523-1535.

Boisnard-Lorig C, Colon-Carmona A, Bauch M, Hodge S, Doerner P, Bancharel E, Dumas C, Haseloff J, Berger F. 2001. Dynamic Analyses of the Expression of the HISTONE::YFP Fusion Protein in Arabidopsis Show That Syncytial Endosperm Is Divided in Mitotic Domains. *Plant Cell* **13**, 495-509.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.
Bräutigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Mass J, Lercher MJ, Westhoff P, Hibberd JM, Weber AP. 2011. An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiol* **155**, 142-156.
Capy P, Gasperi G, Biémont C, Bazin C. 2000. Stress and transposable elements: co-evolution or useful parasites? *Heredity* **85**, 101.
Chitty JA, Furbank RT, Marshall JS, Chen Z, Taylor WC. 1994. Genetic transformation of the C4 plant, Flaveria bidentis. *The Plant Journal* **6**, 949-956.
Clough SJ, Bent AF. 1998. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J* **16**, 735-743.
Edwards GE, Ku MSB. 1987. Biochemistry of C3–C4 Intermediates. *Photosynthesis*: Academic Press, 275-325.
Ehleringer JR, Monson RK. 1993. Evolutionary and Ecological Aspects of Photosynthetic Pathway Variation. *Annual Review of Ecology and Systematics* **24**, 411-439.
Engel N, van den Daele K, Kolukisaoglu Ü, Morgenthal K, Weckwerth W, Pärnik T, Keerberg O, Bauwe H. 2007. Deletion of Glycine Decarboxylase in Arabidopsis Is Lethal under Nonphotorespiratory Conditions. *Plant Physiology* **144**, 1328-1335.
Engelmann S, Wiludda C, Burscheidt J, Gowik U, Schlue U, Koczor M, Streubel M, Cossu R, Bauwe H, Westhoff P. 2008. The gene for the P-subunit of glycine decarboxylase from the C4 species Flaveria trinervia: analysis of transcriptional control in transgenic Flaveria bidentis (C4) and Arabidopsis (C3). *Plant Physiol* **146**, 1773-1785.
Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397-405.
Feschotte C, Zhang X, Wessler SR. 2002. Miniature Inverted-Repeat Transposable Elements and Their Relationship to Established DNA Transposons. *Mobile DNA II*: American Society of Microbiology.
Fior S, Vianelli A, Gerola PD. 2009. A novel method for fluorometric continuous measurement of β-glucuronidase (GUS) activity using 4-methyl-umbelliferyl-β-d-glucuronide (MUG) as substrate. *Plant Science* **176**, 130-135.
Gowik U, Bräutigam A, Weber KL, Weber AP, Westhoff P. 2011. Evolution of c4 photosynthesis in the genus flaveria: how many and which genes does it take to make c4? *Plant Cell* **23**, 2087-2105.
Hanson AD, Roje S. 2001. One-Carbon Metabolism in Higher Plants. *Annu Rev Plant Physiol Plant Mol Biol* **52**, 119-137.
Holaday AS, Lee KW, Chollet R. 1984. C3–C4 Intermediate species in the genus Flaveria: leaf anatomy, ultrastructure, and the effect of O2 on the CO2 compensation concentration. *Planta* **160**, 25-32.
Huang X, Madan A. 1999. CAP3: A DNA Sequence Assembly Program. *Genome Res* **9**, 868-877.
Jefferson R, Kavanagh T, Bevan M. 1987. GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO Journal* **6**, 3901 - 3907.
Keerberg O, Parnik T, Ivanova H, Bassuner B, Bauwe H. 2014. C2 photosynthesis generates about 3-fold elevated leaf CO2 levels in the C3-C4 intermediate species Flaveria pubescens. *J Exp Bot* **65**, 3649-3656.
Kejnovsky E, Hawkins JS, Feschotte C. 2012. Plant Transposable Elements: Biology and Evolution. In: Wendel JF, Greilhuber J, Dolezel J, Leitch IJ, eds. *Plant Genome Diversity*

Volume 1: Plant Genomes, their Residents, and their Evolutionary Dynamics. Vienna: Springer Vienna, 17-34.
Kelly GJ, Latzko E. 1977. Chloroplast Phosphofructokinase: II. Partial Purification, Kinetic and Regulatory Properties. *Plant Physiol* **60**, 295-299.
Khoshravesh R, Lundsgaard-Nielsen V, Sultmanis S, Sage TL. 2017. Light Microscopy, Transmission Electron Microscopy, and Immunohistochemistry Protocols for Studying Photorespiration. In: Fernie AR, Bauwe H, Weber APM, eds. *Photorespiration: Methods and Protocols*. New York, NY: Springer New York, 243-270.
Kim YJ, Lee SH, Park KY. 2004. A leader intron and 115-bp promoter region necessary for expression of the carnation S-adenosylmethionine decarboxylase gene in the pollen of transgenic tobacco. *FEBS Lett* **578**, 229-235.
Kosugi S, Ohashi Y, Nakajima K, Arai Y. 1990. An improved assay for β-glucuronidase in transformed cells: Methanol almost completely suppresses a putative endogenous β-glucuronidase activity. *Plant Science* **70**, 133-140.
Ku MSB, Monson RK, Littlejohn RO, Nakamoto H, Fisher DB, Edwards GE. 1983. Photosynthetic Characteristics of C(3)-C(4) Intermediate Flaveria Species : I. Leaf Anatomy, Photosynthetic Responses to O(2) and CO(2), and Activities of Key Enzymes in the C(3) and C(4) Pathways. *Plant Physiology* **71**, 944-948.
Ku MSB, Wu J, Dai Z, Scott RA, Chu C, Edwards GE. 1991. Photosynthetic and Photorespiratory Characteristics of Flaveria Species. *Plant Physiology* **96**, 518.
Kurihara D, Mizuta Y, Sato Y, Higashiyama T. 2015. ClearSee: a rapid optical clearing reagent for whole-plant fluorescence imaging. *Development* **142**, 4168-4179.
Lazo GR, Stein PA, Ludwig RA. 1991. A DNA transformation-competent Arabidopsis genomic library in Agrobacterium. *Biotechnology (N Y)* **9**, 963-967.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN].
Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659.
Logemann E, Birkenbihl RP, Ulker B, Somssich IE. 2006. An improved method for preparing Agrobacterium cells that simplifies the Arabidopsis transformation protocol. *Plant Methods* **2**, 16.
Lyu M-J, Gowik U, Kelly S, Covshoff S, Mallmann J, Westhoff P, Hibberd J, Stata M, Sage R, Lu H, Wei X, Wong G, Zhu X-G. 2015. RNA-Seq based phylogeny recapitulates previous phylogeny of the genus Flaveria (Asteraceae) with some modifications. *BMC Evol Biol* **15**, 116.
Magallón S, Castillo A. 2009. Angiosperm diversification through time. *Am J Bot* **96**, 349-365.
Mallmann J, Heckmann D, Bräutigam A, Lercher MJ, Weber AP, Westhoff P, Gowik U. 2014. The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. *eLife* **3**, e02478.
McClintock B. 1984. The significance of responses of the genome to challenge. *Science* **226**, 792-801.
McKown AD, Moncalvo JM, Dengler NG. 2005. Phylogeny of Flaveria (Asteraceae) and inference of C4 photosynthesis evolution. *Am J Bot* **92**, 1911-1928.
Morgan CL, Turner SR, Rawsthorne S. 1993. Coordination of the cell-specific distribution of the four subunits of glycine decarboxylase and of serine hydroxymethyltransferase in leaves of C3-C4 intermediate species from different genera. *Planta* **190**, 468-473.
Patel M, Corey AC, Yin LP, Ali S, Taylor WC, Berry JO. 2004. Untranslated regions from C4 amaranth AhRbcS1 mRNAs confer translational enhancement and preferential

bundle sheath cell expression in transgenic C4 Flaveria bidentis. *Plant Physiol* **136**, 3550-3561.
Patel M, Siegel AJ, Berry JO. 2006. Untranslated Regions of FbRbcS1 mRNA Mediate Bundle Sheath Cell-specific Gene Expression in Leaves of a C4 Plant. *Journal of Biological Chemistry* **281**, 25485-25491.
Rawsthorne S. 1992. C3–C4 intermediate photosynthesis: linking physiology to gene expression. *The Plant Journal* **2**, 267-274.
Rawsthorne S, Hylton CM, Smith AM, Woolhouse HW. 1988a. Distribution of photorespiratory enzymes between bundle-sheath and mesophyll cells in leaves of the C3–C4 intermediate species Moricandia arvensis (L.) DC. *Planta* **176**, 527-532.
Rawsthorne S, Hylton CM, Smith AM, Woolhouse HW. 1988b. Photorespiratory metabolism and immunogold localization of photorespiratory enzymes in leaves of C3 and C3-C4 intermediate species of Moricandia. *Planta* **173**, 298-308.
Rebollo R, Romanish MT, Mager DL. 2012. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annu Rev Genet* **46**, 21-42.
Sage RF. 2004. The evolution of C4 photosynthesis. *New Phytologist* **161**, 341-370.
Sage RF. 2016. A portrait of the C4 photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineages, and Hall of Fame. *J Exp Bot* **67**, 4039-4056.
Sage RF, Khoshravesh R, Sage TL. 2014. From proto-Kranz to C4 Kranz: building the bridge to C4 photosynthesis. *Journal of Experimental Botany* **65**, 3341-3356.
Sage RF, Sage TL, Kocacinar F. 2012. Photorespiration and the evolution of C4 photosynthesis. *Annu Rev Plant Biol* **63**, 19-47.
Sage TL, Busch FA, Johnson DC, Friesen PC, Stinson CR, Stata M, Sultmanis S, Rahman BA, Rawsthorne S, Sage RF. 2013. Initial Events during the Evolution of C4 Photosynthesis in C3 Species of Flaveria. *Plant Physiology* **163**, 1266-1276.
Schulze S, Mallmann J, Burscheidt J, Koczor M, Streubel M, Bauwe H, Gowik U, Westhoff P. 2013. Evolution of C4 photosynthesis in the genus flaveria: establishment of a photorespiratory CO2 pump. *Plant Cell* **25**, 2522-2535.
Schulze S, Westhoff P, Gowik U. 2016. Glycine decarboxylase in C3, C4 and C3-C4 intermediate species. *Curr Opin Plant Biol* **31**, 29-35.
Timm S, Giese J, Engel N, Wittmiß M, Florian A, Fernie AR, Bauwe H. 2017. T-protein is present in large excess over the other proteins of the glycine cleavage system in leaves of Arabidopsis. *Planta*.
Vogan PJ, Frohlich MW, Sage RF. 2007. The functional significance of C3–C4 intermediate traits in Heliotropium L. (Boraginaceae): gas exchange perspectives. *Plant, Cell & Environment* **30**, 1337-1345.
Wessler SR. 1996. Plant retrotransposons: Turned on by stress. *Current Biology* **6**, 959-961.
Wiludda C, Schulze S, Gowik U, Engelmann S, Koczor M, Streubel M, Bauwe H, Westhoff P. 2012. Regulation of the photorespiratory GLDPA gene in C(4) flaveria: an intricate interplay of transcriptional and posttranscriptional processes. *Plant Cell* **24**, 137-151.
Zhou Q, Yu Q, Wang Z, Pan Y, Lv W, Zhu L, Chen R, He G. 2013. Knockdown of GDCH gene reveals reactive oxygen species-induced leaf senescence in rice. *Plant, Cell & Environment* **36**, 1476-1489.

Author contribution

Jan Emmerling wrote the manuscript, designed and conducted the experiments with the exception of the transformation of *Flaveria bidentis*, which was carried out by Monika Streubel and Maria Koczor.

Supplemental Information

Supplemental Table S1. Oligonucleotides used in this study.

Supplemental Figure S1. GUS localisation of construct *Ft*-GLDT$_{PRO}$21 in transgenic *F. bidentis*.

Supplemental Data S1. Reconstruction of region 2-like consensus sequence.

Supplemental Data S2. Sequence alignment of region 2-like elements *from Flaveria robusta*, in FASTA format. (Enclosed CD only)

Supplemental Data S3. Sequence alignment of region 2-like elements from *Helianthus annuus*. (Enclosed CD only)

Table S1 Oligonucleotides used in this study. Underscores highlight the attached restriction sites indicated by the identifier.

FraGSP1	GCCGGTACCAGGAGCAAGTCCAGCCACA
FraGSP2	CCAACCAGCAAAAGGTACCATCTTTCCACC
FraGSP3	AAGCCTCAACGAAATGAGATAAGAAACCCC
FraGSP4	GGGTTTAGCCCCACATGAACAAACACTC
FraGSP5	TTGTACGGGGTAATGGAATGAACAAAGG
FraGSP6	GAATGGATGTGAGAATAGAATGGATG
FraGSP7	GCTTGTTGCAGCAATGTTTGATGTGC
FroGSP1	AGCAACCACAAGCTTCTCAAGAAACGGGA
FroGSP2	TGGATTGGCATGCTCCAACCAGCAAAAGGT
FroGSP3	GCCTCAAGGAAATGAGAAACGAATACCC
FroGSP4	AATTCACACCAAACGCAGGTTAAATGGC
FroGSP5	ATGAGAGAGAGAATCCATCATCACTGCC
FroGSP6	CAATGAACTTAGAGCAATATCCACCAAC
FroGSP7	TTTGTGTGTGTAGGTTTATGATGAGG
FraGLDT-FW:SfaAI	*CCTGCGATCGC*CCTTAAGCTAGCGTAAAA
FraGLDT-RV:SgsI	*CAAGGCGCGCC*TGTGCTTTATTCTTTAGAAAC
FroGLDT-FW:SgsI	*GCCGGCGCGCC*TATTAAATTCTTGATAAACAT
FroGLDT-RV:XmaI	*GTGCCCGGG*TGTGCTTTATACTTCAAAAA
FtGSP-FW	GTCTAGTTCAAGTCTCCCGGACAAC
FtGSP-RV	GCCGGTACCAGGAGCAAGTCCAGCCACA
FtGLDT-FW:SalI	*TATGTCGAC*CCGTAAATAGGTCAAAATAGCAGC
FtGLDT-RV:XmaI	*ATGCCCGGG*TGTGCTTTATTCTTTAGAAACAAGC
H2B:YFP-FW:XmaI	*CCCGGG*ATGGCGAAGGCAGATAAG
H2B:YFP-RV:SacI	*GAGCTC*TTAGTGGTGGTGGTGG
FtGLDTR3-FW:SfaAI	*GCGATCGC*ATTGATGTAGGTTTATGG
FtGLDTR3-RV:XhoI	*TATCTCGAG*AATATTTTTCCTGTAAAGT
FtGLDTR2-FW:SfaAI	*GCGATCGC*CACCTACACAGGAATGTTCT
FtGLDTR1-FW:SfaAI	*GCGATCGC*CACCCAGATATGTACAAATT
FtGLDTR1-FW:XhoI	*TATCTCGAG*CACCCAGATATGTACAAATT
FtGLDTR2-FW:XhoI	*CTCGAG*GACGAGGAATCTTAAAAACAC
FtGLDTR2-RV:XhoI	*CTCGAG*CACCTACACAGGAATGTTC
YFP313-FW:XhoI	*CTCGAG*AAAACTACCTGTTCCATGGCC
YFP313-RV:XhoI	*CTCGAG*TCGGCCATGATGTATACGTTG
FtGLDTR1ID-RV:XmaI	*CCCGGGTGTGCTTTATT*AGGTGACCTTAGAGAGC
FtGLDTR3-FW	ATTGATGTAGGTTTATGG
FtGLDTR2-RV	CACCTACACAGGAATGTTC
35S-60-FW	CCCACTATCCTTCGCAAG
35S-60-RV	TCCTCTCCAAATGAAATGAA
35S-60-RV:XmaI	*CCCGGG*TCCTCTCCAAATGAAATGAA

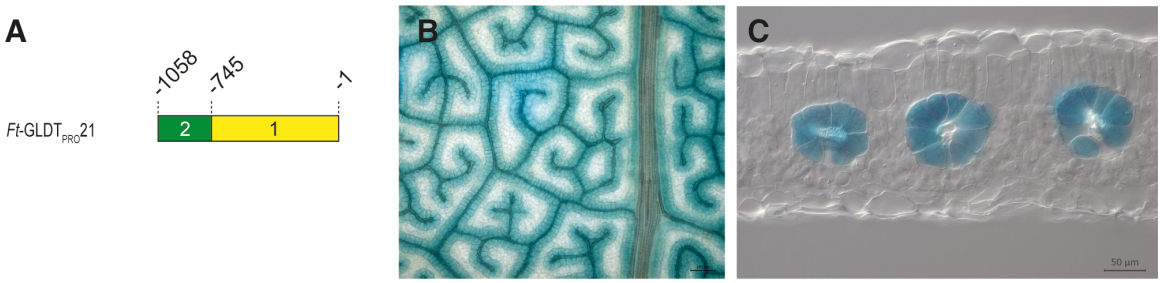


Figure S1 GUS expression of construct *Ft*-GLDT$_{PRO}$21 in transgenic F. bidentis. (A) Schematic representation of construct Ft-GLDT$_{PRO}$21. **(B – C)** Spatial GUS expression of construct Ft-GLDT$_{PRO}$21 in leaves of transgenic *F. bidentis* in top-view and cross section, respectively. Leaf sections were incubated for 16h (B) and 4h (C) and cleared with chloral hydrate. Imaging was conducted using differential interference contrast (DIC) microscopy.

Supplemental Data S1

Reconstruction of region-2-like consensus from transcriptomic data reveals similarity to sunflower transposable element

While there is no genomic data of *Flaveria* available yet, we assumed that region 2, if a transposable element, might also be present in transcriptomic data. However, *de novo* assemblies of available RNA-Seq experiments from *F. robusta* seemed to struggle with the palindromic structure, low complexity and low coverage. Thus, we decided to use the more general-purpose tools BLASTN and CAP3 to first identify 1,783 reads with any similarity to region 2 of *F. ramosissima* (since structural features were less apparent in *F. trinervia*), and assemble those into 151 short *high confidence* contigs, to reduce bias of highly abundant transcripts. Of those contigs, 75 actually mapped back to region 2 to generate a first consensus sequence. This was refined by another round of mapping, resulting in 79 mapped contigs. For a better estimation of abundance: After collapsing contigs with more than 90% sequence similarity we still found 14 and 17 contigs spanning the 5' and 3' insertion site, respectively. Hence we assume there are at least 14 full-length insertions of region-2-like elements in transcribed space of *F. robusta*. The generated region-2-like consensus sequence exhibited TIRs of 83 bp but sequence similarity of TIRs increased from 65% to 84%, in comparison to region 2 from *F. ramosissima*. Additionally, in contrast to region 2 of *F. ramosissima* and *F. trinervia*, we found a preference for insertion into TA-rich regions and distinct G-rich and C-rich patterns at the 5' and 3' termini, respectively (Fig. S2; Supplemental data S2) similar to the *stowaway* MITE (Bureau and Wessler, 1994; Feschotte *et al.*, 2003) but in reverse orientation.

BLAST searches for the reconstructed consensus sequence in available genomes identified a small family of at least 17 interspersed, not annotated elements in *H. annuus* of comparable length and increasing similarity towards the distal ends of region-2-*like* consensus, including the distinct terminal patterns and insertion preference for TA-rich regions (see supplemental data S3 for alignment).

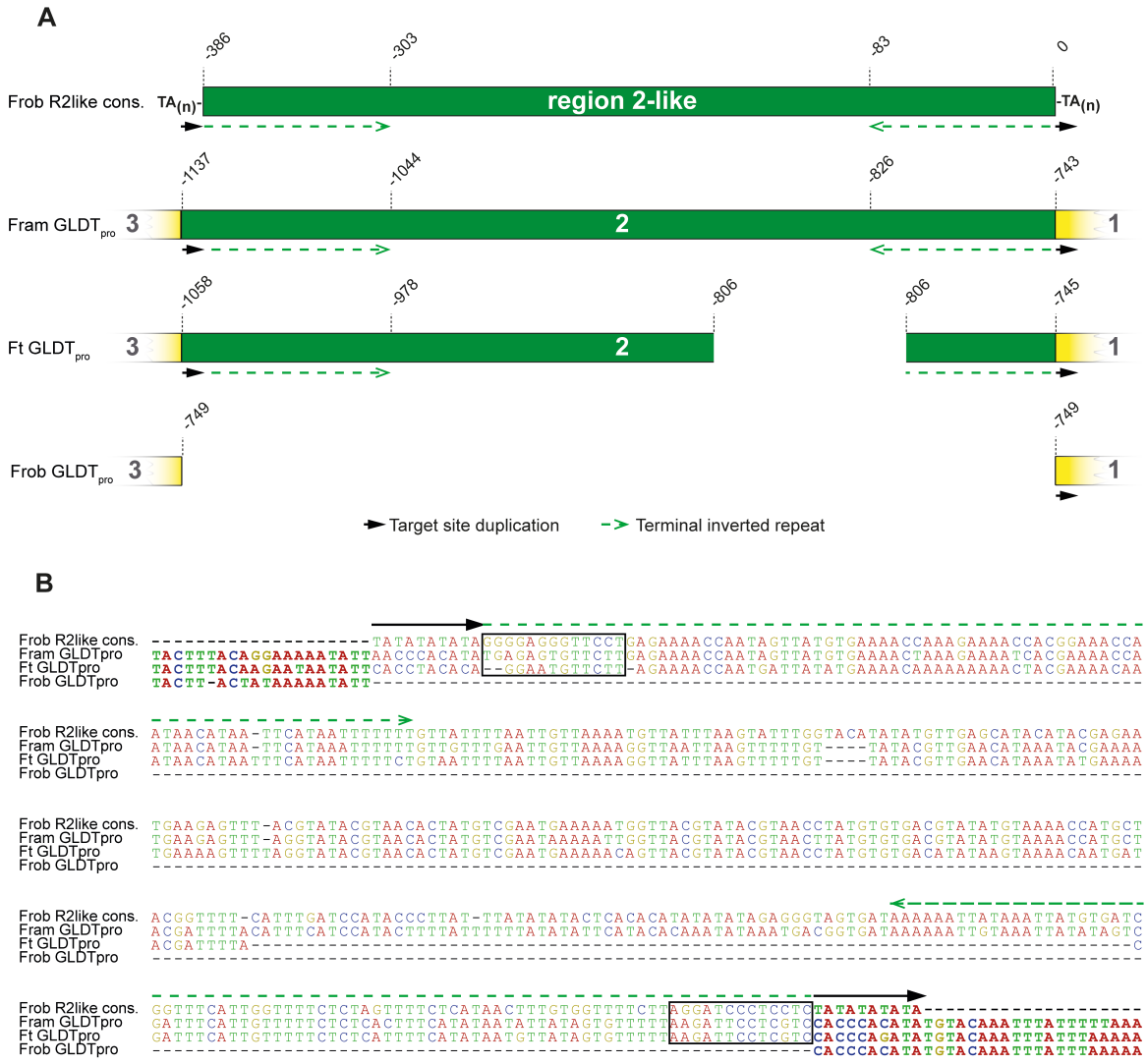


Figure S2 Sequence comparison of region 2 and reconstructed region 2-like consensus sequence. The sequence of the region 2-like putative transposable element was reconstructed from RNA-seq data of *F. robusta* (Frob R2like consensus) and re-aligned to region 2 of *F. ramosissima*, *F. trinervia* and the corresponding position in in the upstream flanking sequence of the *F. robusta GLDT* gene. **(A)** Schematic representation of aligned sequences in (B) and their structural features. **(B)** Sequence alignment of the reconstructed region 2-like consensus sequence and region 2 from *F. ramosissima*, *F. trinervia* and the corresponding position in in the upstream flanking sequence of the *F. robusta GLDT* gene. Black rectangles highlight distinct terminal motifs (see text). Flanking sequences (i.e. regions 3 and 1) are depicted by bold characters. Putative target site duplications are indicated by black arrows and terminal inverted repeats by dashed green arrows.

References

Bureau TE, Wessler SR. 1994. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6**, 907-916.

Feschotte C, Swamy L, Wessler SR. 2003. Genome-Wide Analysis of mariner-Like Transposable Elements in Rice Reveals Complex Relationships With Stowaway Miniature Inverted Repeat Transposable Elements (MITEs). *Genetics* **163**, 747-758.

2. Towards mapping of cis-regulatory elements in the upstream flanking sequence of GLDT from the genus Flaveria

Introduction

C_4 photosynthesis is a carbon concentrating mechanism that evolved to reduce loss of photorespiratory CO_2. On the evolutionary trajectory from C_3 to C_4, glycine decarboxylase (GDC) becomes restricted to the bundle sheath, where the decarboxylation of photorespiratory glycine leads to a local CO_2 enrichment, largely suppressing the futile oxygenation of ribulose-1,5-bisphosphate by ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco).

The C_4 cycle, in higher plants, relies on dedicated expression of photosynthetic enzymes in either bundle sheath cells (BSC) or mesophyll cells (MC). In order to efficiently introduce C_4 photosynthesis into C_3 crop plants, detailed knowledge about how C_4 plants facilitate differential gene expression in MC and BSC is required. Analysis of phosphoenolpyruvate carboxylase (PEPC) (Akyildiz *et al.*, 2007; Gowik *et al.*, 2004; Gowik *et al.*, 2017) and carboanhydrase (CA) (Gowik *et al.*, 2017; Williams *et al.*, 2016) upstream flanking sequences from *Flaveria trinervia* and *Gynandropsis gynandra* revealed discrete sequence motifs mediating mesophyll expression, and systematic approaches identified a plethora of short consensus sequences, enriched in genes that are exclusively expressed in one of both cell types (Burgess *et al.*, 2017; Cao *et al.*, 2016; Sheen, 1999; Wang *et al.*, 2014; Xu *et al.*, 2016). However, so far, only one study reported experimental evidence that linked a *cis*-regulatory motif to expression in BSC: Xu *et al.* (2001) found two elements in the upstream flanking sequence and one in the 3'-UTR of a maize *rbcS* gene that are necessary for suppression of Rubisco activity in mesophyll cells.

For the genus *Flaveria* it was shown that restriction of the GDC subunit P (GLDP) – the actual decarboxylase – to the BSC occurred gradually (Schulze *et al.*, 2013), suggesting an increased evolutionary pressure for other GDC subunits, such as GLDT, to adapt. While it was known that the GLDT protein from the C_4 *Flaveria* species *F. trinervia* localises to the BSC, too (Morgan *et al.*, 1993), analysis of the upstream flanking sequences of *GLDT* from C_3, C_4 and C_3-C_4 intermediate *Flaveria* species suggested that this is predominantly controlled at the transcriptional level (Chapter 1). Dissection of the *GLDT* 5'-flanking sequences of *F. trinervia* (C_4) showed that a 313 bp segment (region 2) contains positive *cis*-regulatory elements (CREs) and is necessary to restrict spatial expression to the bundle sheath and vasculature of transgenic *A. thaliana* (Chapter 1). However, additional substitution

experiments suggested that redundant CREs are present in the conserved upstream region 3 (~600 bp), which is also required to maintain high levels of reporter gene expression (Chapter 1). In contrast, the *GLDT* upstream sequence of *F. robusta* (C_3) lacks region 2, but contains region 3, and mediates expression in all leaf chlorenchyma cells, i.e. also in the mesophyll (Chapter 1).

To further illuminate how BSC specific expression of GLDT in *Flaveria* evolved, additional *GLDT* upstream flanking sequences were isolated and analysed in transgenic *A. thaliana*. Chimeric upstream flanking sequences were generated to confirm the presence of CREs in region 3 and to elucidate dependencies of region 2. Corresponding sequences were compared in context of their phylogeny and the observed expression pattern in *A. thaliana*, to identify sequence polymorphisms that associate with the observed differences in expression patterns.

Results and Discussion

Phylogenetic analysis of the 5' flanking sequences of GLDT genes of C_3, C_3-C_4 and C_4 Flaveria species

To isolate *GLDT* upstream flanking sequences from additional *Flaveria* species two sets of nested PCR primers were designed that should hybridise with well-conserved sequences in i) region 4 and the *GLDT* coding sequence (CDS; set A) or ii) region 5 and the CDS (set B). Genomic DNA of all available *Flaveria* species was isolated and PCR was conducted using both primer sets on all species.

Using primer set A, upstream flanking sequences were obtained from clade A species *F. trinervia*, *F. palmeri, F. ramosissima* and the basal species *F. angustifolia* and *F. pringlei.* However, the *F. pringlei* cultivar used here was shown to be a hybrid of *F. pringlei* and *F. angustifolia* (Lyu *et al.*, 2015). Correspondingly, the sequence amplified from *F. pringlei*, using primer set A, exhibited 99% similarity to the sequence isolated from *F. angustifolia* and hence it was termed *F. angustifolia* (2) in the following. A second *GLDT* upstream flanking sequence was isolated from *F. pringlei* independently by genome walking and due to its similarity to *F. robusta* and *F. cronquistii* (Fig. 1), most likely corresponds to the actual *F. pringlei GLDT* gene.

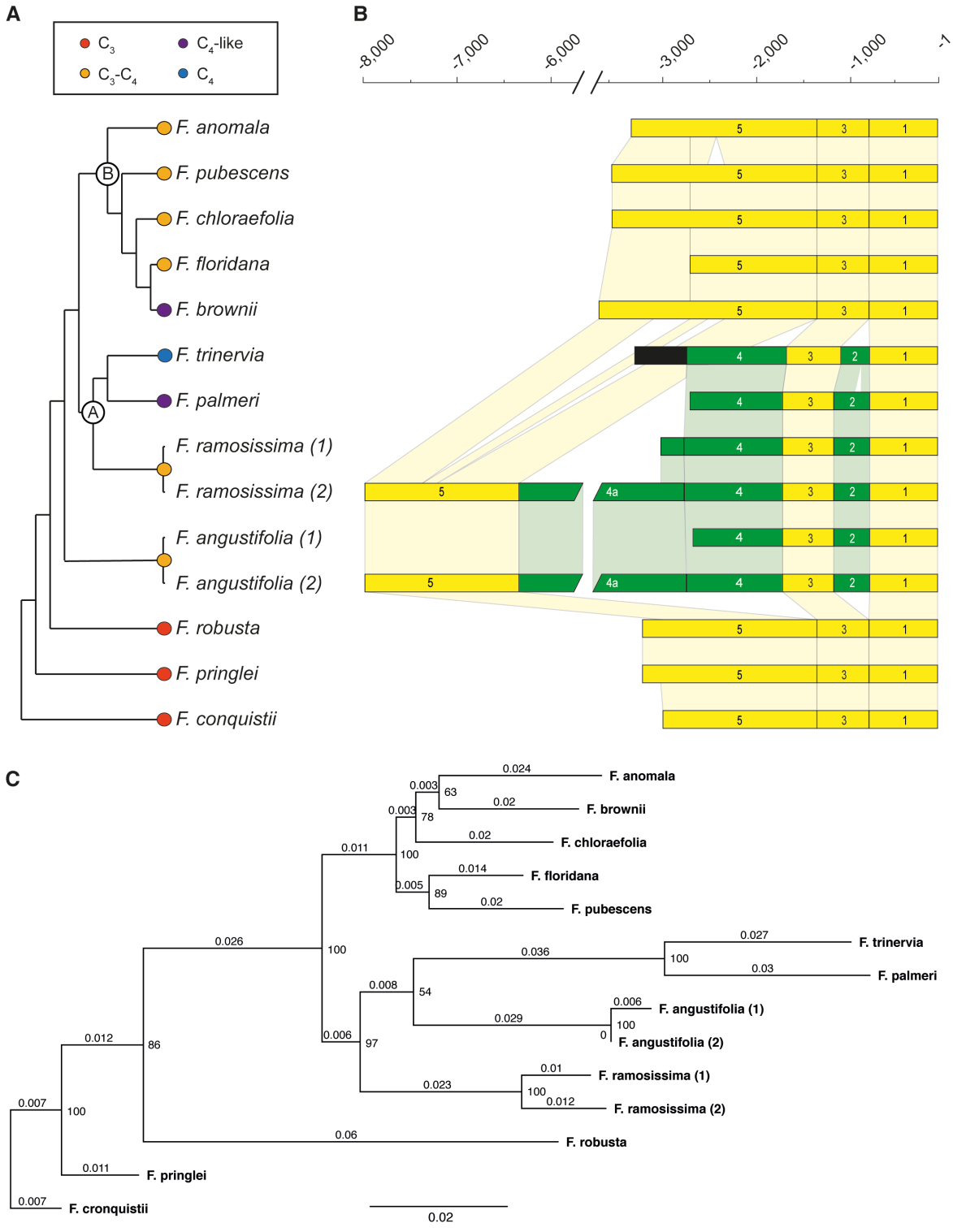


Figure 1 Ancestry and schematic representation of isolated *GLDT* upstream flaking sequences. (A) Composite phylogeny of the genus Flaveria, inferred from McKown *et al.* (2005) and Lyu *et al.* (2015). Clade A and B are depicted by corresponding letters on the respective nodes. Corresponding taxa have been duplicated, if two sequences were isolated, to match with (B). **(B)** Schematic representation of isolated sequences, sorted according to (A). Corresponding numbers indicate conserved regions. Regions present only in clade A and *F. angustifolia* are highlighted in green. Black indicates regions not found in other species. **(C)** Maximum-likelihood tree of isolated sequences. Scale bar and branch labels indicate substitutions per site. Bootstrap support values are displayed at the corresponding nodes.

Using primer set B, upstream flanking sequences were isolated from basal C_3 species *F. robusta* and *F. cronquistii,* as well as from clade B species *F. anomala, F. pubescens, F. chloraefolia, F. floridana* and *F. brownii*.

Surprisingly, usage of primer set B amplified two fragments from *F. pringlei*. While the shorter fragment was identical to the upstream flanking sequence from *F. pringlei,* previously isolated by genome walking, the co-amplified large fragment was identical to *F. angustifolia* (2), but exceeded the sequence 5' by 4.8 kb. The distal 1.5 kb clearly corresponded to region 5 found in clade B and basal species, showing 89% sequence similarity on average. Similar results were also found for a large fragment isolated from *F. ramosissima,* using primer set B. However, sequence similarity to the previously isolated sequence was *only* 97%, hence, the large fragment was termed *F. ramosissima* (2).

The intercalary sequence between region 5 and 4 was termed 4a (see Fig. 1B for summary). However, region 4a does not show any similarity to the region upstream of region 4 found in *F. trinervia*, suggesting multiple independent insertion/deletion events. BLAST search of region 4 and 4a against RepBase database (http://www.girinst.org/repbase) revealed weak similarity of region 4a to *PIF/Harbinger*-like DNA transposons of *Medicago truncatula* and *Vitis vinifera* (Carrier *et al.*, 2012; Grzebelus *et al.*, 2007), but no results were obtained for region 4. Further, neither terminal inverted repeats nor target site duplications were visible. Due to the absence of BLAST hits for region 4, it could not be satisfyingly answered whether region 4 and 4a constitute a single insertion event, but rather render it the most parsimonious scenario. Further, a short non-conserved segment upstream of region 4 from *F. trinervia* suggests, at least, one additional insertion/deletion event for this species.

Notably, BLAST search also identified region 5 as part of a putative *U3 small nucleolar RNA-associated protein 6* gene. However, its coding sequence exhibited only low conservation between isolated sequences and thus region 5 most likely represents part of a pseudo gene.

Intriguingly, the *GLDT* locus of the basal C_3-C_4 species *F. angustifolia* contained regions 2 and 4, which were neither found in other basal nor clade B species, implying that either both regions must have been lost in clade B species or that the *GLDT* locus of *F. angustifolia* does simply not correlate to the species phylogeny. Thus, a maximum likelihood gene tree was calculated from an alignment of all

isolated upstream flanking sequences (Fig 1C). The tree roughly resembled the known phylogeny of *Flaveria* (Fig. 1A; Lyu *et al.*, 2015; McKown *et al.*, 2005), but nodes of clade B and *F. angustifolia* exhibited low statistical support. Notably, this method does not consider insertions and deletions, i.e. regions 2, 4 and 5 were ignored, since they are not present in all sampled species. Nonetheless, maximum likelihood places *F. angustifolia* within clade A (54 % bootstrap support; or at its basis, 46 % bootstrap support, data not shown), contrary to the species phylogeny, where *F. angustifolia* branches before the split of clade A and B (Fig. 1A). This and the absence of regions 2 and 4 from clade B suggest that the *GLDT* locus of *F. angustifolia* may have been subject to hybridisation or incomplete lineage sorting.
Taken together, these results indicate that the upstream flanking sequence of *GLDT* was hit, at least, twice by transposable elements. Intriguingly, those insertions are only present in clade A (including *F. angustifolia*), to which the only true C_4 *Flaveria* species belong. This indicates that evolution of C_4 photosynthesis in the genus *Flaveria* may have been accompanied by bursts of transposable element activity, corresponding to the need of altered expression patterns for several hundred genes (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011) and the potential of transposable elements to convey such (Feschotte, 2008; Rebollo *et al.*, 2012).

Promoter activity analysis of the 5' flanking sequences of GLDT genes of C3, C3-C4 and C4 species in transgenic Arabidopsis

To analyse the expression pattern mediated by the isolated upstream flanking sequences, only regions 3 to 1 were analysed, because previous results demonstrated that in the upstream flanking sequence of *F. trinervia*, these regions were sufficient for correct spatial expression (Chapter 1). The selected upstream sequences were fused to a GUS reporter gene and transformed into *A. thaliana* (Fig. 2).
Regions 3 and 1 of the upstream flanking sequences of the clade B C_3-C_4 species *F. anomala* and *F. brownii,* as well as of the basal C_3 species *F. robusta* and *F. pringlei,* resulted in a more or less uniform expression in the leaves of *A. thaliana*, i.e. the expression of the GUS reporter gene was not confined to the bundle sheath cells and the vasculature (Fig. 2). In all these *GLDT* upstream sequence a region 2 equivalent

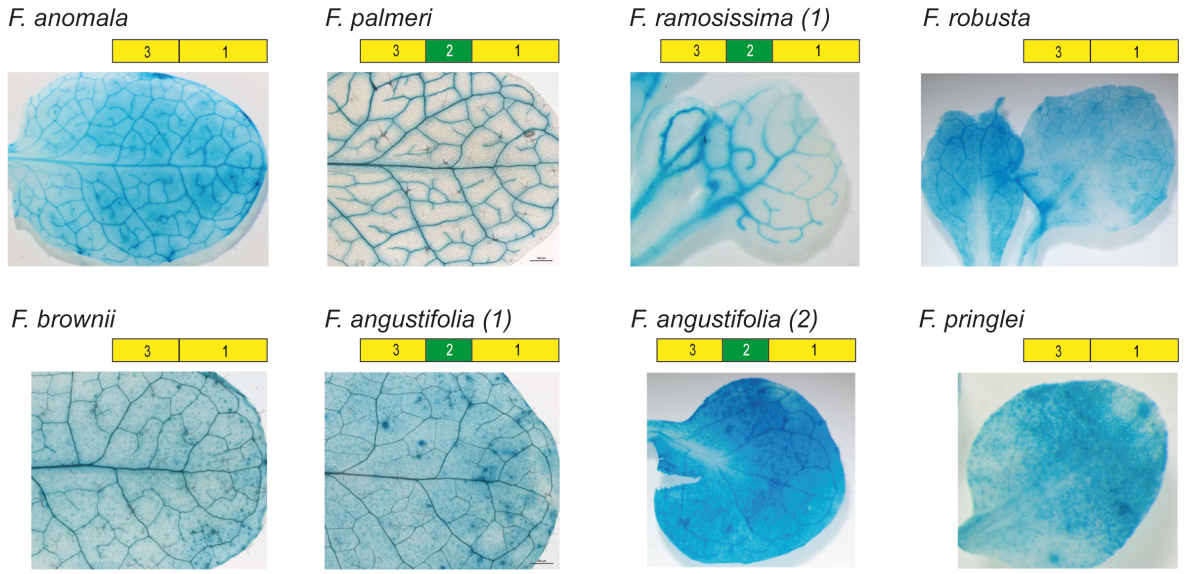


Figure 2 GUS expression of selected *GLDT* upstream flanking sequences in transgenic Arabidopsis. Upstream flanking sequences downstream of region 4 or 5 were fused to a GUS reporter gene in transformed into *A. thaliana*. Species and corresponding sequence topology are depicted above images. Images of *F. palmeri, F. brownii* and *F. angustifolia* (1) were conducted by translight microscopy, others by reflected light microscopy.

is missing, thus reinforcing the previous finding that region 2 is necessary to confine *GLDT* promoter activity to the bundle sheath and the vasculature in transgenic *A. thaliana* (Chapter 1). In support of this conclusion the *GLDT* upstream sequences of the C_3-C_4 species *F. ramosissima* and the C_4 species *F. palmeri,* both of which contain region 2 in addition to regions 3 and 1, mediated GUS expression restricted to the bundle sheath and the vasculature (Fig. 2; Supplemental Fig. S1).

In contrast, the upstream flanking sequences of the two *GLDT* genes of the C_3-C_4 intermediate *F. angustifolia*, although both harbouring a region 2 segment, did not confine GUS expression to the bundle sheath and vasculature, as both were clearly active also in the mesophyll (Fig. 2). This finding suggests that region 2, although necessary for the restriction of GLDT expression to the bundle sheath and the vasculature (Chapter 1), may depend in its activity on CREs, most likely residing in region 3 (Chapter 1). Additionally, the high similarity of the *GLDT* upstream flanking sequence of *F. angustifolia* to the ones of *F. trinervia, F. palmeri* and *F. ramosissima* (see Fig. 1), which mediate confined expression, could also allow a precise localisation of corresponding CREs by sequence comparison.

Impact of cis-regulatory elements in region 3 on spatial expression

Key *cis*-regulatory elements for GLDT transcription in bundle sheath and vasculature appear to be located in region 3, and the cis-regulatory information encoded in this region may differ between the GLDT upstream sequences from C_3, C_3-C_4 and C_4 *Flaveria* species. Moreover, the influence of region 2 may be dependent on the informational content of region 3, as previous results (Chapter 1) and the upstream flanking sequence of *F. angustifolia* indicated. To get a first insight into the dependency of CREs in region 3 and 2 and their distribution within the genus, three chimeric upstream flanking sequences were generated. To elucidate whether region 3 of *F. robusta* mediates the same expression as that of *F. trinervia*, but its activity is masked by the absence of region 2 – similar to construct *Ft*-GLDT$_{PRO}$31 (Chapter 1) – region 2 of *F. trinervia* was inserted between regions 3 and 1 of *F. robusta*, giving rise to construct GLDT$_{PRO}$Frob3-Ft2-Frob1 (Fig. 3A). A second set of constructs exchanged regions 3 of *F. trinervia* and *F. angustifolia* to confirm that the predominant expression is mediated by region 3 in both species (GLDT$_{PRO}$Fang3-Ft21 and GLDT$_{PRO}$Ft3-Fang21, Fig. 3A).

Insertion of region 2 from *F. trinervia* into the upstream flanking sequence of *F. robusta* (GLDT$_{PRO}$Frob3-Ft2-Frob1) was not able to change the expression pattern (cmp. Fig. 2 and 3C). This showed that, although region 2 does harbour corresponding positive CREs and its presence is necessary in *F. trinervia* (Chapter

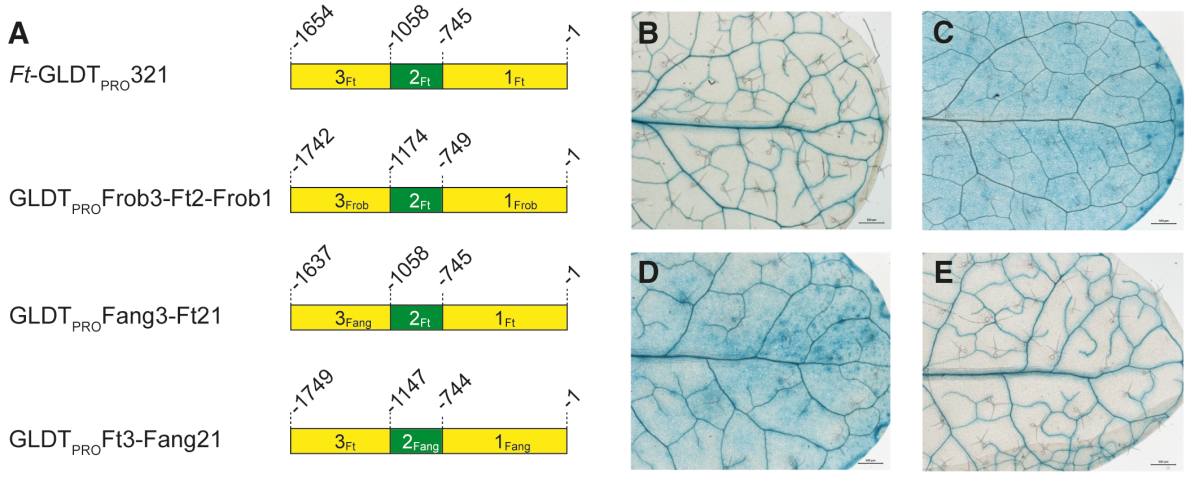


Figure 3 GUS expression of chimeric upstream flanking sequences. (A) Schematic representation of reference (*Ft*-GLDT$_{PRO}$321) and chimeric constructs. For construct GLDT$_{PRO}$Frob3-Ft2-Frob1 region 2 of *F. trinervia* was inserted into the upstream flanking sequence of *F. robusta,* at the corresponding position. For constructs GLDT$_{PRO}$Fang3-Ft21 and GLDT$_{PRO}$Ft3-Fang21 region 3 of *F. angustifolia* and *F. trinervia* were exchanged. **(B)** GUS expression of construct *Ft*-GLDT$_{PRO}$321 in leaves of transgenic *A. thaliana* (taken from Chapter 1 for comparison). **(C – E)** GUS expression in leaves of transgenic *A. thaliana* of constructs GLDT$_{PRO}$Frob3-Ft2-Frob1, GLDT$_{PRO}$Fang3-Ft21 and GLDT$_{PRO}$Ft3-Fang21, respectively.

1), it is not sufficient to generate expression confined to BSC and vasculature. Moreover, this also indicated that region 3 of *F. robusta* does not harbour the same CREs as region 3 of *F. trinervia*. Corresponding results were also obtained with the other chimeric constructs: Fusion of region 3 of *F. trinervia* to regions 2 and 1 of *F. angustifolia* ($GLDT_{PRO}$Ft3-Fang21) mediated the same expression pattern as the full upstream flanking sequence of *F. trinervia* (cmp. Fig 3A and 3E), while the reciprocal construct ($GLDT_{PRO}$Fang3-Ft21) conveyed the same expression as the upstream flanking sequence of *F. angustifolia* (cmp. Fig 2 and 3D).

In summary, these experiments confirmed our previous conclusion that region 3 of the *GLDT* upstream flanking sequences contains the dominant *cis*-regulatory determinants for spatial expression of GLDT and that region 3 from C_4 species has acquired *cis*-regulatory information that convey BSC and vasculature specific expression in *A. thaliana*, which is not present in the *GLDT* genes from the C_3 species (*F. robusta*) and likely all other C_3 *Flaveria* species (see Fig. 1: *F. pringlei, F. cronquistii*). More importantly, this also showed that region 3 of the basal C_3-C_4 *F. angustifolia* does also not contain the relevant *cis*-regulatory information. The sequence of *F. angustifolia* shows higher similarity to the ones of *F. trinervia, F. palmeri* and *F. ramosissima* - which all mediated expression confined to BSC and vasculature – than to those sequences, which also showed expression in MC (e.g. *F. robusta*, *F. pringlei;* see Fig. 1). This dramatically reduces the amount of false-positive positions in a subsequent discriminative sequence comparison.

Discriminative sequence analysis identifies 11 candidate positions for relevant CREs

For sequence analysis of region 3, the most parsimonious scenario was assumed, in which the CREs that confine expression of GLDT to BSC and vasculature evolved only once in a common ancestor and were conserved in the species *F. trinervia*, *F. palmeri* and *F. ramosissima*. Subsequent alignment of these sequences (Fig. 4: sequences 1 – 4) to region 3 of the basal C_3-species and *F. angustifolia* (Fig. 4: sequences 5 – 9) identified 11 conserved polymorphisms that associate with the observed differential expression (Fig. 4). Note that, although the included sequences

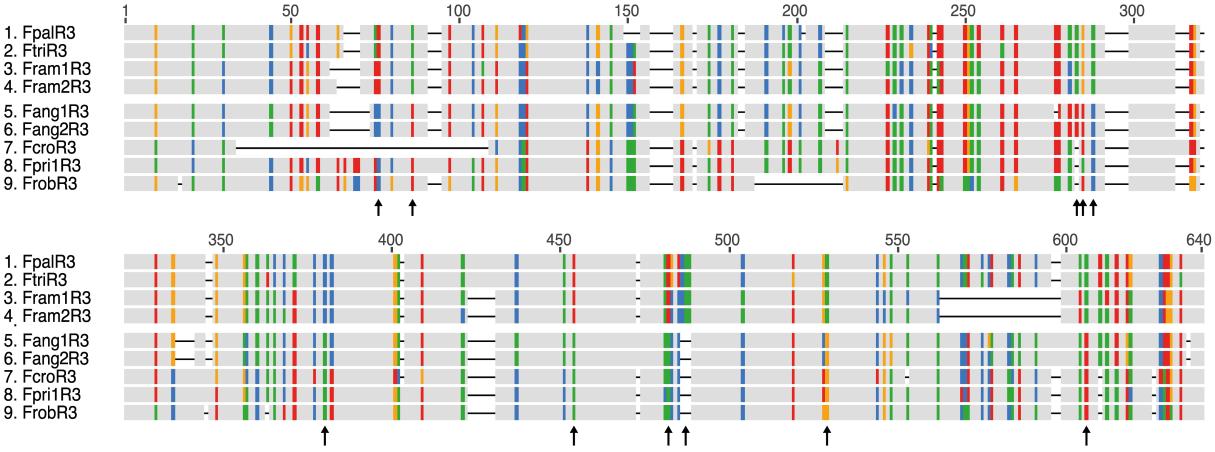


Figure 4 Candidate positions in region 3 that associate with differential expression of *GLDT*. Colour-coded alignment of region 3 from analysed and closely related sequences. Identical nucleotide positions are depicted in grey and polymorphisms in corresponding clustal colours (A = Red, T = Green, G = Yellow, C = Blue). Black arrows highlight conserved polymorphisms that associate with the corresponding expression pattern, i.e. confined expression (sequences 1 to 4) or ubiquitous expression (sequences 5 to 9).

of *F. ramosissima (2)* and *F. cronquistii* were not tested for their activity, all possible candidate positions in these sequences exhibited the same state as their closest relatives, i.e. the analysis was not affected by their inclusion, but they support that the identified positions are actually conserved. Interestingly, some of these polymorphisms are in close proximity to each other and thus generated more complex changes – three positions between 283 to 288 bp of the sequence alignment and two between 482 to 488 bp, one of which a conserved tri-nucleotide insertion/deletion (Fig. 4; see supplemental data S1 for full alignment). If multiple polymorphisms in such close proximity are conserved, it seems more likely that they represent the acquisition of a novel transcription factor binding site. Additional prediction of known CREs showed that of these 11 polymorphisms 6 altered the sequence to putative CREs that were unique, either to those upstream flanking sequences that mediated confined expression or to those, which did not (summarised in Table 1; Supplemental data S1). Notably, besides these CREs that associated with conserved polymorphisms, it was also observed that region 3 of those sequences, which conveyed confined expression were enriched in putative C2H2 zinc finger binding sites, while those which did not were enriched in putative AT-Hook binding sites (Supplemental Data S1).

Although the close relation of the analysed sequences dramatically increased the resolution for sequence comparison, it was not able to pinpoint a particular region of interest, suggesting that further dissection of region 3 is inevitable. Nonetheless, by

confirming the presence of relevant CREs in additional sequences and excluding them from others, the results presented here also provide a reasonable fundament for further analysis.

Table 2 Summary of identified polymorphisms and associated putative binding sites. Sequences that mediated confined reporter gene expression belong to clade A, sequences which did not belong to the basal C_3 species (incl. *F. angustifolia*).

Alignment Position	Nucleotide state clade A	CRE clade A	Nucleotide state basal	CRE basal
76	A	TGA/bZIP	Not conserved	#NA
86	T	#NA	Not conserved	#NA
283	T	#NA	C	#NA
285	G	#NA	Not conserved	#NA
288	T	#NA	Not conserved	#NA
380	C	#NA	T	#NA
454	A	SRS	T	IBOX motif
482	A	C2H2	T	#NA
486-488	CTT	#NA	-	AP2/RAV/B3
529	T	TBOX motif	G	#NA
606	T	#NA	A	AT-Hook

Summary

Isolation of *GLDT* upstream flanking sequences from several *Flaveria* species confirmed that region 2, which was previously shown to likely be an ancient insertion of a transposable element and which is necessary for confined expression of GLDT, is unique to clade A *Flaveria* species and *F. angustifolia*, suggesting incomplete lineage sorting or hybridisation. Further, additional sequences isolated from *F. pringlei* and *F. ramosissima* suggested that region 4 might be part of a larger transposable element that inserted between regions 5 and 3, indicating that high transposable element activity may be associated with the abundant alterations in gene expression encompassing C_4 evolution. Analysis of promoter:GUS fusion constructs in *A. thaliana* showed that the upstream flanking sequence of *F. palmeri* mediates similar expression as that of *F. trinervia*, while the closely related sequence of *F. angustifolia* does not. The expression mediated by chimeric upstream flanking sequences confirmed the previous assumption that the expression pattern is predominantly dictated by CREs in region 3 and verified or excluded their existence in a subset of isolated sequences. Subsequent alignment identified 11 sequence polymorphisms that associated with differential activity of these upstream flanking sequences and may facilitate future analyses.

Material and Methods

Plant transformation

F. bidentis was transformed as described by Chitty *et al.* (1994). *A. thaliana* was transformed by floral dip (Clough and Bent, 1998), as adapted by Logemann *et al.* (2006). For transformation of both species the *Agrobacterium tumefaciens* strain AGL1 (Lazo *et al.*, 1991) was used. T-DNA insertion into the host genome was confirmed by PCR.

Isolation of upstream flanking sequences

The upstream flanking sequence of *F. pringlei* was isolated by genome walking in context of chapter 1. Isolated sequences were aligned to design primer set A and to hybridise with regions 4 and 5, respectively. Primers hybridising with the coding sequence were designed by aligning publically available sequences (acc. No. Z99769, Z71184, Z25858) and contigs of *de novo* assembled RNA-seq experiments (Mallmann *et al.*, 2014). Sequences of *F. palmeri* and *F. angustifolia* (1) were isolated in a nested PCR using primer pairs GLDTA-FW1/GLDT-RV1 and GLDTA-FW2/GLDT-RV3. Sequences of *F. ramosissima* (2), *F. anomala, F. brownii* and *F. cronquistii* were isolated using primer pairs GLDTB-FW1/GLDT-RV1 and GLDTB-FW2/GLDT-RV3. The sequence of *F. angustifolia* (2) was isolated using primer pairs GLDTB-FW1/GLDT-RV1 and GLDTB-FW2/GLDT-RV2. The sequence of *F. floridana* was isolated using primer pairs GLDTB-FW1/GLDT-RV1 and GLDTB-FW3/GLDT-RV3. All primers mentioned here are listed in supplemental table S1 and the isolated sequences in supplemental data S2.

Cloning of reporter gene constructs

All PCR products were purified by gel extraction (QIAquick Gel Extraction Kit, QIAGEN), cloned into cloning vector pJET1.2 (CloneJET PCR Cloning Kit, Fermentas/Thermo Fisher Scientific) and confirmed by sequencing prior to digestion and ligation with an in-house version of pBI121 expression vector (Jefferson *et al.*, 1987; Chapter 1). All primers mentioned in the following are listed in supplementary table S1. Due to the high conservation of the isolated sequences, several primers

hybridise to the corresponding region of multiple species, contrary to the abbreviations used as primer identifier.

Construct depicted in Figure 2 were cloned by PCR amplification from isolated upstream flanking sequences, attaching appropriate restriction sites, using the following primer combinations: *F. anomala* – anoR3-FW:SfaAI/anoR1-RV:SgsI; *F. palmeri* – palR3-FW:SfaAI/palR1-RV:SgsI; *F. ramosissima* (1) – anoR3-FW:SfaAI/anoR1-RV:SgsI; *F. robusta* – robR3-FW:SfaAI/robR1-RV:XmaI; *F. brownii* – anoR3-FW:SfaAI/broR1-RV:SgsI; *F. angustifolia* (1, 2) palR3-FW:SfaAI/angR1-RV:SgsI; *F. pringlei* – priR3-FW:SfaAI/priR1-RV:XmaI. All fragments were cloned into the expression vector under digestion with restriction enzymes indicated in the corresponding primer identifier.

For construct *Frob*-GLDT$_{PRO}$3-Ft2-1 regions 3 and 1 were amplified from construct *Frob*-GLDT$_{PRO}$ (Chapter 1) using primer combinations robR3-FW:SfaAI/robR3-RV:XhoI and robR1-FW:BcuI/robR1-RV:XmaI, respectively. Region 2 was amplified from construct *Ft*-GLDT$_{PRO}$ (Chapter 1) using primer combination triR2-FW:XhoI/triR2-RV:BcuI. After subcloning, region 2 was digested from pJET1.2 backbone using *Xho*I and *Bcu*I. Regions 3 and 1 were digested from pJET1.2 backbone by *Bgl*II/*Xho*I and *Bcu*I/*Bgl*II, respectively. Regions 3, 2 and 1 were triple-ligated, subcloned and amplified using primers robR3-FW:SfaAI and robR1-RV:XmaI. Cloning into the expression vector was conducted as stated above.

Constructs *Ft*-GLDT$_{PRO}$*Fang*3-21 and *Fang*-GLDT$_{PRO}$*Ft*3-21 were cloned by exploitation of the endogenous restriction sites *Pdm*I and *Dra*III in *F. angustifolia* upstream flanking sequence to facilitate seamless ligation. For construct *Ft*-GLDT$_{PRO}$*Fang*3-21 region 3 was amplified from *F. angustifolia* (1) upstream flanking sequence using primers palR3-FW:SfaAI and angR3-RV:PdmI. Regions 2 and 1 were amplified in tandem from construct *Ft*-GLDT$_{PRO}$ (Chapter 1) using primers triR2-FW:PdmI and triR1-RV:XmaI. Both fragments were triple-ligated with the *SfaA*I/*Xma*I *digested* expression vector backbone. For construct *Fang*-GLDT$_{PRO}$*Ft*3-21 region 3 was amplified from construct *Ft*-GLDT$_{PRO}$ (Chapter 1) using primers robR3-FW:SfaAI and triR3-RV:DraIII. Regions 2 and 2 in tandem were amplified from *F. angustifolia* (1) upstream flanking sequence using primers angR2-FW:DraIII and angR1-RV:SgsI. Both fragments were triple-ligated with the *SfaA*I/*Sgs*I *digested* expression vector backbone.

Constructs *Ft*-GLDT$_{PRO}$321 and *Ft*-GLDT$_{PRO}$21 are described in Chapter 1.

In situ detection of GUS activity and fluorometric measurement

In situ detection and fluorometric quantification of GUS were carried out as described in Chapter 1. Statistical confidence was calculated using Mann-Whitney test.

Sequence analysis and tree construction

Sequence alignments were conducted using MAFFT v7.310 (Katoh and Standley, 2013) and refined manually, if necessary. Construction of the maximum likelihood tree was conducted using PhyML3.0 (Guindon *et al.*, 2010), provided as plugin for geneious v6.1.7 (Kearse *et al.*, 2012). The substitution model used was *general time reversible* with estimated gamma distribution. Statistic support was calculated by bootstrapping with 100 replicates.

CRE prediction was carried out using PlantPan (Chow *et al.*, 2016) and CIS-BP (Weirauch *et al.*, 2014). Predicted CREs were compared and filtered using Python3. The general workflow was to identify CREs, which are common to all sequences of a set A, filter those which do not occur anywhere in any sequence of a set B and vice versa.

References

Akyildiz M, Gowik U, Engelmann S, Koczor M, Streubel M, Westhoff P. 2007. Evolution and function of a cis-regulatory module for mesophyll-specific gene expression in the C4 dicot Flaveria trinervia. *Plant Cell* **19**, 3391-3402.

Bräutigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Mass J, Lercher MJ, Westhoff P, Hibberd JM, Weber AP. 2011. An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiol* **155**, 142-156.

Burgess SJ, Reyna-Llorens I, Jaeger K, Hibberd JM. 2017. A transcription factor binding atlas for photosynthesis in cereals identifies a key role for coding sequence in the regulation of gene expression. *bioRxiv* **165787**, https://doi.org/10.1101/165787.

Cao C, Xu J, Zheng G, Zhu X-G. 2016. Evidence for the role of transposons in the recruitment of cis-regulatory motifs during the evolution of C4 photosynthesis. *BMC Genomics* **17**, 1-11.

Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, Audeguin L, Boursiquot J-M, This P. 2012. Transposable elements are a major cause of somatic polymorphism in Vitis vinifera L. *PLoS ONE* **7**, e32973.

Chitty JA, Furbank RT, Marshall JS, Chen Z, Taylor WC. 1994. Genetic transformation of the C4 plant, Flaveria bidentis. *The Plant Journal* **6**, 949-956.

Chow C-N, Zheng H-Q, Wu N-Y, Chien C-H, Huang H-D, Lee T-Y, Chiang-Hsieh Y-F, Hou P-F, Yang T-Y, Chang W-C. 2016. PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res* **44**, D1154-D1160.
Clough SJ, Bent AF. 1998. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J* **16**, 735-743.
Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397-405.
Gowik U, Bräutigam A, Weber KL, Weber AP, Westhoff P. 2011. Evolution of c4 photosynthesis in the genus flaveria: how many and which genes does it take to make c4? *Plant Cell* **23**, 2087-2105.
Gowik U, Burscheidt J, Akyildiz M, Schlue U, Koczor M, Streubel M, Westhoff P. 2004. cis-Regulatory elements for mesophyll-specific gene expression in the C4 plant Flaveria trinervia, the promoter of the C4 phosphoenolpyruvate carboxylase gene. *Plant Cell* **16**, 1077-1090.
Gowik U, Schulze S, Saladié M, Rolland V, Tanz SK, Westhoff P, Ludwig M. 2017. A MEM1-like motif directs mesophyll cell-specific expression of the gene encoding the C4 carbonic anhydrase in Flaveria. *Journal of Experimental Botany* **68**, 311-320.
Grzebelus D, Lasota S, Gambin T, Kucherov G, Gambin A. 2007. Diversity and structure of PIF/Harbinger-like elements in the genome of Medicago truncatula. *BMC Genomics* **8**, 409.
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321.
Jefferson R, Kavanagh T, Bevan M. 1987. GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO Journal* **6**, 3901 - 3907.
Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772-780.
Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649.
Lazo GR, Stein PA, Ludwig RA. 1991. A DNA transformation-competent Arabidopsis genomic library in Agrobacterium. *Biotechnology (N Y)* **9**, 963-967.
Logemann E, Birkenbihl RP, Ulker B, Somssich IE. 2006. An improved method for preparing Agrobacterium cells that simplifies the Arabidopsis transformation protocol. *Plant Methods* **2**, 16.
Lyu M-J, Gowik U, Kelly S, Covshoff S, Mallmann J, Westhoff P, Hibberd J, Stata M, Sage R, Lu H, Wei X, Wong G, Zhu X-G. 2015. RNA-Seq based phylogeny recapitulates previous phylogeny of the genus Flaveria (Asteraceae) with some modifications. *BMC Evol Biol* **15**, 116.
Mallmann J, Heckmann D, Bräutigam A, Lercher MJ, Weber AP, Westhoff P, Gowik U. 2014. The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. *eLife* **3**, e02478.
McKown AD, Moncalvo JM, Dengler NG. 2005. Phylogeny of Flaveria (Asteraceae) and inference of C4 photosynthesis evolution. *Am J Bot* **92**, 1911-1928.
Morgan CL, Turner SR, Rawsthorne S. 1993. Coordination of the cell-specific distribution of the four subunits of glycine decarboxylase and of serine

hydroxymethyltransferase in leaves of C3-C4 intermediate species from different genera. *Planta* **190**, 468-473.
Rebollo R, Romanish MT, Mager DL. 2012. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annu Rev Genet* **46**, 21-42.
Schulze S, Mallmann J, Burscheidt J, Koczor M, Streubel M, Bauwe H, Gowik U, Westhoff P. 2013. Evolution of C4 photosynthesis in the genus flaveria: establishment of a photorespiratory CO2 pump. *Plant Cell* **25**, 2522-2535.
Sheen J. 1999. C4 GENE EXPRESSION. *Annual Review of Plant Physiology and Plant Molecular Biology* **50**, 187-217.
Wang L, Czedik-Eysenberg A, Mertz RA, Si Y, Tohge T, Nunes-Nesi A, Arrivault S, Dedow LK, Bryant DW, Zhou W, Xu J, Weissmann S, Studer A, Li P, Zhang C, LaRue T, Shao Y, Ding Z, Sun Q, Patel RV, Turgeon R, Zhu X, Provart NJ, Mockler TC, Fernie AR, Stitt M, Liu P, Brutnell TP. 2014. Comparative analyses of C4 and C3 photosynthesis in developing leaves of maize and rice. *Nature Biotechnology* **32**, 1158.
Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJM, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443.
Williams BP, Burgess SJ, Reyna-Llorens I, Knerova J, Aubry S, Stanley S, Hibberd JM. 2016. An untranslated cis-element regulates the accumulation of multiple C4 enzymes in Gynandropsis gynandra mesophyll cells. *Plant Cell*.
Xu J, Bräutigam A, Weber APM, Zhu X-G. 2016. Systems analysis of cis-regulatory motifs in C(4) photosynthesis genes using maize and rice leaf transcriptomic data during a process of de-etiolation. *Journal of Experimental Botany* **67**, 5105-5117.
Xu T, Purcell M, Zucchi P, Helentjaris T, Bogorad L. 2001. TRM1, a YY1-like suppressor of rbcS-m3 expression in maize mesophyll cells. *Proc Natl Acad Sci U S A* **98**, 2295-2300.

Author contribution

Jan Emmerling wrote the manuscript, designed and conducted all experiments.

Supplemental Information

Supplemental Figure S1. Comparison of *F. trinervia* and *F. palmeri* upstream flanking sequence GUS expression.

Supplemental Table S1. Oligonucleotides used in this study.

Supplemental Data S1. Sequences of region 3 with annotated features in genbank format. (Enclosed CD only)

Supplemental Data S2. Isolated upstream flanking sequences in genbank format. (Enclosed CD only)

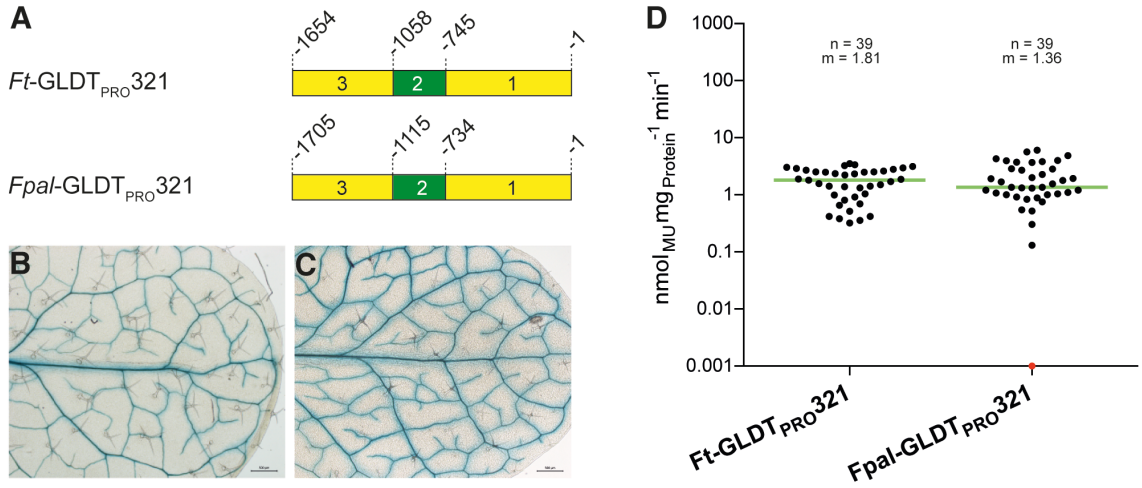


Figure S1 Comparison of GUS expression from *F. trinervia* and *F. palmeri GLDT* upstream flanking sequences. (A) Schematic representation of transformed constructs. **(B – C)** Gus expression of *Ft*-GLDT$_{PRO}$321 (taken from Chapter 1) and *Fpal*-GLDT$_{PRO}$321, respectively, in leaves of transgenic *A. thaliana*. **(D)** Fluorometric measurement of GUS activity. Green bars correspond to median values. Red dots indicate measurements below detection limit. Quantity (n) and median (m) are depicted above the corresponding scatter plot.

Table S3 Oligonucleotides used in this study. Underscores highlight the attached restriction sites indicated by the identifier.

GLDTA-FW1	**ATTTAAGATATAAAATGCAACATTTTA**
GLDTA-FW2	**TTATTTTGTGATATGGTTTGTT**
GLDTB-FW1	**TCATCTGACAATTACATAGATCC**
GLDTB-FW2	**TGGCAGAAAAACTTTGCAATGG**
GLDTB-FW3	**GATGGGTTCTCCCTCTCATC**
GLDT-RV1	**GGAAATAGCCATCACAACTT**
GLDT-RV2	**GATTATGCTGGCTTGTAGTA**
GLDT-RV3	**AAGGGCATTTTGGTAACCGA**
anoR3-FW:SfaAI	**GCGATCGCATTGATGTAGGTTTATGGGA**
anoR1-RV:SgsI	**GGCGCGCCTGTGCTTTATTCTTTAGAAACA**
broR1-RV:SgsI	**GGCGCGCCTGTGCTTTATTCTTTAGGAAAA**
palR3-FW:SfaAI	**GCGATCGCATTGATGTAGGTTTATGGGATGTG**
palR1-RV:SgsI	**GGCGCGCCTGTGCTTTATTCTTTAGAAACAAGC**
angR1-RV:SgsI	**GGCGCGCCTATGCCTTATTCTTTAGAAACAAAC**
robR3-FW:SfaAI	**GCGATCGCATTGATGTAGGTTTATGG**
robR1-RV:XmaI	**CCCGGGTGTGCTTTATACTTCAAAAA**
priR3-FW:SfaAI	**GCGATCGCATTGATGTATGTTTATGG**
priR1-RV:XmaI	**CCCGGGTGTGCTTTATGCTTCAGAAAC**
robR3-RV:XhoI	**CTCGAGAATATTTTTATAGTAAGTAA**
triR2-FW:XhoI	**CTCGAGCACCTACACAGGAATGTTCT**
triR2-RV:BcuI	**ACTAGTGACGAGGAATCTTAAAAACA**
robR1-FW:BcuI	**ACTAGTCACCCACATATGTACAAATT**
angR3-RV:PdmI	**GAATATTTTCTTGTAAAGTAAACATTTAAAAG**
triR2-FW:PdmI	**GAAAATATTCACCTACACAGGAATGTTCTTAGAAAACC**
triR1-RV:XmaI	**CCCGGGTGTGCTTTATTCTTTAGAAACAAGC**
triR3-RV:DraIII	**CACAGGGTGAATATTATTCTTGTAAAGTAACTATTTAAATG**
angR2-FW:DraIII	**CACCCTGTGCCCGCACATGAGAGGG**

3. Dissection of the phosphoenolpyruvate carboxykinase upstream flanking sequence from the C_4 grass *Zoysia japonica*

Introduction

With an expected world population of nine to ten billion people by 2050 and stagnating progress in conventional crop yield improvement, the world is facing a looming food crisis (Baulcombe *et al.*, 2009; Zhu *et al.*, 2010). Correspondingly, new efforts are underway to improve yields of the world's most important crop – rice. By transgenic introduction of C_4 photosynthesis, yields could theoretically be improved by ~50 % (Zhu *et al.*, 2010). C_4 photosynthesis heavily relies on strong, differential gene expression between the leaf's mesophyll cells (MC) and bundle sheath cells (BSC). However, rice lacks closely related C_4 species, which is probably the reason why, so far, no upstream flanking sequence was identified that directs strong expression exclusively in BSC of transgenic rice. In a broad survey for rice BSC promoters, upstream flanking sequences of 27 monocotyledonous candidate genes (composed of endogenous rice genes and C_4-cycle genes of other species) were fused to a reporter gene and expressed in transgenic rice (Karki *et al*., unpublished). Of these candidates only six showed expression at all, mostly in either the whole vascular bundle or vascular tissue alone. Here, in context of this survey, the previously analysed phosphoenolpyruvate carboxykinase (PCK) upstream flanking sequence of *Zoysia japonica* (Nomura *et al.*, 2005) was reassessed. *Zoysia japonica* is a C_4 grass of the Chloridoideae subfamily and presumably operates a rather strong PCK-subtype C_4 cycle, as only low activity of other C_4 decarboxylases were found (Gutierrez *et al.*, 1974). Dissection of the *PCK* upstream flanking sequence and fusion to a ß-glucuronidase (GUS) reporter gene was conducted to identify potential *cis*-regulatory elements that mediate BSC expression in transgenic rice.

Results and Discussion

The upstream flanking sequence of Z. japonica PCK mediates bundle sheath specific expression in rice

Previous analysis of the *Z. japonica PCK* upstream flanking sequence (Acc.-No. AB199899) revealed transcriptional activity in BSC and vascular tissue of transgenic rice (Nomura *et al.*, 2005). Based on this finding, the corresponding sequence was

re-isolated from *Z. japonica* ($ZjPCK_{PRO}$) and used as a control for a broader survey, aiming to identify upstream flanking sequences that confer MC and BSC specific gene expression (Karki *et al.*, unpublished). $ZjPCK_{PRO}$ harbours a substantial amount of polymorphisms compared to the previously analysed sequence (~1.2 %; Supplemental Data S1). Surprisingly, in contrast to the results from Nomura *et al.* (2005), this sequence was found to reliably mediate GUS expression in the BSC of transgenic rice, but not the vasculature (Fig. 1C). Additionally, quantitative analysis also distinguished from Nomura *et al.* (2005), as $ZjPCK_{PRO}$ mediated lower GUS activity (app. one order of magnitude; cmp. Fig. 1B and Nomura *et al.*, 2005). However, direct inference of a regulatory impact on spatial or quantitative expression was not possible, since differences in experimental procedure and particularly the plant transformation vector might have affected the observed expression pattern and activity.

Further comparison with the recently sequenced genome of *Z. japonica* (Tanaka *et al.*, 2016) identified a second PCK gene (*PCK2*), which was not recognized so far (Christin *et al.*, 2009; Nomura *et al.*, 2005). *PCK2* resides approximately 20 kb downstream of PCK1 and exhibits 96 % sequence similarity between the mature transcripts, potentially explaining, why it was not recognized before. *PCK2* is also conserved in the closely related species *Z. matrella* and *Z. pacifica*. However, the high degree of conservation, suggests a rather recent gene duplication event.

Intriguingly, mapping of publicly available RNA-seq data to both copies suggested slightly stronger transcription of *PCK2* compared to *PCK1*, as *PCK2* exhibited ~20 % higher read count (Supplemental Data S2). This is particularly interesting, since the upstream flanking sequence also exhibits a substantial degree of conservation, assuming a similar expression pattern and potentially allowing inference of putative conserved *cis*-regulatory elements (CREs) or the absence of such. Corresponding efforts to isolate PCK2 are underway. Notably, mapping of RNA-seq data also confirmed the existence of a short leader intron in the 5'-UTR of *PCK1* (-45 to -186 bp) and *PCK2* (-39 to -177 bp). Leader introns are known to have an enhancing effect on gene expression (Gallegos and Rose, 2015). However, its effect on PCK1 expression was not addressed in this study.

Additional sequence comparison with PCK genes from other Chloridoid C_4 species revealed only moderate conservation of the upstream flanking sequence (Supplemental Data S3), mainly restricted to the first ~350 bp, i.e. 5'-UTR, leader

intron and the putative minimal promoter around the TATA-Box. Unfortunately, besides the very similar sequences of *Z. matrella* and *Z. pacifica*, no other *PCK* upstream flanking sequences from Chloridoid species of the PCK-subtype are currently publicly available, averting the identification of C_4-related CREs by sequence conservation.

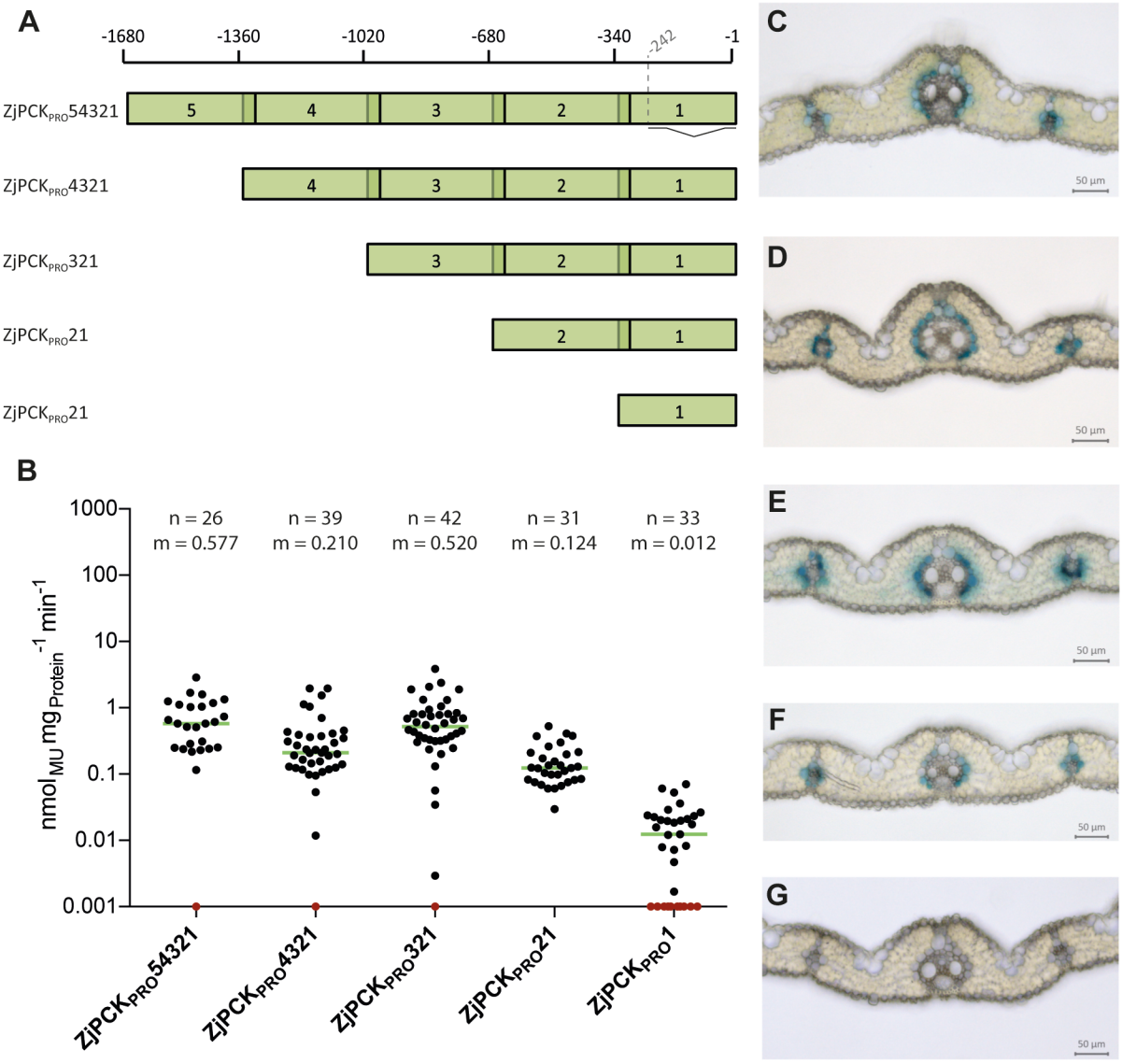


Figure 1 Consecutive 5'-deletion of the PCK1 upstream flanking sequence from *Z. japonica*. The upstream flanking sequence was subdivided into five overlapping regions, which were subsequently deleted and transformed into rice. **(A)** Schematic representation of truncation constructs. Thin line indicates the spliced 5'-UTR and the corresponding transcriptional start site. **(B)** Fluorometric measurements of GUS activity in transgenic rice. Median values are indicated by green lines and stated above (m). Red dots mark values below detection limit. **(C – G)** *In situ* GUS localisation in transversal leaf sections of rice transformed with constructs ZjPCK$_{PRO}$54321, ZjPCK$_{PRO}$4321, ZjPCK$_{PRO}$321, ZjPCK$_{PRO}$21 and ZjPCK$_{PRO}$1, respectively.

5' deletion analysis of the PCK1 upstream flanking sequence

Due to the lack of annotatable conserved regions, the upstream flanking sequence of the *Z. japonica* PCK1 gene was subdivided into five arbitrary overlapping fragments of ~340 to 380 bp, termed regions 1 to 5 in 3'->5' direction. Constructs harbouring consecutive 5'-deletions of these regions were fused to a GUS reporter and transformed into rice, to delimit the position of relevant *cis*-elements (Fig. 1).
In situ GUS localisation and fluorometric quantification showed that delimiting the upstream flanking sequence to ~1020 bp had no effect on spatial GUS expression (Fig. 1B, C - E). However, results of quantitative analysis exhibited significant deviation of construct ZjPCK$_{PRO}$4321, potentially due to differences in the experimental setup (see Material and Methods). Deletion of region 3 (-1017 to -643 bp) led to a ~5-fold decrease of GUS activity, while the spatial expression pattern was maintained (Fig. 1B, 1F), indicating the presence of enhancing CREs in this region. Further deletion of region 2 (-679 to -343 bp) caused a loss of visually detectable GUS expression (Figure 1G), while residual activity was still measurable by fluorometric quantification (Figure 1B). Additional sequence comparison with other grass PCK genes suggested that this loss of cell specific expression was not generated by disruption of the core promoter, as the putative TATA-Box (-286 bp) and conserved flanking sequences reside exclusively in region 1 (Supplemental Figure 1).
These results indicate that region 2 is necessary for BSC specific expression of PCK1.

Excision of region 2 reveals existence of putative mesophyll repressor

To further characterise the function of region 2 on BSC expression, the non-overlapping part of region 2 (-637 to -343 bp) was excised from construct ZjPCK$_{PRO}$321, generating construct ZjPCK$_{PRO}$31 (Figure 2A).
Intriguingly, excision of region 2 not only led to a loss of BSC expression but also to a gain of mesophyll expression (Figure 2B), while GUS activity was significantly reduced to levels between constructs ZjPCK$_{PRO}$321 and ZjPCK$_{PRO}$21 (Figure 2C).

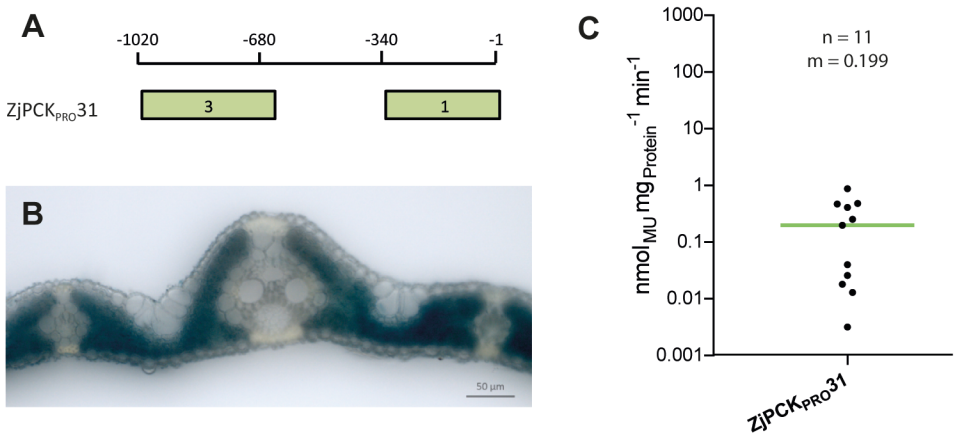


Figure 2 Excision of region 2 of the *PCK1* upstream flanking sequence from *Z. japonica*. Regions 3 and 1 were fused in tandem, deleting the non-overlapping part of region 2. **(A)** Schematic representation of the transformed fusion construct- **(B)** *In situ* GUS localisation of construct ZjPCK$_{PRO}$31 in leaf cross sections of transformed rice plants. **(C)** Fluorometric quantification of FUS activity in transgenic rice. Median values are indicated by green lines and stated above (m).

Since region 1 alone did not produce visibly detectable GUS expression (Figure 1G), further constructs were generated, where region 1 was substituted with the *Cauliflower Mosaic Virus* (*CaMV*) 35S minimal promoter (-60 to -1 bp). Although this minimal promoter was successfully used in rice before (Jeong *et al.*, 2002; Wu *et al.*, 1998; Wu *et al.*, 2000; Yanhai *et al.*, 1997), it did not generate stable expression patterns, when combined with regions 2 and/or 3 of the *Z. japonica PCK1* upstream flanking sequence (Supplemental Figure 2).

These results demonstrate that region 2 must harbour CREs for BSC expression, but apparently also for MC suppression. However, substitution with the *CaMV* 35S minimal promoter indicated that stable functionality depends on presence of region 1 and could be associated with the PCK TATA-Box region, which seemed to be highly conserved throughout the Poaceae (Supplemental Figure 1).

Combination of CRE prediction and C_4 expression profiles identify putative cis-elements for BSC expression

To identify potential CREs, which mediate BSC expression, the upstream flanking sequence from -1017 to -1 bp was searched for matches to known and derived binding sites of rice transcription factors (TFs) using PlantPan2.0 (Chow *et al.*, 2016). This database mainly consists of positional weight matrices (PWMs) from Weirauch

et al. (2014), who used protein-binding microarrays to identify sequence preferences of several thousand plant TFs and extrapolated these on other TFs based on sequence homology of the binding domain. Due to the striking difference of GUS expression patterns from constructs ZjPCK$_{PRO}$321 and ZjPCK$_{PRO}$31, candidate PWMs were identified, which were unique to that part of region 2 excised in construct ZjPCK$_{PRO}$31. Interestingly, only binding sites that corresponded to *Golden2*-like (GLK) and NAC TFs were found. While NAC TFs (comprised of NAM, ATAF and CUC TFs), one of the largest groups of plant TFs, are mainly associated with abiotic stress signalling (Nuruzzaman *et al.*, 2013). GLK TFs belong to the GARP Family and are often associated with C_4 gene expression, since *Golden2* and GLK1 are differentially localised in maize BSC and MC, where they regulate chloroplast development/maintenance and expression of several photosynthesis related genes (Rossini *et al.*, 2001; Waters *et al.*, 2009). Notably, none of the identified binding sites was directly affected by the polymorphisms found between the upstream flanking sequence used by Nomura *et al.* (2005) and the one used here.

Table 1 Candidate transcription factor binding sites unique to excised part of region 2. Putative binding sites were filtered by conserved expression pattern in maize and *S. viridis.* Green and red colour indicates preferential expression in BSC and MC, respectively.

Family	Pos.	Strnd.	Sequence	Rice_ID	Maize orthologs	Maize exp.ratio (BS/M)	Setaria orthologs	Setaria exp.ratio (BS/M)
Myb/SANT	-485	-	taGATTCact	LOC_Os01g08160	GRMZM2G348238 /GLK14	23,50	Si001230m.g	70,68
Myb/SANT	-485	-	taGATTCact					
Myb/SANT	-469	+	aagGAATCaa	LOC_Os02g07770	GRMZM2G052544 /GLK53	158,35	Si017608m.g	41,96
				LOC_Os06g45410	GRMZM2G081671 /GLK34	119,59	Si006865m.g	24,46
				LOC_Os08g25799	GRMZM2G113742 /GLK38	0,32	Si014513m.g	0,37
Myb/SANT	-468	+	taagGAATCa	LOC_Os03g20900	GRMZM2G009060 /GLK26	467,49	Si036390m.g	48,56
NAC;NAM	-457	-	taaCGCAAagt	LOC_Os03g21030	GRMZM2G031001 /NACTF11	0,20	Si030526m.g	0,07
							Si035639m.g	0,62
NAC;NAM	-458	-	ttaACGCAaa	LOC_Os02g36880	GRMZM5G898290 /NACTF40	3,36	Si017567m.g	5,71
				LOC_Os03g21030	GRMZM2G031001 /NACTF11	0,20	Si030526m.g	0,07
							Si035639m.g	0,62

Based on the result that the regulatory mechanism, which directs PCK1 expression to BSC in *Z. japonica,* is obviously conserved in the distantly related C_3 species rice,

it was assumed that the expression pattern of corresponding TFs might also be conserved in other grass species. Hence, the number of putatively binding TFs was further narrowed down by comparison with BSC and MC RNA-seq data of maize (Chang *et al.*, 2012) and *Setaria viridis* (John *et al.*, 2014), reducing the candidates to orthologues, which exhibited a conserved preference for either cell type in both species (Table 1). Notably, all candidate motifs, which can be summarised into three core loci, are in tight vicinity, suggesting competitive or cooperative DNA binding of NAC and GLK TFs.

Additionally, putative binding sites in region 3 and 1 were identified, corresponding to 12 core loci, potentially bound by several TFs of six major classes (Supplemental Table S1). Most abundant binding sites belonged to GATA, bHLH and SBP TFs. Unfortunately, all three are rather large TF classes involved non-preferentially in regulation of most processes in plants (Behringer and Schwechheimer, 2015; Toledo-Ortiz *et al.*, 2003; Wang *et al.*, 2009), not allowing any inference why construct ZjPCK_{PRO}31 mediated MC specific expression.

Conclusion

Nomura *et al.* (2005) found that the upstream flanking sequence of *Z. japonica* PCK1 gene mediates GUS expression in BSCs and vascular tissue of transgenic rice. Re-isolation and transformation of the upstream flanking sequence revealed stable GUS expression in BSCs only, but also substantially lower activity. These differences might be related to either a number of polymorphisms between sequences, experimental procedure or vector backbone. The latter largely affected the distance and nucleotide composition between upstream flanking sequence (including the transcriptional start site) and reporter coding sequence and thus may influence RNA stability and transcriptional/translational efficacy.

Nonetheless, dissection of the upstream flanking sequence delimited the effective promoter to ~1 kb and identified region 2 (~300 bp), which is essential for BSC expression in transgenic rice, but might be dependent on additional downstream elements. Search for known CREs revealed that putative binding sites for NAC and GLK transcription factors are unique to this part of the upstream flanking sequence. The latter are highly associated with differential expression of photosynthesis related

genes in BSC and MC. Consequently, those CREs are priority targets for further dissection of the PCK1 upstream flanking sequence.

Intriguingly, excision of region 2 did not only lead to a loss of GUS expression in BSC, but instead exerted tight restriction to MCs. This suggests that spatial expression of PCK1 might not only be mediated by region 2, but may be subject to combinatorial regulation of several factors.

Material and Methods

Plant transformation and growth

Constructs were transformed into *Agrobacterium tumefaciens* strain LBA4404 by freeze thaw method (Wise *et al.*, 2006). Freshly harvested immature embryos (8-12 days after anthesis) of rice (*Oryza sativa* L.) Nipponbare, a japonica rice variety, were used as explants. *Agrobacterium* mediated transformation of immature embryos was performed following the method described by Hiei and Komari (2006). After one week of co-cultivation and successive resting for 5 days, emerging resistant callus was selected with 50mg/L of hygromycin B added in the tissue culture medium. Regenerated rice plantlets were transferred to hydroponics (Yoshida culture solution, Yoshida et al., 1971) for 2 weeks to acclimatise before transplanting into soil. Two weeks after regeneration transgene integration was confirmed by PCR.

Positively tested plants were transplanted into 0.5 L pots filled with soil. All plants were cultivated within a transgenic containment screen house facility with a day and night time temperature of 25-32 °C and 70-90 % relative humidity. The screen house is located in the grounds of the International Rice Research Institute (IRRI, Los Baños, Philippines-14°9′53.58″S and 121°15′32.19″E). Plants were watered daily and grown in sterilized garden soil containing 0.4 g/L of Osmocote plus 15-9-12 (The Scotts Company Ltd., Thorne UK).

Cloning of reporter gene constructs

In general, all fragments were amplified by PCR, attaching corresponding cloning sites, if required. Fragments were subsequently purified by gel extraction (QIAquick Gel Extraction Kit, QIAGEN) and cloned into pJET1.2 cloning vector (CloneJET PCR Cloning Kit, Thermo Fisher Scientific) or pDONR221 (Invitrogen) for sequencing,

prior to downstream processing. Final fragments were cloned by Gateway cloning (Invitrogen) into a modified pMDC164 expression vector (Curtis and Grossniklaus, 2003), harbouring a maize ubiquitin promoter, instead of the *CaMV* 35S promoter, driving the hygromycin resistance gene. Primers used in this study are summarised in Supplemental Table S2.

Primers Zj.PCK-FW and Zj.PCK-RV were designed on basis of the publicly available sequence of *Z. japonica* PCK gene (Acc.-No. AB199899) and used to amplify the upstream flanking sequence and part of the first exon from genomic DNA, isolated by DNeasy Plant Mini Kit (QIAGEN).

Fragments for consecutive deletion constructs ($ZjPCK_{PRO}54321$ to $ZjPCK_{PRO}1$) were amplified from the isolated PCK clone, using corresponding forward primers (Zj.pPCKR5-attB1 to Zj.pPCKR1-attB1) in combination with the reverse primer Zj.pPCKR1-attB2 and attaching Gateway cloning sites.

Construct $ZjPCK_{PRO}31$ was generated by prior amplification of regions 3 and 1, with primer pairs Zj.pPCKR3-FW:BL/ Zj.pPCKR3-RV:XhoI and Zj.pPCKR1-FW:XhoI/ Zj.pPCKR1-RV:BL, respectively. Both fragments were digested with *Xho*I, ligated and purified. The ligation product was used as template for PCR-attachment of Gateway cloning sites, using primers Zj.pPCKR3-attB1 and Zj.pPCKR1-attB2.

Construct $ZjPCK_{PRO}32\text{-}60$ was generated by prior amplification of regions 3 and 2 in tandem and the *CaMV* 35S minimal promoter (-60 to -1 bp) from non-modified pMDC164 vector, using primer combinations Zj.pPCKR3-FW:BL/Zj.pPCKR2-RV:XhoI and 35Sm-FW:XhoI/ 35Sm-RV:BL, respectively. Both amplicons were digested with *Xho*I, ligated and purified. The ligation product was used as template for PCR-attachment of Gateway cloning site, using primers Zj.pPCKR3-attB1 and 35Sm-attB2. Supplementary constructs $ZjPCK_{PRO}3\text{-}60$ and $ZjPCK_{PRO}2\text{-}60$ were generated accordingly. Using primer pairs Zj.pPCKR3-FW:BL/Zj.pPCKR3-RV:XhoI and Zj.pPCKR2-FW:BL/Zj.pPCKR2-RV:XhoI, regions 3 and 2 were amplified, prior to *Xho*I digestion and ligation with 35S minimal promoter fragment. Primer pairs Zj.pPCKR3-attB1/35Sm-attB2 and Zj.pPCKR2-attB1/35Sm-attB2 were used to attach Gateway cloning sites to the corresponding fragments.

In situ detection of GUS activity and fluorometric measurement

For *in situ* GUS detection, transversal sections of the 5th youngest mature leaf were prepared, focussing around the centre of the proximo-distal axis. GUS staining was conducted as described in Chapter 1.

Fluorometric quantification of GUS activity was conducted of the 6th youngest leaf as described in Chapter 1, with the exceptions of construct ZjPCK$_{PRO}$4321, which was measured in Synergy MX multi-mode microplate reader (BioTek) and construct ZjPCK$_{PRO}$54321, which was measured by discontinuous measurement, as described in Engelmann *et al.* (2008).

Bioinformatic analyses

For RNA-seq read mapping, 10 million paired end reads of the publicly available short read archive DRR016092 were clipped and trimmed using Trimmomatic v0.36 (Bolger *et al.*, 2014). Reads were mapped on extracted loci of PCK1 and PCK2 from *Z. japonica* genomic scaffolds (r1.1; http://zoysia.kazusa.or.jp) using STAR v2.6.0c (Dobin *et al.*, 2013) in two-pass mode.

Prediction of CREs was conducted using PlantPan2.0 (Chow *et al.*, 2016). Candidate CREs were filtered using an in-house Python3 script. Comparison with BSC and MC RNA-seq data of Chang *et al.* (2012) and John *et al.* (2014) was conducted using Python3 and PhytoMine portal (phytozome.jgi.doe.gov/phytomine/).

Validation of statistical significance was performed by using Mann-Whitney-U test, as implemented in Prism v7.0a, with $p \leq 0.05$.

References

Baulcombe D, Crute I, Davies B, Dunwell J, Gale M, Jones J, Pretty J, Sutherland W, Toulmin C. 2009. Reaping the benefits: science and the sustainable intensification of global agriculture.

Behringer C, Schwechheimer C. 2015. B-GATA transcription factors – insights into their structure, regulation, and role in plant development. *Frontiers in Plant Science* **6**, 90.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Chang Y-M, Liu W-Y, Shih AC-C, Shen M-N, Lu C-H, Lu M-YJ, Yang H-W, Wang T-Y, Chen SC-C, Chen SM, Li W-H, Ku MSB. 2012. Characterizing Regulatory and Functional Differentiation between Maize Mesophyll and Bundle Sheath Cells by Transcriptomic Analysis. *Plant Physiology* **160**, 165-177.

Chow C-N, Zheng H-Q, Wu N-Y, Chien C-H, Huang H-D, Lee T-Y, Chiang-Hsieh Y-F, Hou P-F, Yang T-Y, Chang W-C. 2016. PlantPAN 2.0: an update of plant promoter

analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res* **44**, D1154-D1160.
Christin P-A, Petitpierre B, Salamin N, Büchi L, Besnard G. 2009. Evolution of C4 Phosphoenolpyruvate Carboxykinase in Grasses, from Genotype to Phenotype. *Mol Biol Evol* **26**, 357-365.
Curtis MD, Grossniklaus U. 2003. A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol* **133**, 462-469.
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21.
Engelmann S, Wiludda C, Burscheidt J, Gowik U, Schlue U, Koczor M, Streubel M, Cossu R, Bauwe H, Westhoff P. 2008. The gene for the P-subunit of glycine decarboxylase from the C4 species Flaveria trinervia: analysis of transcriptional control in transgenic Flaveria bidentis (C4) and Arabidopsis (C3). *Plant Physiol* **146**, 1773-1785.
Gallegos JE, Rose AB. 2015. The enduring mystery of intron-mediated enhancement. *Plant Sci* **237**, 8-15.
Gutierrez M, Gracen VE, Edwards GE. 1974. Biochemical and cytological relationships in C4 plants. *Planta* **119**, 279-300.
Hiei Y, Komari T. 2006. Improved protocols for transformation of indica rice mediated by Agrobacterium tumefaciens. *Plant Cell, Tissue and Organ Culture* **85**, 271.
Jeong D-H, An S, Kang H-G, Moon S, Han J-J, Park S, Lee HS, An K, An G. 2002. T-DNA Insertional Mutagenesis for Activation Tagging in Rice. *Plant Physiology* **130**, 1636-1644.
John CR, Smith-Unna RD, Woodfield H, Covshoff S, Hibberd JM. 2014. Evolutionary Convergence of Cell-Specific Gene Expression in Independent Lineages of C(4) Grasses. *Plant Physiology* **165**, 62-75.
Nomura M, Higuchi T, Ishida Y, Ohta S, Komari T, Imaizumi N, Miyao-Tokutomi M, Matsuoka M, Tajima S. 2005. Differential expression pattern of C4 bundle sheath expression genes in rice, a C3 plant. *Plant Cell Physiol* **46**, 754-761.
Nuruzzaman M, Sharoni AM, Kikuchi S. 2013. Roles of NAC transcription factors in the regulation of biotic and abiotic stress responses in plants. *Frontiers in Microbiology* **4**, 248.
Rossini L, Cribb L, Martin DJ, Langdale JA. 2001. The Maize Golden2 Gene Defines a Novel Class of Transcriptional Regulators in Plants. *Plant Cell* **13**, 1231-1244.
Tanaka H, Hirakawa H, Kosugi S, Nakayama S, Ono A, Watanabe A, Hashiguchi M, Gondo T, Ishigaki G, Muguerza M, Shimizu K, Sawamura N, Inoue T, Shigeki Y, Ohno N, Tabata S, Akashi R, Sato S. 2016. Sequencing and comparative analyses of the genomes of zoysiagrasses. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* **23**, 171-180.
Toledo-Ortiz G, Huq E, Quail PH. 2003. The Arabidopsis Basic/Helix-Loop-Helix Transcription Factor Family. *Plant Cell* **15**, 1749-1770.
Wang Y, Hu Z, Yang Y, Chen X, Chen G. 2009. Function Annotation of an SBP-box Gene in Arabidopsis Based on Analysis of Co-expression Networks and Promoters. *International Journal of Molecular Sciences* **10**, 116-132.
Waters MT, Wang P, Korkaric M, Capper RG, Saunders NJ, Langdale JA. 2009. GLK Transcription Factors Coordinate Expression of the Photosynthetic Apparatus in Arabidopsis. *Plant Cell* **21**, 1109-1128.
Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJM, Bouget FY, Ratsch G, Larrondo LF, Ecker

JR, Hughes TR. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443.
Wise AA, Liu Z, Binns AN. 2006. Three Methods for the Introduction of Foreign DNA into Agrobacterium. In: Wang K, ed. *Agrobacterium Protocols*. Totowa, NJ: Humana Press, 43-54.
Wu C-Y, Suzuki A, Washida H, Takaiwa F. 1998. The GCN4 motif in a rice glutelin gene is essential for endosperm-specific gene expression and is activated by Opaque-2 in transgenic rice plants. *The Plant Journal* **14**, 673-683.
Wu C-Y, Washida H, Onodera Y, Harada K, Takaiwa F. 2000. Quantitative nature of the Prolamin-box, ACGT and AACA motifs in a rice glutelin gene promoter: minimalcis-element requirements for endosperm-specific gene expression. *The Plant Journal* **23**, 415-421.
Yanhai Y, Lili C, Roger B. 1997. Promoter elements required for phloem‐specific gene expression from the RTBV promoter in rice. *The Plant Journal* **12**, 1179-1188.
Yoshida S, Forno DA, Cock JH. 1971. *Laboratory manual for physiological studies of rice*. Los Baños, Philippines: The International Rice Research Institute.
Zhu X-G, Long SP, Ort DR. 2010. Improving Photosynthetic Efficiency for Greater Yield. *Annual Review of Plant Biology* **61**, 235-261.

Author contribution

Jan Emmerling – wrote the manuscript, designed and conducted the experiments, with the exceptions stated below.
Shanta Karki – conducted the transformation of rice and wrote the corresponding paragraph in methods section.
Stefanie Schulze – isolated the sequence of PCK1; designed, cloned and analysed construct ZjPCK$_{PRO}$54321; modified the pMDC expression vector.
Janine Meyer – analysed constructs ZjPCK$_{PRO}$31, ZjPCK$_{PRO}$32-60, ZjPCK$_{PRO}$3-60 and ZjPCK$_{PRO}$2-60.

Supplemental Information

Supplemental Figure S1 Sequence alignment of *PCK1* region 1 from different grass species.

Supplemental Figure S2. GUS localisation of regions 3 and 2 in combination with 35S minimal promoter.

Supplemental Data S1. Sequence alignment of AB199899 and the re-isolated sequence (ZjPCK) in FASTA format. (Enclosed CD only)

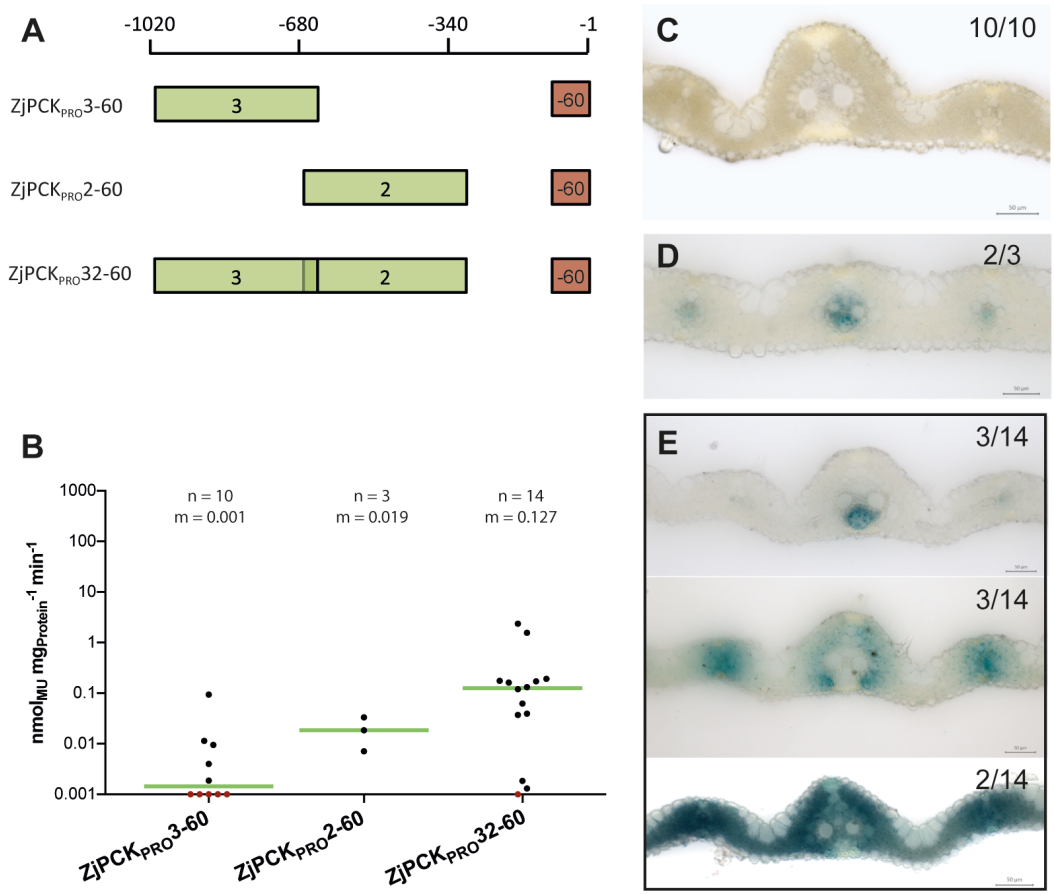
Supplemental Data S2. RNA-seq read-mapping for the extracted PCK1 and PCK2 loci, in BAM format. (Enclosed CD only)

Supplemental Data S3. Alignment of PCK1 upstream flanking sequences from different chloridoid grass species in FASTA format. (Enclosed CD only)

Supplemental Table S1. Candidate CREs and putative orthologues unique to regions 3 and 1. (Enclosed CD only)

Supplemental Table S2. Oligonucleotides used in this study.

Supplemental Figure S1 Alignment of region 1 from different grass species. Available *PCK1* sequences from BEP and PACMAD grass species were aligned to identify conserved elements. Positions conserved across all sequences are highlighted in clustal colours.



Supplemental Figure S2 GUS localisation of regions 3 and 2 in combination with 35S minimal promoter. **(A)** Schematic representation of transformed constructs. Red area represents the *CaMV* 35S minimal promoter (-60 to -1 bp). **(B)** Fluorometric quantification of GUS activity. Median values are indicated by green lines and stated above (m). Red dots mark values below detection limit. **(C – E)** *In situ* localisation of GUS activity in transversal leaf sections of rice transformed with constructs ZjPCK$_{PRO}$3-60 ZjPCK$_{PRO}$2-60 and ZjPCK$_{PRO}$32-60 respectively. Numbers in the upper right corner indicate, how many of the transgenic lines exhibited the displayed expression pattern. Unmentioned lines showed no expression. Note that construct ZjPCK$_{PRO}$32-60 did generate a variety of expression patterns (E).

Supplemental Table S2 Oligonucleotides used in this study. Underscores highlight attached cloning sites as indicated in the corresponding identifiers.

Zj.PCK-FW	GTCGACAACTTATTTTTGAGACCGGAG
Zj.PCK-RV	CAGCACGCCCAAGGACAGCGCCGCC
Zj.pPCKR1-attB2	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCGCCGGCGCGCGTGCGGCACG
Zj.pPCKR5-attB1	GGGGACAAGTTTGTACAAAAAAGCAGGCTGACAACTTATTTTTGAGAC
Zj.pPCKR4-attB1	GGGGACAAGTTTGTACAAAAAAGCAGGCTTTTCCTATGGTGTGTTCTTCG
Zj.pPCKR3-attB1	GGGGACAAGTTTGTACAAAAAAGCAGGCTATTGGCCAAGTTGCAAGATC
Zj.pPCKR2-attB1	GGGGACAAGTTTGTACAAAAAAGCAGGCTTTTGATCATTCCAGAGAGTT
Zj.pPCKR1-attB1	GGGGACAAGTTTGTACAAAAAAGCAGGCTACCGTCACCGCCGGATGG
Zj.pPCKR1-FW:XhoI	AAACTCGAGACCGTCACCGCCGGATGG
Zj.pPCKR1-RV:BL	CTCGCCGGCGCGCGTGC
Zj.pPCKR3-RV:XhoI	AAACTCGAGTAGATTTTGGGCTGACCCAA
Zj.pPCKR3-FW:BL	TTGGCCAAGTTGCAAGATCTAAAACTTCC
Zj.pPCKR2-RV:XhoI	AAACTCGAGCCTAACCGTGGTTACATTC
Zj.pPCKR2-FW:BL	TTTGATCATTCCAGAGAGTTTT
35Sm-FW:XhoI	AAACTCGAGCCCACTATCCTTCGCAAG
35Sm-RV:BL	TCCTCTCCAAATGAAATGAACTTCCT
35Sm-attB2	GGGGACCACTTTGTACAAGAAAGCTGGGTATCCTCTCCAAATGAAATGAA

4. Knockdown of potential negative Kranz anatomy regulators in rice

Introduction

Approximately 2 % of all known plant species conduct C_4 photosynthesis, bypassing photorespiration and thus maximising CO_2 salvage. For this most C_4 plants conduct a two-celled C_4-cycle, spatially confining CO_2 prefixation by phosphoenolpyruvate carboxylase (PEPC) and final assimilation by 1,5-ribulosebisphosphate carboxylase/oxygenase (Rubisco) to the leaf mesophyll cells (MCs) and bundle sheath cells (BSCs), respectively. To operate high fluxes of these processes, most C_4 species rely on a specialised leaf morphology, called Kranz anatomy, in which both cell types are tightly connected in an approximate 1:1 ratio, forming repetitions of the characteristic transversal pattern Vein–BSC–MC–MC–BSC–Vein. Typically, these BSCs are larger than their C_3 cognates and comprise several likewise enlarged chloroplasts and mitochondria, depending on the C_4 subtype (Edwards and Voznesenskaya, 2011). Evidence has accumulated that the SCARECROW/SHORTROOT (SCR; SHR) pathway (Slewinski *et al.*, 2014; Slewinski *et al.*, 2012), GOLDEN2 (G2) and G2-like (GLK) transcription factors (TFs) are involved in determining Kranz anatomy (Rossini *et al.*, 2001; Waters *et al.*, 2008), but the entire process is poorly understood. However, its recurrent evolution in almost 60 plant lineages (Sage, 2016) suggests that i) only few genetic leverage points for evolution exist to adapt the leaf anatomy and ii) these leverage points are already existent in C_3 lineages and need not to evolve *de novo*.

To cope with the emerging food crisis that the world is facing, efforts are undertaken to introduce C_4 photosynthesis into C_3 crops. In order of this effort, Wang *et al.* (2013) conducted transcriptome sequencing on maize foliar and husk leaf primordia. The latter are leaf like organs that surround the female inflorescence and are distinguished from foliar leaves by the absence of Kranz anatomy features and cell specific C_4 gene expression (Langdale *et al.*, 1988). The authors found differential expression of 494 genes that were further filtered by gene ontology terms associated with regulatory functions and their expression pattern across a maize leaf gradient. This analysis resulted in 71 genes that were exclusively upregulated in foliar leaf primordia (i.e. potential positive regulators of Kranz anatomy) and 23 that were exclusively upregulated in husk leaf primordia (potential negative regulators).

The assumption of negative regulators, at a first glance, implies the unlikely event that Kranz anatomy would have been the default state, from which C_3 anatomy arose

several hundred times (Fouracre *et al.*, 2014). However, Kranz anatomy is not a novel trait but a complex one, formed by several individual traits, such as size and number of veins, cells and organelles. Traits, which also exhibit plasticity in C_3 plant lineages and are likely dependent on tight regulation of hormonal signalling, developmental programs and even environmental cues (Krogan and Long, 2009). Consequently, any gene with a suppressive function in these processes can also be considered a negative regulator of Kranz anatomy.

To analyse the impact of these potential regulators on C_3 leaf anatomy, they were further filtered by orthology and expression in rice, resulting in 60 positive candidates that were ectopically expressed in rice (Wang *et al.*, 2017) and 18 negative candidates. It was the aim of this study to knock down orthologues of these 18 negative candidates (Table 1) by RNA interference (Fire *et al.*, 1998) in rice.

Table 1 List of putative negative Kranz anatomy regulators. Candidates were identified by a comparative study on maize foliar and husk leaf primordia (Wang *et al.*, 2013).

LAB ID	Maize Gene ID	Rice Orthologue ID	AA simil-arity (%)	RPKM Rice Seedling	Gene Family
JL83	GRMZM2G328742	LOC_Os12g06080	66	318.262	AP2/RAV/B3 TF (RAV2 like)
JL84	GRMZM2G137541	LOC_Os09g29830	67	685.855	bHLH TF
JL85	GRMZM2G180406	LOC_Os01g68700	84	25.585	bHLH TF
JL86	GRMZM5G851485	LOC_Os03g19780	63	181.626	bHLH TF
JL87	GRMZM2G077124	LOC_Os07g48660	49	363.435	bZIP TF (similar to ABA-insensitive)
JL88	GRMZM2G171600	LOC_Os01g69910	54	218.064	CAMTA TF (ethylene induced)
JL89	GRMZM2G421033	LOC_Os04g55520	50	189.556	DRE binding protein (AP2 TF)
JL91	GRMZM2G140694	LOC_Os02g47810	68	265.575	Dof Zinc Finger
JL92	GRMZM2G062244	LOC_Os07g39320	50	251.148	HD-Zip TF
JL93	GRMZM2G132367	LOC_Os03g08960	68	669.957	HD-Zip TF
JL94	GRMZM2G412430	LOC_Os03g27390	57	993.335	bHLH TF
JL95	GRMZM2G005155	LOC_Os04g52410	80	0.898	MADS-box TF
JL96	GRMZM2G181030	LOC_Os04g49450	46	211.799	Myb TF
JL97	GRMZM2G003715	LOC_Os08g06140	63	889.887	NAC/NAM TF
JL98	GRMZM2G065451	LOC_Os02g04680	49	126.547	SBP TF
JL99	GRMZM2G371033	LOC_Os01g69830	66	288.534	SBP TF
JL100	GRMZM2G000842	LOC_Os05g37170	69	130.076	bZIP TF (TGA6)
JL110	GRMZM5G897473	LOC_Os09g37520	82	134.589	bZIP TF

Results

To determine whether the identified genes have a negative effect on Kranz anatomy features, 500 bp fragments of the mature transcripts were synthesized and cloned into the plasmid pANIC 8B (Mann *et al.*, 2012), utilising the *ZmUbi1* promoter (Christensen *et al.*, 1992; Cornejo *et al.*, 1993) to express hairpin-RNAs (hpRNAs) of the corresponding fragment and thus induce post-transcriptional gene silencing (PTGS) of the target genes (Fire *et al.*, 1998; Fusaro *et al.*, 2006). As a proof of concept, quantitative real-time PCR (qRT-PCR) was conducted on total leaf cDNA from seven independent lines of the first construct transformed (JL97), resulting in ~25 to 55 % residual target gene mRNA, compared with transgenic lines harbouring no hpRNA (Fig. 1).

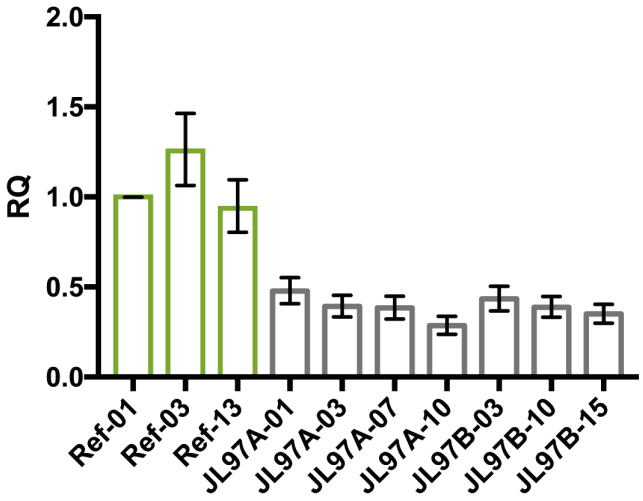


Figure 1 Relative transcript abundance of JL97 in independent knock down lines. Real-time PCR was conducted on cDNA from mature leaves of seven independent knock down lines and three independent reference lines, transformed with selection markers only (Ref-01, Ref-03 and Ref-13). Transcript levels were calculated as relative quantity (RQ) of Ref-01, as it represents the median mRNA abundance of the three reference lines. Whiskers correspond to the minimal and maximal transcript levels over three technical replicates.

In order to allow a rapid macroscopic assessment of several Kranz anatomy related features in the transgenic lines, the first subset of candidates was transformed with an additional chloroplast-targeted yellow fluorescent protein (cTP:YFP), under control of the ZjPCK1 promoter (Chapter 3; Nomura *et al.*, 2005), to label the bundle sheath. However, a fluorescent signal was neither detectable in young nor mature leaves (data not shown). Similarly, cytosolic expression of a fluorescent protein, utilising the ZjPCK1 promoter, did also not generate visable labeling of the bundle sheath (Sarah Covshoff, personal communication). Thus, the initial characterisation of knock down

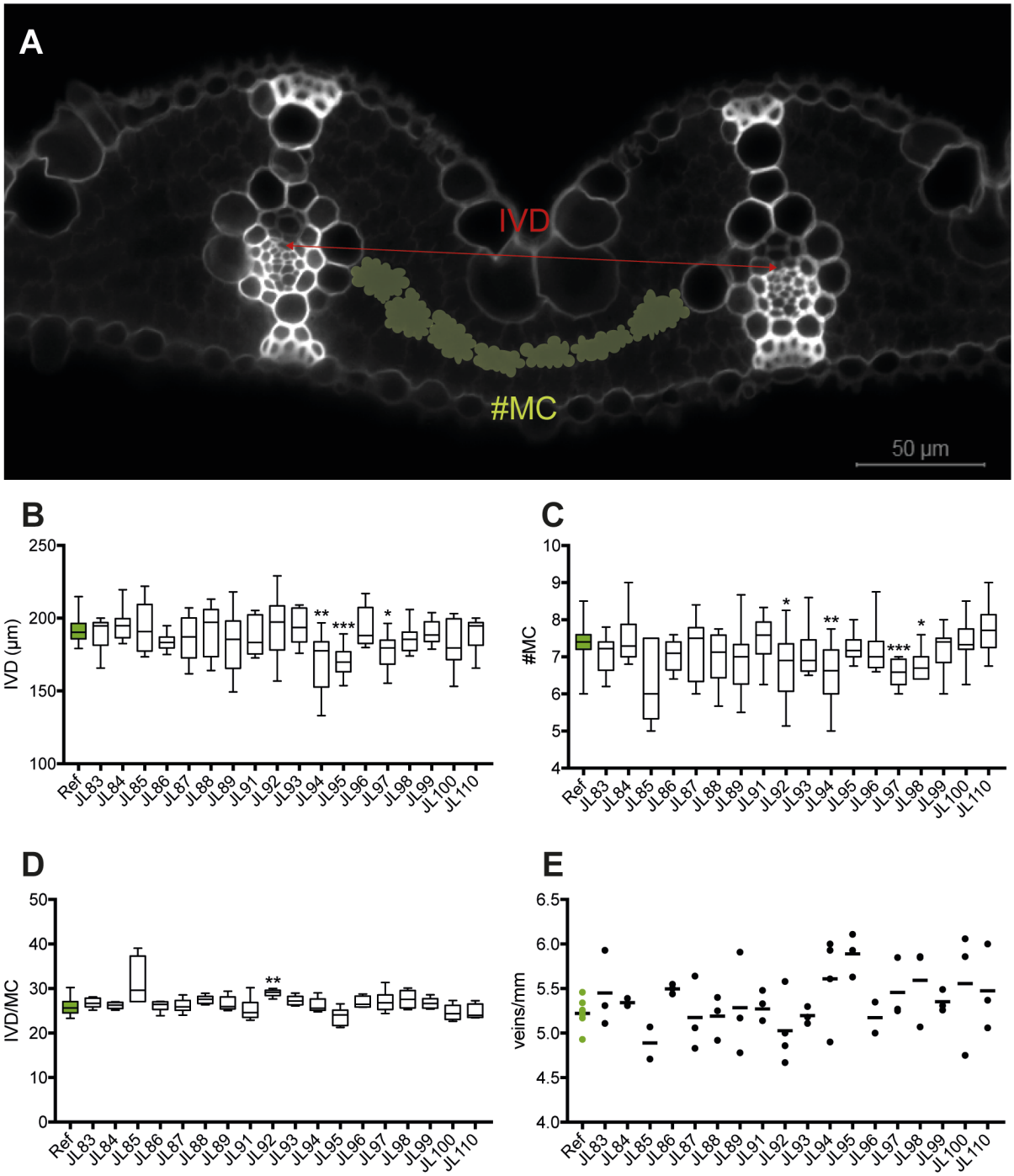


Figure 2 Measurement of Kranz anatomy related parameters in T0 rice knock down lines. (A) Example for the measurement of interveinal distance (IVD) and interveinal mesophyll cells (#MC). **(B – C)** Measurements of IVD and #MC. Data points represent averages over all minor veins between the 1st and 2nd lateral vein on both sides of the leaf. For each candidate at least 2 transgenic lines were measured in two technical replicates. **(D)** Ratio of IVD per MC. Data points were acquired from (B) and (C) and averaged for each replicate. Boxes correspond to the 25th to 75th percentile and internal horizontal lines to median values. Whiskers mark the minima and maxima. **(E)** Ratio of veins per millimetre. Leaf width was measured once for each transgenic line, prior to cross sectioning. The number of veins was acquired from full width cross sections. All data sets were compared to a set of T0 reference lines, transformed with the same vector backbone, but depleted of the hpRNA expression unit (highlighted in green). Horizontal lines indicate mean values. Significant deviation from the reference was assessed by Kruskal-Wallis-Test and marked by one to four asterisks, which correspond to p-values of ≤ 0.0332, ≤ 0.0021, ≤ 0.0002 and ≤ 0.0001, respectively.

lines was limited to measuring interveinal distance (IVD), the number of MC between veins (#MC; Fig. 2A) and veins/mm of individual plants.

For each candidate two cross sections of at least two independent T0 plants were analysed. The average IVD and #MC were acquired over all minor veins between the 1st and 2nd lateral vein for both sides of the leaf. The obtained data were compared to a reference set of eight independent T0 plants, transformed with a control vector, harbouring only selection markers.

All transgenic lines analysed, including the reference, exhibited substantial variation over IVD and #MC, hampering clear identification of knock down phenotypes (Fig. 2B and C). However, statistical analysis indicated that IVD was significantly lower in knock down lines of candidate JL94 ($p = 0.002$), JL95 ($p < 0.001$) and JL97 ($p = 0.015$). A corresponding decrease in #MC was only observed for candidates JL94 ($p = 0.001$) and JL97 ($p < 0.001$), but otherwise also found for candidates JL92 ($p = 0.029$) and JL98 ($p = 0.027$).

To address whether knock down of candidate genes affected the size of the vascular bundle (VB) or MC, the ratio of IVD per MC was calculated (Fig. 2D). IVD and #MC correlated well for all candidates except JL85 and JL92. Both showed an increased IVD/MC ratio, but only JL92 significantly distinguished from the reference ($p = 0.006$), suggesting that either MC or VB size was increased.

Assessment of the number of veins per mm indicated high deviation from the reference for several candidates, but non were statistically significant, due to the low

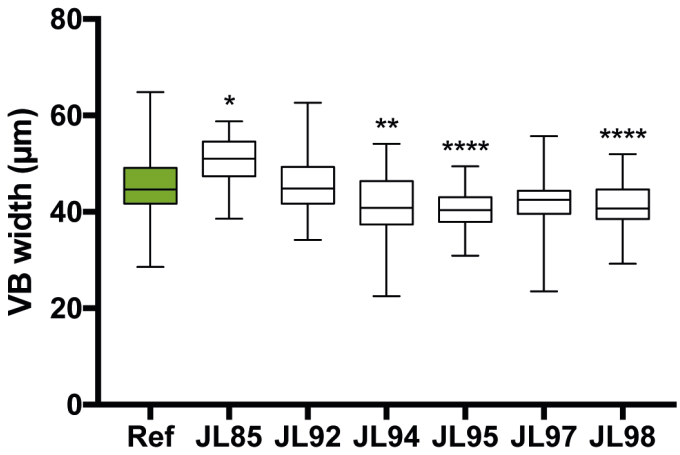


Figure 3 Width of minor vein vascular bundles (VB). VB width was measured of all minor veins between the 1st and 2nd lateral vein from ≥ 2 transgenic lines. Data was compared to the set of T0 reference lines (highlighted in green). Significant deviation from the reference was assessed by Kruskal-Wallis-Test and marked by one to four asterisks, which correspond to p-values of ≤ 0.0332, ≤ 0.0021, ≤ 0.0002 and ≤ 0.0001, respectively. Boxes correspond to the 25th to 75th percentile and internal horizontal lines to median values. Whiskers mark the minima and maxima.

sampling rate. Nonetheless, from all candidates JL85 exhibited the lowest and JL95 the highest vein density (Fig. 2E).

Additionally, the width of vascular bundles was measured for all candidates, which exhibited significantly reduced IVD or #MC and JL85, (Fig. 3). While JL85 lines showed a 12 % increase in VB width ($p = 0.033$) JL94, JL95 and JL98 exhibited significantly decreased VB width ($p = 0.002$ for JL94 and $p < 0.0001$ for JL95 and JL98) by approximately 10 %.

Knock down of some candidates also induced macroscopic phenotypes, but they were mostly uncorrelated to leaf anatomy. As such, JL86 and JL95 exhibited early floral induction, while JL87 flowered very late. The latter further distinguished by its small, bushy stature and pale colour. JL100 showed impaired regeneration from tissue culture and maintained severe necrosis in most leaves.

All significant results and the macroscopic phenotypes are summarised in Table 2.

Table 2 Summary of significant Kranz anatomy related measurements and macroscopic phenotypes. Asterisks mark significant deviation (see Figures 2 and 3 for detail). Median values that deviated from the reference by more than 10 % are indicated by '+' or '–'.

	#MC	IVD	IVD/MC	V/mm	VBw	# Lines analysed	Other
JL85	–	0	+	–	+*	2	None
JL86	–	–	0	+	n/a	3	Early flowering
JL87	0	0	0	0	n/a	3	Late flowering; Small stature; Pale; Bushy
JL92	–*	0	+**	–	0	4	None
JL94	–**	–**	0	+	–**	4	None
JL95	0	–***	–	+	–****	3	Early flowering
JL97	–***	–*	+	+	0	3	None
JL98	–*	0	+	+	–****	3	None
JL100	0	–	–	+	n/a	3	Maintained necrotic leaves

Discussion

This study aimed to identify the major negative regulators of Kranz anatomy in maize leaves by hpRNA-induced PTGS of candidate orthologues in rice. The list of candidate genes originated from a comprehensive comparative RNA-seq study of maize husk and foliar leaf primordia (Wang *et al.*, 2013). The developmental trajectory of maize husk and foliar leaf venation is well described (Bosabalidis *et al.*,

1994; Esau, 1943; Langdale *et al.*, 1989; Langdale *et al.*, 1988; Nelson and Dengler, 1997; Sharman, 1942; Wang *et al.*, 2013). In both the central vein is initiated in plastochron 1 (P1), lateral veins form in P2, while minor veins, which are responsible for the high density in foliar leaves, initiate in P4 and already exhibit typical Kranz structures in P5. Since husk leaves show only very few minor veins, negative regulators of Kranz anatomy are expected to be significantly stronger expressed during early husk leaf plastochrons, while positive regulators would be increased during corresponding foliar leaf plastochrons. As major regulators of early leaf development both are assumed to show decreased expression in later developmental stages (Wang *et al.*, 2013).

As a proof of concept, qRT-PCR showed that hpRNA expression successfully reduced the transcript levels of JL97 by ~45 to 75 % (Fig. 1). Similarly, gel blot and qRT-PCR analysis confirmed ectopic expression of positive regulator candidates (Wang *et al.*, 2017). In both studies the UBIQUITIN1 promoter of maize (*ZmUbi1*) was used for transgene expression (Christensen *et al.*, 1992; Cornejo *et al.*, 1993).

Correspondingly, both approaches were able to induce phenotypes. In knock down lines of JL100, maintained leaf necrosis was observed, after the plants were transferred to soil. JL100 is an orthologue of the TF TGA6, which positively regulates systemic acquired resistance in Arabidopsis (Zhang *et al.*, 2003). Although plants from all lines exhibited leaf necrosis after transfer to soil, they usually recovered within two weeks and produced healthy leaves. Contrarily, knock down lines of JL100 consistently developed new necrosis in younger leaves. This indicates that plants were unable to defend against secondary pathogen infections, acquired after transfer from sterile tissue culture to non-sterile soil. Further, JL87 encodes an ABSCISIC ACID-INSENSITIVE 5-like protein and its knock down produced a small, bushy, late flowering phenotype, characteristic for high levels of abscisic acid (ABA). This suggests that JL87 is a repressor of ABA signalling or biosynthesis.

Like the knock down of negative regulators, ectopic expression of positive regulator candidates also induced phenotypes in some cases, but only affected shoot development, root development or secondary cell wall formation (Wang *et al.*, 2017).

However, in both studies the majority of candidates did not induce severe phenotypes, particularly none that affected vein patterning. Thus, raising the question, why knock down or ectopic expression of candidate genes, which are

primarily expressed during the stage of leaf venation patterning, was not able to induce severe changes to this process?

Although the most parsimonious assumption, it seems unlikely that all candidates were falsely identified. The developmental trajectory of Kranz anatomy in maize has been comprehensively analysed over the last century and filtration of candidate genes was designed accordingly, but loose enough to co-identify early and late regulatory genes. Samples were pooled from corresponding plastochrons of several hundred primordia, to prevent artificial bias by PCR amplification steps and reduce background from individual samples. Consequently, the filtration of candidates also identified known regulators of leaf venation. Among the positive regulator candidates were SCARECROW1 (SCR1), one orthologue of DEFECTIVELY ORGANIZED TRIBUTARIES 5 (DOT5) and SHORTROOT orthologues. All are already known to be involved in the regulation of vein development (Petricka *et al.*, 2008; Slewinski *et al.*, 2014; Slewinski *et al.*, 2012). However, the only known negative regulator, AtMYC2 was not identified. AtMYC2 was recently shown to negatively regulated the tryptophan-dependent auxin synthesis in Cleomaceae and thus control vein-density (Huang *et al.*, 2017). The orthologue of AtMYC2 (GRMZM2G001930) showed the expected expression profile, but exhibited high deviation between technical replicates of one primordial stage, hence it did not pass the selection criteria (Wang *et al.*, 2013). Several of the potential regulators were further supported by at least two other RNA-seq studies (summarised in Huang and Brutnell, 2016), including JL93, which was identified in three other studies of maize and *Gynandropsis gynandra* (Aubry *et al.*, 2014; Li *et al.*, 2010; Tausta *et al.*, 2014). Of course, selection criteria could have been further loosened to identify all genes involved, but not without increasing false-positives to numbers unfeasible for individual phenotypic assessment. Again, this emphasises the benefit of a markable rice bundle sheath to allow rapid macroscopic phenotyping in large-scale.

Since none of the candidates was able to induce severe perturbation of leaf venation, although they are supported by multiple RNA-seq studies and co-identification of known regulators, the likeliest explanation suggests functional redundancy between candidates or high levels of post-transcriptional regulation. Quantitative real-time PCR of JL97 already indicated that even in the reference lines' endogenous mRNA levels fluctuate from ~80 to 145 % (Fig. 1). This and the lack of phenotypic effects by

hpRNA-induced PTGS, down to levels of ~25 to 55 %, indicate that the regulatory system controlling vein patterning is highly robust against variation in mRNA levels and may be primarily regulated by translational or post-translational mechanisms. Further, functional redundancy between several candidates and maybe even mutual regulation might additionally protect the regulatory system against fluctuating mRNA levels to secure proper leaf development. Consequently, the impact of PTGS may be further reduced and thus induce only subtle phenotypes that cannot be detected in the T0 generation. Plants of T0 often deviate in recovery from tissue culture, i.e. some plants handle the transfer to soil better than other. Additionally, tissue culture (which involves application of auxin, the major phytohormone in vascular development) and antibiotic selection itself might induce phenotypes, not observed in soil grown wild type plants. Hence, a broad set of transgenic lines was used as reference, compensating potential artefacts from tissue culture but also increasing phenotypic variance and thus hampering clear identification of subtle phenotypes.

One additional factor might have had an impact on the outcome of this study and of the ectopic expression of positive candidates by Wang *et al.* (2017). Both studies utilised the widely used *ZmUbi1* promoter for ectopic expression or knock down of candidate genes. *ZmUbi1* is preferably used for transgene expression in rice, since it was reported to mediate substantially higher expression in monocot species than the *Cauliflower Mosaic Virus* 35S promoter (Christensen *et al.*, 1992; Cornejo *et al.*, 1993; Gallo-Meagher and Irvine, 1993; Gupta *et al.*, 2001). While true for callus-derived protoplasts of rice, spatial GUS expression from *ZmUbi1* in whole rice plants showed that the expression pattern is far from being ubiquitous. Its activity in leaves is generally weaker than in protoplasts and only detectable in vascular tissue and stomata (Cornejo *et al.*, 1993). Although reduction of JL97 mRNA levels in mature leaves (Fig. 1) indirectly showed that the promoter is active, its extent was never assessed in leave primordia, where venation patterning actually takes place. Additionally, candidate mRNA levels are significantly higher in primordia, often by several orders of magnitude (Wang *et al.*, 2013). Assuming that activity of the *ZmUbi1* promoter in primordia is as low as in the rest of the leaf, reduction of candidate mRNA levels in primordia might be substantially less than indicated in Figure 1 and thus could explain the subtle phenotypes.

Conclusion and outlook

The knock down of 18 potential Kranz anatomy regulator orthologues in rice did not induce severe changes to relevant measurements in the T0 generation, similar to the ectopic expression of 60 potential positive regulators. This indicates that either the regulatory network controlling venation patterning is buffered against mRNA fluctuation or that the *ZmUbi1* promoter used for target gene and hpRNA expression is not sufficiently active in leaf primordia. Consequently, qRT-PCR on leaf primordia of knock down lines will be conducted to address the extent of PTGS during venation patterning. However, subtle changes to interveinal distance, number of interveinal mesophyll cells and vascular bundle size were observed, but closer characterisation suffered from high deviation of tissue culture-grown plants and has to be assessed in T1. As such, candidates summarised in Table 2 will be of particular interest. To address potential redundancy between candidates, re-assessment of the candidate's expression profiles might reveal tightly co-regulated genes, indicating a redundant role, which could be analysed by crossing of corresponding knock down lines or co-transformation. This study indicates that, despite its convergent evolution, genetic regulation of Kranz anatomy development might be more complex than previously thought and, in the long term, methods to identify or prioritise candidates and subsequently assess their phenotypic output might need further improvement.

Material and Methods

Generation of transformation constructs

For each rice orthologue in Table 1, the first 500 bp of the coding sequence were synthesised at GenScript USA Inc. with flanking Gateway attL cloning sites, cloned by Gateway LR reaction (Invitrogen) into pANIC 8B plant transformation vector (Mann *et al.*, 2012) and confirmed by digestion with *Hind*III and *Xho*I. The pANIC 8B vector contains two Gateway cloning cassettes in inverse orientation, separated by a short linker sequence, to allow hpRNA formation of the transcribed sequence. Notably, due to the double-sided LR reaction, inversion of the whole hpRNA transcriptional unit was observed on some occasions, resulting in an alternative restriction pattern.

For the production of rice reference lines, a control construct (pANIC 8B-C) was generated, by deletion of the hpRNA expression cassette. Therefore, pANIC 8B was digested with *Sac*I and *Asc*I. The 5'- and 3'-ends were blunted with the blunting enzyme comprised in the CloneJet PCR Cloning Kit (Thermo Fischer Scientific) and re-ligated.

Plant transformation and growth

Transformation of the rice cultivar Kitaake was performed on callus derived from mature rice seeds, using *Agrobacterium tumefaciens* strain AGL1 (Lazo *et al.*, 1991). Callus induction, subsequent seedling selection and regeneration was conducted according to a modified protocol from Toki *et al.* (2006), available from https://langdalelab.files.wordpress.com/2015/07/kitaake_transformation_2015.pdf.
Hygromycin resistant T0 plantlets were transferred to the green house, but kept in liquid culture (Yoshida culture solution, Yoshida *et al.*, 1971) for two weeks to allow acclimatisation. Afterwards, plantlets were checked by PCR for transgene integration and transferred to soil (John Innes Compost No. 2). The green house is located in Düsseldorf, Germany. Day/night temperature was maintained at 30/22 ± 3 °C with a diurnal light cycle of 16 h light (supplemented to ~300 µM m^{-2} sec^{-1}) and 8 h dark.
To distinguish PTGS-induced phenotypes from generally impaired plant regeneration, each batch of knock down constructs was accompanied by transformation of the hpRNA-depleted control construct (pANIC 8B-C). Plantlets generated from this control were also used as reference lines for the phenotypic analysis.

Quantitative real-time PCR

For the quantitative assessment of PTGS, total RNA was isolated from mature leaves of JL97 and reference lines, using RNeasy Plant Mini Kit (QIAGEN). Complementary DNA was synthesised using QuantiTect Reverse Transcription Kit (QIAGEN). Quantitative real-time PCR was conducted using KAPA SYBR FAST qPCR Kit (KAPABIOSYSTEMS Inc.) on a 7500 Fast Real-Time PCR System (Applied Biosystems) according to the manufacturers recommendations. A 100 bp amplicon was generated from JL97 coding sequence using primers JL97-qRT-FW (TTGATGGCAATATGAATCCTCG) and JL97-qRT-RV (TACTGAGCCTTTGAT-GTTGTTG). As endogenous control UBIQUITIN5 was chosen and amplified

according to Jain *et al.* (2006). ΔΔCt values were calculated as described in Livak and Schmittgen (2001) and primer efficiencies were included accordingly.

Phenotypic analysis

Phenotypic analysis was conducted on T0 plants approximately four weeks after transfer from liquid culture to soil. For each candidate the expanded 6th leaf of the first tiller from 2 – 4 plants was analysed. Width was measured at the centre of the leaf blade and cross sections were prepared by hand and stained with methylene blue. Two full-width sections were analysed for each plant and total number of veins was counted and averaged. Interveinal distance and the number of interveinal mesophyll cells were assessed of all minor veins between the first and second lateral vein on both sides of the leaf and averaged. Photographs and measurements were conducted using Zen 2012 software (Zeiss) and an Axio Imager.M2 microscope (Zeiss). Measurements were compared to a dataset of eight reference lines by non-parametric Kruskal-Wallis test, as implemented in Prism 7 (GraphPad Software).

References

Aubry S, Kelly S, Kumpers BM, Smith-Unna RD, Hibberd JM. 2014. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis. *PLoS Genet* **10**, e1004365.

Bosabalidis AM, Evert RF, Russin WA. 1994. Ontogeny of the Vascular Bundles and Contiguous Tissues in the Maize Leaf Blade. *Am J Bot* **81**, 745-752.

Christensen AH, Sharrock RA, Quail PH. 1992. Maize polyubiquitin genes: structure, thermal perturbation of expression and transcript splicing, and promoter activity following transfer to protoplasts by electroporation. *Plant Mol Biol* **18**, 675-689.

Cornejo MJ, Luth D, Blankenship KM, Anderson OD, Blechl AE. 1993. Activity of a maize ubiquitin promoter in transgenic rice. *Plant Mol Biol* **23**, 567-581.

Edwards GE, Voznesenskaya EV. 2011. Chapter 4 C4 Photosynthesis: Kranz Forms and Single-Cell C4 in Terrestrial Plants. In: Raghavendra AS, Sage RF, eds. *C4 Photosynthesis and Related CO2 Concentrating Mechanisms*. Dordrecht: Springer Netherlands, 29-61.

Esau K. 1943. Ontogeny of the vascular bundle in Zea mays. *Hilgardia* **15**, 325-368.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806.

Fouracre JP, Ando S, Langdale JA. 2014. Cracking the Kranz enigma with systems biology. *J Exp Bot* **65**, 3327-3339.

Fusaro AF, Matthew L, Smith NA, Curtin SJ, Dedic-Hagan J, Ellacott GA, Watson JM, Wang MB, Brosnan C, Carroll BJ, Waterhouse PM. 2006. RNA interference-inducing hairpin RNAs in plants act through the viral defence pathway. *EMBO Rep* **7**, 1168-1175.

Gallo-Meagher M, Irvine JE. 1993. Effects of tissue type and promoter strength on transient GUS expression in sugarcane following particle bombardment. *Plant Cell Rep* **12**, 666-670.
Gupta P, Raghuvanshi S, K Tyagi A. 2001. Assessment of the Efficiency of Various Gene Promoters via Biolistics in Leaf and Regenerating Seed Callus of Millets, Eleusine coracana and Echinochloa crusgalli. *Plant Biotechnology* **18**, 275-282.
Huang C-F, Yu C-P, Wu Y-H, Lu M-YJ, Tu S-L, Wu S-H, Shiu S-H, Ku MSB, Li W-H. 2017. Elevated auxin biosynthesis and transport underlie high vein density in C4 leaves. *Proceedings of the National Academy of Sciences* **114**, E6884-E6891.
Huang P, Brutnell TP. 2016. A synthesis of transcriptomic surveys to dissect the genetic basis of C4 photosynthesis. *Current Opinion in Plant Biology* **31**, 91-99.
Jain M, Nijhawan A, Tyagi AK, Khurana JP. 2006. Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. *Biochem Biophys Res Commun* **345**, 646-651.
Krogan NT, Long JA. 2009. Why so repressed? Turning off transcription during plant growth and development. *Curr Opin Plant Biol* **12**, 628-636.
Langdale JA, Lane B, Freeling M, Nelson T. 1989. Cell lineage analysis of maize bundle sheath and mesophyll cells. *Developmental Biology* **133**, 128-139.
Langdale JA, Zelitch I, Miller E, Nelson T. 1988. Cell position and light influence C4 versus C3 patterns of photosynthetic gene expression in maize. *EMBO J* **7**, 3643-3651.
Lazo GR, Stein PA, Ludwig RA. 1991. A DNA transformation-competent Arabidopsis genomic library in Agrobacterium. *Biotechnology (N Y)* **9**, 963-967.
Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR. 2010. The developmental dynamics of the maize leaf transcriptome. *Nat Genet* **42**.
Livak KJ, Schmittgen TD. 2001. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* **25**, 402-408.
Mann DG, Lafayette PR, Abercrombie LL, King ZR, Mazarei M, Halter MC, Poovaiah CR, Baxter H, Shen H, Dixon RA, Parrott WA, Neal Stewart C, Jr. 2012. Gateway-compatible vectors for high-throughput gene functional analysis in switchgrass (Panicum virgatum L.) and other monocot species. *Plant Biotechnol J* **10**, 226-236.
Nelson T, Dengler N. 1997. Leaf Vascular Pattern Formation. *Plant Cell* **9**, 1121-1135.
Petricka JJ, Clay NK, Nelson TM. 2008. Vein patterning screens and the defectively organized tributaries mutants in Arabidopsis thaliana. *Plant J* **56**, 251-263.
Rossini L, Cribb L, Martin DJ, Langdale JA. 2001. The Maize Golden2 Gene Defines a Novel Class of Transcriptional Regulators in Plants. *Plant Cell* **13**, 1231-1244.
Sage RF. 2016. A portrait of the C4 photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineages, and Hall of Fame. *J Exp Bot* **67**, 4039-4056.
Sharman BC. 1942. Developmental Anatomy of the Shoot of Zea mays L. *Ann Bot* **6**, 245-282.
Slewinski TL, Anderson AA, Price S, Withee JR, Gallagher K, Turgeon R. 2014. Short-Root1 Plays a Role in the Development of Vascular Tissue and Kranz Anatomy in Maize Leaves. *Molecular Plant* **7**, 1388-1392.
Slewinski TL, Anderson AA, Zhang C, Turgeon R. 2012. Scarecrow Plays a Role in Establishing Kranz Anatomy in Maize Leaves. *Plant and Cell Physiology* **53**, 2030-2037.
Tausta SL, Li P, Si Y, Gandotra N, Liu P, Sun Q, Brutnell TP, Nelson T. 2014. Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C(4)-related processes. *Journal of Experimental Botany* **65**, 3543-3555.

Toki S, Hara N, Ono K, Onodera H, Tagiri A, Oka S, Tanaka H. 2006. Early infection of scutellum tissue with Agrobacterium allows high-speed transformation of rice. *Plant J* **47**, 969-976.
Wang P, Karki S, Biswal AK, Lin H-C, Dionora MJ, Rizal G, Yin X, Schuler ML, Hughes T, Fouracre JP, Jamous BA, Sedelnikova O, Lo S-F, Bandyopadhyay A, Yu S-M, Kelly S, Quick WP, Langdale JA. 2017. Candidate regulators of Early Leaf Development in Maize Perturb Hormone Signalling and Secondary Cell Wall Formation When Constitutively Expressed in Rice. *Scientific Reports* **7**, 4535.
Wang P, Kelly S, Fouracre JP, Langdale JA. 2013. Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz anatomy. *Plant J* **75**, 656-670.
Waters MT, Moylan EC, Langdale JA. 2008. GLK transcription factors regulate chloroplast development in a cell‐autonomous manner. *The Plant Journal* **56**, 432-444.
Yoshida S, Forno DA, Cock JH. 1971. *Laboratory manual for physiological studies of rice*. Los Baños, Philippines: The International Rice Research Institute.
Zhang Y, Tessaro MJ, Lassner M, Li X. 2003. Knockout Analysis of Arabidopsis Transcription Factors TGA2, TGA5, and TGA6 Reveals Their Redundant and Essential Roles in Systemic Acquired Resistance. *Plant Cell* **15**, 2647-2653.

Author contribution

Jan Emmerling wrote the manuscript and conducted all experiments, with exception of the rice transformation, which was performed by Monika Streubel.

Acknowledgments / Danksagung

Ich danke…

… Prof. Dr. Peter Westhoff für die Möglichkeit meine Doktorarbeit an seinem Institut durchzuführen und die hilfreiche Unterstützung während dieser Zeit.

... Prof. Dr. Maria von Korff Schmising für die Übernahme des Koreferats.

... Maria und **Monika** für die unermüdliche Transformation meiner Konstrukte in Flaveria und Reis.

... den **Gärtnern** des Dachgewächshauses für die Betreuung meiner Pflanzen.

... Udo für Ratschläge, Diskussionen und neue Blickwinkel.

... Sandra für Gespräche, Kurzweil und Kreuzworträtsel.

... Kumari for company, discussions and being my paragon of working morale ;)

... Steffi für die letzten acht Jahre, unzähligen Diskussionen und unerbittliche Unterstützung, Zuversicht und Nachsicht, Rückhalt und Rückenfreihalten, Korrekturlesen und Pflege meiner Pflanzen. Ich schulde dir mehr als ich hier aufzählen könnte... Danke!

... dem Rest der **BOTIV** für das angenehme Arbeitsklima, Hilfsbereitschaft, die gemeinsame Zeit und gemeinsame Feste.

... meiner **Familie** und **Freunden** für die bedingungslose Unterstützung, die nötige Zerstreuung und endlose Nachsicht mit einem Workaholic^^

... Eva dafür, dass du mich seit drei Jahren begleitest, deine Geduld, Fürsorge und Vertrauen :*