Establishment of C_4 Photosynthesis in Ontogeny and Evolution

Kumulative Dissertation

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

vorgelegt von
Alisandra Kaye Denton
aus White Salmon, WA

Düsseldorf, Januar 2015

Aus dem Institut für Biochemie der Pflanzen
der Heinrich-Heine-Universität Düsseldorf

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

Referent: Prof. Dr. Andreas P. M. Weber
Korreferent: Prof. Dr. Laura E. Rose

Tag der mündlichen Prüfung: 22.07.2015

Eidesstattliche Versicherung und Selbstständigkeitserklärung

Ich versichere an Eides statt, dass ich die vorliegende Dissertation eigenständig und ohne unerlaubte Hilfe unter Beachtung der Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf angefertigt habe. Die Dissertation habe ich in dieser oder ähnlicher Form noch bei keiner anderen Institution vorgelegt. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Ort, Datum Alisandra Denton

"The woods are lovely, dark, and deep,
But I have promises to keep,
And miles to go before I sleep,
And miles to go before I sleep."

Robert Frost

HEINRICH-HEINE-UNIVERSITÄT DÜSSELDORF

Zusammenfassung

Mathematisch-Naturwisselschaftliche Fakultät
Institut für Biochemie der Pflanzen

Establishment of C_4 Photosynthesis in Ontogeny and Evolution

Alisandra Kaye Denton

In vielen Pflanzenarten findet sich das adaptive Merkmal, das als C_4-Syndrom bekannt ist. Der C_4-Zyklus beinhaltet eine biochemische Pumpe, die CO_2 in der Nähe des wesentlichen Kohlenstoff-fixierenden Enzyms Rubisco anreichert und dadurch die Fixierung von O_2 und somit die Photorespieration unterdrückt. Dies ist von großem Vorteil für C_4-Pflanzen, da die Photorespiration energieaufwendig ist und in einem Netto-Verlust von Kohlenstoff resultiert. Darüberhinaus weisen C_4-Pflanzen eine erhöhte Wassernutzungseffizienz auf, da sie besser dazu in der Lage sind, das Öffnen und Schließen der Stomata zu regulieren und eine ausreichende CO_2-Konzentration um Rubisco aufrechtzuerhalten. Außerdem verfügen sie über eine effizientere Stickstoffnutzung, da weniger Stickstoff in die Produktion von Rubisco investiert werden muss. Diese Eigenschaften schlagen sich in einem starken selektiven Vorteil für C_4-Spezies in heißen und trockenen Umgebungen nieder.

Das C_4-Syndrom kommt in vielen Nutzpflanzen vor, die große Mengen an Biomasse produzieren, darunter Mais, Sorghum und Zuckerrohr. Aus diesem Grund besteht ein großes Interesse daran, Nutzpflanzen mit dem ancestralen C_3-Typ der Photosynthese zur Nutzung der C_4-Photosynthese zu modifizieren. Zur vollständigen Integration des C_4-Photosyntheseweges bedarf es komplexer Modifikationen, um den CO_2-Konzentrations-Zyklus zu unterstützen.

In den meisten C_4-Spezies wird CO_2 zuerst im äußeren Mesophyll-Gewebe (M) fixiert und anschließend in die inneren Bündelscheidenzellen (BS) gepumpt, wo CO_2 freigesetzt und durch Rubisco re-fixiert wird. Umfangreiche Änderungen in der Anatomie sind nötig, um Diffusionswege der Metaboliten des C_4-Zyklus zu reduzieren und um Nutzen aus der CO_2-Konzentration ziehen zu können. Diese Änderungen beinhalten eine erhöhte Blattaderdichte, vergrößerte Bündelscheidenzellen, eine erhöhte Anzahl von Organellen in den Bündelscheidenzellen und Modifikationen der BS-Zellwand, die den Austritt von CO_2 durch Diffusion reduzieren. C_4 Photosynthese erfordert die funktionelle Spezialisierung von M- und BS-Zellen, insbesondere eine Beschränkung des Calvin-Benson-Bassham-Zyklus (CBB) und des photorespirativen Zyklus primär auf das BS-Gewebe. Weitere verbreitete Anpassungen beinhalten die Beschränkung des Photosystem II zum

M-Gewebe und die Etablierung von Redox-Shuttles, um Energie zwischen den beiden Geweben auszugleichen.

Die hohe Komplexität der C_4-Photosynthese führt sowohl zu Fragen ihrer Evolution als auch zu technischen Herausforderungen. Die Evolution der C_4-Photosynthese ist dadurch besonders faszinierend, dass sie – trotz ihrer so hohen Komplexität und Abwesenheit eines Master-Regulator-Gens – mehrfach unabhängig evolutionär hochgradig konvergent entstanden ist. Ein schrittweises Modell fasst einen oft beobachteten Weg zur C_4-Photosynthese zusammen, ausgehend von genetischer und anatomischer Präkonditionierung, über die Etablierung der photorespiratorischen Pumpe und anschließender Hochregulierung und Optimierung des Zyklus.

Drei der Manuskripte dieser Arbeit beschäftigen sich mit den Voraussetzungen der Evolution der C_4-Photosynthese. Die hier gewonnen Erkenntnisse decken sich mit den bestehenden Modellen und ergänzen sie um zusätzliche Details. In Denton et al. (in preparation) erläutern wir, wie Genduplikationen, über die Haupt-C_4-Gene hinaus, zum C_4-Syndrom in Mais beitragen. Paraloge, die eine für die anatomische Spezialisierung wichtige Funktion haben, wie etwa Zellwand- oder Auxin-Response-Funktion, zeigten spezifische Divergenzmuster in jungen Geweben. Drei der vier ATP-verbrauchenden Enzyme des CBB- und des photorespiratorischen Zyklus sind Paraloge mit Funktionen die für den Energieausgleich wichtig sind und, zeigten komplementäre Expression in voll entwickeltem M- und BS-Gewebe. Darüberhinaus hing die BS- bzw. M-Spezifität mit dem Duplikationsgrad auf genomweiter Ebene zusammen.

In Denton et al. (2013) haben wir die jüngsten Fortschritteund Erkenntisse aus dem Bereich der Präkonditionierung, wie etwa BS-Zellgröße, hohe Blattaderdichte, und die Vorteile in heißen und trockenen Umgebungen, analysiert.

Abschließend modellierten und überprüften wir in Heckmann et al. (2013) den evolutionären Verlauf ausgehend von einem C_3-Zustand zur vollständig integrierten C_4-Biochemie und fanden.

Die Errichtung der C_4-photosynthetischen Anatomie findet nicht in vollentwickelten, sondern in sich entwickelnden Geweben statt; zu einem vollen mechanistischen Verständnis sind vergleichende Studien der Ontogenese erforderlich. Zwei der Manuskripte dieser Arbeit generierten und analysierten solche vergleichende Ontogenese-Daten. Denton et al. (in preparation) vergleicht BS- und M-Gewebe während der Entwicklung des Mais-Blattes und zeigte, zusätzlich zu gewebespezifischen Paralogen, Transkriptionsregulatoren mit früher Gewebe-Spezifität. Külahoglu et al. (2014) vergleicht die Blatt-Ontogenese zwischen zwei nahe verwandten C_3- und C_4-Cleomaceae-Spezies und findet eine Verbindung zwischen Transkription und Anatomie für vergrößerte BS und hohe Blattaderdichte in den C_4-Spezies. Die vergrößerten BS in den C_4-Spezies korrelierten mit höheren BS-Ploiditätsstufen und der Herunterregulation eines Transkriptionsfaktors, der eine Schlüsselrolle in der Inhibition der Endoreduplikation spielt. Die vergrößerte

Blattaderdichte scheint durch eine Verzögerung der Gewebedifferenzierung ermöglicht zu sein, welche auf transkriptionaler und anatomischer Ebene beobachtet werden konnte.

Zusammengenommen tragen die Manuskripte in dieser Arbeit zum Verständnis über die für die Entwicklung der C_4-Photosynthese nötigen Schritte bei und bieten Einblicke in die Mechanismen und Details des vollständig integrierten C_4-Syndroms.

HEINRICH HEINE UNIVERSITY

Abstract

Faculty of Mathematics and Natural Sciences
Institute of Plant Biochemistry

Establishment of C_4 Photosynthesis in Ontogeny and Evolution

by Alisandra Kaye Denton

Many plant species harbor an adaptive photosynthetic trait known as C_4 photosynthesis. The C_4 cycle is a biochemical pump that concentrates CO_2 in the vicinity of the central carbon fixing enzyme Rubisco, suppressing the fixation of O_2 and thereby photorespiration. This is highly advantageous for C_4 plants because photorespiration is energetically costly and results in a net loss of carbon. Further, C_4 plants show increased water-use efficiency, as they are more able to modulate stomatal opening and closing and maintain a sufficient CO_2 concentration near Rubisco; and increased nitrogen use efficiency, as they can reduce the amount of nitrogen that must be invested in the extremely abundant Rubisco protein. These characteristics result in a strong selective advantage for C_4 species in hot and arid environments.

The C_4 trait is found in many high-biomass producing crop plants, including maize, sorghum, and sugar cane. Therefore, there is strong interest in engineering C_4 photosynthesis into crop plants of the ancestral C_3 photosynthetic type. A fully integrated C_4 photosynthetic trait requires complex modifications to support the CO_2 concentrating C_4 cycle.

In most species with C_4 photosynthesis, CO_2 is initially fixed in the exterior mesophyll (M) tissue and then pumped into interior bundle sheath (BS) tissue, where the CO_2 is released and then re-fixed by Rubisco. Extensive changes in anatomy are required, both to reduce diffusional distances for the metabolites of the C_4 cycle and to take advantage of the concentrated CO_2. These changes include an increased vein density, enlarged bundle sheath cells, increased organelle content in bundle sheath cells, and modifications to the BS cell wall that reduce diffusive escape of CO_2. C_4 photosynthesis requires specialization of function between M and BS cells, notably with the Calvin-Benson-Bassham (CBB) cycle and the photorespiratory cycle restricted primarily to the BS. Further common changes include the restriction of photosystem II to M tissue, and the establishment of redox shuttles to balance energy between the two tissue types.

The high complexity of the C_4 trait leads to both evolutionary questions and engineering challenges. The evolution of C_4 photosynthesis is particularly intriguing, because despite the high complexity and lack of master regulator, C_4 photosynthesis evolves in a

highly convergent fashion. A step wise model summarizes a commonly observed path to C_4 photosynthesis, starting with genetic and anatomical preconditioning, and proceeding to the establishment of a photorespiratory pump and later the up-regulation and optimization of the cycle.

Three manuscripts examine what facilitates the evolution of C_4 photosynthesis, with findings consistent with, but providing additional detail to the standing model for C_4 evolution. In Denton et al. (in preparation), we elucidated how duplication contributes to the C_4 trait in maize, beyond the core C_4 genes. Paralogs with functions relevant to anatomical specialization, including cell wall and auxin response, showed specific patterns of divergence in immature tissue. Paralogs with functions relevant to energy balance, namely 3 out of the 4 ATP consuming enzymes in the CBB and photorespiratory cycles, showed complementary expression in mature M and BS tissue. Further BS or M tissue specificity was related to duplication level on a genome wide scale.

In Denton et al. (2013) we reviewed recent progress in understanding anatomical preconditioning factors, such as BS cell size and dense vein spacing, and their advantages in hot and arid environments. Finally, in Heckmann et al. (2013) we modeled and cross checked the evolutionary progression from C_3 to fully integrated C_4 biochemistry.

Establishment of the C_4 photosynthetic anatomy occurs not in mature but in developing tissues, and a full mechanistic understanding of the C_4 trait requires comparative ontogenies. Two manuscripts in this thesis generate and analyze comparative ontogeny data. Denton et al. (in preparation) compares BS and M tissues in maize leaf development, and showed, in addition to tissue specific paralogs, transcriptional regulators with early tissue specificity. Külahoglu et al. (2014), compares leaf ontogeny in closely related C_3 and C_4 Cleomaceae species, and finds a link between transcription and anatomy for both enlarged BS and dense vein spacing in the C_4 species. The enlarged BS correlated with a higher BS ploidy level and down-regulation of a key endoreduplication inhibiting transcription factor in the C_4 species. The increased vein formation in the C_4 species appears to be facilitated by a delay in tissue differentiation observed at both the transcriptional and anatomical level.

Taken together, the manuscripts in this thesis have contributed to understanding the natural evolutionary path towards C_4 photosynthesis and provided insight into the mechanisms and details of a fully integrated C_4-trait.

Acknowledgements

Among the many things I have learned while working on my PhD is just how valuable a little of someone else's time and expertise can be, and I'd like to sincerely thank those who have helped and supported me through this time.

I am greatful to my supervisor Prof. Dr. Andreas P.M. Weber not just for the opportunity to pursue my PhD in his lab and scientific support, but for always believing in my capabilities.

My post-doctoral supervisor Dr. Andrea Bräutigam helped me beyond her critical evaluation of my research through her interest, encouragement and repeatedly bailing me out when I got in a bind.

I've had the priveledge of receiving extra supervision and support from other experienced researchers. In particular, I'd like to thank my committee members: my mentor Prof. Dr. Laura Rose for her advice and the perspective she brought to my research; and to thank Prof. Dr. Shin-Han Shiu for the warm welcome in MSU and continued honesty in regards to my project. Further, I appreciate the community support from the iGRADplant and other HHU experienced researchers. I'm especially thankful to Dr. Christian Eßer, for first teaching me to program, and Dr. Udo Gowik for both supervision and collaboration. I appreciate the time and effort that Prof. Dr. Martha Ludwig and Terry MacFarlane put into helping me plan the collection trip in Western Australia. Even though it didn't work out in the end; I learned a lot.

I'd like to thank my iGRADplant peers for the welcoming and fun learning environment at iGRADplant events. I owe an extra thanks to those who collaborated with or simply helped me: Dr. Thea Pick, Dr. David Heckman, Dr. Canan Külahoglu and Sarah Richards.

It was a great benefit to me to work in my lab with so many smart, driven people who still found time to answer questions, share a protocol and generally help someone out, while maintaining a friendly atmosphere. Specifically, I'd like to thank Samantha Kurz for teaching me how to clone, Anja Nöcker for compensating my lack of organization, and Lance Valls for helping with plant work all summer 2012. My colleagues–Simon Schliesky, Dr. Freddy Breuers, Dr. Sarah Keßel-Vigelius, Dr. Jan Wiese, Manuel Sommer, Dominik Brillhaus, Thomas Wrobel, and Angelo Agossou Yao–I thank for making the hard days a little easier.

Thank you to all of my collaborators and co-authors who are not mentioned by name.

The amazing coordinators of iGRADplant have made the transitions and duration of my doctoral work unbelievably easier. I would like to offer a heart-felt thank you to Dr. Sigrun Wegener-Feldbrügge for helping me with everything from paperwork to arranging an apartment with food in the fridge on the day of my arrival; and another heart-felt

thank you to Prof. Dr. Barb sears for sharing her wonderful home (and pets!) with us during transitions at Michigan State University.
For all the enjoyable lunches and jokes I would like to thank the bioinformatics institute. The last stretch of my PhD has been both easier and more enjoyable thanks to the Chaosdorf, where I have had encouragement to program more and more reusably, and a great place to learn and work.
I cannot sufficiently write out how much assistance and support I've received from Janina Maß. I am very greatful for her help with work and continuing of my computer education, and even more grateful for everything else.
Finally, I could not have made it this far without my parents, Jim Denton and Avalon Totten-Denton, my sister, Elsie Denton, my brothers in heart, Loehn and Brynden Rawdin-Morris, and all the other friends and family who I don't have space to name here. Thank You.

In memory of the dinosaurs

Contents

Abbreviations

2-PG	**2-P**hospho**g**lycolate
3-PG	**3-P**hospho**g**lycerate
ALAAT	**A**lanine **a**mino**t**ransferase
ASPAT	**A**spartate **a**mino**t**ransferase
ATP	**A**denine **t**ri**p**hosphate
BEP	**B**ambusoideae, **E**hrhartoideae, **P**ooideae
BS	**B**undle **s**heath
CA	**C**arbonic **a**nhydrase
CBB cycle	**C**alvin-**B**enson-**B**assham cycle
CO_2	**C**arbon **dio**xide
GDC	**G**lycene **d**ecarboxylase **c**omplex
M	**M**esophyll
MDH	**M**alate **d**e**h**ydrogenase
NAD	**N**icotinamide**a**denine **d**inucleotide
NADP	**N**icotinamide **a**denine **d**inucleotide**p**hosphate
NADPH	**N**icotinamide **a**denine **d**inucleotide**p**hosphate
NADME	**NAD**-dependent **m**alic **e**nzyme
NADPME	**NADP**-dependent **m**alic **e**nzyme
O_2	**O**xygen
OAA	**O**x**a**lo**a**cetate
PACMAD	**P**anicoideae, **A**rundinoideae, **C**hloridoideae, **M**icrairoideae, **A**ristidoideae, **D**anthonioideae
PS	**P**hoto**s**ystem
PEP	**P**hospho***e****nol*pyruvate
PEPCK	**P**hospho***e****nol*pyruvate **c**arboxy**k**inase
Rubisco	**R**ib**u**lose-1,5-**bis**phosphate **c**arboxylase/**o**xygenase
RuBP	**R**ib**u**lose-1,5-**bis**phosphate
TP	**T**riose **p**hosphate

Chapter 1

Introduction

Many complex biological traits are of high interest to humankind. There is strong agricultural interest in improving traits such as pest-resistance, drought-resistance, and ultimately yield. While some traits, like pest-resistance, can frequently be modified by changing a single gene (Hammond-Kosack and Jones, 1997), and are therefore amenable to breeding or genetic engineering; complex traits like drought resistance and photosynthetic type involve many genes and are inherently harder to understand or recreate (Xu et al., 2014a; von Caemmerer et al., 2012).

This thesis focuses on understanding the development – both evolution and ontogenesis – of the complex trait C_4 photosynthesis to obtain a better understanding of the trait itself.

C_4 photosynthesis is an evolutionary add-on to C_3 photosynthesis that helps plants thrive in hot or arid conditions and helps crop plants to achieve exceptionally high yield. It includes changes to leaf tissue architecture, cellular architecture, and to the leaf biochemistry (Sage, 2004).

1.1 Motivation

The world around us is filled with morphological and biochemical diversity, which is in turn made up of many different traits, many of them complex. While some complex traits, such as the synthetic pathway for various medicines (Chang and Keasling, 2006), are simple enough to be fully understood and recreated, other traits, such as C_4 photosynthesis, remain partially understood and beyond the range of current engineering, despite their potential benefits. To try and understand a complex trait in sufficient detail to ultimately recreate it, we examine how the complex trait C_4 photosynthesis develops in nature on both the evolutionary and the leaf developmental time scale.

Despite its complexity C_4 photosynthesis evolves in a highly convergent manner (Sage et al., 2011). Thus, we examine which factors and mechanisms facilitate the evolution of C_4 photosynthesis, and help it repeatedly evolve such complexity.

To identify simpler mechanisms underlying the complexity of the C_4 trait in a mature leaf we look to leaf development. The anatomical modifications found in C_4 plants are established during leaf ontogeny, and these developmental modifications are ultimately inseperable from the trait itself. Therefore to more fully understand the trait, we deeply characterize tissues at different stages of development. Successful engineering of the mature trait will have to include modifications to the developmental process to achieve the fine-tuned C_4 state.

1.2 Complex traits

Complex traits are major contributors to the morphological diversity found in living creatures; and despite the many beneficial mutations they require, some complex traits evolve convergently. By definition, a complex trait is any trait that does not have mendelian inheritance because it consists of more than one gene. In practice, complex traits can vary from a handful of genes to substantial portions of the genome that are involved in traits like yield of fitness (Falke et al., 2013). All of the mutations required to evolve a trait must arise without inccuring a fitness penalty, even when the advantages of the trait are not necessarily realized until the trait is fully established. For instance, for a complex trait like flight feathers to evolve, the intermediate stages have to be beneficial, or at least nearly selectively neutral in their own right. While this is an evolutionary necessity, it is not always simple to determine how the sub-traits contributed an evolutionary advantage based soley on the species alive today. Thus, the debate continues on whether proto-feathers may have been more beneficial to dinosaurs for thermal regulation, colorful communication and mating displays, or both (Xu et al., 2014b). Generally, as the complexity of a trait increases, the more of the total change required for its evolution is contained in the sum of many small-effect mutations (Falke et al., 2013). This is part of the reason that a complex trait such as yield is so hard to improve.

Despite the potentially long and twisted evolutionary path to establishment or modification of a complex trait, some traits have evolved more than once. This is called convergent evolution. Some convergent traits have only evolved a few times, such as wings, which have evolved in bats, pterosaurs, birds, and insects; while other traits, such as pigmentation patterns (Kronforst et al., 2012), viviparity (Blackburn, 2005), and the photosynthetic adaptation known as C_4 photosynthesis (Sage et al., 2012) have evolved again and again.

1.3 C_4 photosynthesis, Rubisco and photorespiration

C_4 photosynthesis is a complex trait of high agricultural importance because it increases the efficiency of photosynthesis in hot and arid conditions.

C_4 photosynthesis was first discovered based on its distinct anatomy – also termed Kranz anatomy – in which the spacing between veins is reduced, and the bundle sheath layer encircling the vein is enlarged and packed with organelles (Haberlandt, 1904). Since then, we have learned that these anatomical changes provide an appropriate setup for a biochemical cycle that concentrates CO_2 in the interior bundle sheath tissue.

The high concentration of CO_2 suppresses a side reaction of the carbon fixing enzyme, Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco), and thereby reduces the wasteful process of photorespiration (Andrews et al., 1971; Sage, 2004). This allows plants to have a higher photosynthetic efficiency under appropriate conditions, and many high yielding crop plants such as maize, sorghum and sugar cane utilize C_4 photosynthesis (Sage and Zhu, 2011). Thus, there is an interest in engineering the C_4 cycle in other crop plants, with some suggesting it could bring about a second green revolution (von Caemmerer et al., 2012).

The C_4 cycle provides the ancestral-like, high-CO_2 microenvironment to Rubisco, which reduces photorespiration. When photosynthesis first evolved, the concentration of CO_2 relative to O_2 was much higher than it is today, and the relatively low specificity of the key carbon fixing enzyme Rubisco for CO_2 over O_2 was not important (Badger and Andrews, 1987). Under these conditions, Rubisco and the rest of the Calvin-Benson-Bassham (CBB) cycle became central to plant metabolism.

In the CBB cycle, Rubisco fixes one molecule of CO_2 onto the 5-carbon ribulose-1,5-bisphosphate (RuBP) producing two molecules of the 3-carbon 3-Phosphoglycerate (3-PG). This is followed by an extensive reduction and recycling process with a net gain of one molecule of triose-phosphate (TP) per three turnovers (Bassham et al., 1954). As the primary output, TP is used for sugar or starch synthesis (Melis, 2013).

Plants bearing just the ancestral CBB cycle are referred to as C_3 plants, as the first stable product of carbon fixation has three carbons. Similarly, C_4 plants derive their name from their four carbon initial product of carbon fixation.

The Achilles heel of the C_3 cycle is the limited substrate specificity of Rubisco for CO_2 over O_2, which leads to photorespiration. In today's atmosphere, Rubisco frequently fixes a molecule of O_2, instead of CO_2, onto RuBP, resulting in one molecule of 3-PG and one molecule of the toxic 2-phosphoglycolate (2-PG). To recycle 2-PG, plants use the photorespiratory cycle, which has an energetic cost and releases 25% of the CO_2 entering the cycle. After reactions in three different organelles – chloroplasts, mitochondria, and peroxyzomes – the photorespiratory cycle returns one 3-PG per two turnovers (reviewed in Bauwe et al., 2010).

Photorespiration cannot be circumvented by increasing specificity as there is a direct trade-off in Rubisco's catalytic mechanism between specificity and turnover rate (Tcherkez et al., 2006) and Rubisco is already extremely slow. High temperature and any condition that decreases CO_2 concentrations in the leaf, exacerbate the fixation of O_2 by Rubisco and increase photorespiration (Carmo-Silva et al., 2012; Bowes, 1991). Thus, under many stress conditions, C_4 plants achieve a much higher photosynthetic efficiency than C_3 plants as they can provide a high-CO_2 microenvironment to Rubisco and suppress photorespiration (Ehleringer et al., 1997). The concentrated CO_2 allows C_4 plants to reduce investment in Rubisco, which increases their nitrogen use efficiency, and allows plants to reduce stomatal opening, which increases their water use efficiency (Sage and Pearcy, 1987).

1.4 The biochemistry of C_4 photosynthesis

The C_4 cycle concentrates CO_2 through a biochemical pump that fixes CO_2 onto an organic acid, transports this acid to- and releases CO_2 in- the vicinity of Rubisco. In the C_4 cycle, CO_2 is converted to HCO_3^- by carbonic anhydrase (CA) in the mesophyll (M). The carbon from HCO_3^- is fixed onto phospho*enol*pyruvate (PEP) by the oxygen insensitive enzyme PEP carboxylase (PEPC) to form oxaloacetate (OAA). OAA is converted to either aspartate by aspartate aminotransferase (ASPAT) or malate by malate dehydrogenase (MDH) before diffusing to the bundle sheath where the four-carbon acid is converted to the substrate of the decarboxylating enzyme and decarboxylated, releasing CO_2 (reviewed in Furbank, 2011).

The decarboxylation can take place through any of three enzymes: NADME, NADPME, or PEPCK. These decarboxylation enzymes give name to the major subtypes of C_4; however these subtypes are not exclusive, and many C_4 plants use more than one decarboxylation enzyme (Furbank, 2011; Pick et al., 2011; Sommer et al., 2012; Bräutigam et al., 2014; Wang et al., 2014).

NADME and NADPME produce the 3-carbon molecule pyruvate, while PPDK directly produces PEP. The 3-carbon acid remaining after decarboxylation diffuses back to the M, with or without intermediate conversion to alanine by alanine aminotransferase (ALAAT), and if necessary, is regenerated to PEP by PPDK (reviewed in Furbank, 2011).

C_4 cycles have higher energy uses and specialized energy balance between cell types. The C_4 cycle requires ATP where PEP is produced. When PEP is regenerated from pyruvate by PPDK in the M chloroplast, ATP is converted to AMP. Whereas, the decarboxylation of OAA to PEP by PEPCK in the bundle sheath (BS) requires half the energy from ATP, as ATP is converted only to ADP. Additionally, when NADPME is used for decarboxylation, the cycle is thought to consume reducing equivalents in

the M, while producing them in the BS. The energy intensive CBB cycle, is primarily localized to the BS, but NADPH consuming conversion of 1,3-bisphosphoglycerate to TP is localized to the M (reviewed in Hatch, 1987; Furbank, 2011). Despite or even due to the reducing equivalent shuttles, maintaining a constant balance of energy between the two tissue types is non-trivial. Some C_4 plants deplete photosystem II and thereby linear electron transport from the BS, which increases the ratio of ATP to reducing equivalent produced in the BS (Romanowska et al., 2008). Modeling attempts have indicated that the mixing of C_4 cycles may help with providing sufficient energy to the BS (Wang et al., 2014), and with maintaining energy balance in fluctuating light conditions (Bellasio and Griffiths, 2014).

1.5 Modifications to support the C_4 cycle

The core C_4 cycle does not work in isolation, but requires additional specialization to yield its benefits. The enzymes of the C_4 cycle are split between two cell types and various subcellular compartments. Therefore, additional measures are necessary to speed the transfer of metabolites between enzymes, including reduced diffusion distances and the up-regulation of metabolite transporters.

C_4 plants have tight vein spacing, with just two layers of mesophyll cells separating vascular bundles allowing direct diffusion from a mesophyll to adjacent BS cell (McKown and Dengler, 2010). To get metabolites in and out of subcellular compartments, metabolite transporters are expressed at levels comparable to the C_4 enzymes. However, not all transporters are known (Weber and von Caemmerer, 2010).

The BS tissue must undergo massive changes both to support the C_4 cycle and serve as the primary tissue for incorporation of CO_2 into organic molecules. The BS cell wall is heavily lignified to reduce diffusive loss of CO_2 back to the M. However, the anatomy must support the high flux of metabolites of the C_4 cycle and BS and M cells are connected by dense plasmodesmata. In accordance with its new function as the primary tissue for fixation and incorporation of carbon into organic molecules, the BS shows a drastic increase in organelle number. This increase allows for the localization of Rubisco, and the majority of both CBB and photorespiratory cycle to the bundle sheath. All these anatomical and molecular changes require specialized regulation and development, however, there is comparatively little known on how this is achieved.

1.6 Evolution of C_4 photosynthesis

Despite its complexity, C_4 photosynthesis has evolved repeatedly, but in a clustered fashion, throughout the angiosperms. The hundreds of genes that are differentially regulated between closely related C_3 and C_4 species (Bräutigam et al., 2011; Gowik

et al., 2011a) provide an estimate of the total complexity of the C_4 trait. Thus, it is impressive that C_4 photosynthesis has evolved no less than 66 different times (Sage et al., 2012).

C_4 lineages are spread throughout the angiosperms and occur in 19 different families (Sage et al., 2012). However, there is a strong clustering of C_4 origins in some taxonomic groups, such as the sedges and grasses with 6 and 22-24 independent lineages, respectively (Grass Phylogeny Working Group II, 2011; Sage et al., 2012). While the common ancestor of sedges and grasses was presumably primed for C_4 photosynthesis when the lineages split, there is still a strong clustering of C_4 evolutions within the grasses.

The grasses can be divided into two major clades: BEP with the subfamilies Bambusoideae, Ehrhartoideae, and Pooideae and PACMAD with the subfamilies Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae, and Danthonioideae. These subfamilies are known to have radiated prior to 65 mya, as they were ingested by Dinosaurs (Piperno and Sues, 2005). Of the two major clades, all the C_4 lineages occur in PACMAD; further there is strong variation within the subfamilies with C_4 evolving once at the base of the Chlorodoideae; C_4 evolving in 19 independent lineages in the Panicoideae, but no C_4 species in the Danthonioideae. (Grass Phylogeny Working Group II, 2011).

One potential factor that could relate to differences in tendency to evolve C_4 is the difference in environmental preferences of these lineages (Edwards and Smith, 2010). Several environmental factors are thought to relate to and encourage the evolution of C_4 photosynthesis, and while the environmental context of C_4 evolution varies across so many origins, there are some consistent themes. The most obvious is that all C_4 origins occurred around or after the Oligocene drop in atmospheric CO_2 (Edwards et al., 2010; Christin et al., 2008; Pagani et al., 2005). For less epochal environmental changes, the environments of extant species can be used to reconstruct the likely ancestral environment. Using this, Edwards et al. (2010) found that there is a tendency for a switch towards open habitats at C_4 origins in the grasses. This could relate to the extra ATP required for the C_4 cycle. Finally, many C_4 species grow in hot, arid, and saline environments, which are therefore also thought of as driving forces for the evolution of C_4 photosynthesis.

Several genera, most notably Flaveria in the dicots, contain multiple C_4 origins and or intermediates between C_3 and C_4 photosynthesis. Careful examination of evolutionary patterns with a focus on these genera have led to a pyramid like model of C_4 evolution (Sage, 2004; Gowik and Westhoff, 2010). In this model, preconditioning factors accumulate that are beneficial for other reasons. These preconditioning factors include gene duplication, narrowing of vein spacing, and increase in BS size. In the next step, which is often referred to as C_2 photosynthesis (Vogan et al., 2007), the Glycene Decarboxylase Complex (GDC) – and thereby the release of CO_2 from the photorespiratory cycle

– is localized specificaly to the BS This provides selective pressure towards increasing Rubisco in the bundle sheath, and additional CO_2 pumping, leading to the evolution of the full C_4 cycle. Finally, the cycle is fully incorporated with the fine-tuning of details such as enzyme kinetics (Sage, 2004; Paulus et al., 2013; Christin et al., 2008).

1.7 New techniques and old questions

Advancing techniques and resources allow us to test both this evolutionary model and observe the ontogeny of C_4 photosynthesis on a scale not previously possible. Recent sequencing projects have increased the phylogenetic resolution and identified new clades where C_4 photosynthesis has evolved (Grass Phylogeny Working Group II, 2011). This allows environmental and anatomical changes to be evaluated on a more general scale (Edwards et al., 2010; Christin et al., 2012; Griffiths et al., 2013).

While for projects focusing on a few species, decreasing cost of transcriptome sequencing allow not just the comparison between two end states but the comparison of leaf development or many different tissue types. This allows not just for a snap shot of transcriptional investment, but for a slide show of changes over time, space or condition. Recent large transcriptional studies have been successful in answering outstanding questions about C_4 development, for instance clarifying that there is no phase of C_3 photosynthesis in maize leaf ontogeny during the sink to source transition (Pick et al., 2011).

This thesis focuses on understanding the evolution and ontogeny of C_4 photosynthesis using modern, high throughput techniques. The first manuscript reviews recent progress in elucidating the anatomical preconditioning of C_4 photosynthesis (Denton et al., 2013). Two of the research manuscripts evaluate different parts of the model for the evolution of C_4 photosynthesis. Manuscript H evaluates the feasibility and fitness of the path from a preconditioned C_3 plant to a full fledged C_4 plant (Heckmann et al., 2013). Manuscript DM evaluates how gene duplication contributes to the evolution of C_4 photosynthesis on a genome wide scale (Denton et al., in preparation). Two of the manuscripts examine the ontogeny of the C_4 photosynthetic transcriptome. Manuscript DM looks at the differences in the ontogeny of BS and M cells, while manuscript KD compares transcriptome atlases with leaf developmental gradients between a C_3 and C_4 species in the Cleomaceae (Külahoglu et al., 2014). Finally, manuscript K takes a different approach by looking at the transcriptome of a sedge that can either develop C_4 photosynthesis on land or an intermediate between C_3 and C_4 photosynthesis underwater (Külahoglu et al., submitted).

Chapter 2

Discussion

C_4 photosynthesis is a complex trait of high adaptive advantage for plants growing under photorespiration inducing conditions. The central function of C_4 photosynthesis is carried out by a fairly simple biochemical cycle that concentrates CO_2 and thereby suppresses photorespiration. However, reaping the full benefits of the C_4 cycle requires further adaptations such as specialized anatomy and energy balance. To achieve the ambitious goal of engineering C_4 photosynthesis into C_3 crop plants, both a full understanding of the trait and how to establish it is required.

This thesis focuses on understanding how the C_4 trait is established in nature, both in terms of evolution and ontogeny. Investigating the evolution of the C_4 trait elucidates what steps occur in C_4 evolution, when they occur, and how they are beneficial in their own right. Examining the ontogeny of C_4 plants provides key information on how the specialized anatomy is established.

While there is much versatility in the details, the road from C_3 to C_4 photosynthesis starts with intrinsically beneficial preconditioning steps, with no apparent turn offs after the establishment of a photorespiratory CO_2 pump. Sage (2004) outlined a path from C_3 to C_4 photosynthesis from general preconditioning to integration and optimization of the final trait. The general preconditioning was proposed to include high levels of gene duplication (Sage, 2004), which can result in reduced evolutionary constraint on paralogs (Chain et al., 2008; Hellsten et al., 2007), facilitating the evolution of new functions.

However, supporting evidence for the benefit of duplication in the evolution of C_4 photosynthesis has been limited to studies on the core-C_4 genes (Wang et al., 2009). Further, as C_4 species do not necessarily have more heavily duplicated genomes (van den Bergh et al., 2014) nor even more heavily duplicated core-C_4 genes (Williams et al., 2012) than C_3 species, the advantage of gene duplication has been recently questioned.

In contrast, we find that duplicated genes show patterns of expression divergence consistent with a role in C_4 photosynthesis. In particular, we found BS and M specific paralogs of ATP-consuming photosynthetic enzymes. Maize has an elaborate scheme

for balancing energy between BS and M cells (Kramer and Evans, 2011), to which independently regulated ATP consuming enzymes could add robustness. Further, we found both BS tissue specific and general paralogs for cell wall functions, and for auxin regulators and response. Both of these functions are thought to be specialized in developing C_4 anatomy (Eastman et al., 1988; McKown and Dengler, 2007)). Finally, we provide evidence that ancient gene duplications preconditioned development of tissue specificity, a key feature of C_4 photosynthesis (Sheen, 1999).

Anatomical preconditioning follows general preconditioning in the path laid out by Sage (2004). The manuscript (Denton et al., 2013) summarizes recent progress in understanding anatomical preconditioning and its relationship with the environment. Two different studies found that the heat adapted PACMAD grass clade shows an overall enlargement of BS cell size, even in C_3 species (Christin et al., 2012; Griffiths et al., 2013). Other environmental factors, such as aridity, salinity and high-light were associated with, but do not predate, C_4 origins.

Finally, progress has been made in understanding how anatomical preconditioning factors are beneficial for environmental adaptation. BS cells have been characterized as a *smart-pipe* that regulates flow in and out of the vasculature and provides both structural rigidity and anti-cavitation response important in hot or arid environments (Griffiths et al., 2013). Denser vein spacing has been shown to be linked to photosynthetic capacity throughout the evolution of land plants, and is necessary to supply sufficient water and avoid dessication of M tissue with stomata open in hot and arid conditions (Brodribb and Feild, 2010). From a preconditioned state, the establishment of a photorespiratory CO_2 pump is, based on current evidence, a committed step to evolution of C_4 photosynthesis. Many C_3 plants in photorespiration inducing environments have proto-Kranz anatomy with a close association between organelles promoting the scavenging of respired CO_2. From this state, establishment of a photorespiratory CO_2 pump is mechanistically simple, requiring only loss in expression of one GDC subunit in the M tissue (Morgan et al., 1993). Modeling in (Heckmann et al., 2013) indicates a continuous increase in fitness during the biochemical transition from a preconditioned C_3 state to a fully integrated C_4 state. This is consistent with the occurrence of C_3-C_4 intermediate species only in young C_4 evolving lineages. Further, the simulations indicate a preferred and modular order for changes in biochemistry that traces observed evolutionary paths.

Taken together, the three evolutionary-focused manuscripts both provide empirical support for and clarify details of the standing model of C_4 evolution (Sage, 2004). They help us understand what allows the C_4 trait to evolve so convergently despite its complexity.

While the evolutionary path is ultimately too slow for human interest in synthetic

recreation of the C_4 trait, these analyses have highlighted points that should be considered in any attempt to recreate the trait. These key points include the advantages of starting with a species with proto-Kranz anatomy, such as rice (Heckmann et al., 2013; Denton et al., 2013), the attention to detail required to achieve C_4-caliber energy balance (Denton et al., in preparation), and the potential to exploit existing regulatory mechanisms such as auxin signaling (Denton et al., in preparation, 2013).

A mature C_4-photosynthetic state does not occur without specialization during leaf ontogeny. Anatomical C_4-characteristics such as enlarged BS cells and denser vein spacing are already established in a mature photosynthetic leaf, and the underlying mechanisms are thus hidden from detection. Developmental transcriptomes, however, provide a window towards understanding these traits (Pick et al., 2011; Li et al., 2010).

Separation of BS and M tissue along a developmental gradient showed not just the specialization between paralogs, but general differential expression in immature tissues, including 274 transcription factors in the youngest section. While all current separation methods for M and BS are subject to some form of bias, contrasting results from different studies using different techniques (Chang et al., 2012; Li et al., 2010; Tausta et al., 2014), (Denton et al., in preparation) allows narrowing down a core set of regulators associated with M and BS tissue specificity. These studies, and other inventive experiments, such as comparison of expression in different photosynthetic modes of Eleocharis species (Külahoglu et al., submitted), comparing primordial development between the maize leaf and maize husk, which lacks kranz-anatomy (Wang et al., 2013), and comparative studies between C_3 and C_4 species (Külahoglu et al., 2014; Bräutigam and Weber, 2011; Gowik et al., 2011b), help to compile a list of transcription factors important for understanding and ultimately engineering C_4 photosynthesis.

Key to engineering C_4 photosynthesis is the comparison between C_3 and C_4 development. Külahoglu et al. (2014) linked transcriptional to anatomical changes in development between a C_3 and C_4 Cleomacea species. Compared to the C_3 species, the C_4 species showed a delay in tissue differentiation and increased vein formation. The anatomical delay in tissue differentiation was matched by a delay in expression changes during leaf development, including prolonged expression of the COP9 signalosome, which helps repress photomorphogenesis (Chamovitz et al., 1996; Dohmann et al., 2008). As differentiation of photosynthetic tissues has been found to limit vein formation (Scarpella and Meijer, 2004; Kang et al., 2007), this delay in differentiation likely allows for the increased vein formation observed in the C_4 species.

Developmental comparison of the Cleomacea species indicated a relationship between enlarged BS cells and endoreduplication. Many large cell types harbor increased amounts of DNA, whether in the form of additional nuclei as in muscle cells, or increased ploidy as in trichomes (Lee et al., 2009). The C_4 Cleomacea species showed an increase in BS ploidy, and down regulation of a suppressor of endoreduplication, GTL (Külahoglu et al.,

submitted). While endoreduplication does not necessarily drive cellular enlargement, it is likely beneficial, or even necessary, to support sufficient transcriptional levels for a large cytoplasm (Lee et al., 2009).

Taken together, these studies of leaf ontogeny provide insight into the natural mechanisms of achieving C_4 specific anatomy. Some of these adaptations appear mechanistically simple, and were putatively associated with regulatory genes. A few C_4-related regulatory elements are known to be conserved across angiosperms (Kajala et al., 2012). If simple, but high-impact mechanisms can be transfered between species, this opens the possibility of engineering fairly drastic changes in anatomy with fairly few genes. Before this can be achieved, more characterization of transcriptional regulators and their targets will be necessary.

Bibliography

Andrews, T. J., Lorimer, G. H., and Tolbert, N. E. (1971). Incorporation of molecular oxygen into glycine and serine during photorespiration in spinach leaves. *Biochemistry*, 10(25):4777–4782.

Badger, M. and Andrews, T. (1987). Co-evolution of rubisco and co2 concentrating mechanisms. In Biggins, J., editor, *Progress in Photosynthesis Research*, pages 601–609. Springer Netherlands.

Bassham, J. A., Benson, A. A., Kay, L. D., Harris, A. Z., Wilson, A. T., and Calvin, M. (1954). The path of carbon in photosynthesis. xxi. the cyclic regeneration of carbon dioxide acceptor1. *Journal of the American Chemical Society*, 76(7):1760–1770.

Bauwe, H., Hagemann, M., and Fernie, A. R. (2010). Photorespiration: players, partners and origin. *Trends in plant science*, 15(6):330–336.

Bellasio, C. and Griffiths, H. (2014). The Operation of Two Decarboxylases, Transamination, and Partitioning of C4 Metabolic Processes between Mesophyll and Bundle Sheath Cells Allows Light Capture To Be Balanced for the Maize C4 Pathway. *Plant Physiology*, 164(1):466–80.

Blackburn, D. (2005). Amniote perspectives on the evolutionary origins of viviparity and placentation. In Grier, H. and Uribe, M., editors, *Viviparous Fishes*, pages 301–322. New Life Publications, Homestead, Florida.

Bowes, G. (1991). Growth at elevated co2: photosynthetic responses mediated through rubisco. *Plant, Cell & Environment*, 14(8):795–806.

Bräutigam, A., Kajala, K., Wullenweber, J., Sommer, M., Gagneul, D., Weber, K. L., Carr, K. M., Gowik, U., Mass, J., Lercher, M. J., Westhoff, P., Hibberd, J. M., and Weber, A. P. M. (2011). An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiology*, 155(1):142–56.

Bräutigam, A., Schliesky, S., Külahoglu, C., Osborne, C. P., and Weber, A. P. M. (2014). Towards an integrative model of C4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C4 species. *Journal of Experimental Botany*, 65(13):3579–93.

Bräutigam, A. and Weber, A. P. M. (2011). Do metabolite transport processes limit photosynthesis? *Plant physiology*, 155(1):43–8.

Brodribb, T. J. and Feild, T. S. (2010). Leaf hydraulic evolution led a surge in leaf photosynthetic capacity during early angiosperm diversification. *Ecology letters*, 13(2):175–83.

Carmo-Silva, A. E., Gore, M. A., Andrade-Sanchez, P., French, A. N., Hunsaker, D. J., and Salvucci, M. E. (2012). Decreased co¡ sub¿ 2¡/sub¿ availability and inactivation of rubisco limit photosynthesis in cotton plants under heat and drought stress in the field. *Environmental and Experimental Botany*, 83:1–11.

Chain, F. J. J., Ilieva, D., and Evans, B. J. (2008). Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evolutionary Biology*, 8:43.

Chamovitz, D. A., Wei, N., Osterlund, M. T., von Arnim, A. G., Staub, J. M., Matsui, M., and Deng, X.-W. (1996). The cop9 complex, a novel multisubunit nuclear regulator involved in light control of a plant developmental switch. *Cell*, 86(1):115–121.

Chang, M. C. and Keasling, J. D. (2006). Production of isoprenoid pharmaceuticals by engineered microbes. *Nature chemical biology*, 2(12):674–681.

Chang, Y.-M., Liu, W.-Y., Shih, a. C.-C., Shen, M.-N., Lu, C.-H., Lu, M.-Y. J., Yang, H.-W., Wang, T.-Y., Chen, S. C.-C., Chen, S. M., Li, W.-H., and Ku, M. S. B. (2012). Characterizing Regulatory and Functional Differentiation between Maize Mesophyll and Bundle Sheath Cells by Transcriptomic Analysis. *Plant Physiology*.

Christin, P.-A., Besnard, G., Samaritani, E., Duvall, M. R., Hodkinson, T. R., Savolainen, V., and Salamin, N. (2008). Oligocene CO2 decline promoted C4 photosynthesis in grasses. *Current Biology : CB*, 18(1):37–43.

Christin, P.-A., Edwards, E. J., Besnard, G., Boxall, S. F., Gregory, R., Kellogg, E. a., Hartwell, J., and Osborne, C. P. (2012). Adaptive Evolution of C(4) Photosynthesis through Recurrent Lateral Gene Transfer. *Current Biology : CB*, 22(5):445–449.

Denton, A. K., Maß, J., Külahoglu, C., Lercher, M., Shiu, S.-H., Bräutigam, A., and Weber, A. P. (in preparation). Expression divergence following gene duplication contributes to the evolution of the complex trait C_4 photosynthesis.

Denton, A. K., Simon, R., and Weber, A. P. (2013). C_4 photosynthesis: From eeolutionary analyses to strategies for synthetic reconstruction of the trait. *Current Opinion in Plant Biology*, 16(3):315–321.

Dohmann, E. M., Levesque, M. P., De Veylder, L., Reichardt, I., Jürgens, G., Schmid, M., and Schwechheimer, C. (2008). The arabidopsis cop9 signalosome is essential for g2 phase progression and genomic stability. *Development*, 135(11):2013–2022.

Eastman, P. A. K., Dengler, N. G., and Peterson, C. A. (1988). Suberized Bundle Sheaths in Grasses (Poaceae) of Different Photosynthetic Types I. Anatomy, Ultrastructure and Histochemistry. *Protoplasma*, 142:92–111.

Edwards, E. J. and Smith, S. a. (2010). Phylogenetic analyses reveal the shady history of C4 grasses. *Proceedings of the National Academy of Sciences of the United States of America*, 107(6):2532–7.

Edwards, E. J., Smith, S. A., and Consortium, C. G. (2010). Ecosystem Science. 587.

Ehleringer, J. R., Cerling, T. E., and Helliker, B. R. (1997). C4 photosynthesis, atmospheric co2, and climate. *Oecologia*, 112(3):285–299.

Falke, K. C., Glander, S., He, F., Hu, J., de Meaux, J., and Schmitz, G. (2013). The spectrum of mutations controlling complex traits and the genetics of fitness in plants. *Current Opinion in Genetics & Development*, 23(6):665–71.

Furbank, R. T. (2011). Evolution of the C(4) photosynthetic mechanism: are there really three C(4) acid decarboxylation types? *Journal of Experimental Botany*, 62(9):3103–8.

Gowik, U., Bräutigam, A., Weber, K. L., Weber, A. P. M., and Westhoff, P. (2011a). Evolution of C4 photosynthesis in the genus Flaveria: how many and which genes does it take to make C4? *The Plant Cell*, 23(6):2087–105.

Gowik, U., Bräutigam, A., Weber, K. L., Weber, A. P. M., and Westhoff, P. (2011b). Evolution of C4 photosynthesis in the genus Flaveria: how many and which genes does it take to make C4? *The Plant cell*, 23(6):2087–105.

Gowik, U. and Westhoff, P. (2010). The Path from C3 to C4 Photosynthesis. *Plant Physiology*, 155(January):56–63.

Grass Phylogeny Working Group II (2011). New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *New Phytologist*, pages 304–312.

Griffiths, H., Weller, G., Toy, L. F., and Dennis, R. J. (2013). You're So Vein: Bundle Sheath Physiology, Phylogeny and Evolution in C3 and C4 Plants. *Plant, Cell & Environment*, 36(2):249–261.

Haberlandt, G. (1904). *Physiologische Pflanzenanatomie*. W. Engelmann.

Hammond-Kosack, K. E. and Jones, J. D. G. (1997). Plant disease resistance genes. *Annual Review of Plant Physiology and Plant Molecular Biology*, 48(1):575–607. PMID: 15012275.

Hatch, M. D. (1987). C¡ sub¿ 4¡/sub¿ photosynthesis: a unique elend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta (BBA)-Reviews on Bioenergetics*, 895(2):81–106.

Heckmann, D., Schulze, S., Denton, A., Gowik, U., Westhoff, P., Weber, A. P. M., and Lercher, M. J. (2013). Predicting C4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell*, 153(7):1579–88.

Hellsten, U., Khokha, M. K., Grammer, T. C., Harland, R. M., Richardson, P., and Rokhsar, D. S. (2007). Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. *BMC Biology*, 5:31.

Kajala, K., Brown, N. J., Williams, B. P., Borrill, P., Taylor, L. E., and Hibberd, J. M. (2012). Multiple Arabidopsis genes primed for recruitment into C photosynthesis. *The Plant journal : for cell and molecular biology*, 69(1):47–56.

Kang, J., Mizukami, Y., Wang, H., Fowke, L., and Dengler, N. G. (2007). Modification of cell proliferation patterns alters leaf vein architecture in arabidopsis thaliana. *Planta*, 226(5):1207–1218.

Kramer, D. M. and Evans, J. R. (2011). The importance of energy balance in improving photosynthetic productivity. *Plant physiology*, 155(1):70–78.

Kronforst, M. R., Barsh, G. S., Kopp, A., Mallet, J., Monteiro, A., Mullen, S. P., Protas, M., Rosenblum, E. B., Schneider, C. J., and Hoekstra, H. E. (2012). Unraveling the thread of nature's tapestry: the genetics of diversity and convergence in animal pigmentation. *Pigment Cell & Melanoma Research*, 25(4):411–33.

Külahoglu, C., Denton, a. K., Sommer, M., Maß, J., Schliesky, S., Wrobel, T. J., Berckmans, B., Gongora-Castillo, E., Buell, C. R., Simon, R., De Veylder, L., Bräutigam, a., and Weber, a. P. M. (2014). Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C3 and C4 Plant Species. *The Plant Cell*, 26(August):3243–3260.

Külahoglu, C., Schliesky, S., Sommer, M., Alisandra K. Denton, A. H., Buell, C. R., Bräutigam, A., and Weber, A. P. M. (submitted). Plasticity of C_4 photosynthesis in the amphibious sedge *Eleocharis retroflexa*.

Lee, H. O., Davidson, J. M., and Duronio, R. J. (2009). Endoreplication: polyploidy with purpose. *Genes & Development*, 23(21):2461–2477.

Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S. L., Kebrom, T. H., Provart, N., Patel, R., Myers, C. R., Reidel, E. J., Turgeon, R., Liu, P., Sun, Q., Nelson, T., and Brutnell, T. P. (2010). The developmental dynamics of the maize leaf transcriptome. *Nature Genetics*, 42(12):1060–1067.

McKown, A. D. and Dengler, N. G. (2007). Key innovations in the evolution of Kranz anatomy and C4 vein pattern in Flaveria (Asteraceae). *American Journal of Botany*, 94(3):382–99.

McKown, A. D. and Dengler, N. G. (2010). Vein patterning and evolution in C4 plants. *Botany*, 88(9):775–786.

Melis, A. (2013). Carbon partitioning in photosynthesis. *Current Opinion in Chemical Biology*, 17(3):453 – 456. Next generation therapeutics Energy.

Morgan, C., Turner, S., and 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(4):468–473.

Pagani, M., Zachos, J., Freeman, K., Tipple, B., and Bohaty, S. (2005). Marked decline in atmospheric carbon dioxide concentrations during the Paleogene. *Science*, (July):600–604.

Paulus, J. K., Schlieper, D., and Groth, G. (2013). Greater efficiency of photosynthetic carbon fixation due to single amino-acid substitution. *Nature Communications*, 4:1518.

Pick, T. R., Bräutigam, A., Schlüter, U., Denton, A. K., Colmsee, C., Scholz, U., Fahnenstich, H., Pieruschka, R., Rascher, U., Sonnewald, U., and Weber, A. P. M. (2011). Systems analysis of a maize leaf developmental gradient redefines the current C4 model and provides candidates for regulation. *The Plant Cell*, 23(12):4208–20.

Piperno, D. and Sues, H. (2005). Dinosaurs Dined on Grass. *Science(Washington)*, 310(November):1126–1128.

Romanowska, E., Kargul, J., Powikrowska, M., Finazzi, G., Nield, J., Drozak, A., and Pokorska, B. (2008). Structural organization of photosynthetic apparatus in agranal chloroplasts of maize. *Journal of Biological Chemistry*, 283(38):26037–26046.

Sage, R. F. (2004). The evolution of c4 photosynthesis. *New Phytologist*, 161(2):341–370.

Sage, R. F., Christin, P.-A., and Edwards, E. J. (2011). The C(4) plant lineages of planet Earth. *Journal of experimental botany*, 62(9):3155–69.

Sage, R. F. and Pearcy, R. W. (1987). The nitrogen use efficiency of c3 and c4 plants ii. leaf nitrogen effects on the gas exchange characteristics of chenopodium album (l.) and amaranthus retroflexus (l.). *Plant Physiology*, 84(3):959–963.

Sage, R. F., Sage, T. L., and Kocacinar, F. (2012). Photorespiration and the evolution of C4 photosynthesis. *Annual Review of Plant Biology*, 63(January):19–47.

Sage, R. F. and Zhu, X.-G. (2011). Exploiting the engine of C4 photosynthesis. *Journal of Experimental Botany*, 62(9):2989–3000.

Scarpella, E. and Meijer, A. H. (2004). *Pattern formation in the vascular system of monocot and dicot plant species*, volume 164.

Sheen, J. (1999). C4 Gene Expression. *Annual Review of Plant Physiology and Plant Molecular Biology*, 50:187–217.

Sommer, M., Bräutigam, A., and Weber, A. P. M. (2012). The Dicotyledonous NAD-malic Enzyme C4 Plant *Cleome gynandra* Displays Age-Dependent Plasticity of C4 Decarboxylation Biochemistry. *Plant Bbiology*, 14(4):621–9.

Tausta, S. L., Li, P., Si, Y., Gandotra, N., Liu, P., Sun, Q., Brutnell, T. P., and Nelson, T. (2014). Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C4-related processes. *Journal of Experimental Botany*, 65(13):3543–55.

Tcherkez, G. G. B., Farquhar, G. D., and Andrews, T. J. (2006). Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proceedings of the National Academy of Sciences of the United States of America*, 103(19):7246–51.

van den Bergh, E., Külahoglu, C., Bräutigam, A., Hibberd, J. M., Weber, A. P., Zhu, X.-G., and Eric Schranz, M. (2014). Gene and genome duplications and the origin of C4 photosynthesis: Birth of a trait in the Cleomaceae. *Current Plant Biology*, pages 1–8.

Vogan, P. J., Frohlich, M. W., and Sage, R. F. (2007). The functional significance of C3-C4 intermediate traits in Heliotropium L. (Boraginaceae): gas exchange perspectives. *Plant, Cell & Environment*, 30(10):1337–45.

von Caemmerer, S., Quick, W. P., and Furbank, R. T. (2012). The development of c4 rice: Current progress and future challenges. *Science*, 336(6089):1671–1672.

Wang, P., Kelly, S., Fouracre, J. P., and Langdale, J. a. (2013). Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz anatomy. *The Plant Journal : For Cell and Molecular Biology*, 75(4):656–70.

Wang, X., Gowik, U., Tang, H., Bowers, J. E., Westhoff, P., and Paterson, A. H. (2009). Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome biology*, 10(6):R68.

Wang, Y., Bräutigam, A., Weber, A. P. M., and Zhu, X.-G. (2014). Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis. *Journal of Experimental Botany*, 65(13):3567–78.

Weber, A. P. M. and von Caemmerer, S. (2010). Plastid transport and metabolism of C3 and C4 plants–comparative analysis and possible biotechnological exploitation. *Current opinion in plant biology*, 13(3):257–65.

Williams, B. P., Aubry, S., and Hibberd, J. M. (2012). Molecular evolution of genes recruited into C4 photosynthesis. *Trends in Plant Science*, 17(4):213–20.

Xu, J., Yuan, Y., Xu, Y., Zhang, G., Guo, X., Wu, F., Wang, Q., Rong, T., Pan, G., Cao, M., et al. (2014a). Identification of candidate genes for drought tolerance by whole-genome resequencing in maize. *BMC plant biology*, 14(1):83.

Xu, X., Zhou, Z., Dudley, R., Mackem, S., Chuong, C.-M., Erickson, G. M., and Varricchio, D. J. (2014b). An integrative approach to understanding bird origins. *Science*, 346(6215).

Chapter 3

First Author Manuscripts

3.1 Manuscript D: C_4 photosynthesis: From evolutionary analyses to strategies for synthetic reconstruction of the trait

Overview

Title: C_4 photosynthesis: From evolutionary analyses to strategies for synthetic reconstruction of the trait
Authors: Alisandra K Denton, Rüdiger Simon and Andreas PM Weber
Published in Current Opinion in Plant Biology, June 2013
Impact factor: 9.385
First authorship

Main Findings

This review summarizes the advantages and characteristics of C_4 photosynthesis, with a focus on recent progress in understanding what facilitates C_4 evolution and how this might be exploited in engineering C_4 photosynthesis. By concentrating CO_2 around Rubisco, the C_4 cycle suppresses the costly photorespiratory cycle. This comes with additional benefits such as higher nitrogen and water use efficiency, as C_4 plants require less Rubisco and can maintain more efficient photosynthesis when stomata close to conserve water. Several photorespiration-inducing environmental conditions have been linked to the evolution of C_4 photosynthesis. These environmental conditions can predate C_4 evolution; for instance C_4 plants have evolved more frequently in hot environments. Alternatively, C_4 plants can show a shift towards an environment at C_4 evolution; for instance C_4 plants occupy more arid habitats than their C_3 neighbors. Recently progress has helped link these environmental features to known steps in C_4 preconditioning, including finding an increase in BS cell size in the heat-adapted and

C_4-evolving grass clade. Of interest to engineering C_4, the close vein spacing required for C_4 photosynthesis may be achievable by increasing concentrations of the growth hormone Auxin. Finally, there appear to be many viable options to acquire a fine tuned C_4 cycle, including lateral gene transfer.

Contributions

- Wrote section on *Drivers of C_4 Preconditioning*
- Prepared *Box 2: The Distribution of C_4 origins with the Poaceae*
- Edited full manuscript

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Plant Biology

C_4 photosynthesis: from evolutionary analyses to strategies for synthetic reconstruction of the trait

Alisandra K Denton[1], Rüdiger Simon[2] and Andreas PM Weber[1]

C_4 photosynthesis represents the most productive modes of photosynthesis in land plants and some of the most productive crops on the planet, such as maize and sugarcane, and many ecologically important native plants use this type of photosynthesis. Despite its ecological and economic importance, the genetic basis of C_4 photosynthesis remains largely unknown. Even many fundamental aspects of C_4 biochemistry, such as the molecular identity of solute transporters, and many aspects of C_4 plant leaf development, such as the Kranz anatomy, are currently not understood. Here, we review recent progress in gaining a mechanistic understanding of the complex C_4 trait through comparative evolutionary analyses of C_3 and C_4 species.

Addresses
[1] Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, Universitätsstrasse 1, D-40225 Düsseldorf, Germany
[2] Institute of Developmental Genetics, Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, Universitätsstrasse 1, D-40225 Düsseldorf, Germany

Corresponding author: Weber, Andreas PM (aweber@hhu.de, andreas.weber@uni-duesseldorf.de)

Current Opinion in Plant Biology 2013, **16**:315–321

This review comes from a themed issue on **Physiology and metabolism**

Edited by **John Browse** and **Edward Farmer**

For a complete overview see the Issue and the Editorial

Available online 16th March 2013

1369-5266/$ – see front matter,

http://dx.doi.org/10.1016/j.pbi.2013.02.013

Introduction

Although the C_4 mode of photosynthesis was discovered more than 40 years ago by researchers in Australia, Canada, and Russia [1], the molecular mechanisms and the genetic basis of the C_4 trait remain largely unknown. Recently, progress in C_4 research is being boosted by the completion of the genome sequences of several C_4 grasses (e.g., *Zea mays*, *Sorghum bicolor*, *Setaria italica*) and the establishment of solid molecular phylogenies of C_3 and C_4 grasses [2–6]. Additional momentum is generated by a recent surge in systems and synthetic approaches to understand the C_4 trait, and several coordinated international research consortia that aim to introduce C_4 traits into C_3 plants [7–9]. The rationale for this renewed interest in C_4 photosynthesis is rather straightforward — above a temperature threshold of 21–23 °C and in open canopies with sufficient photosynthetically active radiation, the efficiency of C_4 photosynthesis can surpass C_3 by as much as 50% [10•]. This high efficiency is based on suppression of photorespiration in C_4 plants due to a biochemical CO_2 concentrating mechanism that increases the concentration of CO_2 in the vicinity of RubisCO, and so reduces the rate of the oxygenation reaction and thereby photorespiration. The more efficient use of RubisCO allows the enormous and one-sided investment of N in this protein to be reduced, and the high affinity of PEP carboxylase (PEPC) for bicarbonate allows rapid photosynthesis to occur at lower internal carbon dioxide concentrations and, thence, lower stomatal conductivity. An important consequence is that C_4 plants display higher nitrogen and water use efficiencies [11]. Putting it in a nutshell, C_4 plants produce more biomass with less input of scarce resources. It is thus not surprising that C_4 grasses have gained great ecological importance over the past 10 million years, as atmospheric CO_2 concentrations fell to present day levels [12•]. They currently represent approximately 20% of the land plant vegetation [13], in particular in open habitats of the tropics and subtropics [12•].

In comparison to C_3, C_4 photosynthesis requires defined and coordinated alterations to leaf anatomy and biochemistry and thus changes to the expression patterns of several hundred genes [14,15•,16•] when compared to a C_3 leaf. Despite the requirement for such complex and coordinated changes to structure and biochemistry and to the expression patterns of many genes, C_4 photosynthesis has frequently and concurrently evolved in at least 62 different plant species between 30 and 15 million years ago, both in the mono- and in the eu-dicotyledonous lineages [17], which indicates that the underlying genetic mechanisms might be rather simple. Several features are common to most multi-cellular C_4 plant species. They revolve around a unique division of the photosynthetic labor that is shared between two cell types: the mesophyll (MC) and the bundle sheath cells (BSCs):

(i) The BSCs contain RubisCO and the entire Calvin–Benson cycle, whereas the MCs contain less or no RubisCO and harbor predominantly the reductive part of the Calvin–Benson cycle. The MCs serve to assimilate and concentrate CO_2 in form of a C_4 (amino-)acid for transport to the BSC. There, CO_2 is released from the C_4 acid by one of three possible decarboxylation reactions [18], depending on C_4 subtype, resulting

in enrichment of CO_2 in close vicinity to RubisCO, whereby photorespiration is repressed. Importantly, recent work emphasizes that the canonical text book view of three distinct decarboxylation schemes is too simple — reality is more complex, with different decarboxylation reactions co-occurring, depending on leaf age, developmental stage, and environmental cues [18,19•,20,21•].

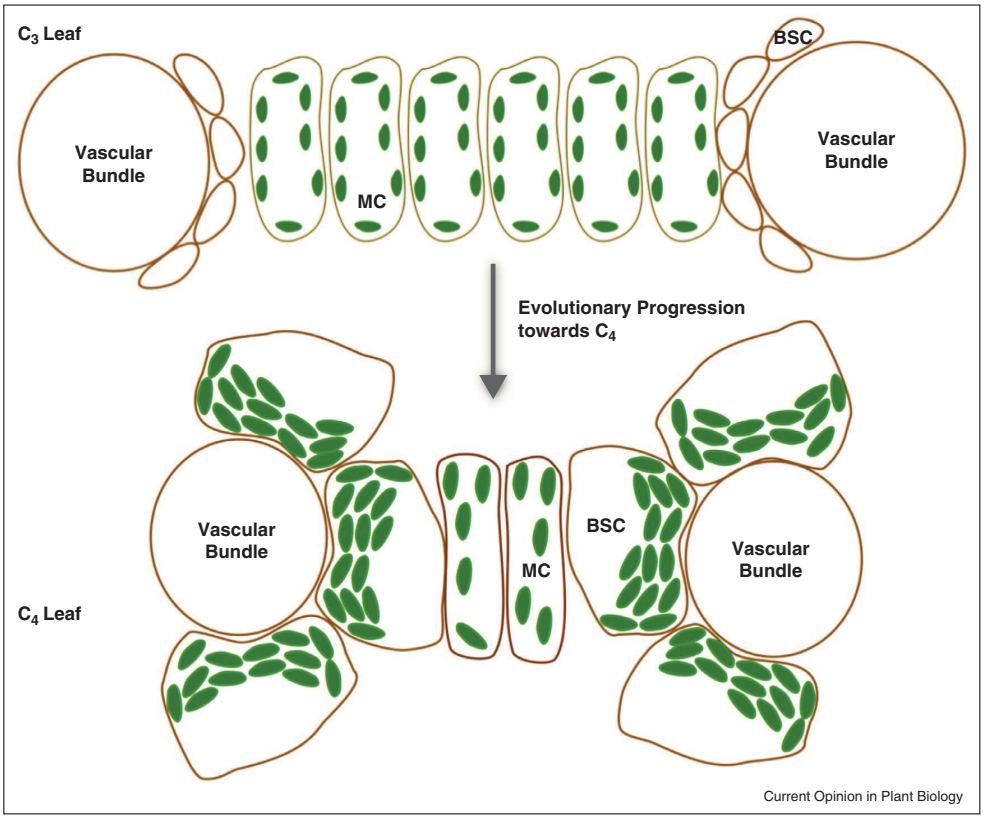
(ii) Enzyme activity and regulation, steady state metabolite pool sizes and metabolites fluxes, as well as transport capacities [22] are altered to accommodate the C_4 cycle.

(iii) The BSCs, a layer of cells surrounding the vascular bundle (Kranz anatomy), are much larger and contain many more photosynthetically active chloroplasts than BSCs in C_3 plants.

(iv) The number of mesophyll cells between two adjacent veins in the leaf tissue is much lower than in C_3 plants. This leads to closer vein spacing (increased venation density) in C_4 leaves (Figure 1).

The frequent concurrent evolution of the highly complex C_4 trait in distantly related plant genera prompted the hypothesis that a global environmental change, such as a steep decline in the atmospheric CO_2concentration to less than 500 ppm, provided a strong selective pressure, which favored the evolution of C_4 [23]. In all documented cases of C_4 evolution, the C_4 species have evolved from C_3 ancestors [17]. That is, C_3 represents the ancestral state and C_4 is the derived state. This indicates the existence of common genetic mechanisms for the evolution of this complex trait and possibly a predisposition for evolving the C_4 trait in some C_3 genera. Comparative evolutionary analyses within and between genera that have evolved C_4 photosynthesis should afford the reconstruction of the progression from C_3 to C_4, thereby eventually identifying the key genes required for the evolution of this complex trait. A mechanistic understanding of the evolutionary progression from C_3 to C_4 is crucial for synthetic approaches aiming at engineering C_4 into C_3 backgrounds and we thus focus on recent work in this direction.

Figure 1



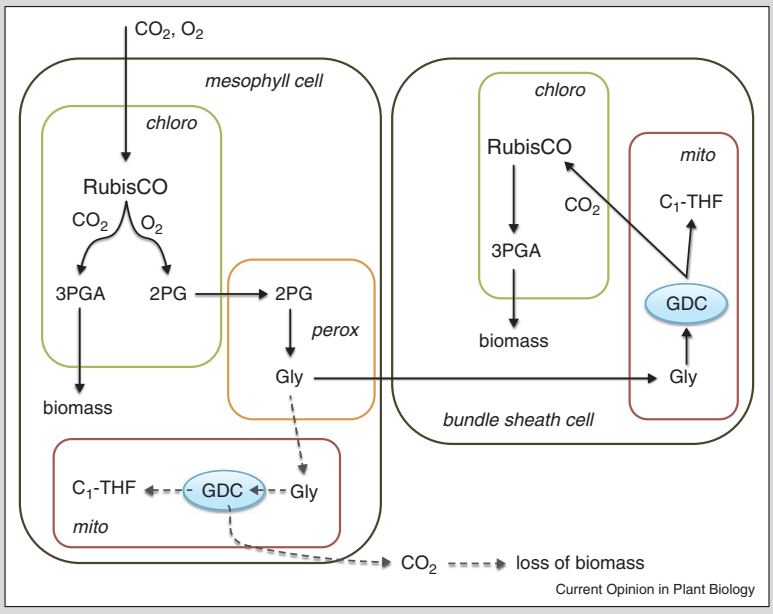
Schematic representation of several major differences between C_3 (upper panel) and C_4 (lower panel) photosynthesis. In the C_4 leaf, vein density is increased due to a lower number of mesophyll cells (MCs) between veins; photosynthesis and organellar function in bundle sheath cells (BSCs) is amplified and photosynthetic labor is shared between MC and BSC in C_4 plants.

Box 1 C_2-photosynthesis as a driver for C_4 evolution

In C_3 plants, the oxygenation reaction of RubisCO leads to photosynthetic inefficiency and loss of biomass in the form of CO_2 during photorespiration. Mechanism for scavenging some of the otherwise lost photorespiratory CO_2 have already evolved in some C_3 plants and it has been hypothesized that these scavenging mechanisms contributed to pre-conditioning for the evolution of C_4 photosynthesis. Key to the understanding of this process is that photorespiration in most C_3 plants is confined to a single cell type in the leaf, the mesophyll cells. 2-phosphoglycolate (2PG) that is produced through the oxygenation reaction of RubisCO is exported from the chloroplast (*chloro*), converted to glycine (Gly) in the peroxisomes (*perox*) and then transported to mitochondria (*mito*), were Gly is converted to CO_2 and methylated tetrahydrofolate (C_1-THF) by the multi-enzyme system glycine decarboxylase (GDC). Some of the CO_2 produced in this reaction is escaping out of the mesophyll cells into the leaf airspace and is thus lost from plant biomass. This pathway is indicated by dashed arrows in the figure below.

Recovery of photorespiratory CO_2 is achieved by splitting photorespiration between two cell types — mesophyll and bundle sheath cells. In this scenario that has been termed C_2-photosynthesis, the CO_2-releasing step of photorespiration (GDC) is confined to mitochondria of the bundle sheath cells. Released CO_2 is captured by bundle sheath localized chloroplasts, and/or by chloroplasts in the mesophyll cells, into 3-phosphoglyceric acid (3PGA) and thus into biomass before it can escape to the leaf air space. In addition to preventing CO_2 loss and thereby increasing photosynthetic efficiency, this pathway also permits the dissipation of excess excitation energy under conditions of limited CO_2 availability, such as closed stomata during periods of limited water supply. In the absence of CO_2 but presence of O_2, the transitory plastidial starch pool becomes accessible and is converted into Calvin–Benson cycle intermediates that are oxidized by RubisCO, yielding 2PG [49••]. Recycling of 2PG resulting from starch oxidation using the C_2 photosynthetic mode permits efficient recovery of photorespiratory CO_2 while at the same time efficiently dissipating excess excitation energy.

Obviously, an efficient C_2 photosynthetic mode requires the diffusion path from mesophyll to bundle sheath cells to be short. That is, under conditions were C_2 photosynthesis provides a selective advantage, the reduction of inter-vein distance and of mesophyll cell number between veins would be advantageous, too. Also the increase of bundle sheath cell size would be beneficial by increasing diffusion distance and generating space for additional organelles. Hence, photorespiration, C_2 photosynthesis, and anatomical preconditioning for C_4 photosynthesis are tightly connected.



Drivers of C_4 preconditioning

A variety of anatomical changes are important as preconditioning for, and during the evolution of, C_4 photosynthesis. Notably proto-Kranz anatomy, which helps to scavenge photorespired CO_2 [24] (see Box 1 for details), increased vein density [25], and BSC enlargement [26,27••] have all been shown to predate evolution of the C_4 cycle in some instances. Recent phylogenetically and ecologically informed studies help to shed light on the driving factors for the evolution not just of C_4 photosynthesis but also of the preconditioning factors, thereby improving our understanding of the clustered evolutionary pattern [2,28,29].

In addition to declining atmospheric CO_2, hot, arid, and saline conditions have long been implicated as driving factors in the evolution of C_4 photosynthesis, and certainly also play a role in early preconditioning. As a

convergent trait, it is likely that there is flexibility in not just the path to C_4 photosynthesis; but also in the relative contribution of environmental drivers, depending upon the lineage.

Reconstruction of ancestral ecological niches shows that C_4 photosynthesis evolved under hot conditions, but that the C_4 branches shifted markedly towards drier and more open environments than their C_3 cousins [28,30]. This indicates high temperatures may be key for promoting C_4 preconditioning factors.

Controlling the extent and timing of cell divisions during early stages of leaf development are the key regulatory events that confine the overall plasticity of vein density patterns. Generation of a dense vascular system largely requires a sufficient number of cells that can respond to the phytohormone auxin, which is the main trigger for development and differentiation of a continuous vascular system. In *Arabidopsis thaliana*, high temperatures directly affect overall auxin levels in the developing leaf by promoting auxin synthesis [31]. Experimentally increasing cell division rates in ground tissues can provide the necessary undifferentiated cells, and in combination with local increases in auxin concentration (by blocking auxin transport from the leaf) can result in the formation of a surplus vasculature [32]. This might indicate a path how the increased vein densities found in extant C_4 plants originated during evolution.

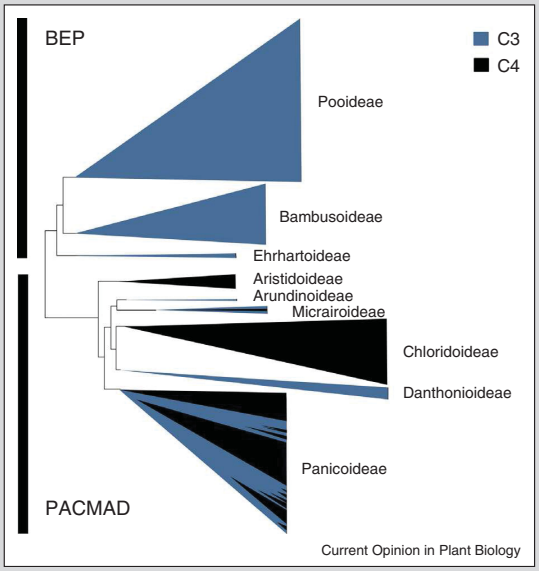
Because of the associated decrease in the specificity of RubisCO, increasing temperatures provide a selective advantage for photorespiration-limiting innovations, from (photo-)respiratory carbon scavenging to fully developed C_4 photosynthesis. Proto-Kranz anatomy, a single-celled, photorespiratory-CO_2 scavenging system, has been described in *Heliotropium* [24], and is seen in other genera with closely related C_3 and C_4 species [33•]. Further evidence for the role of high temperature is found in rice, which is phylogenetically distant from C_4 species, but a very successful tropical grass. Interestingly, rice appears to harbor a CO_2 scavenging system based on the close association of mitochondria and chloroplasts, and highly developed stromules that likely impede the escape of CO_2 [34,35•].

Photorespiration is, however, also strongly induced by aridity and salinity, as limiting stomatal conductivity to conserve water limits the diffusion of CO_2 into the leaf. A phylogenetic study of the eudicot *Chenopodiaceae* clade indicated that succulence and salt adaptation were significant as preconditions for C_4 evolution in this lineage [36]. In the grasses, increased BSC size is found in the PACMAD [26,27••] — the heat-adapted clade of grasses, in which all C_4 origins are clustered — compared to BEP, its C_3-only sister clade (see Box 2 for details). However within C_3 PACMAD, species with a higher %BSC inhabited more arid, but not warmer, environments than species with a lower %BSC [26].

The increase in water use efficiency associated even with minimal CO_2 scavenging [33•,34,35•], and relation of vein density with hydraulic conductivity [37,38] gives a mechanistic explanation for the selective advantage of C_4 preconditioning steps in physiological drought conditions. Insufficient hydraulic conductivity for the climate would result in either the desiccation or starvation of the leaf, and ultimately sets a cap on photosynthetic performance through stomatal conductivity [37,38]. Finally, the role of BSC in controlling water flow from the xylem to the mesophyll [39], and supporting or even repairing the xylem after cavitation [38] may indicate a direct link between BSC size and aridity. In summary, phylogenetically informed comparative evolutionary studies showed

Box 2 The distribution of C_4 origins within the Poaceae

The Poaceae, or true grass family, is very rich in C_4 origins, and C_4 species. The subfamilies have all been unambiguously determined, and the family splits into two clades: PACMAD, containing Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae and Danthonioideae; and BEP, containing Bambusoideae, Ehrhartoideae, and Pooideae. By the current estimate there are 22–24 C_4 origins within the Poaceae [50], and they all occur in the PACMAD clade where they cluster more strongly in some lineages than others (see below). This makes the Poaceae ideal for comparisons, both between the BEP and PACMAD clades, and between C_3 and C_4 sister species or intermediates. Further, the presence of high-resolution phylogenetic data, six fully sequenced genomes, other sequence data, and primary crop species lead to much C_4 research focusing on the Poaceae.



that increase of BSC size and decrease of vein distance are the major factors that precondition (grass) leaves for C_4 photosynthesis. Thus, future studies should focus on the mechanisms controlling leaf cell size and the initiation of major and minor veins.

Establishing C_4 biochemistry

As outlined above, comparative evolutionary analyses indicate that anatomical preconditioning and C_2 photosynthesis (see Box 1) predate a fully functional C_4 photosynthetic metabolism. For C_4 photosynthetic metabolism to function, in comparison to C_3 metabolism massive alterations to steady state metabolite levels and of metabolic fluxes within and between photosynthetic leaf cells are required. For example, the steady state levels of malate, aspartate, pyruvate, and alanine are much higher in C_4 leaves as compared to C_3 leaves [19•,20]. In addition, the fluxes of specific metabolites across the chloroplast and other cellular membranes (e.g., mitochondria) and between cells via plasmodesmata are at least one order of magnitude higher than in C_3 plants [40]. Obtaining an altered state of metabolic homeostasis and maintaining the metabolic fluxes associated with it requires alterations to enzyme biochemistry (e.g., allosteric and post-translational regulation, enzyme affinity, and velocity), to the abundance and regulation of solute transporters, and to the connections between mesophyll and BSCs through plasmodesmata. While some data is available on altered enzyme biochemistry, such as the posttranslational regulation and allosteric properties of C_4-type PEPC [41], we currently do not have a full mechanistic understanding of how the steady state metabolic pools and the fluxes between these pools are established in C_4 photosynthesis. Certainly, increased expression levels of genes encoding enzymes and transporters of C_4 biochemistry [15•,16•] and consequently higher protein amounts [42,43] contribute to achieving higher velocities and flux. While *cis*-regulatory elements controlling mesophyll-cell-specific C_4 gene expression have been reported for *PEPC* [44], *trans* regulators are unknown to date [14]. In maize, the expression of *PEPC* is further regulated in a cell-specific and light dependent manner by methylation of 4 cytosine residues in the *PEPC* promoter [45•]: methylation is high in roots and BSCs in the light and in the dark, whereas it is low in mesophyll cells in the light. Bundle-sheath-specific expression of NAD-malic enzyme (NAD-ME) in *Cleome gynandra* and of NADP-ME in maize was recently shown to be controlled by a 240 bp element in the 5′ of the transcribed region [46••]. That is, post-transcriptional mechanisms are apparently involved in controlling the bundle-sheath-specific accumulation of these transcripts, although the exact mechanism is not yet understood.

Excitingly, it was recently shown that C_4-adapted genes encoding the C_4 versions of PEPC and PEP carboxykinase (PCK) were acquired by independent lateral gene transfers from distantly related C_4 PACMAD species in the C_4 grass *Alloteropsis semialata* [47••]. A similar finding was reported for the grass subtribe Neurachnine [48••]. This indicates that enzymes with modified allosteric and kinetic properties optimized for function in C_4 metabolism can be transferred over significant phylogenetic distance and integrated into a novel metabolic network, given the anatomical preconditioning has previously been established.

Conclusions

Comparative evolutionary analyses of C_3 and C_4 species in a phylogenetically informed context showed that increased BSC size and decreased BSC distance are crucial anatomical enablers that precondition C_3 species for the evolution of C_4 photosynthesis. Similar comparisons at the genomic and transcriptomic levels are expected to provide candidate genes controlling these traits. Several distinct mechanisms, such as *cis* elements, cell-specific DNA methylation, and post-transcriptional regulation are controlling the cell-specific expression of C_4 enzymes. Modifications of allosteric and kinetic properties, together with post-translational regulation govern C_4-specific enzymic properties. Thus, engineering the metabolic aspects of C_4 metabolism might prove to be more complex than establishing the anatomical features.

Acknowledgements

Work in the authors' laboratory is supported by the Deutsche Forschungsgemeinschaft (grants WE2231/8-2, IRTG 1525, EXC 1028), the Federal Ministry of Education and Research, and the 7th Framework of the European Union (3to4, http://3to4.org).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Hatch MD: **C_4 photosynthesis: discovery and resolution**. *Photosynth Res* 2002, **73**:251-256.

2. Besnard G, Christin P-A: **Evolutionary genomics of C_4 photosynthesis in grasses requires a large species sampling**. *C R Biol* 2010, **333**:577-581.

3. Christin P-A, Samaritani E, Petitpierre B, Salamin N, Besnard G: **Evolutionary insights on C_4 photosynthetic subtypes in grasses from genomics and phylogenetics**. *Genome Biol Evol* 2009, **1**:221-230.

4. Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J *et al.*: **Reference genome sequence of the model plant Setaria**. *Nat Biotechnol* 2012, **30**:555-561.

5. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al.*: **The *Sorghum bicolor* genome and the diversification of grasses**. *Nature* 2009, **457**:551-556.

6. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al.*: **The B73 maize genome: complexity, diversity, and dynamics**. *Science (New York)* 2009, **326**:1112-1115.

7. Kajala K, Covshoff S, Karki S, Woodfield H, Tolley BJ, Dionora MJA, Mogul RT, Mabilangan AE, Danila FR, Hibberd JM

et al.: **Strategies for engineering a two-celled C_4 photosynthetic pathway into rice**. *J Exp Bot* 2011, **62**:3001-3010.

8. Weissmann S, Brutnell TP: **Engineering C_4 photosynthetic regulatory networks**. *J Biosci Bioeng* 2012, **23**:298-304.

9. Leegood RC: **Strategies for engineering C_4 photosynthesis**. *J Plant Physiol* 2013, **170**:378-388.

10. • Amthor JS: **From sunlight to phytomass: on the potential efficiency of converting solar radiation to phyto-energy**. *New Phytol* 2010, **188**:939-959.
An excellent review on energy conversion in plants, including C_3 and C_4 crops.

11. Osborne CP, Sack L: **Evolution of C_4 plants: a new hypothesis for an interaction of CO_2 and water relations mediated by plant hydraulics**. *Philos Trans R Soc B* 2012, **367**:583-600.

12. • Edwards EJ, Osborne CP, Strömberg CAE, Smith SA, C_4 Grasses Consortium, Bond WJ, Christin P-A, Cousins AB, Duvall MR, Fox DL *et al.*: **The origins of C_4 grasslands: integrating evolutionary and ecosystem science**. *Science (New York)* 2010, **328**:587-591.
This manuscript summaries the evolutionary scenarios and drivers that contributed to the expansion of C_4 grasslands.

13. Matthews E: **Global vegetation and land use: new high-resolution data bases for climate studies**. *J Appl Meteorol* 1983, **22**:474-487.

14. Hibberd JM, Covshoff S: **The regulation of gene expression required for C_4 photosynthesis**. *Annu Rev Plant Biol* 2010, **61**:181-207.

15. • Bräutigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Maß J, Lercher MJ *et al.*: **An mRNA blueprint for C_4 photosynthesis derived from comparative transcriptomics of closely related C_3 and C_4 species**. *Plant Physiol* 2011, **155**:142-156.
The first report of a transcriptomic comparison of related C_3 and C_4 species. More than 800 genes were identified as differentially expressed between leaves of a C_4 and a C_3 species.

16. • Gowik U, Bräutigam A, Weber KL, Weber APM, Westhoff P: **Evolution of C_4 photosynthesis in the genus *Flaveria*: how many and which genes does it take to make C_4?** *Plant Cell* 2011, **23**:2087-2105.
In this work mRNA-Seq was used to quantify leaf transcriptomes of C_3, C_3–C_4 intermediate and C_4 species within the genus *Flaveria*. C_3–C_4 intermediate species show higher expression of photorespiratory genes than C_3, which emphasizes the importance of C_2 photosynthesis for the evolution of C_4.

17. Sage RF, Christin PA, Edwards EJ: **The C_4 plant lineages of planet Earth**. *J Exp Bot* 2011, **62**:3155-3169.

18. Furbank RT: **Evolution of the C_4 photosynthetic mechanism: are there really three C_4 acid decarboxylation types?** *J Exp Bot* 2011, **62**:3103-3108.

19. • Pick TR, Bräutigam A, Schlüter U, Denton AK, Colmsee C, Scholz U, Fahnenstich H, Pieruschka R, Rascher U, Sonnewald U *et al.*: **Systems analysis of a maize leaf developmental gradient redefines the current C_4 model and provides candidates for regulation**. *Plant Cell* 2011, **23**:4208-4220.
Amongst other findings, this study shows that maize operates a branched C_4 decarboxylation chemistry that involves both NADP malic enzyme and PEP carboxykinase.

20. Sommer M, Bräutigam A, Weber APM: **The dicotyledonous NAD malic enzyme C_4 plant *Cleome gynandra* displays age-dependent plasticity of C_4 decarboxylation biochemistry**. *Plant Biol (Stuttg)* 2012, **14**:621-629.

21. • Muhaidat R, McKown AD: **Significant involvement of PEP-CK in carbon assimilation of C_4 eudicots**. *Ann Bot* 2013 http://dx.doi.org/10.1093/aob/mct017.
PEP-CK C_4 decarboxylation was believed being confined to monocotyledonous C_4 species. This work shows that this is not the case.

22. Weber APM, Caemmerer von S: **Plastid transport and metabolism of C_3 and C_4 plants — comparative analysis and possible biotechnological exploitation**. *Curr Opin Plant Biol* 2010, **13**:257-265.

23. Christin P-A, Besnard G, Samaritani E, Duvall MR, Hodkinson TR, Savolainen V, Salamin N: **Oligocene CO_2 decline promoted C_4 photosynthesis in grasses**. *Curr Biol* 2008, **18**:37-43.

24. Muhaidat R, Sage TL, Frohlich MW, Dengler NG, Sage RF: **Characterization of C_3–C_4 intermediate species in the genus *Heliotropium* L. (Boraginaceae): anatomy, ultrastructure and enzyme activity**. *Plant Cell Environ* 2011, **34**:1723-1736.

25. Gowik U, Westhoff P: **The path from C_3 to C_4 photosynthesis**. *Plant Physiol* 2011, **155**:56-63.

26. Griffiths H, Weller G, Toy LFM, Dennis RJ: **You're so vein: bundle sheath physiology, phylogeny and evolution in C_3 and C_4 plants**. *Plant Cell Environ* 2013, **36**:249-261.

27. •• Christin P-A, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ: **Anatomical enablers and the evolution of C_4 photosynthesis in grasses**. *Proc Natl Acad Sci U S A* 2012 http://dx.doi.org/10.1073/pnas.1216777110.
Anatomical and physiological traits were mapped onto a phylogeny of grasses. Bundle sheath cell size and intervein distance were identified as anatomical enablers for evolving C_4 photosynthesis. Apparently, the BEP clade of grasses lost the capacity to evolve large bundle sheath cells.

28. Edwards EJ, Smith SA: **Phylogenetic analyses reveal the shady history of C_4 grasses**. *Proc Natl Acad Sci U S A* 2010, **107**:2532-2537.

29. Christin P-A, Freckleton RP, Osborne CP: **Can phylogenetics identify C_4 origins and reversals?** *Trends Ecol Evol* 2010, **25**:403-409.

30. Osborne CP, Freckleton RP: **Ecological selection pressures for C_4 photosynthesis in the grasses**. *Philos Trans R Soc B* 2009, **276**:1753-1760.

31. Sun J, Qi L, Li Y, Chu J, Li C: **PIF4-mediated activation of YUCCA8 expression integrates temperature into the auxin pathway in regulating *Arabidopsis* hypocotyl growth**. *PLoS Genet* 2012, **8**:e1002594.

32. Wenzel CL, Marrison J, Mattsson J, Haseloff J, Bougourd SM: **Ectopic divisions in vascular and ground tissues of *Arabidopsis thaliana* result in distinct leaf venation defects**. *J Exp Bot* 2012, **63**:5351-5364.

33. • Sage RF, Sage TL, Kocacinar F: **Photorespiration and the evolution of C_4 photosynthesis**. *Annu Rev Plant Biol* 2012, **63**:19-47.
Authoritative review of the tight connection between photorespiration and the evolution of C_4 photosynthesis.

34. Sage TL, Sage RF: **The functional anatomy of rice leaves: implications for refixation of photorespiratory CO_2 and efforts to engineer C_4 photosynthesis into rice**. *Plant Cell Physiol* 2009, **50**:756-772.

35. • Busch FA, Sage TL, Cousins AB, Sage RF: **C_3 plants enhance rates of photosynthesis by reassimilating photorespired and respired CO_2**. *Plant Cell Environ* 2013, **36**:200-212.
A clear demonstration that re-assimilation of photorespiratory CO_2 can be achieved through proto-Kranz-anatomy.

36. Kadereit G, Ackerly D, Pirie MD: **A broader model for C_4 photosynthesis evolution in plants inferred from the goosefoot family (Chenopodiaceae s.s.)**. *Philos Trans R Soc B* 2012, **279**:3304-3311.

37. Brodribb TJ, Feild TS: **Leaf hydraulic evolution led a surge in leaf photosynthetic capacity during early angiosperm diversification**. *Ecol Lett* 2010, **13**:175-183.

38. Brodribb TJ, Feild TS, Sack L: **Viewing leaf structure and evolution from a hydraulic perspective**. *Funct Plant Biol* 2010, **37**:488-498.

39. Shatil-Cohen A, Attia Z, Moshelion M: **Bundle-sheath cell regulation of xylem-mesophyll water transport via aquaporins under drought stress: a target of xylem-borne ABA?** *Plant J* 2011, **67**:72-80.

40. Weber APM, Bräutigam A: **The role of membrane transport in metabolic engineering of plant primary metabolism**. *Curr Opin Biotechnol* 2012 http://dx.doi.org/10.1016/j.copbio.2012.09.010.

41. Jacobs B, Engelmann S, Westhoff P, Gowik U: **Evolution of C_4 phosphoenolpyruvate carboxylase in *Flaveria*: determinants for high tolerance towards the inhibitor L-malate**. *Plant Cell Environ* 2008, **31**:793-803.

42. Friso G, Majeran W, Huang M, Sun Q, van Wijk KJ: **Reconstruction of metabolic pathways, protein expression, and homeostasis machineries across maize bundle sheath and mesophyll chloroplasts: large-scale quantitative proteomics using the first maize genome assembly**. *Plant Physiol* 2010, **152**:1219-1250.

43. Majeran W, Friso G, Ponnala L, Connolly B, Huang M, Reidel E, Zhang C, Asakura Y, Bhuiyan NH, Sun Q *et al.*: **Structural and metabolic transitions of C_4 leaf development and differentiation defined by microscopy and quantitative proteomics in maize**. *Plant Cell* 2010, **22**:3509-3542.

44. Akyildiz M, Gowik U, Engelmann S, Koczor M, Streubel M, Westhoff P: **Evolution and function of a *cis*-regulatory module for mesophyll-specific gene expression in the C_4 dicot *Flaveria trinervia***. *Plant Cell* 2007, **19**:3391-3402.

45. Tolley BJ, Woodfield H, Wanchana S, Bruskiewich R, Hibberd JM:
• **Light-regulated and cell-specific methylation of the maize PEPC promoter**. *J Exp Bot* 2012, **63**:1381-1390.
This work shows that mesophyll-cell-specific-expression of PEPC in maize is co-controlled by methylation of specific cytosine residues in *cis*-elements of the PEPC promoter.

46. Brown NJ, Newell CA, Stanley S, Chen JE, Perrin AJ, Kajala K,
•• Hibberd JM: **Independent and parallel recruitment of preexisting mechanisms underlying C_4 photosynthesis**. *Science (New York)* 2011, **331**:1436-1439.
The authors demonstrate that the abundant bundle sheath cell-specific accumulation of NAD-ME in *Cleome gyandra* and NADP-ME in maize is mediated by post-transcriptional mechanisms.

47. Christin P-A, Edwards EJ, Besnard G, Boxall SF, Gregory R,
•• Kellogg EA, Hartwell J, Osborne CP: **Adaptive evolution of C_4 photosynthesis through recurrent lateral gene transfer**. *Curr Biol* 2012, **22**:445-449.
Demonstration of lateral transfer of genes encoding the C_4 isozymes of PEPC and PCK between distantly related grass species, which constitutes a new paradigm for the acquisition of C_4 subtraits.

48. Christin PA, Wallace MJ, Clayton H, Edwards EJ, Furbank RT,
•• Hattersley PW, Sage RF, Macfarlane TD, Ludwig M: **Multiple photosynthetic transitions, polyploidy, and lateral gene transfer in the grass subtribe Neurachninae**. *J Exp Bot* 2012, **63**:6297-6308.
Further evidence for lateral transfer of C_4 genes, similar to Ref. [47].

49. Weise SE, Schrader SM, Kleinbeck KR, Sharkey TD: **Carbon
•• balance and circadian regulation of hydrolytic and phosphorolytic breakdown of transitory starch**. *Plant Physiol* 2006, **141**:879-886.
This work clearly shows that the plastidial transitory starch pool is accessible in the light and under photorespiratory conditions can be used to replenish acceptor molecules for the oxygenation reaction of RubisCO.

50. Grass Phylogeny Working Group II: **New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins**. *New Phytol* 2011, **193**:304-312.

3.2 Manuscript KD: Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C_3 and C_4 plant species

Overview

Title: Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C_3 and C_4 plant species

Authors: Canan Külahoglu, Alisandra K. Denton, Manuel Sommer, Janina Maß, Simon Schliesky, Thomas J. Wrobel, Barbara Berckmans, Elsa Gongora-Castillo, C. Robin Buell, Rüdiger Simon, Lieven De Veylder, Andrea Bräutigam and Andreas P.M. Weber

Published in Plant Cell, August 2014

Impact factor: 9.575

Co-first authorship

Main Findings

This manuscript performed a comprehensive analysis of differences between two C_3 and C_4 Cleomaceae species, with a focus on differences in transcription and leaf development. Transcriptional differences during leaf development combined with anatomical measurements increased our understanding of two features of C_4 photosynthesis: close vein spacing and enlarged BS cells. Of interest to vein spacing, the C_4 species showed a delay in tissue differentiation, and a matching delay in many transcriptional changes along the leaf gradient. This included a delay in the up-regulation of photosynthetic genes, and prolonged expression of the light-sensing Constitutive Photomorphogenesis 9 signalosome. As differentiation has been reported to limit vein formation, this delay could allow for higher vein density in the C_4 species. Of interest to enlarged BS, the C_4 species showed an enlargement of BS nuclei under microscopy and an overall increased level of ploidy consistent with 4N or 8N nuclei in the BS. At the level of transcription, the transcription factor GT-2-LIKE1, which suppresses endoreduplication, was down regulated in the C_4 species.

Contributions

- Assistance with enzyme assays
- Organization and analysis of transcriptional abundance data
- Clustering, differential expression testing, enrichment testing, etc...
- Data interpretation and analysis

- Assistance in writing manuscript
- Comprehensive editing of manuscript

The Plant Cell, Vol. 26: 3243–3260, August 2014, www.plantcell.org

RESEARCH ARTICLES

Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C_3 and C_4 Plant Species[C][W][OPEN]

Canan Külahoglu,[a,1] Alisandra K. Denton,[a,1] Manuel Sommer,[a] Janina Maß,[b] Simon Schliesky,[a] Thomas J. Wrobel,[a] Barbara Berckmans,[c] Elsa Gongora-Castillo,[d] C. Robin Buell,[d] Rüdiger Simon,[c] Lieven De Veylder,[e,f] Andrea Bräutigam,[a,1] and Andreas P.M. Weber[a,2]

[a] Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences, Heinrich-Heine-University, 40225 Düsseldorf, Germany
[b] Institute of Informatics, Cluster of Excellence on Plant Sciences, Heinrich-Heine University, 40225 Düsseldorf, Germany
[c] Institute of Developmental Genetics, Cluster of Excellence on Plant Sciences, Heinrich-Heine-University, 40225 Düsseldorf, Germany
[d] Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824
[e] Department of Plant Systems Biology, VIB, B-9052 Gent, Belgium
[f] Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Gent, Belgium

C_4 photosynthesis outperforms the ancestral C_3 state in a wide range of natural and agro-ecosystems by affording higher water-use and nitrogen-use efficiencies. It therefore represents a prime target for engineering novel, high-yielding crops by introducing the trait into C_3 backgrounds. However, the genetic architecture of C_4 photosynthesis remains largely unknown. To define the divergence in gene expression modules between C_3 and C_4 photosynthesis during leaf ontogeny, we generated comprehensive transcriptome atlases of two Cleomaceae species, *Gynandropsis gynandra* (C_4) and *Tarenaya hassleriana* (C_3), by RNA sequencing. Overall, the gene expression profiles appear remarkably similar between the C_3 and C_4 species. We found that known C_4 genes were recruited to photosynthesis from different expression domains in C_3, including typical housekeeping gene expression patterns in various tissues as well as individual heterotrophic tissues. Furthermore, we identified a structure-related module recruited from the C_3 root. Comparison of gene expression patterns with anatomy during leaf ontogeny provided insight into genetic features of Kranz anatomy. Altered expression of developmental factors and cell cycle genes is associated with a higher degree of endoreduplication in enlarged C_4 bundle sheath cells. A delay in mesophyll differentiation apparent both in the leaf anatomy and the transcriptome allows for extended vein formation in the C_4 leaf.

INTRODUCTION

C_4 photosynthesis has evolved concurrently and convergently in angiosperms more than 65 times from the ancestral C_3 state (Sage et al., 2011) and provides fitness and yield advantages over C_3 photosynthesis under permissive conditions, such as high temperatures (Hatch, 1987; Sage, 2004). In brief, C_4 photosynthesis represents a biochemical CO_2 pump that supercharges photosynthetic carbon assimilation through the Calvin-Benson-Bassham cycle (CBBC) by increasing the concentration of CO_2 at the site of its assimilation by the enzyme Rubisco (Andrews and Lorimer, 1987; Furbank and Hatch, 1987). Rubisco is a bifunctional enzyme that catalyzes both the productive carboxylation and the futile oxygenation of ribulose 1,5-bisphosphate. The oxygenation reaction produces a toxic byproduct, 2-phosphoglycolic acid (Anderson, 1971), which is removed by an energy-intensive metabolic repair process called photorespiration. By concentrating CO_2 through the C_4 cycle, the oxygenation of ribulose 1,5-bisphosphate and thereby photorespiration is massively reduced. However, the C_4 cycle requires input of energy to drive the CO_2 pump. Photorespiration increases with temperature and above ~23°C, the energy requirements of metabolic repair become higher than the energy cost of the C_4 cycle (Ehleringer and Björkman, 1978; Ehleringer et al., 1991). Hence, operating C_4 photosynthesis is beneficial at high leaf temperatures, whereas C_3 photosynthesis prevails in cool climates (Ehleringer et al., 1991; Zhu et al., 2008).

With a few exceptions, C_4 photosynthesis requires specialized Kranz anatomy (Haberlandt, 1896), in which two distinct cell types share the photosynthetic labor, namely, mesophyll cells (MCs) and bundle sheath cells (BSCs). MCs surround the BSCs in a wreath-like manner and both cell types form concentric rings around the veins. This leads to a stereotypic vein-BSC-MC-MC-BSC-vein pattern (Brown, 1975). MCs serve as carbon pumps that take in CO_2 from the leaf intercellular air space, convert it into a C_4 carbon compound, and load it into the BSCs. Here, CO_2 is released from the C_4 compound and assimilated

[1] These authors contributed equally to this work.
[2] Address correspondence to andreas.weber@uni-duesseldorf.de.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Andreas P.M. Weber (andreas.weber@uni-duesseldorf.de).
[C] Some figures in this article are displayed in color online but in black and white in the print edition.
[W] Online version contains Web-only data.
[OPEN] Articles can be viewed online without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.114.123752

into biomass by the CBBC, and the remaining C_3-compound is returned to the MC to be loaded again with CO_2. The carbon pump runs at a higher rate than the CBBC (overcycling), which leads to an increased concentration of CO_2 in the BSCs. Our understanding of the different elements required for C_4 photosynthesis varies, with many components of the metabolic cycle known, while their interplay and regulation remain mostly enigmatic, and very little is known about their anatomical control (Sage and Zhu, 2011).

C_4 photosynthesis can be considered a complex trait, since it requires changes to the expression levels of hundreds or perhaps thousands of genes (Bräutigam et al., 2011, 2014; Gowik et al., 2011). While complex traits are typically dissected by measuring the quantitative variation across a polymorphic population, this approach is not promising for C_4 photosynthesis, due to lack of known plasticity in "C_4-ness" (Sage and McKown, 2006). Historical crosses between C_3 and C_4 plants (Chapman and Osmond, 1974) are no longer available and would have to be reconstructed before they can be analyzed with molecular tools.

Alternatively, closely related C_3 and C_4 species provide a platform for studying C_4 photosynthesis. In the Cleomaceae and Asteraceae, comparative transcriptomic analyses have identified more than 1000 genes differentially expressed between closely related C_3 and C_4 species (Bräutigam et al., 2011; Gowik et al., 2011). These studies, however, compared the end points of leaf development, i.e., fully matured photosynthetic leaves. Therefore, they do not provide insight into the dynamics of gene expression during leaf ontogeny, which is important for understanding the establishment of C_4 leaf anatomy. Systems analyses of maize (*Zea mays*) leaf gradients have provided a glimpse into developmental gene expression modules (Li et al., 2010; Pick et al., 2011; Wang et al., 2013); however, maize lacks a close C_3 relative and has simple parallel venation making any generalizations to dicot leaf development difficult.

Tarenaya hassleriana, previously known as *Cleome hassleriana* (Iltis and Cochrane, 2007; Iltis et al., 2011), which is a C_3 plant, and *Gynandropsis gynandra* (previously known as *Cleome gynandra*), which is a derived C_4 plant, represent an ideal pair for a comparative analysis of the complex trait of C_4 photosynthesis (Bräutigam et al., 2011). Both species belong to the family of Cleomaceae, are closely related to each other and to the well-annotated C_3 plant model species *Arabidopsis thaliana* (Brown et al., 2005; Marshall et al., 2007; Inda et al., 2008), and both Cleome sister lineages share many traits (Iltis et al., 2011). In addition, the genome of *T. hassleriana* has been recently sequenced and serves as a reference for expression profiling via RNA sequencing (Cheng et al., 2013).

In this study, we take advantage of the phylogenetic proximity between *G. gynandra* and *T. hassleriana* to compare the dynamic changes in gene expression during leaf development (Inda et al., 2008). We generated a transcriptome atlas for each species, consisting of three biological replicates of six different stages of leaf development, three different stages of each seed and seedling development, reproductive organs (carpels, stamen, petals, and sepals), stems, and roots. In parallel, we performed microscopy analysis of the leaf anatomy. Finally, we measured leaf cell ploidy levels by flow cytometry and measurements of nuclear size in different leaf cell types by confocal laser scanning microscopy.

RESULTS

Selection of Tissues Featured in the Comparative Atlases

For high-resolution characterization of photosynthetic development between a dicotyledonous C_3 and C_4 species, a leaf developmental gradient was defined. Stage 0 was the youngest sampled leaf, 2 mm in length, and not yet emerged from the apex. The stage 0 leaves are the first to show a discernible palmate shape and contain the first order vein (midrib vein) in both species (Figure 1A; Supplemental Figure 1A). New leaves emerged from the apex every 2 d (plastochron = 2 d) in both species and were numbered sequentially from the aforementioned stage 0 to stage 5 (Figure 1A). The leaves emerge and initiate secondary vein formation at stage 1 (Supplemental Figure 1B) and fully mature by stages 4 and 5 (Supplemental Figures 1E and 1F). The mature leaf of the C_4 species has more minor veins (up to 7°) than that of the C_3 species (up to 6°; Supplemental Figure 1F). The leaf expansion rate is initially indistinguishable and never significantly different between the species (Figure 1B). The sampled leaf gradient covered the development from non-light-exposed sink tissues to fully photosynthetic source tissues.

Complementary to this and to provide a broader comparison between C_3 and C_4 plants, seedlings, minor photosynthetic, and

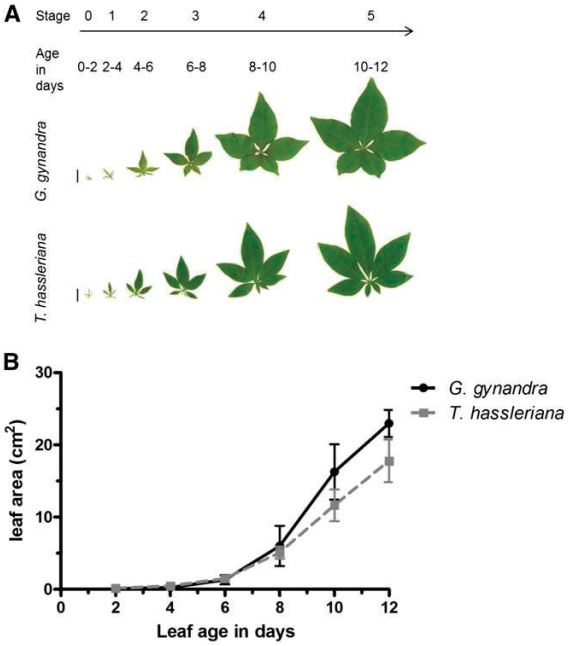


Figure 1. Overview of Leaf Shape and Expansion Rate in *G. gynandra* and *T. hassleriana*.

(A) Image of each leaf category sequenced (bar = 1 cm). Each category is 2 d apart from the other.

(B) Leaf expansion rate of each leaf category in cm^2 over 12 d (n = 5; ±SD)

[See online article for color version of this figure.]

heterotrophic tissues were selected for further characterization. The aerial portion of seedlings (cotyledon and hypocotyl) was sampled 2, 4, and 6 d after germination to cover early cotyledon maturation (Supplemental Figure 2). The full root system and stem tissue were sampled from plants after 6 to 8 weeks of growth before inflorescence emergence (Supplemental Figure 3A); floral organs (petals, carpels, stamen, and sepals) were harvested during flowering of 10- to 14-week-old plants as well as three different stages of seed development (Supplemental Figure 3B). In total, 10 phototrophic and 8 heterotrophic tissues per species were included in the atlases (Table 1).

The C_3 and C_4 Transcriptomes Are of High Quality and Comparable between Species

Cross-species mapping provided a more reliable data set than de novo transcriptome assembly. Between 1.4 and 67 million high-quality reads were generated per replicate (Supplemental Data Set 2). Initially, paired-end reads from each tissue were assembled by VELVET/OASES (Supplemental Table 1). Comparing the resulting contigs to reference data, including the *T. hassleriana* genome (Cheng et al., 2013), revealed several quality issues. These include excessive numbers of contigs mapping to single loci, fused and fragmented contigs, and the absence of C_4 transcripts known to be highly expressed in *G. gynandra* (Supplemental Figures 4A to 4C and Supplemental Data Set 3). As an alternative, we aligned single-end reads from both species to the recently sequenced *T. hassleriana* genome (Cheng et al., 2013). Albeit slightly lower, the mapping efficiency and specificity remained comparable between both species with 60 to 70% of reads mapped for both leaf gradients (Supplemental Data Set 1). To define an upper boundary for any artifacts caused by cross-species mapping, three *T. hassleriana* samples (mature leaf stage 5, stamen, and young seed) were mapped to *Arabidopsis*. The correlation between replicates was equivalent in reads mapped to the cognate genome and across species with an average $r = 0.98$. Furthermore, there was a strong correlation between both mappings, reaching an average Pearson correlation of $r = 0.86$ after collapsing expression data to *Arabidopsis* identifiers to minimize bias from different genome duplication histories (Supplemental Table 2 and Supplemental Figure 5). Cross-species mapping has been successfully used for inter species comparisons before (Bräutigam et al., 2011, 2014; Gowik et al., 2011), and in this study mapping of both species to the *T. hassleriana* genome provided a quality data set with a limited degree of artifacts.

The generated transcriptome atlases were reproducible and comparable between species. To reduce noise, downstream analyses focused on genes expressed above 20 reads per mappable million (RPKM; Supplemental Figure 6), unless otherwise noted. Biological replicates of each tissue clustered closely together and were highly correlated (mean $r = 0.92$, median $r = 0.97$; Figure 2A; Supplemental Figures 7A and 7B and Supplemental Table 3). On average, 4686 and 5308 genes displayed significantly higher expression values in *G. gynandra* and *T. hassleriana*, respectively, with the greatest differences observed in seed and stem tissue (Supplemental Table 4). In contrast, the transcriptome patterns were highly similar between the sister species (Figure 2A; Supplemental Figure 7C). Principle component analysis (PCA) showed that the first component separated the species and accounted for only 15% of the total variation (Supplemental Figure 8A).

Table 1. Sequencing and Mapping Stats for Each Averaged Tissue Sample in *T. hassleriana* and *G. gynandra*

		T. hassleriana			*G. gynandra*		
		Total No. of Reads in Three Replicates	No. of Genes Expressed > 1 RPKM	No. of Genes Expressed > 1000 RPKM	Total No. of Reads in Three Replicates	No. of Genes Expressed > 1 RPKM	No. of Genes Expressed > 1000 RPKM
Leaf gradient	0	58,874,878	23,238	64	75,895,556	22,357	104
	1	59,389,701	23,134	74	66,822,298	22,021	133
	2	63,590,283	23,104	81	55,247,053	22,143	129
	3	90,654,684	23,004	90	75,944,275	21,854	144
	4	36,572,303	22,844	106	69,951,930	21,734	119
	5	102,018,867	22,905	106	69,639,670	21,039	119
Floral organs	Sepal	103,721,357	23,656	74	77,430,418	23,145	83
	Petal	21,754,853	21,379	86	10,872,686	21,322	77
	Stamen	57,929,412	22,642	140	55,748,506	22,489	133
	Carpel	28,021,839	23,910	67	4,929,824	23,577	76
	Stem	30,932,633	23,292	75	59,516,389	22,508	98
	Root	88,911,824	24,255	68	86,879,963	23,430	89
Seedling	2 DAG	90,777,012	23,306	120	89,262,140	21,960	130
	4 DAG	89,517,055	23,041	116	112,658,149	22,036	130
	6 DAG	71,271,739	22,877	138	64,470,699	21,910	136
Seed maturation	1	52,229,844	23,708	118	32,763,383	22,991	118
	2	31,872,067	22,969	145	29,958,720	22,262	148
	3	53,271,349	21,737	138	56,453,325	20,082	152

Reads were normalized as RPKM ($n = 3$). DAG, days after germination.

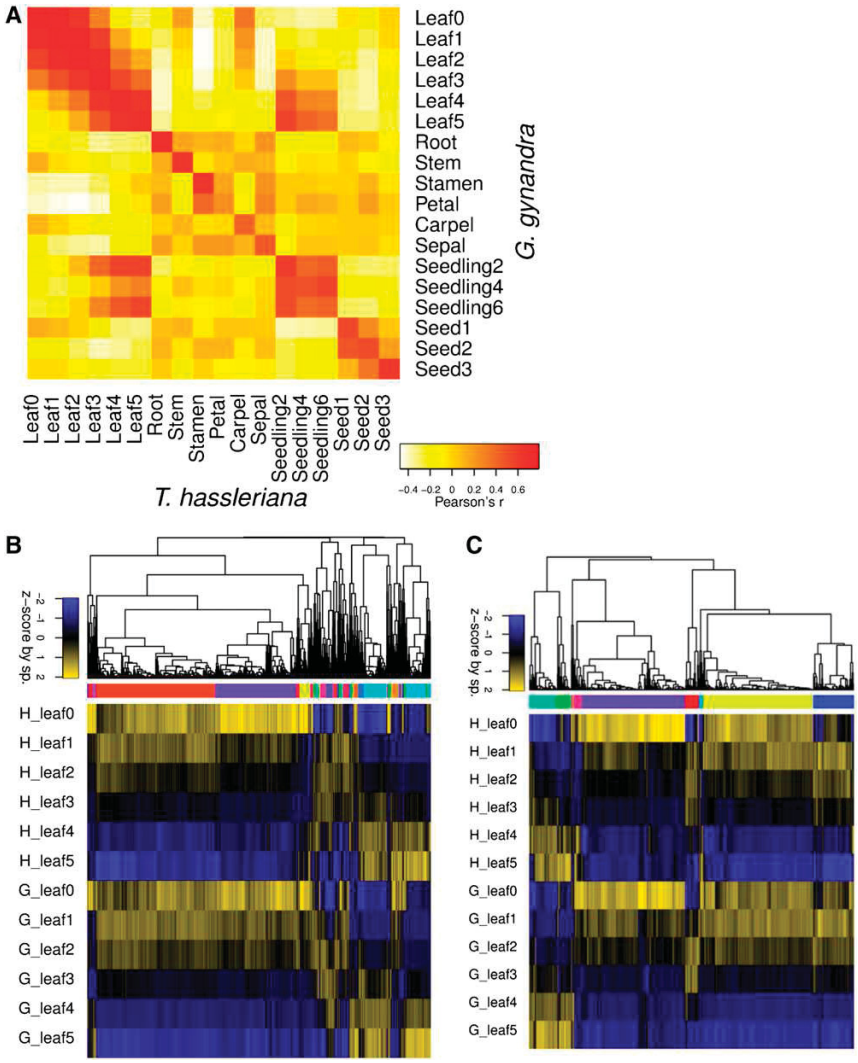


Figure 2. Comparative Tissue Dynamics and Gene Expression Pattern between *G. gynandra* and *T. hassleriana*.

(A) Pearson's correlation heat map of the expression of tissue-specific signature genes (RPKM) of all leaf gradient sample averages ($n = 3$) per species. Yellow, low expression; red, high expression. G, *G. gynandra*; H, *T. hassleriana*.

(B) Pearson's correlation hierarchical cluster of all leaf gradient sample averages as Z-scores. Blue is the lowest expression and yellow the highest expression.

(C) Expression patterns of transcriptional regulators in both species within the leaf gradient. Pearson's correlation hierarchical cluster of all sample averages as Z-scores. Blue is the lowest expression and yellow the highest expression.

Gene expression patterns and dynamics are conserved between species. The number of genes expressed above 20 RPKM varied by tissue from 6900 to 12,000, with the fewest in the mature leaf and most in the stem and youngest leaf in both species (Table 1; Supplemental Data Set 2). Hierarchical clustering revealed major modules with increasing and decreasing expression along the leaf gradient (Figure 2B), a large overlap of peak expression between seedlings and mature tissue, and distinct gene sets for the other sampled tissues (Supplemental Figure 9A). In leaves, the genes with decreasing expression split into two primary clusters, of which the smaller cluster maintained higher expression longer in the C_4 than the C_3 species (Figure 2B). Clustering of the tissues with 10,000 bootstrap replications confirmed the visual similarity of mature leaves and seedlings and showed further major branches consisting of (1) carpel, stem, and root; (2) a seed gradient and remaining floral

organs; and (3) young leaves (Supplemental Figure 9A). Limiting the clustering to transcription factors (TFs) showed equivalent results (Supplemental Figure 9B; Figure 2C), except that in leaves, a higher proportion of the TFs with decreasing expression maintained expression longer in the C_4 species. Notably, this delay impacted the clustering of the tissues and older C_4 leaves tended to cluster with younger C_3 leaves by TF expression (Supplemental Figures 9A and 9B). The delay was further reflected in a PCA of the leaf gradient where stage 0 and 1 show much less separation in *G. gynandra* than in *T. hassleriana* (Supplemental Figure 8B).

The functional categories with dominant expression showed distinct patterns across the tissues and high conservation between the species. As in the hierarchical clustering, the species showed similar profiles when examining the number of signature genes (expressed over 1000 RPKM; Figure 3) or the total RPKM (Supplemental Figure 9) in each functional category. As expected, in mature leaves and seedlings, transcriptional activity is dominated by photosynthesis, which is almost entirely lacking from roots, seeds, stamens, and petals (Figure 3; Supplemental Figure 9). Younger leaf tissues of the C_3 species show higher expression of genes in the photosynthetic category, displayed as signature genes (Figure 3) or as cumulative RPKM per category (Supplemental Figure 9). In all floral tissues, roots, and stems, transcriptional activity is comparatively balanced between categories. In seeds, a major portion of the total expression is allocated to a few, extremely highly expressed lipid transfer protein type seed storage proteins (Supplemental Figure 9). The differences between the two species lie in the details, especially within the developmental leaf gradient. In young *G. gynandra* leaves, more signature genes encode DNA and protein-associated MapMan terms than in *T. hassleriana* (Figure 3). A close examination of secondary MapMan categories shows that specifically histone proteins (34 genes with $P < 0.05$ in stage 1, enriched with Fisher's exact test $P = 2.6 \cdot 10^{-13}$) and protein synthesis (222 genes with $P < 0.05$ in stage 1, enriched with Fisher's exact test $P = 1.8 \cdot 10^{-17}$) are upregulated in *G. gynandra* and that these categories have a larger dynamic range in *G. gynandra* than *T. hassleriana* (Supplemental Figure 10).

In summary, transcriptomic analysis indicates the tissues are well paired and comparable between species and although there are differences in expression level, there is conservation of expression patterns between species. Within the leaf gradient, there is a subset of genes that shows a delay in the onset of expression changes in *G. gynandra*.

The Comparative Transcriptome Atlases Revealed Diverse Recruitment Patterns from the C_3 Plant *T. hassleriana* to C_4 Photosynthesis

The expression patterns of the core C_4 cycle genes were compared in *G. gynandra* and *T. hassleriana* to gain insight into the evolutionary recruitment of C_4 cycle genes to photosynthesis. During convergent evolution of C_4 photosynthesis, these genes

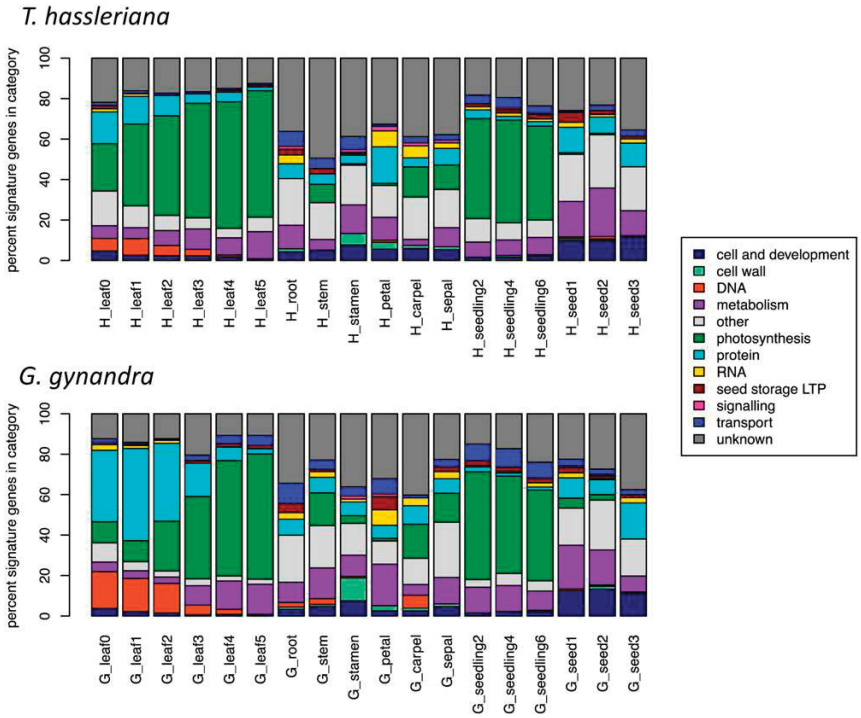


Figure 3. Distribution of Signature Genes in Each Tissue in *G. gynandra* and *T. hassleriana*.

Percentage of signature genes expressed over 1000 RPKM falling in each basal MapMan category for every averaged tissue.

were recruited from ancestral C_3 genes (Sage, 2004; Edwards et al., 2010; Sage et al., 2011). To contextualize the change in expression of the C_4 cycle genes, the between species Euclidean (absolute) and Pearson (pattern) distances were calculated and compared from the leaf developmental gradients (Figure 4A). All known C_4 cycle genes showed a large Euclidean distance (844 to 9156 RPKM), while they split between a correlated and an inversely correlated pattern. In addition to the known C_4 genes, histones, lipid transfer proteins, protein synthesis, and DNA synthesis are functional categories found among genes with greater than 844 RPKM differences in absolute expression (Supplemental Data Set 6).

To identify ancestral C_3 expression domains from which C_4 genes were recruited, the expression of the core C_4 cycle genes was compared between species. In *G. gynandra*, all core C_4 cycle genes increase in expression along the leaf gradient and are high in seedlings (Figures 4C and 4D; Supplemental Figures 12A to 12F); this pattern matches that of other photosynthetic genes (Figure 4B). For each C_4 cycle gene, the *T. hassleriana* sequence to which most *G. gynandra* reads mapped was taken as the most likely closest putative ortholog (Supplemental Figures 13 and 14). The putative orthologs of core C_4 genes are expressed at comparatively low levels in C_3 (Supplemental Figures 13 and 14). Activity measurements of the core C_4 cycle enzymes match the observed gene expression profiles (Supplemental Figure 15). In contrast to leaves and seedlings, the remaining tissues show a variety of expression patterns of C_4 cycle genes in both species (Figures 4C to 4E; Supplemental Figures 12A to 12G). Of the C_4 cycle genes, *NAD-MALIC ENZYME* (*NAD-ME*) and the *SODIUM: HYDROGEN ANTIPORTER* (*NHD*) show a fairly constitutive expression pattern in C_3, while the others have a small number of tissues where the expression peaks (Figure 4C; Supplemental Figure 12A). The expression of *PYRUVATE PHOSPHATE DIKINASE* (*PPDK*), the *PHOSPHO*ENOL*PYRUVATE TRANSLOCATOR* (*PPT*), and *DICARBOXYLATE CARRIER* (*DIC*) peaks in floral organs (Supplemental Figures 12B and 12C; Figure 4D); the expression of *ASPARTATE AMINO TRANSFERASE* (*AspAT*) and *ALANINE AMINOTRANSFERASE* (*AlaAT*) peaks in seed (Figure 4E; Supplemental Figure 12D); and the expression of the pyruvate transporter *BILE ACID:SODIUM SYMPORTER FAMILY PROTEIN2* (*BASS2*) peaks in the young leaf (Supplemental Figure 12E). Albeit erroneous identification of the closest C_3 ortholog in some cases (e.g., *BASS2* and *PHOSPHO*ENOL*PYRUVATE CARBOXYLASE* [*PEPC*]) impedes identification of the ancestral C_3 expression domain (Supplemental Figures 12 and 13), the majority of known C_4 cycle genes were recruited to a photosynthetic expression pattern from a variety of expression domains (Figure 4B).

To assess the possibility of small modular recruitment from other tissues to the C_4 leaf, we searched for evidence of an expression shift between the C_3 root and the C_4 leaf. This shift is expected, if the bundle sheath tissue is partially derived from the regulatory networks of root endodermis, as proposed previously (Slewinski, 2013). Expression pattern filters were used to identify 37 genes that were expressed primarily in the C_3 root and the C_4 leaf (C_3 leaf/root < 0.3; C_4/C3 leaf > 1; C_4 leaf4-5/root > 0.5; C_4 leaf5 > 30 RPKM; leaf5/root enrichment 6-fold greater in C_4), significantly more than in a randomized data set (P value < 10^{-29}; Supplemental Table 5). This set of genes showed a very similar expression pattern to photosynthetic genes along the C_4 leaf gradient (Figure 5A).

The functions encoded by the genes that were apparently recruited to the leaf from a root expression domain were consistent with structural modifications and C_4 photosynthesis. In *Arabidopsis*, 29 of the corresponding homologs are heterogeneously expressed across different root tissues with their highest expression in either the endodermis or cortex, analogous to bundle sheath and mesophyll cells, respectively (Slewinski, 2013). Three functional groups could be identified in the cluster. The first is related to tissue structure, i.e., cell wall modification and plasmodesmata, the second to metabolic flux and redox balance, and the third to signaling (Figure 5B). Among these genes are two C_4 cycle genes, namely, *DIC1*, and a carbonic anhydrase. The group contains three TFs, one of which is involved in auxin response stimulation. Coexpression network analysis of the *Arabidopsis* homologs (ATTED-II) shows 11 genes from the cluster occur in a shared regulatory network. In summary, a set of genes related to cell wall, metabolic/redox flux, and signaling was recruited from the C_3 root to the C_4 leaf, many of which are coexpressed in *Arabidopsis* and found in leaf tissues analogous to BSC and MC.

Changes in the Leaf Transcriptomes Reveal Differences in Cellular Architecture and Leaf Development in the C_4 Species

Altered expression of cell cycle genes and enlarged BSC nuclei in *G. gynandra* suggest the occurrence of endoreduplication within this cell type. During early leaf development, *G. gynandra* leaf samples clustered together with younger samples in *T. hassleriana* (Supplemental Figures 8A and 8B), indicating a delay in leaf maturation. We hypothesized this delay in *G. gynandra* leaf maturation is manifested through alterations of cell cycle gene expression during leaf development. Hierarchical clustering of absolute expression values showed that the majority of known core cell cycle genes (Vandepoele et al., 2002; Beemster et al., 2005) have comparable expression patterns between both species (Supplemental Figure 16 and Supplemental Data Set 7). However, two distinct groups of genes were identified, which are either upregulated in *G. gynandra* between stage 0 to 2 (group 1: 9 of 18 genes with P value < 0.05) or show a delayed decrease during C_4 leaf development (group 2: 9 of 12 genes with P value < 0.05 between stage 0 and 3; Supplemental Figure 16 and Supplemental Data Set 7). Interestingly, *GT-2-LIKE1* (*GTL1*), a key cell cycle regulator, was not correlated between *G. gynandra* and *T. hassleriana* during leaf development. *GTL1* is upregulated in later stages of leaf development in *T. hassleriana* but not in *G. gynandra* (P value < 0.001 in stage 5; Supplemental Figure 16 and Supplemental Data Set 7).

As GTL1 has been demonstrated to operate as an inhibitor of endoreduplication and ploidy-dependent cell growth (Breuer et al., 2009, 2012), we examined whether nuclei were enlarged in any *G. gynandra* leaf tissues. First, both leaf developmental gradients were subjected to flow cytometry. Polyploidy (DNA content > 2C) was observed in both species, but clearly enriched in C_4 compared with C_3, especially in the more mature leaves (5% versus 1% ≥ 8C, 16% versus 4% ≥ 4C; Figure 6A).

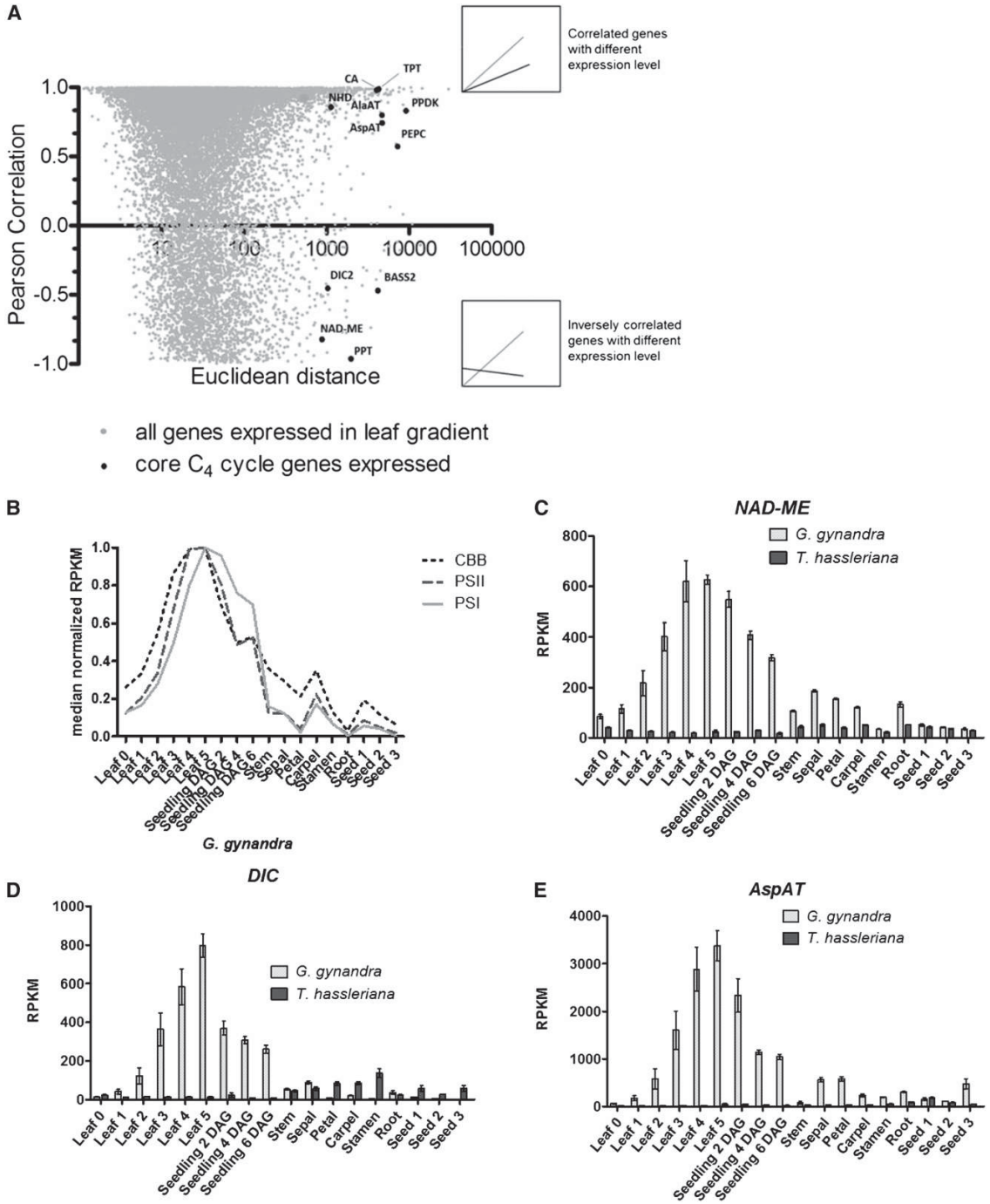


Figure 4. Comparison of Gene Expression Dynamics within the Leaf Gradient of Both Species.

(A) Euclidean distance versus Pearson's correlation of average RPKM (n = 3) of genes expressed (>20 RPKM) in both leaf developmental gradients. Comparison of gene expression by similarity of expression pattern and expression level in *T. hassleriana* and *G. gynandra*. Relevant highly expressed C_4

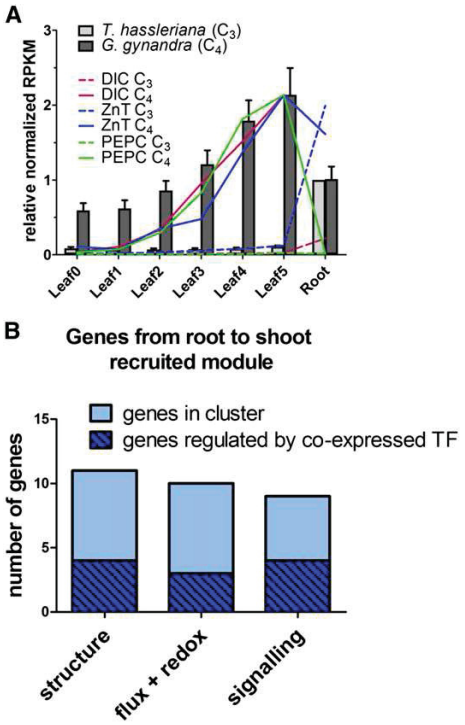


Figure 5. Recruitment of Genes from the Root to Leaf Expression Domain in the C_4 Plant *G. gynandra*.

(A) Relative average RPKM normalized to expression in *G. gynandra* leaf 5 (gray bars). Bars represent the arithmetic means of all 37 genes; lines show expression patterns of a reference C_4 cycle gene (*PEPC*) and of two genes found in the shifted module.
(B) Genes in the module displayed as functional groups. Light blue: absolute number of genes in the group. Dark blue overlay: portion of genes controlled by a transcription factor of the module.

In the *G. gynandra* C_4 leaf, the BSC nuclei were 2.9-fold larger than those in the MC ($P < 0.001$; Figures 6B and 6C). In contrast, the C_3 *T. hassleriana* nuclei of both cell types were similar sizes with a size ratio of 1.0 (Figures 6B and 6C). The proportion of BSC in the leaf was estimated from transversal sections as 15% in *G. gynandra* and 6% *in T. hassleriana* (Figures 7A to 7L). This number fits with the subpopulation of cells with higher ploidy observed in *G. gynandra* in the mature leaf. In summary, the extended expression of a subgroup of cell cycle genes and downregulation of *GTL1* correlate with higher ploidy levels in the *G. gynandra* mature leaf based on BSC nuclei area and flow cytometry measurements.

The C_4 Species Shows Delayed Differentiation of Mesophyll Tissue, Coinciding with Increased Vein Formation

The transcriptional delay in a large subset of *G. gynandra* genes (Figures 2B, 2C, and 3) reflects a later differentiation of the C_4 leaf. The delayed pattern of this large subset of genes indicated that there might be a delay in the differentiation of leaf internal anatomy, although leaf growth rates and shape are similar between species (Figure 1A). Thus, the leaves were examined microscopically. Since dicotyledonous leaves differentiate in a wave from tip toward petiole (Andriankaja et al., 2012), leaves were cross-sectioned at the midpoint (50% leaf length) for comparison. The cross sections revealed that in C_4 leaves, cell differentiation was delayed in the transition from undifferentiated ground tissue toward fully established palisade parenchyma (Figures 7A to 7L). Both species start undifferentiated at leaf stage 0 with only the primary vein distinctly visible in cleared leaves (Figures 7A and 7G; Supplemental Figure 1A). In stage 1, the C_3 leaf starts to differentiate its palisade parenchyma, while the C_4 leaf shows dividing undifferentiated cells (Figures 7B and 7H). Mesophyll differentiation has finished by stage 2 in the C_3 leaf (Figure 7I), but not until stage 4 in the C_4 leaf (Figure 7D). Classical mature C_4 leaf architecture appears in stage 4 in *G. gynandra* (Figure 7E). C_4 leaves ultimately develop more veins and open veinlets leading to Kranz anatomy (Supplemental Figure 1). Leaf mesophyll tissue of the C_3 species differentiates faster and develops fewer veins than the C_4 species.

The expression of genes related to vein development was consistent with greater venation in the C_4 leaf but failed to explain the larger delay in expression patterns and mesophyll differentiation in the C_4 leaf. Hierarchical clustering indicated that most known leaf and vasculature developmental factors (reviewed in Ohashi-Ito and Fukuda, 2010) showed similar expression patterns in the two species (Supplemental Figure 17 and Supplemental Table 6). However, two clusters with distinct expression patterns were detected. In the C_4 species, seven genes were upregulated (P value < 0.05), including vasculature facilitators *PIN-FORMED* (*PIN1*), *HOMEOBOX GENE8* (*HB8*), and *XYLOGEN PROTEIN1* (*XYP1*) (Motose et al., 2004; Scarpella et al., 2006; Donner et al., 2009), while five genes were downregulated (P value < 0.05), among those the negative regulators *KANADI1* and *2*, as well as *HOMEOBOX GENE15* (Supplemental Figure 17 and Supplemental Table 6; Ilegems et al., 2010).

To further elucidate the magnitude and nature of the delayed expression changes on the transcriptional level, the leaf gradient data were clustered with the *K*-means algorithm (Supplemental

Figure 4. (continued).

cycle genes are marked in plot. Above inset shows an example of two highly correlated genes by expression trend and strength. Lower inset shows an example of two genes inversely correlated with different expression level.
(B) Expression pattern across the atlas of averaged relative expression of transcripts encoding for photosystem I (PSI), photosystem II (PSII), and soluble enzymes of the Calvin-Benson-Bassham (CBB) cycle in *G. gynandra*.
(C) to **(E)** Average expression pattern of highest abundant ortholog of C_4 cycle genes (*NAD-ME*, *DIC*, and *AspAT*) in photo- and heterotrophic tissues in *G. gynandra* (light gray) and *T. hassleriana* (dark gray); ±SE, $n = 3$.

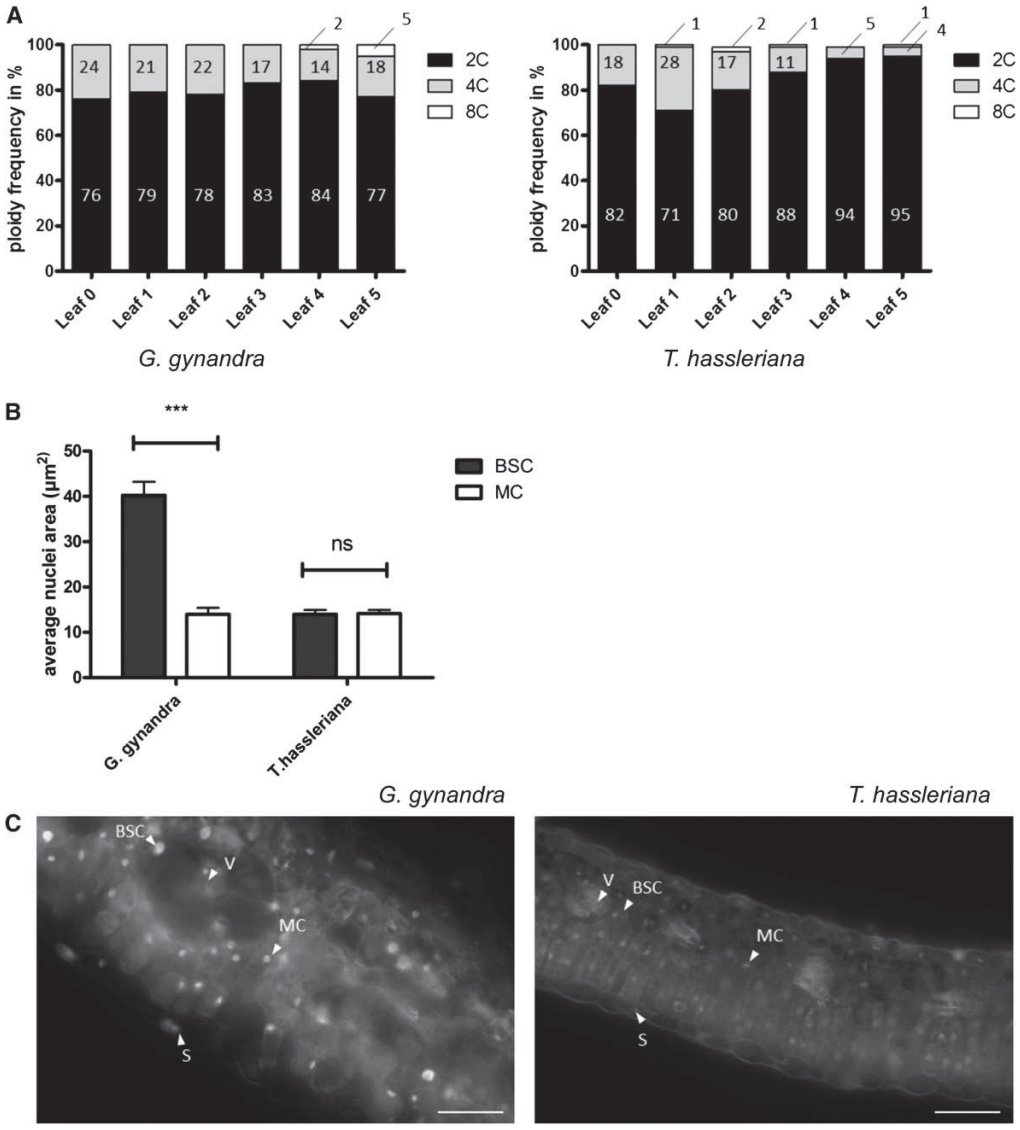


Figure 6. Distribution of Ploidy Levels during Leaf Development and Nuclei Area of BSC and MC between *G. gynandra* and *T. hassleriana*.

(A) Ploidy distribution of developing leaf (category 0 till 5) in percentage in *G. gynandra* and *T. hassleriana*. Measurements performed in $n = 3$ (except G0 = 1 replicate). For each replicate, at least 2000 nuclei were measured by flow cytometry.

(B) Quantification of BSC and MC nuclei area in cross sections ($n = 3 \pm$ SE) of mature *G. gynandra* and *T. hassleriana* leaves (stage 5). Area of nuclei in μm^2 with at least 150 nuclei analyzed per cell type per species per replicate. Asterisks indicate statistically significant differences between BSC and MC (***P value < 0.001); n.s., not significant.

(C) Fluorescence microscopy images of propidium iodide-stained leaf cross sections (stage 5) of *T. hassleriana* (left) and *G. gynandra* (right). Arrowheads point to nuclei of the indicated cell type. V, vein; S, stomata. Bar = 50 μm.

Figures 17A and 17B and Supplemental Data Set 9). Of 16 clusters, six were divergent (1 to 3, 8, 9, and 15; 1270 genes). The remaining clusters were similar; however, four showed a transcriptional delay (4, 5, 13, and 16; 3361 genes), while six did not (6, 7, 10 to 12, and 14; 5162 genes). Of all clustered genes, 87% belonged to highly conserved clusters, 34% with a delay and 53% without. Thus, the transcriptional delay cannot be explained by general slower development.

All of the *K*-means clusters were functionally characterized by testing for enrichment in MapMan categories (Supplemental

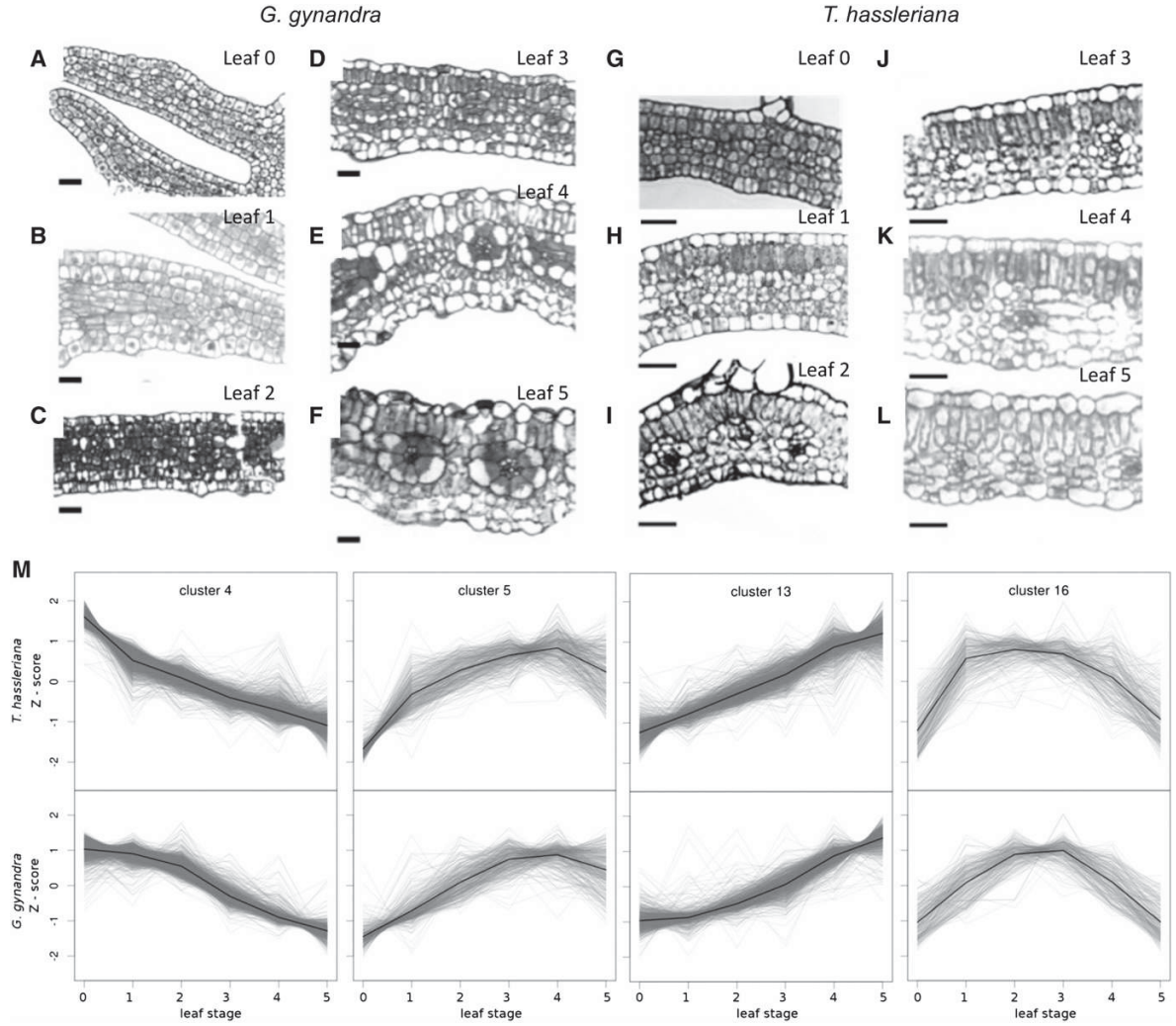


Figure 7. Analysis of Shifted Gene Expression Pattern and Leaf Anatomy during Leaf Ontogeny.

(A) to **(L)** Leaf anatomy development along the gradient in *G. gynandra* and *T. hassleriana* depicted by cross sections stained with toluidine blue. Bar = 20 μm.

(M) Selected clusters from *K*-means clustering of gene expression shown as *Z*-scores, which show a phase shift between *G. gynandra* and *T. hassleriana* during leaf development.

Data Set 10). The visually "shifted" patterns were: later onset of increase in clusters 13 and 5 (1058 and 395 genes, respectively), delayed decrease in cluster 4 (1644 genes), and a later peak in cluster 16 (264 genes; Figure 7M). The "late decrease" cluster 4 is enriched in genes related to mitochondrial electron transfer, *CONSTITUTIVE PHOTOMORPHOGENESIS9* (*COP9*) signalosome, and protein degradation by the proteasome (Figure 7M; Supplemental Data Set 10). The "late onset" cluster 13 is enriched in all major photosynthetic categories: N-metabolism, and chlorophyll, isoprenoid, and tetrapyrrole biosynthesis (P value < 0.05; Supplemental Figures 17C and 17D and Supplemental Data Sets 9 and 10). The smaller "late onset" cluster 5 is enriched in the categories protein synthesis, tetrapyrrole synthesis, carotenoids, and peroxiredoxin. Cluster 16 peaks earlier in *T. hassleriana* than *G. gynandra* and is enriched in lipid metabolism (e.g., *ACYL CARRIER PROTEIN4*, *CHLOROPLASTIC ACETYLCOA CARBOXYLASE1*, *3-KETOACYL-ACYL CARRIER PROTEIN SYNTHASE1*, and *3-KETOACYL-ACYL CARRIER PROTEIN SYNTHASE III*) and plastid division genes, such as the *FILAMENTATION TEMPERATURE-SENSITIVE* genes *FtsZ2*, *FtsH*, and *FtsZ*, as well as *ACCUMULATION AND REPLICATION OF CHLOROPLASTS11* (Figure 7M; Supplemental Data Sets 9 and 10).

The core of the phase-shifted clusters, defined as genes with Pearson's correlation coefficient of $r > 0.99$ to the cluster center, contained candidate regulators for the observed delayed patterns. The core of cluster 13 contained 17 TFs and genes involved in chloroplast maintenance (Supplemental Data Set 11). The core of cluster 4 contained 30 transcriptional regulators, including *PROPORZ1* (*PRZ1*), and eight other chromatin-remodeling genes. Nineteen cell cycle genes were found in the core of cluster 4 (Supplemental Figures 19A and 19B), including *CELL DIVISION CYCLE20* (*CDC20*), *CDC27*, and *CELL CYCLE SWITCH PROTEIN52* (*CCS52*), which are key components of cell cycle progression from M-phase to S-phase (Pérez-Pérez et al., 2008; Mathieu-Rivet et al., 2010b).

Our data were quantitatively compared with data from *Arabidopsis* leaf development to test if the observed phase shift related to a switch from proliferation to differentiation (Andriankaja et al., 2012). This study identified genes that were significantly up- or downregulated during the shift from proliferation to expansion (Andriankaja et al., 2012). Putative orthologs of these genes were clustered by the *K*-means algorithm (without prior expression filtering), producing seven clusters for the upregulated genes (containing 483 genes in total) and five clusters for the downregulated genes (1112 genes in total; Supplemental Figure 20). The trend was well conserved across species, with 75% of the upregulated and 96% of the downregulated genes falling into clusters with a matching trend. The genes showed a higher proportion of delay in *G. gynandra* than in the total data set, with 60 and 68% falling in delayed up- and downregulated clusters, respectively (Supplemental Figure 20).

In summary, about a third of all gene expression patterns show a delay in the *G. gynandra* leaf (Figure 7M; Supplemental Figure 18). Delayed genes include major markers of leaf maturity such as the upregulation of photosynthetic gene expression and downregulation of mitochondrial electron transport (Supplemental Figures 19C and 19D and Supplemental Data Set 10). This delay was more common in putative orthologs of genes differentially regulated during the shift from cell proliferation to expansion (Supplemental Figure 19; Andriankaja et al., 2012). The slow maturation can be seen on the anatomical level as a delayed differentiation that coincides with increased vein formation in the C_4 species (Figures 7A to 7L).

DISCUSSION

Comparative Transcriptome Atlases Provide a Powerful Tool for Understanding C_4 Photosynthesis

Two transcriptome atlases were generated to allow the analysis of gene recruitment to photosynthesis and to detect differences related to C_4 leaf anatomy. Two Cleomaceae species were chosen for this study due to their phylogenetic proximity to the model species *Arabidopsis* (Marshall et al., 2007). The sampled leaf tissues covered development from sink tissue to fully mature source tissue (Figures 1 and 3), and all higher order vein development (Supplemental Figure 1). Since C_4 genes are recruited from genes already present in C_3 ancestors, where they carry out housekeeping functions (Sage, 2004; Besnard et al., 2009; Christin and Besnard, 2009; Christin et al., 2009), seed, stem, floral, and root tissues were included in the atlases in addition to leaves and seedlings.

The high similarity in expression pattern between the species maximizes our ability to detect differences related to C_4 photosynthesis. While PCA analysis showed that the first principle component separated the data set by species, this accounted for only 15% of the variation (Supplemental Figure 8A). Excluding floral organs and stem, all tissues correlated with $r > 0.7$ between species (Supplemental Figure 7C and Supplemental Table 3). Hierarchical and *K*-means clustering showed the vast majority of genes had a similar pattern between species, and tissue types clustered closely with the same tissue in the other species. Specific groups of highly expressed genes exclusively expressed in one tissue type, such as root, stamen, and petal, are shared between *G. gynandra* and *T. hassleriana*, suggesting that these genes might represent drivers for the respective tissue identity (Supplemental Figure 9). A subset of genes showed a consistent adjustment to their expression pattern, namely, a delay in the leaf gradient of *G. gynandra* relative to *T. hassleriana* (Figure 7M). Thus, organ identity is highly conserved between *G. gynandra* and *T. hassleriana*, but the rate at which organ identity, especially the leaf, is established can differ.

Expression Patterns of C_3 Putative Orthologs Support Small-Scale or Modular Recruitment to Photosynthesis, Implying That a General C_4 Master Regulator Is Unlikely

Ancestral expression patterns can be compared with assess whether a master regulator could have facilitated recruitment of genes to C_4 photosynthesis. The patterns of gene expression in *T. hassleriana* provide a good proxy for the ancestral C_3 expression pattern due to its phylogenetic proximity to *G. gynandra* (Inda et al., 2008; Cheng et al., 2013). Genes active in the C_4 cycle were recruited from previously existing metabolism (Matsuoka, 1995; Chollet et al., 1996; Streatfield et al., 1999; Wheeler et al., 2005; Tronconi et al., 2010). Expression patterns in *T. hassleriana* reflect known metabolism and expression; for instance, *PPDK* is expressed in seeds, stamens, and petals (Supplemental Figure 12B), which is similar to the expression domain reported by Chastain et al. (2011). Furthermore, *PPT* is highly expressed in stamens and during seed development (Supplemental Figure 12C; Knappe et al., 2003a, 2003b), since it is required for fatty acid production (Hay and Schwender, 2011).

The C_3 putative orthologs of C_4 cycle genes show a variety of expression patterns within the atlas, providing strong evidence they could not have been recruited by a single master regulator. All C_4 cycle genes are expressed to a low degree in *T. hassleriana*, either constitutively or in defined tissues such as stamens, seeds, or young leaves (Figures 4C to 4E). Expression of *NHD*, *AlaAT*, *AspAT*, and *PPDK* increased along the leaf gradient in both C_3 and C_4 species, but in C_3, the expression was highest in tissues other than the leaf (Figure 4E; Supplemental Figures 12A, 12B, and 12D). In contrast, *DIC*, *BASS2*, *NAD-ME*, and *PPT* are expressed in inverse patterns between C_3 and C_4 along the leaf gradient (Figures 4C and 4D; Supplemental Figures 12C and 12E), and *PEPC* is expressed only in mature leaves in the C_3 species (Supplemental Figure 12F). Except for *DIC* and *PPDK*, the expression level of the C_4 cycle genes was higher in *G. gynandra* across all tissues (Figure 4; Supplemental

Figures 12 to 14). Thus, most of the C_4 cycle genes may still maintain their ancestral functions in addition to the acquired C_4 function. The correct ortholog in C_3 may not have been conclusively determined by cross species read mapping in all cases reported here. However, the main conclusion—that C_4 cycle genes are recruited from a variety of C_3 expression patterns—holds regardless of which putative C_3 paralog is selected (Supplemental Figures 13 and 14).

A set of genes shifted from a root to leaf expression domain during C_4 evolution provides an example of small-scale modular recruitment. The proposed analogy between root endodermis and bundle sheath and between root cortex and mesophyll (Slewinski, 2013) has been linked to cooption of the *SCARECROW* (*SCR*) and *SHORTROOT* (*SHR*) regulatory networks into developing leaves (Slewinski et al., 2012; Wang et al., 2013). A set of 37 genes consistent with such a recruitment module was identified. For this gene set, the C_3 species *T. hassleriana* (Figure 5; Supplemental Table 5) and *Arabidopsis* (Brady and Provart, 2009) showed conserved root expression, while the C_4 species showed an expression pattern similar to photosynthesis. Much of the root to leaf gene set was coregulated in *Arabidopsis*, and it contained TFs, including *ETHYLENE RESPONSE FACTOR1* (Mantiri et al., 2008), as well as an AUX/IAA regulator (Pérez-Pérez et al., 2010) and *VND-INTERACTING2* (Yamaguchi et al., 2010). Functionally, the majority of the gene set is involved in processes related to cell wall synthesis and modification. The set contains the cell wall-plasma membrane linker protein (Stein et al., 2011) and the xyloglucan endotransglycosylase *TOUCH4* (Xu et al., 1995), the tonoplast intrinsic protein involved in cell elongation (Beebo et al., 2009), and a plasmodesmata-located protein (Bayer et al., 2008). The observed coregulation and structural functions support an underlying structural relationship between the root tissues endodermis and cortex, and the leaf tissues bundle sheath and mesophyll.

It is still unresolved whether expression level recruitment of genes to the C_4 cycle was facilitated by the action of one or a few master switches controlling C_4 cycle gene expression and/or by changes to promoter sequences of C_4 genes (Westhoff and Gowik, 2010). The diverse transcriptional patterns of the core C_4 cycle genes in *T. hassleriana* provide strong evidence that they were not recruited as a single transcriptional module facilitated by one or a few master regulators. However, the identified root to leaf module indicates that small-scale corecruitment occurs, and this may help bring about the 3 to 4% overall transcriptional changes occurring during C_4 evolution (Bräutigam et al., 2011, Gowik et al., 2011). The similarities in expression pattern between photosynthetic genes and C_4 cycle genes are evident (Figure 4B), and light-dependent induction of C_4 genes has been reported (Christin et al., 2013), leading us to hypothesize that C_4 cycle genes may use the same light-induced regulatory circuits employed for the photosynthetic genes, possibly through acquisition of *cis*-regulatory elements or modification of chromatin structure, as has been shown for the *PEPC* gene promoter in maize (Tolley et al., 2012).

Cell Size in *G. gynandra* Coincides with Nuclei Size and Ploidy

In addition to the biochemical C_4 cycle genes, transcriptional changes related to cell and tissue architecture are required for C_4 leaf development (Westhoff and Gowik, 2010). The comparative atlases were contextualized with anatomical data to better understand BSC size.

G. gynandra has generally larger cells (Figures 7A to 7L), which might be attributed to a larger genome. After divergence from *T. hassleriana*, the *G. gynandra* lineage has undergone a putative whole-genome duplication (Inda et al., 2008). Cell size has been tied to genome ploidy status previously (Sugimoto-Shirasu and Roberts, 2003; Lee et al., 2009b; Chevalier et al., 2011). A relationship between ploidy and cell size could explain the generally larger cells in *G. gynandra* leaves (Figures 7A to 7L) or relate to the upregulation of DNA and histone-associated genes in developing leaves (Figure 3; Supplemental Figures 10 and 11).

Changes in the expression of key cell cycle genes indicated endoreduplication may be increased in *G. gynandra*, and follow-up nuclear size measurements indeed indicate BSCs have undergone endoreduplication. Enlargement of BSC is a common feature of C_4 plants (Sage, 2004; Christin et al., 2013) including *G. gynandra* (Figures 7D to 7F), but the genetic mechanism is unknown. During leaf development, key cell cycle genes showed changes in expression pattern and expression level between *G. gynandra* and *T. hassleriana* (Supplemental Figure 16). *CDC20* and *CCS52A*, which are closely linked with cell cycle M-to-S-phase progression or endocycle onset (Lammens et al., 2008; Larson-Rabin et al., 2009; Kasili et al., 2010; Mathieu-Rivet et al., 2010a), exhibit prolonged expression during C_4 leaf development, whereas the expression of the master endoreduplication regulator *GTL1* (Breuer et al., 2009, 2012; Caro et al., 2012) is suppressed in the older leaf stages (Supplemental Figure 16). Although a comparison of the more distantly related species *Arabidopsis* and *G. gynandra* discounted endoreduplication as a factor in bundle sheath cell size (Aubry et al., 2013), the BSC and MC nuclei area measurements of mature *G. gynandra* and *T. hassleriana* leaves revealed that the BSC nuclei are 2.9-fold enlarged compared with MC nuclei in *G. gynandra* (Figures 6B and 6C). At the same time, *T. hassleriana* BSC and MC cells do not differ significantly in nuclei size (Figures 6A and 6C).These results are supported by a flow cytometry analysis of both leaf developmental gradients, where the proportion of endoreplicated cells in the mature C_4 leaf (Figures 6A) matches the number of BSCs present in *G. gynandra* (Figures 6A and 7A to F). Interestingly, we also find significant ($P > 0.001$) enlarged BSC nuclei in other C_4 species (e.g., *Flaveria trinervia*, *Megathyrsus maximum*, and maize; Supplemental Figure 22), indicating that larger nuclei size in BSC compared with the MC could be a general phenomenon in C_4 plants conserved across mono- and dicotyledons. Whether endoreplication is the cause of increased cell size in C_4 BSC, as found for trichomes and tomato (*Solanum lycopersicum*) karyoplasm (Traas et al., 1998; Chevalier et al., 2011) or whether endoreplication only occurs to support the high metabolic activity and large size of the BSCs (Sugimoto-Shirasu and Roberts, 2003) remains to be determined.

Late Differentiation of Mesophyll Tissue Allows Denser Venation

General regulators of leaf anatomy and shape (reviewed in Byrne, 2012) are expressed in very similar patterns between the two species (Supplemental Figure 17), reflecting the very similar

palmate five-fingered leaf shape and speed of leaf expansion (Figures 1A and 1B). However, anatomical studies of leaf development show that differentiated palisade parenchyma is already observed at the midpoint of stage 1 leaves in *T. hassleriana* (Figure 7H) but can only be detected in the middle of the leaf in stages 3 and 4 in *G. gynandra* (Figures 7D to 7F). Hierarchical clustering of transcriptome data indicates a similarity between younger *T. hassleriana* and older *G. gynandra* tissues (Supplemental Figure 9), which we attribute to a delay in *G. gynandra* leaf expression changes observed in the hierarchical clusters (Figures 2B and 2C) and observed for *K*-means clustering involving about a third of clustered genes (Figure 7M; Supplemental Figure 18). Analysis of the delayed clusters for significant enrichment of functional categories indicated that the metabolic shift from sink to source tissue was delayed (Figures 3 and 7M; Supplemental Figure 18 and Supplemental Data Set 10). Furthermore, the "delayed decrease" cluster 4 was enriched in *COP9* signalosome and marker genes of the still developing heterotrophic leaf.

Cell cycle and cell differentiation regulators show a delayed expression pattern in *G. gynandra*. The expression of *PRZ1*, which switches development from cell proliferation to differentiation in *Arabidopsis* (Sieberer et al., 2003; Anzola et al., 2010), is prolonged in the C_4 leaf (Figure 7M, cluster 4), as is the expression of chromatin remodeling factor *GRF1-INTERACTING FACTOR3* implicated in the control of cell proliferation upstream of cell cycle regulation (Lee et al., 2009a). Plastid division genes peak around leaf stage 1 in *T. hassleriana* and leaf stage 2 in *G. gynandra* (Figure 7M, cluster 16). It has recently been shown that chloroplast development and division precedes photosynthetic maturity in *Arabidopsis* leaves and retrograde signaling from the chloroplasts affects cell cycle exit from proliferation (Andriankaja et al., 2012). Quantitative comparison of differentially regulated genes during the shift from cell proliferation to cell expansion found in *Arabidopsis* (Supplemental Figure 20; Andriankaja et al., 2012) to the expression patterns of the putatively orthologous genes along leaf developmental gradients in Cleome, reveals a strong conservation of expression pattern between *Arabidopsis* and Cleome during development. A higher proportion of delay of *G. gynandra* genes is observed in this gene set. This supports the idea that the transcriptional delay is directly linked to the anatomical delay in differentiation observed in *G. gynandra* (Supplemental Figure 19).

The delay in cell differentiation allows for increased vein formation in the C_4 leaf. Mesophyll differentiation has already been shown to limit minor vein formation in *Arabidopsis* (Scarpella et al., 2004; Kang et al., 2007). *G. gynandra* and *T. hassleriana* have altered vein densities, which result from more minor vein orders in *G. gynandra* (Supplemental Figure 1), similar to results for the dicot *Flaveria* species (McKown and Dengler, 2009). Given that differentiation of photosynthetic mesophyll cells limits minor vein formation in *Arabidopsis* (Scarpella et al., 2004; Kang et al., 2007) and that mesophyll differentiation is delayed in the C_4 species compared with the C_3 species (Figure 7), dense venation may indeed be achieved by delaying mesophyll differentiation.

Genes related to vascular patterning are expressed in a manner consistent with higher venation in the C_4 leaf. The high expression of vascular pattern genes such as *PIN1*, *HB8*, *ARF3*, and *XYP1* in the C_4 leaf (Supplemental Figure 17) is similar to that observed for Kranz patterned leaves in maize (Wang et al., 2013). However, these genes may be a consequence, rather than a cause, of higher venation, especially since some of these markers are only expressed after pre-procambial or procambial identity is introduced (Ohashi-Ito and Fukuda, 2010). Once procambial fate is established, cellular differentiation of vein tissues proceeds through positional cues and localized signaling, possibly via the SCR/SHR pathway (Langdale and Nelson, 1991; Nelson and Langdale, 1992; Nelson and Dengler, 1997; Griffiths et al., 2013; Wang et al., 2013; Lundquist et al., 2014). Interestingly, in accordance with the delay in leaf differentiation in *G. gynandra*, we could monitor a delay in higher expression for *SHR* peaking around leaf stage 1 to 3 (Supplemental Figure 21A). *SCR* transcript abundance is clearly divided in both *G. gynandra* and *T. hassleriana* between two homologs, one of which is more abundant in the C_4 leaf and the other in the C_3 leaf (Supplemental Figure 21B). *SCR* expression in *G. gynandra* follows the *SHR* pattern with a delayed upregulation. This is in accordance with earlier studies conducted in maize, where *SHR* transcript highly accumulates in the BSC to activate *SCR* expression (reviewed in Slewinski et al., 2012)

The identification of mesophyll differentiation as the proximate cause for fewer minor vein orders in *T. hassleriana* raises the question of how mesophyll differentiation is controlled. In both C_4 and C_3 species, vascular patterning precedes photosynthetic tissue differentiation (Sud and Dengler, 2000; Scarpella et al., 2004; McKown and Dengler, 2010). Light is one of the most important environmental cues that regulate leaf development, including its cellular differentiation and onset of photosynthesis (Tobin and Silverthorne, 1985; Nelson and Langdale, 1992; Fankhauser and Chory, 1997). The *COP9* signalosome, which plays a central role in repression of photomorphogenesis and G2/M cell cycle progression (Chamovitz et al., 1996; Dohmann et al., 2008), showed a delayed decrease in *G. gynandra* compared with *T. hassleriana* (Supplemental Figure 19B). The delay and earlier vein formation termination induced by excess light in *Arabidopsis* (Scarpella et al., 2004) suggest that light perception and its signal transduction may be differentially regulated in species with denser venation patterns.

Conclusions

In this study, we report a detailed comparison of the transcriptomes and the leaf development of two Cleomaceae species with different modes of photosynthetic carbon assimilation, i.e., C_3 and C_4 photosynthesis. The gene expression patterns are quite similar between both species, which facilitates the identification of differences related to C_4 photosynthesis. We could link two key features of Kranz anatomy to developmental processes through integration of expression and anatomical data. First, we show that the larger size of the bundle sheath cells in the C_4 species is associated with a higher ploidy in these cells, which might be controlled by delayed repression of the endocycle via the transcription factor *GTL1*. Second, a prominent difference between C_3 and C_4 leaf development is the delayed differentiation of the leaf cells in C_4, which is associated with a delayed onset of photosynthetic gene expression, chloroplast proliferation and development, and altered expression of a few

distinct cell cycle genes. Delayed mesophyll differentiation allows for increased initiation of vascular tissue and thus contributes to the higher vein density in C_4. We hypothesize that delayed onset of mesophyll and chloroplast differentiation is a consequence of the prolonged expression of the *COP9* signalosome and, hence, a delayed derepression of photomorphogenesis.

METHODS

Plant Material and Growth Conditions

Gynandropsis gynandra and *Tarenaya hassleriana* plants for transcriptome profiling by Illumina Sequencing were grown in standard potting mix in a greenhouse between April and August 2011. Internal transcribed spacer sequences of *G. gynandra* and *T. hassleriana* were analyzed and plant identity confirmed according to Inda et al. (2008). Leaves were harvested from 4- to 6-week-old plants, prior to inflorescence initiation. All samples were harvested during midday. Flowers, stamens, sepals, and carpels were harvested after induction of flowering. Green tissues from seedlings were harvested 2, 4, and 6 d after germination. Root material was harvested from plants grown in vermiculite for 6 weeks and supplemented with Hoagland solution. Leaf material for the ontogeny analysis was selected by the order of leaf emergence from the apex in leaf stages from 0 to 5. Up to 40 plants were pooled for each biological replicate.

Leaf Expansion Rate

Leaves from stage 0 to 5 were analyzed in five biological replicates for each *G. gynandra* and *T. hassleriana*. Leaves were scanned on a flat bed scanner (V700 Photo; Epson), and the area was analyzed with free image analysis software ImageJ.

Leaf Cross Sections for Anatomical Studies

Leaves from stage 0 to 5 were analyzed in biological triplicates. Leaf material (2 × 2 mm) was cut next to the major first order vein at 50% of the whole leaf length. Leaf material was fixed in 4% paraformaldehyde solution overnight at 4°C, transferred to 0.1% glutaraldehyde in phosphate buffer, and vacuum infiltrated three times for 5 min. The leaf material was then dehydrated with an ascending ethanol series (70, 80, 90, and 96%) with a 1-h incubation in each solution. Samples were incubated twice in 100% ethanol and twice in 100% acetone, each for 20 min, and infiltrated with an acetone:araldite (1:1) mixture overnight at 4°C. After acetone evaporation, fresh araldite was added to the leaf samples until samples were covered and incubated for 3 to 4 h. Samples were transferred to fresh araldite in molds and polymerized at 65°C for 48 h. Cross sections were stained with toluidine blue for 15 s and washed with H_2O_{dest}. Cross sections were imaged with bright-field settings using an Eclipse Ti-U microscope (Nikon).

Flow Cytometry

Three biological replicate samples were chopped with a razor blade in 200 μL of Cystain UV Precise P Nuclei extraction buffer followed by the addition of 800 μL of staining buffer (buffers from Partec). The chopped leaves in buffer were filtered through a 50-μm mesh. The distribution of the nuclear DNA content was analyzed using a CytoFlow ML flow cytometer and FLOMAX software (Partec) as described (Zhiponova et al., 2013).

Measurement of Nuclei from Mature Leaves

Fresh mature leaves (leaf stage 5, three biological replicates) of *G. gynandra* and *T. hassleriana* were cut transversally, fixed in 1× PBS buffer (1% Tween 20 and 3% glutaraldehyde) overnight at room temperature, and stained with propidium iodide solution directly on the microscopic slide. Cross sections were imaged by fluorescence microscopy using an Axio Imager M2M fluorescence microscope (Zeiss) with an HE DS-Red Filter. Images were processed with ZEN10 software (Zeiss), and the nuclear area of at least 200 nuclei per cell type per species was measured with ImageJ.

RNA Extraction, Library Construction, and Sequencing

Plant material was extracted using the Plant RNeasy extraction kit (Qiagen). RNA was treated on-column (Qiagen) and in solution with RNA-free DNase (New England Biolabs). RNA integrity, sequencing library quality, and fragment size were checked on a 2100 Bioanalyzer (Agilent). Libraries were prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina), and library quantification was performed with a Qubit 2.0 (Invitrogen). Single-end sequenced samples were multiplexed with six libraries per lane with ~20 million reads per library. For paired-end sequencing, RNA of all photosynthetic and nonphotosynthetic samples was pooled equally for each species and prepared as one library per species. Paired end libraries were run on one lane with ~175 million clean reads for *T. hassleriana* and 220 million clean reads for *G. gynandra*. All libraries were sequenced on the HISEQ2000 Illumina platform. Libraries were sequenced in the single-end or paired-end mode with length ranging from 80 to 100 nucleotides. The paired-end library of *G. gynandra* had an average fragment size of 304 bp; *T. hassleriana* had an average fragment size of 301 bp.

Gene Expression Profiling

Reads were checked for quality with FASTQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/), subsequently cleaned and filtered for quality scores greater than 20 and read length greater than 50 nucleotides using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit). Expression abundances were determined by mapping the single-end read libraries (each replicate for each tissue) independently against *T. hassleriana* representative coding sequences (Cheng et al., 2013) using BLAT V35 (Kent, 2002) in protein space and counting the best mapping hit based on e-value for each read uniquely. Default BLAT parameters were used for mapping both species. Expression was normalized to reads per kilobase *T. hassleriana* coding sequence per million mappable reads (RPKM). *T. hassleriana* coding sequences were annotated using BLASTX searches (cutoff $1e^{-10}$) against the TAIR10 proteome database. The best BLAST hit per read was filtered by the highest bit score. A threshold of 20 RPKM per coding sequence in at least one species present in at least one tissue was chosen to discriminate background transcription (Supplemental Figure 14). Differential expression between *T. hassleriana* and *G. gynandra* was determined by EdgeR (Robinson et al., 2010) in R (R Development Core Team, 2009). A significance threshold of 0.05 was applied after the P value was adjusted with false discovery rate via Bonferroni-Holms correction (Holm, 1979).

Data Analysis

Data analysis was performed with the R statistical package (R Development Core Team, 2009) unless stated otherwise. For Pearson's correlation and PCA analysis, *Z*-scores were calculated by gene across both species. For all other analyses, *Z*-scores were calculated by gene within each species, to focus on comparing expression patterns. For *K*-means and hierarchical clustering, genes were filtered to those with more than 20 RPKM in at least one of the samples used in each species. To determine the number of centers for *K*-means clustering, the sum of SE within clusters was plotted against cluster number and compared with randomized data (Supplemental Figures 18B, 20C, and 20D). A total of 16 centers was chosen, and *K*-means clustering was performed 10,000 times and the best solution, as defined by the minimum sum of SE of genes in the cluster, was taken for downstream analyses (Peeples, 2011). Multiscale bootstrap resampling of the hierarchical clustering was

performed for samples with 10,000 repetitions using the pvclust R package (Suzuki and Shimodaira, 2006).

Stage enrichment was tested for all *K*-means clusters and for tissue "signature genes" with expression of over 1000 RPKM in each tissue using TAIR10 MapMan categories (from http://mapman.gabipd.org) for the best *Arabidopsis thaliana* homolog. Categories with more than five members in the filtered (*K*-means) or complete (signature genes) data set were tested for enrichment by Fisher's exact test, and P values were adjusted to false discovery rates via Benjamini-Yekutieli correction, which is tolerant of dependencies (Yekutieli and Benjamini, 1999).

Accession Numbers

Sequence data from this article can be found in NCBI GenBank under the following accession numbers: SRP036637 for *G. gynandra* and SRP036837 for *T. hassleriana*.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Venation Patterning during Leaf Development of *G. gynandra* and *T. hassleriana*.

Supplemental Figure 2. *G. gynandra* Cotyledon Anatomy 2, 4, and 6 d after germination (DAG).

Supplemental Figure 3. Images of Tissues Harvested for Atlases in *G. gynandra* and *T. hassleriana*.

Supplemental Figure 4. Quality Assessment of Velvet/OASES Assembled *T. hassleriana* Contigs against Predicted Corresponding CDS from *T. hassleriana* Genome

Supplemental Figure 5. Quality Assessment of the Biological Replicates of *T. hassleriana* Libraries Mapped to *A. thaliana* and Mapping Similarity of *T. hassleriana* Libraries Mapped to *A. thaliana* and to Its Own CDS.

Supplemental Figure 6. Determination of Baseline Gene Expression via a Histogram of Photosystem (PS) I and II Transcript Abundances (RPKM) in the *G. gynandra* Root.

Supplemental Figure 7. Quality Assessment of the Biological Replicates within Each Species and Tissue Similarity between *G. gynandra* and *T. hassleriana*.

Supplemental Figure 8. Principle Component Analysis between *G. gynandra* and *T. hassleriana.*

Supplemental Figure 9. Hierarchical Cluster Analysis with Bootstrapped Samples of *G. gynandra* and *T. hassleriana.*

Supplemental Figure 10. Transcriptional Investment of Each Tissue Compared in Both Species.

Supplemental Figure 11. Transcriptional Investment at Secondary MapMan Category Level of Each Tissue Compared in Both Species.

Supplemental Figure 12. Comparison of Gene Expression Dynamics within the Leaf Gradient of Both Species.

Supplemental Figure 13. Plot of the Expression Pattern (RPKM) of all C_4 Gene Orthologs Expression Pattern in *G. gynandra.*

Supplemental Figure 14. Plot of the Expression Pattern of all C_4 Gene Putative Orthologs Expression Pattern (RPKM) in *T. hassleriana.*

Supplemental Figure 15. Enzyme Activity Measurement of Soluble C_4 Cycle Enzymes.

Supplemental Figure 16. Hierarchical Clustering of Average RPKM with Euclidean Distance of Core Cell Cycle Genes.

Supplemental Figure 17. Hierarchical Clustering with Pearson's Correlation of Leaf Developmental Factors.

Supplemental Figure 18. *K*-Means Clustering of Leaf Gradient Expression Data and Quality Assessment.

Supplemental Figure 19. Z-Score Plots of Enriched MapMan Categories in the Shifted Clusters.

Supplemental Figure 20. *K*-Means Clustering of Genes Differentially Regulated during the Transition from Proliferation to Enlargement.

Supplemental Figure 21. Transcript Abundances of *SCARECROW* and *SHORTROOT* Homologs in *G. gynandra* and *T. hassleriana* Leaf and Root.

Supplemental Figure 22. Nuclei Area and Images of C_4 and C_3 Species.

Supplemental Table 1. Velvet/OASES Assembly Stats from *G. gynandra* and *T. hassleriana* Paired-End Reads.

Supplemental Table 2. Cross-Species Mapping Results.

Supplemental Table 3. Pearson's Correlation between G. *gynandra* and *T. hassleriana* Individual Tissues.

Supplemental Table 4. Number of Significantly Up- or Downregulated Genes in *G. gynandra* Compared with *T. hassleriana* within the Different Tissues.

Supplemental Table 5. List of Genes Present in Root-to-Shoot Recruitment Module.

Supplemental Table 6. List of Clustered General Leaf Developmental and Vasculature Regulating Genes along Both Leaf Gradients.

Supplemental Methods.

The following materials have been deposited in the DRYAD repository under accession number http://dx.doi.org/10.5061/dryad.8v0v6.

Supplemental Data Set 1. Annotated Transcriptome Expression Data of Both Atlases in RPKM.

Supplemental Data Set 2. Sequencing and Mapping Statistics for All Single-End Libraries Sequenced.

Supplemental Data Set 3. Quality Assessment of Representative Contigs against Predicted CDS within *T. hassleriana*.

Supplemental Data Set 4. MapMan Categories of Highly Expressed Genes in Each Tissue.

Supplemental Data Set 5. Transcriptional Investment of Each Enriched Basal MapMan Categories in Percentage for Each Tissue.

Supplemental Data Set 6. List of All Genes with Euclidean Distance over 800 RPKM Expressed within Both Leaf Gradients.

Supplemental Data Set 7. List of Core Cell Cycle Genes Selected for Clustering.

Supplemental Data Set 8. Statistical Analysis of Differential Transcript Abundances between *G. gynandra* and *T. hassleriana* for Each Tissue.

Supplemental Data Set 9. Genes Assigned by *K*-Means Clustering to Each Cluster.

Supplemental Data Set 10. MapMan Enrichment Analysis of *K*-Means Clustering.

Supplemental Data Set 11. List of Genes Highly Correlated with Cluster Centers of Shifted Clusters.

ACKNOWLEDGMENTS

Work in our laboratory was supported by grants from the Deutsche Forschungsgemeinschaft (EXC 1028, IRTG 1525, and WE 2231/9-1 to A.P.M.W.). We thank the HHU Biomedical Research Center (BMFZ) for support with RNA-seq analysis and the MSU High Performance Computing Cluster for support with computational analysis of RNA-seq

data. We thank Stefanie Weidtkamp-Peters and the HHU Center for Advanced Imaging for expert advice and support with confocal microscopy and image analysis.

AUTHOR CONTRIBUTIONS

C.K. performed experimental work, analyzed data, and wrote the article. A.K.D. analyzed data and cowrote the article. M.S. assisted in data analysis, identified the root-to-shoot shift, and cowrote the article. J.M., S.S., T.J.W., and E.G.-C. assisted in data analysis. B.B. assisted in design of ploidy experiments. C.R.B assisted in data analysis and experimental design. R.S. assisted in data discussion. L.D.V. assisted in ploidy determination. A.B. analyzed data and wrote the article. A.P.M.W. designed the study and wrote the article.

Received January 30, 2014; revised June 20, 2014; accepted July 6, 2014; published August 8, 2014.

REFERENCES

Anderson, L.E. (1971). Chloroplast and cytoplasmic enzymes. II. Pea leaf triose phosphate isomerases. Biochim. Biophys. Acta **235:** 237–244.

Andrews, T.J., and Lorimer, G.H. (1987). Rubisco: Structure, mechanisms, and prospects for improvement. In The Biochemistry of Plants, Vol. 10, Photosynthesis, M.D. Hatch and N.K. Boardman, eds (San Diego, CA: Academic Press), pp. 131–218.

Andriankaja, M., Dhondt, S., De Bodt, S., Vanhaeren, H., Coppens, F., De Milde, L., Mühlenbock, P., Skirycz, A., Gonzalez, N., Beemster, G.T.S., and Inzé, D. (2012). Exit from proliferation during leaf development in *Arabidopsis thaliana*: a not-so-gradual process. Dev. Cell **22:** 64–78.

Anzola, J.M., Sieberer, T., Ortbauer, M., Butt, H., Korbei, B., Weinhofer, I., Müllner, A.E., and Luschnig, C. (2010). Putative Arabidopsis transcriptional adaptor protein (PROPORZ1) is required to modulate histone acetylation in response to auxin. Proc. Natl. Acad. Sci. USA **107:** 10308–10313.

Aubry, S., Knerová, J., and Hibberd, J.M. (2013). Endoreduplication is not involved in bundle-sheath formation in the C_4 species *Cleome gynandra*. J. Exp. Bot. **65:** 3557–3566.

Bayer, E., Thomas, C., and Maule, A. (2008). Symplastic domains in the Arabidopsis shoot apical meristem correlate with PDLP1 expression patterns. Plant Signal. Behav. **3:** 853–855.

Beebo, A., et al. (2009). Life with and without AtTIP1;1, an Arabidopsis aquaporin preferentially localized in the apposing tonoplasts of adjacent vacuoles. Plant Mol. Biol. **70:** 193–209.

Beemster, G.T.S., De Veylder, L., Vercruysse, S., West, G., Rombaut, D., Van Hummelen, P., Galichet, A., Gruissem, W., Inzé, D., and Vuylsteke, M. (2005). Genome-wide analysis of gene expression profiles associated with cell cycle transitions in growing organs of Arabidopsis. Plant Physiol. **138:** 734–743.

Besnard, G., Baali-Cherif, D., Bettinelli-Riccardi, S., Parietti, D., and Bouguedoura, N. (2009). Pollen-mediated gene flow in a highly fragmented landscape: consequences for defining a conservation strategy of the relict Laperrine's olive. C. R. Biol. **332:** 662–672.

Brady, S.M., and Provart, N.J. (2009). Web-queryable large-scale data sets for hypothesis generation in plant biology. Plant Cell **21:** 1034–1051.

Bräutigam, A., et al. (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, C.P., and Weber, A.P.M. (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. J. Exp. Bot. **65:** 3579–3593.

Breuer, C., Morohashi, K., Kawamura, A., Takahashi, N., Ishida, T., Umeda, M., Grotewold, E., and Sugimoto, K. (2012). Transcriptional repression of the APC/C activator CCS52A1 promotes active termination of cell growth. EMBO J. **31:** 4488–4501.

Breuer, C., Kawamura, A., Ichikawa, T., Tominaga-Wada, R., Wada, T., Kondou, Y., Muto, S., Matsui, M., and Sugimoto, K. (2009). The trihelix transcription factor GTL1 regulates ploidy-dependent cell growth in the Arabidopsis trichome. Plant Cell **21:** 2307–2322.

Brown, N.J., Parsley, K., and Hibberd, J.M. (2005). The future of C4 research—maize, Flaveria or Cleome? Trends Plant Sci. **10:** 215–221.

Brown, W.V. (1975). Variations in anatomy, associations, and origins of Kranz tissue. Am. J. Bot. **62:** 395–402.

Byrne, M.E. (2012). Making leaves. Curr. Opin. Plant Biol. **15:** 24–30.

Caro, E., Desvoyes, B., and Gutierrez, C. (2012). GTL1 keeps cell growth and nuclear ploidy under control. EMBO J. **31:** 4483–4485.

Chamovitz, D.A., Wei, N., Osterlund, M.T., von Arnim, A.G., Staub, J.M., Matsui, M., and Deng, X.W. (1996). The COP9 complex, a novel multisubunit nuclear regulator involved in light control of a plant developmental switch. Cell **86:** 115–121.

Chapman, E.A., and Osmond, C.B. (1974). The effect of light on the tricarboxylic acid cycle in green leaves: III. A Comparison between some C(3) and C(4) plants. Plant Physiol. **53:** 893–898.

Chastain, C.J., Failing, C.J., Manandhar, L., Zimmerman, M.A., Lakner, M.M., and Nguyen, T.H.T. (2011). Functional evolution of C(4) pyruvate, orthophosphate dikinase. J. Exp. Bot. **62:** 3083–3091.

Cheng, S., et al. (2013). The Tarenaya hassleriana genome provides insight into reproductive trait and genome evolution of crucifers. Plant Cell **25:** 2813–2830.

Chevalier, C., Nafati, M., Mathieu-Rivet, E., Bourdon, M., Frangne, N., Cheniclet, C., Renaudin, J.-P., Gévaudant, F., and Hernould, M. (2011). Elucidating the functional role of endoreduplication in tomato fruit development. Ann. Bot. (Lond.) **107:** 1159–1169.

Chollet, R., Vidal, J., and O'Leary, M.H. (1996). Phosphoenolpyruvate carboxylase: A ubiquitous, highly regulated enzyme in plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. **47:** 273–298.

Christin, P.-A., and Besnard, G. (2009). Two independent C4 origins in Aristidoideae (Poaceae) revealed by the recruitment of distinct phosphoenolpyruvate carboxylase genes. Am. J. Bot. **96:** 2234–2239.

Christin, P.-A., Osborne, C.P., Chatelet, D.S., Columbus, J.T., Besnard, G., Hodkinson, T.R., Garrison, L.M., Vorontsova, M.S., and Edwards, E.J. (2013). Anatomical enablers and the evolution of C4 photosynthesis in grasses. Proc. Natl. Acad. Sci. USA **110:** 1381–1386.

Christin, P.A., Salamin, N., Kellogg, E.A., Vicentini, A., and Besnard, G. (2009). Integrating phylogeny into studies of C4 variation in the grasses. Plant Physiol. **149:** 82–87.

Dohmann, E.M.N., Levesque, M.P., De Veylder, L., Reichardt, I., Jürgens, G., Schmid, M., and Schwechheimer, C. (2008). The Arabidopsis COP9 signalosome is essential for G2 phase progression and genomic stability. Development **135:** 2013–2022.

Donner, T.J., Sherr, I., and Scarpella, E. (2009). Regulation of preprocambial cell state acquisition by auxin signaling in Arabidopsis leaves. Development **136:** 3235–3246.

Edwards, E.J., et al.; C4 Grasses Consortium (2010). The origins of C4 grasslands: integrating evolutionary and ecosystem science. Science **328:** 587–591.

Ehleringer, J.R., and Björkman, O. (1978). A comparison of photosynthetic characteristics of encelia species possessing glabrous and pubescent leaves. Plant Physiol. **62:** 185–190.

Ehleringer, J.R., Sage, R.F., Flanagan, L.B., and Pearcy, R.W. (1991). Climate change and the evolution of C(4) photosynthesis. Trends Ecol. Evol. (Amst.) **6:** 95–99.

Fankhauser, C., and Chory, J. (1997). Light control of plant development. Annu. Rev. Cell Dev. Biol. **13:** 203–229.

Furbank, R.T., and Hatch, M.D. (1987). Mechanism of c(4) photosynthesis: the size and composition of the inorganic carbon pool in bundle sheath cells. Plant Physiol. **85:** 958–964.

Gowik, U., Bräutigam, A., Weber, K.L., Weber, A.P.M., and 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.

Griffiths, H., Weller, G., Toy, L.F., and Dennis, R.J. (2013). You're so vein: bundle sheath physiology, phylogeny and evolution in C3 and C4 plants. Plant Cell Environ. **36:** 249–261.

Haberlandt, G. (1896). Physiologische Pflanzenanatomie. (Leipzig, Germany: Verlag von Wilhelm Engelmann).

Hatch, M.D. (1987). C-4 photosynthesis - a unique blend of modified biochemistry, anatomy and ultrastructure. Biochim. Biophys. Acta **895:** 81–106.

Hay, J., and Schwender, J. (2011). Computational analysis of storage synthesis in developing *Brassica napus* L. (oilseed rape) embryos: flux variability analysis in relation to ^{13}C metabolic flux analysis. Plant J. **67:** 513–525.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6:** 65–70.

Ilegems, M., Douet, V., Meylan-Bettex, M., Uyttewaal, M., Brand, L., Bowman, J.L., and Stieger, P.A. (2010). Interplay of auxin, KANADI and Class III HD-ZIP transcription factors in vascular tissue formation. Development **137:** 975–984.

Iltis, H.H., and Cochrane, T.S. (2007). Studies in the Cleomaceae V: A new genus and ten new combinations for the flora of North America. Novon **17:** 447–451.

Iltis, H.H., Hall, J.C., Cochrane, T.S., and Sytsma, K.J. (2011). Studies in the Cloemaceae I. On the separate recognition of Capparaceae, Cleomaceae, and Brassicaceae. Annals Miss. Bot. Gard. **98:** 28–36.

Inda, L.A., Torrecilla, P., Catalán, P., and Ruiz-Zapata, T. (2008). Phylogeny of Cleome L. and its close relatives Podandrogyne Ducke and Polanisia Raf. (Cleomoideae, Cleomaceae) based on analysis of nuclear ITS sequences and morphology. Plant Sys. Evol. **274:** 111–126.

Kang, J., Mizukami, Y., Wang, H., Fowke, L., and Dengler, N.G. (2007). Modification of cell proliferation patterns alters leaf vein architecture in *Arabidopsis thaliana*. Planta **226:** 1207–1218.

Kasili, R., Walker, J.D., Simmons, L.A., Zhou, J., De Veylder, L., and Larkin, J.C. (2010). SIAMESE cooperates with the CDH1-like protein CCS52A1 to establish endoreplication in *Arabidopsis thaliana* trichomes. Genetics **185:** 257–268.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. **12:** 656–664.

Knappe, S., Flügge, U.I., and Fischer, K. (2003a). Analysis of the plastidic phosphate translocator gene family in Arabidopsis and identification of new phosphate translocator-homologous transporters, classified by their putative substrate-binding site. Plant Physiol. **131:** 1178–1190.

Knappe, S., Löttgert, T., Schneider, A., Voll, L., Flügge, U.I., and Fischer, K. (2003b). Characterization of two functional phosphoenolpyruvate/phosphate translocator (PPT) genes in Arabidopsis—AtPPT1 may be involved in the provision of signals for correct mesophyll development. Plant J. **36:** 411–420.

Lammens, T., Boudolf, V., Kheibarshekan, L., Zalmas, L.P., Gaamouche, T., Maes, S., Vanstraelen, M., Kondorosi, E., La Thangue, N.B., Govaerts, W., Inzé, D., and De Veylder, L. (2008). Atypical E2F activity restrains APC/CCCS52A2 function obligatory for endocycle onset. Proc. Natl. Acad. Sci. USA **105:** 14721–14726.

Langdale, J.A., and Nelson, T. (1991). Spatial regulation of photosynthetic development in C_4 plants. Trends Genet. **7:** 191–196.

Larson-Rabin, Z., Li, Z., Masson, P.H., and Day, C.D. (2009). FZR2/CCS52A1 expression is a determinant of endoreduplication and cell expansion in Arabidopsis. Plant Physiol. **149:** 874–884.

Lee, B.H., Ko, J.-H., Lee, S., Lee, Y., Pak, J.-H., and Kim, J.H. (2009a). The Arabidopsis GRF-INTERACTING FACTOR gene family performs an overlapping function in determining organ size as well as multiple developmental properties. Plant Physiol. **151:** 655–668.

Lee, H.O., Davidson, J.M., and Duronio, R.J. (2009b). Endoreplication: polyploidy with purpose. Genes Dev. **23:** 2461–2477.

Li, P., et al. (2010). The developmental dynamics of the maize leaf transcriptome. Nat. Genet. **42:** 1060–1067.

Lundquist, P.K., Rosar, C., Bräutigam, A., and Weber, A.P. (2014). Plastid signals and the bundle sheath: mesophyll development in reticulate mutants. Mol. Plant **7:** 14–29.

Mantiri, F.R., Kurdyukov, S., Lohar, D.P., Sharopova, N., Saeed, N.A., Wang, X.-D., Vandenbosch, K.A., and Rose, R.J. (2008). The transcription factor MtSERF1 of the ERF subfamily identified by transcriptional profiling is required for somatic embryogenesis induced by auxin plus cytokinin in *Medicago truncatula*. Plant Physiol. **146:** 1622–1636.

Marshall, D.M., Muhaidat, R., Brown, N.J., Liu, Z., Stanley, S., Griffiths, H., Sage, R.F., and Hibberd, J.M. (2007). Cleome, a genus closely related to Arabidopsis, contains species spanning a developmental progression from C(3) to C(4) photosynthesis. Plant J. **51:** 886–896.

Mathieu-Rivet, E., Gévaudant, F., Cheniclet, C., Hernould, M., and Chevalier, C. (2010a). The anaphase promoting complex activator CCS52A, a key factor for fruit growth and endoreduplication in tomato. Plant Signal. Behav. **5:** 985–987.

Mathieu-Rivet, E., Gévaudant, F., Sicard, A., Salar, S., Do, P.T., Mouras, A., Fernie, A.R., Gibon, Y., Rothan, C., Chevalier, C., and Hernould, M. (2010b). Functional analysis of the anaphase promoting complex activator CCS52A highlights the crucial role of endo-reduplication for fruit growth in tomato. Plant J. **62:** 727–741.

Matsuoka, M. (1995). The gene for pyruvate, orthophosphate dikinase in C_4 plants: structure, regulation and evolution. Plant Cell Physiol. **36:** 937–943.

McKown, A.D., and Dengler, N.G. (2009). Shifts in leaf vein density through accelerated vein formation in C4 Flaveria (Asteraceae). Ann. Bot. (Lond.) **104:** 1085–1098.

McKown, A.D., and Dengler, N.G. (2010). Vein patterning and evolution in C-4 plants. Botany **88:** 775–786.

Motose, H., Sugiyama, M., and Fukuda, H. (2004). A proteoglycan mediates inductive interaction during plant vascular development. Nature **429:** 873–878.

Nelson, T., and Langdale, J.A. (1992). Developmental genetics of C-4 photosynthesis. Annu. Rev. Plant Physiol. Plant Mol. Biol. **43:** 25–47.

Nelson, T., and Dengler, N. (1997). Leaf vascular pattern formation. Plant Cell **9:** 1121–1135.

Ohashi-Ito, K., and Fukuda, H. (2010). Transcriptional regulation of vascular cell fates. Curr. Opin. Plant Biol. **13:** 670–676.

Peeples, M.A. (2011). R Script for K-Means Cluster Analysis. http://www.mattpeeples.net/kmeans.html.

Pérez-Pérez, J.M., Candela, H., Robles, P., López-Torrejón, G., del Pozo, J.C., and Micol, J.L. (2010). A role for AUXIN RESISTANT3 in the coordination of leaf growth. Plant Cell Physiol. **51:** 1661–1673.

Pérez-Pérez, J.M., Serralbo, O., Vanstraelen, M., González, C., Criqui, M.C., Genschik, P., Kondorosi, E., and Scheres, B. (2008). Specialization of CDC27 function in the *Arabidopsis thaliana* anaphase-promoting complex (APC/C). Plant J. **53:** 78–89.

Pick, T.R., Bräutigam, A., Schlüter, U., Denton, A.K., Colmsee, C., Scholz, U., Fahnenstich, H., Pieruschka, R., Rascher, U., Sonnewald, U., and Weber, A.P.M. (2011). Systems analysis of a maize leaf developmental gradient redefines the current C4 model and provides candidates for regulation. Plant Cell **23:** 4208–4220.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. (Vienna, Austria: R Foundation for Statistical Computing).

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26:** 139–140.

Sage, R.F. (2004). The evolution of C-4 photosynthesis. New Phytol. **161:** 341–370.

Sage, R.F., and McKown, A.D. (2006). Is C4 photosynthesis less phenotypically plastic than C3 photosynthesis? J. Exp. Bot. **57:** 303–317.

Sage, R.F., Christin, P.A., and Edwards, E.J. (2011). The $C_{(4)}$ plant lineages of planet Earth. J. Exp. Bot. **62:** 3155–3169.

Sage, R.F., and Zhu, X.G. (2011). Exploiting the engine of $C_{(4)}$ photosynthesis. J. Exp. Bot. **62:** 2989–3000.

Scarpella, E., Francis, P., and Berleth, T. (2004). Stage-specific markers define early steps of procambium development in Arabidopsis leaves and correlate termination of vein formation with mesophyll differentiation. Development **131:** 3445–3455.

Scarpella, E., Marcos, D., Friml, J., and Berleth, T. (2006). Control of leaf vascular patterning by polar auxin transport. Genes Dev. **20:** 1015–1027.

Sieberer, T., Hauser, M.T., Seifert, G.J., and Luschnig, C. (2003). PROPORZ1, a putative Arabidopsis transcriptional adaptor protein, mediates auxin and cytokinin signals in the control of cell proliferation. Curr. Biol. **13:** 837–842.

Slewinski, T.L. (2013). Using evolution as a guide to engineer kranz-type c4 photosynthesis. Front. Plant Sci. **4:** 212.

Slewinski, T.L., Anderson, A.A., Zhang, C., and Turgeon, R. (2012). Scarecrow plays a role in establishing Kranz anatomy in maize leaves. Plant Cell Physiol. **53:** 2030–2037.

Stein, H., Honig, A., Miller, G., Erster, O., Eilenberg, H., Csonka, L.N., Szabados, L., Koncz, C., and Zilberstein, A. (2011). Elevation of free proline and proline-rich protein levels by simultaneous manipulations of proline biosynthesis and degradation in plants. Plant Sci. **181:** 140–150.

Streatfield, S.J., Weber, A., Kinsman, E.A., Häusler, R.E., Li, J., Post-Beittenmiller, D., Kaiser, W.M., Pyke, K.A., Flügge, U.I., and Chory, J. (1999). The phosphoenolpyruvate/phosphate translocator is required for phenolic metabolism, palisade cell development, and plastid-dependent nuclear gene expression. Plant Cell **11:** 1609–1622.

Sud, R.M., and Dengler, N.G. (2000). Cell lineage of vein formation in variegated leaves of the C-4 grass *Stenotaphrum secundatum*. Ann. Bot. (Lond.) **86:** 99–112.

Sugimoto-Shirasu, K., and Roberts, K. (2003). "Big it up": endoreduplication and cell-size control in plants. Curr. Opin. Plant Biol. **6:** 544–553.

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics **22:** 1540–1542.

Tobin, E.M., and Silverthorne, J. (1985). Light regulation of gene-expression in higher plants. Annu. Rev. Plant Biol. **36:** 569–593.

Tolley, B.J., Woodfield, H., Wanchana, S., Bruskiewich, R., and Hibberd, J.M. (2012). Light-regulated and cell-specific methylation of the maize PEPC promoter. J. Exp. Bot. **63:** 1381–1390.

Traas, J., Hülskamp, M., Gendreau, E., and Höfte, H. (1998). Endoreduplication and development: rule without dividing? Curr. Opin. Plant Biol. **1:** 498–503.

Tronconi, M.A., Gerrard Wheeler, M.C., Maurino, V.G., Drincovich, M.F., and Andreo, C.S. (2010). NAD-malic enzymes of *Arabidopsis thaliana* display distinct kinetic mechanisms that support differences in physiological control. Biochem. J. **430:** 295–303.

Vandepoele, K., Raes, J., De Veylder, L., Rouzé, P., Rombauts, S., and Inzé, D. (2002). Genome-wide analysis of core cell cycle genes in Arabidopsis. Plant Cell **14:** 903–916.

Wang, P., Kelly, S., Fouracre, J.P., and Langdale, J.A. (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.

Westhoff, P., and Gowik, U. (2010). Evolution of C4 photosynthesis—looking for the master switch. Plant Physiol. **154:** 598–601.

Wheeler, M.C.G., Tronconi, M.A., Drincovich, M.F., Andreo, C.S., Flügge, U.I., and Maurino, V.G. (2005). A comprehensive analysis of the NADP-malic enzyme gene family of Arabidopsis. Plant Physiol. **139:** 39–51.

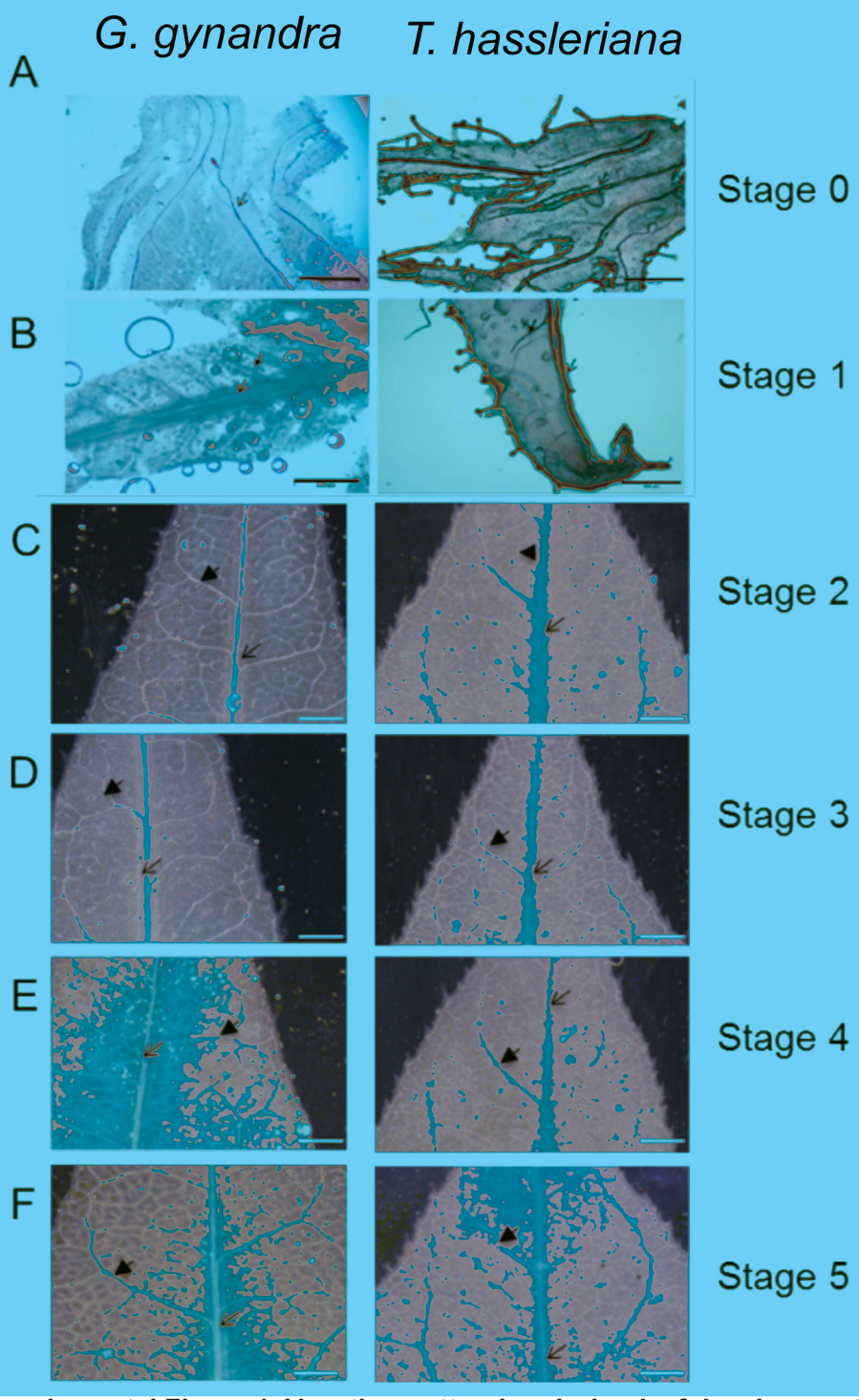
Xu, W., Purugganan, M.M., Polisensky, D.H., Antosiewicz, D.M., Fry, S.C., and Braam, J. (1995). Arabidopsis TCH4, regulated by hormones and the environment, encodes a xyloglucan endotransglycosylase. Plant Cell **7:** 1555–1567.

Yamaguchi, M., Ohtani, M., Mitsuda, N., Kubo, M., Ohme-Takagi, M., Fukuda, H., and Demura, T. (2010). VND-INTERACTING2, a NAC domain transcription factor, negatively regulates xylem vessel formation in Arabidopsis. Plant Cell **22:** 1249–1263.

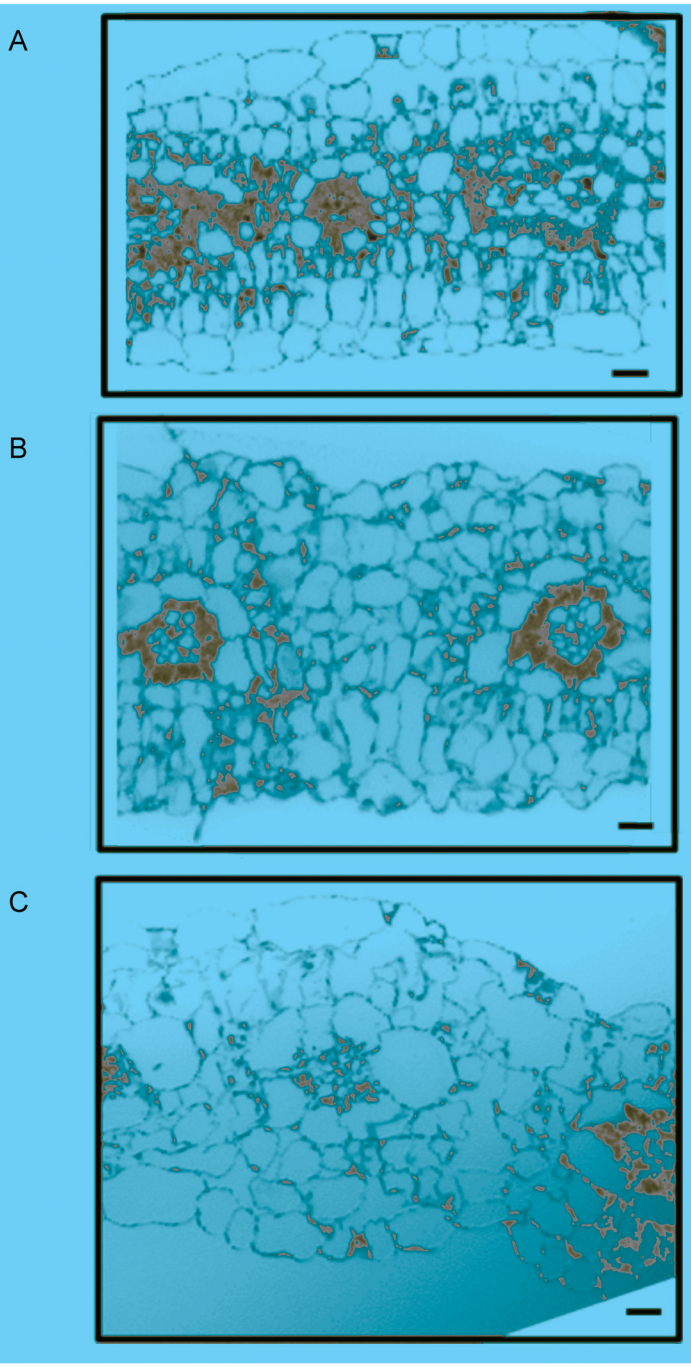
Yekutieli, D., and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J. Stat. Plan. Inference **82:** 171–196.

Zhiponova, M.K., et al. (2013). Brassinosteroid production and signaling differentially control cell division and expansion in the leaf. New Phytol. **197:** 490–502.

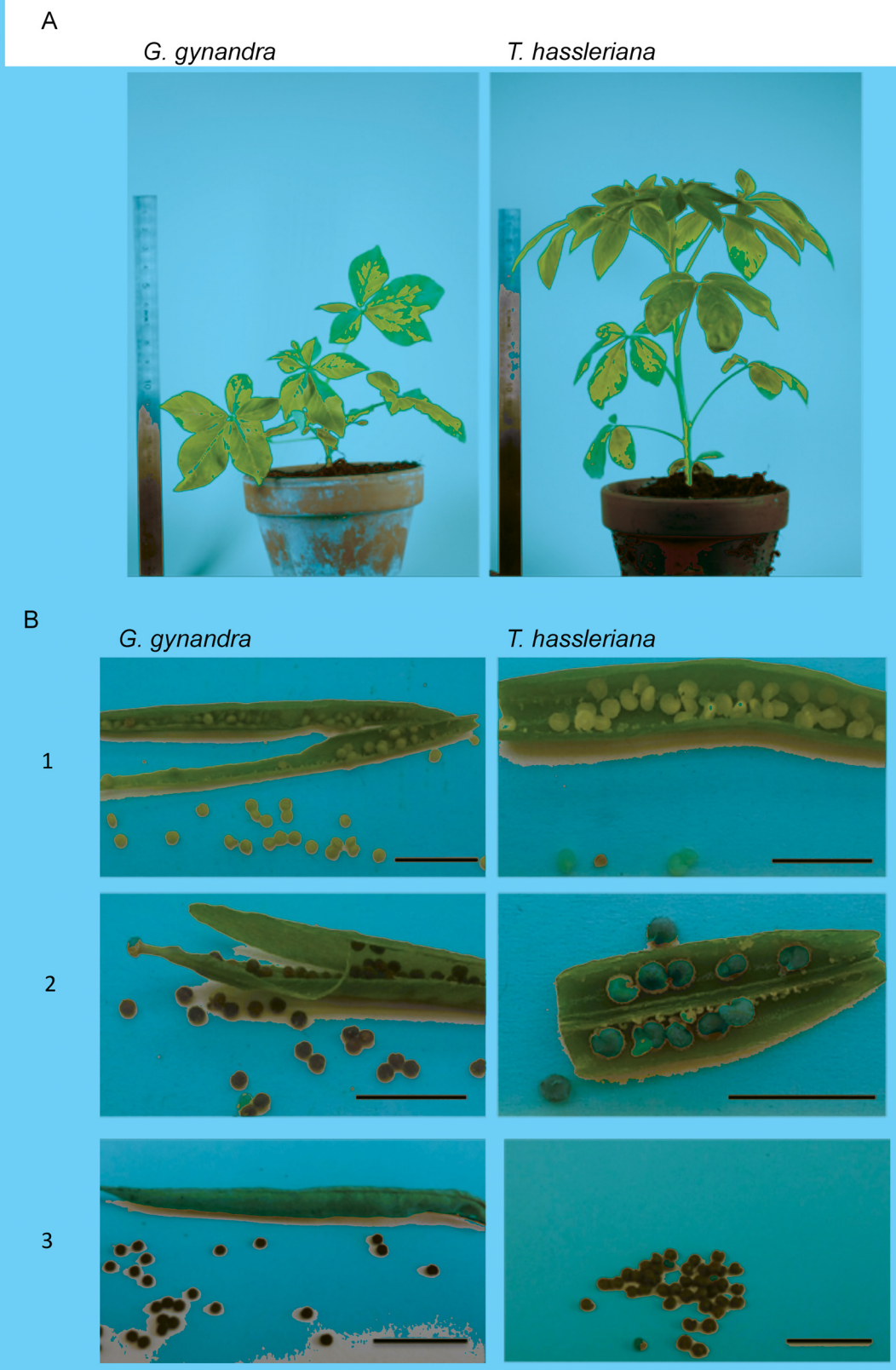
Zhu, X.G., Long, S.P., and Ort, D.R. (2008). What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? Curr. Opin. Biotechnol. **19:** 153–159.



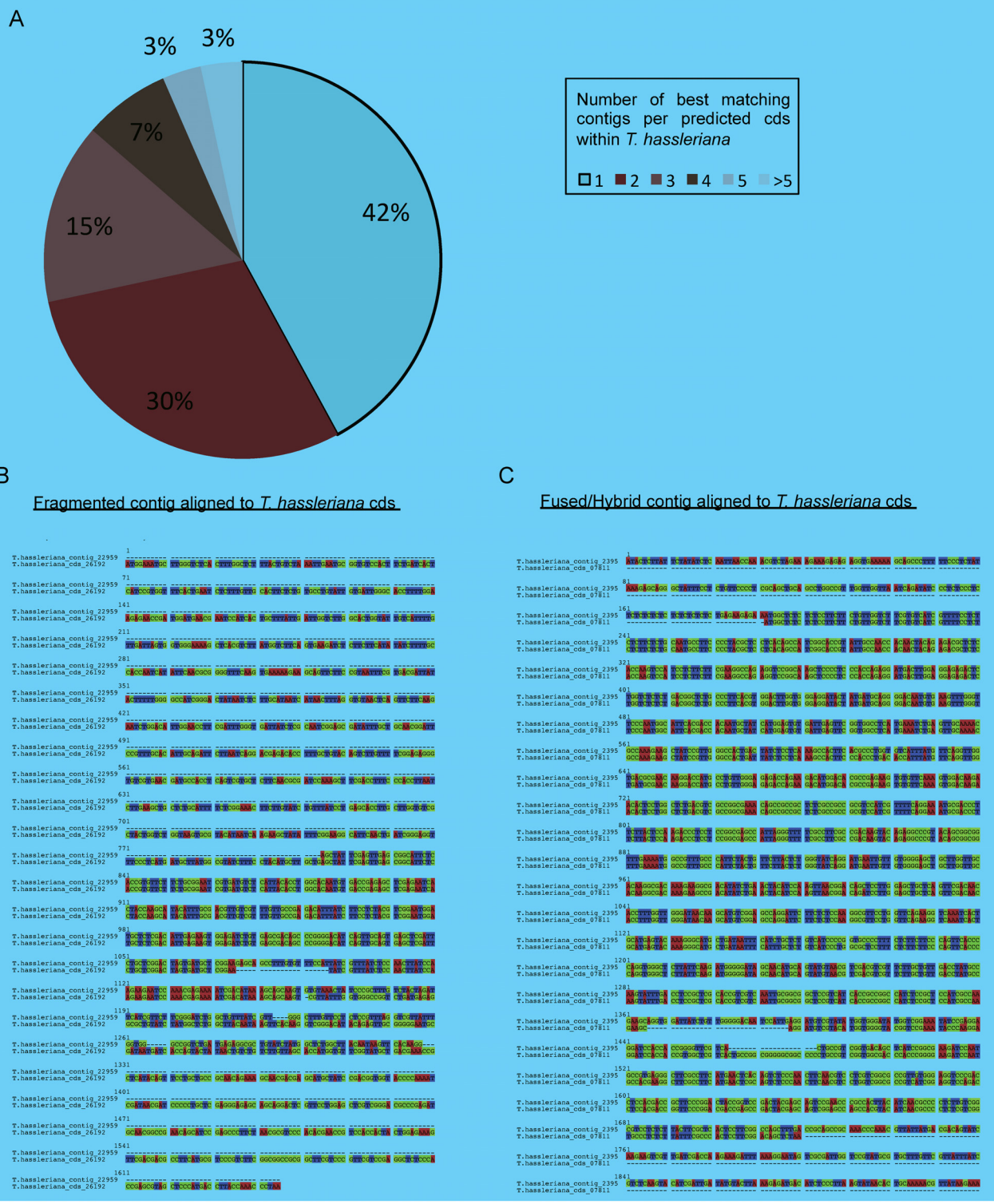
Supplemental Figure 1. Venation patterning during leaf development of *G. gynandra* and *T. hassleriana*.
(A-B) Cleared safranine stained leaves of stage 0 and 1 (n=3; scale bar 0.5 mm)
(C-F) Cleared leaves of stage 2, 3, 4 and 5 respectively (n=3; scale bar 1 mm)
Open arrows indicate the midvein (1°) and closed arrows the secondary vein (2°) localization



Supplemental Figure 2. *G. gynandra* cotyledon anatomy two, four and six days after germination (DAG). Semi-thin cross sections (3 µm) of *G. gynandra* cotyledons after two **(A)**; four **(B)**; six **(C)** DAG. Cross sections were stained with Toluidine Blue. (Scale bar 10 µm, n=3)

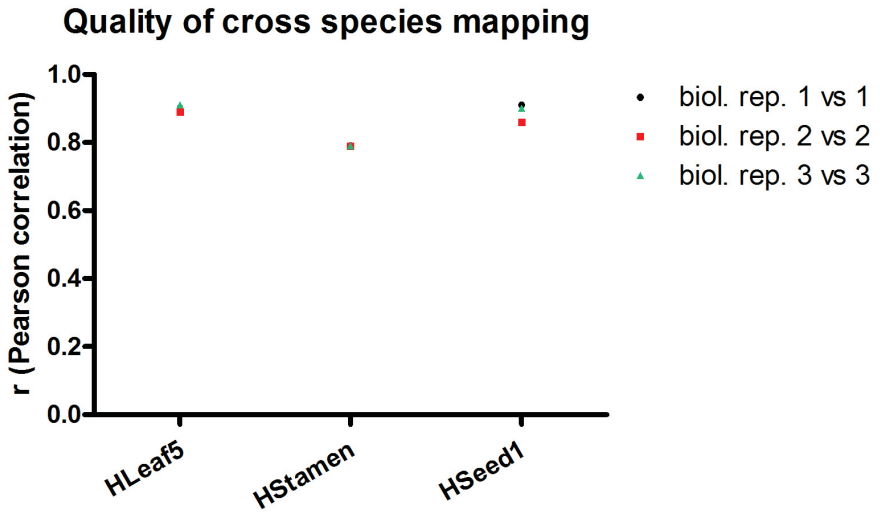


Supplemental Figure 3. Images of tissues harvested for RNA-seq in *G. gynandra* and *T. hassleriana*. (A) Photographic image of *G. gynandra* and *T. hassleriana* 8-week old plants, from which leaf gradient, stem and root system were harvested **(B) Seed coat development from harvested developmental seed gradient. (1)** young seed **(2)** semi-mature seed **(3)** mature seed. (Scale bar = 1cm)

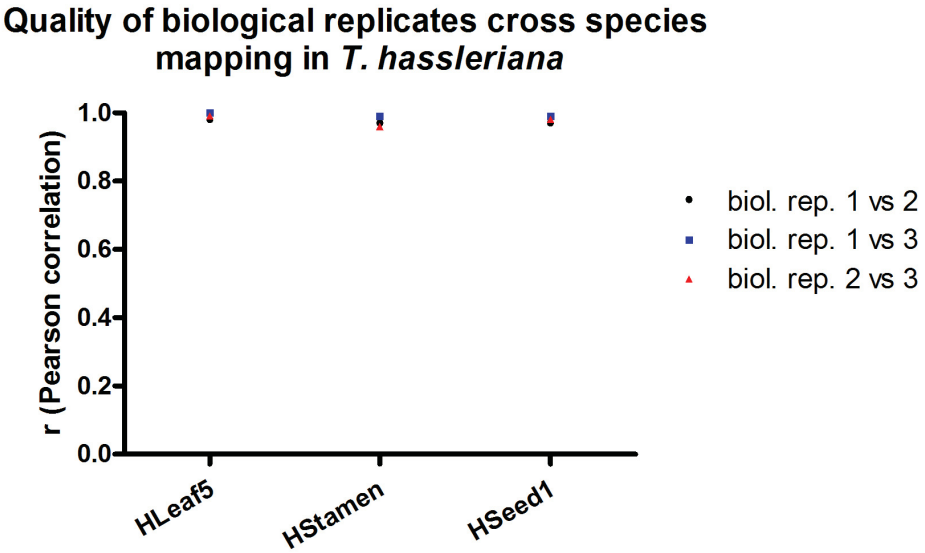


Supplemental Figure 4. Quality assessment of Velvet/OASES assembled *T. hassleriana* contigs against predicted corresponding cds from *T. hassleriana* genome.
(A) Percentage of contig number per predicted cds (Cheng et al., 2013) showing redundancy in assembled contigs.
(B) ClustalW alignment of fragmented contig (top) with corresponding cds (below).
(C) ClustalW alignment of fused contig (top) with corresponding cds (below).

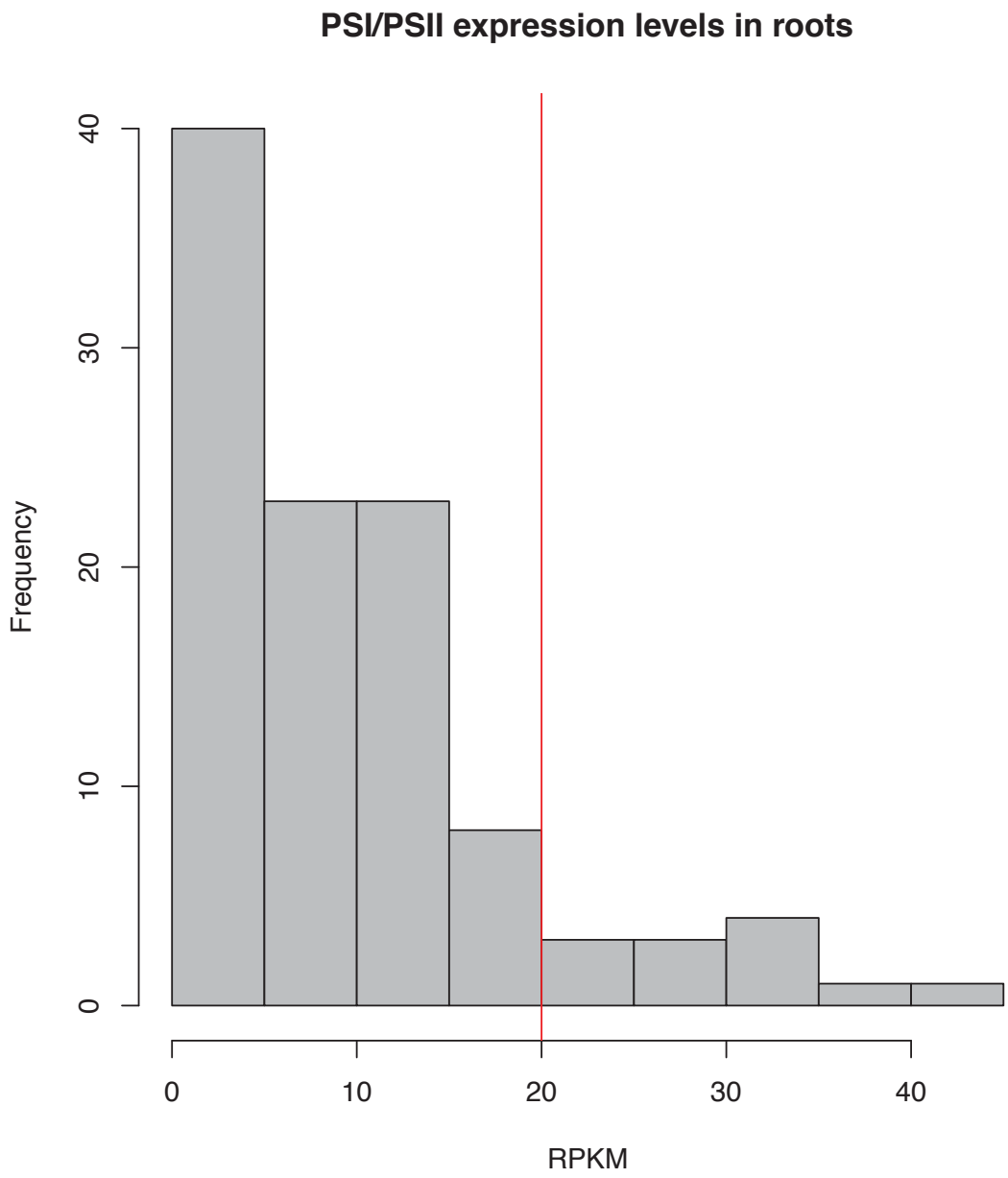
A



B



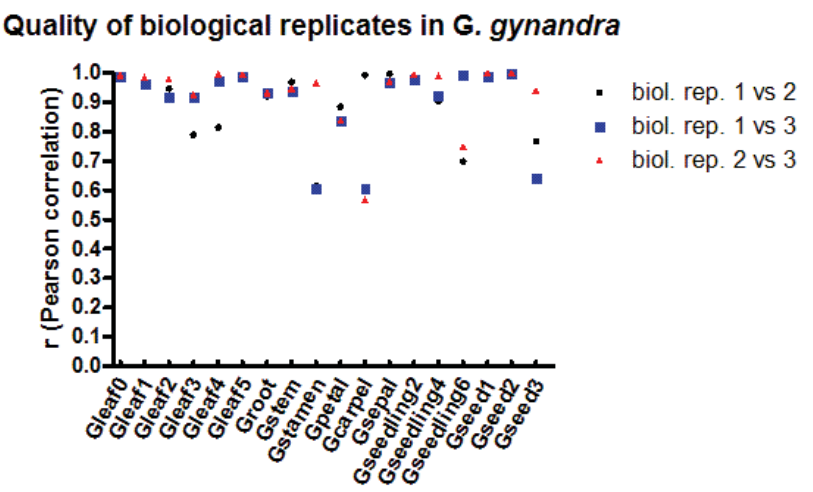
Supplemental Figure 5. Quality assessment of the biological replicates of *T. hassleriana* libraries mapped to *A. thaliana* and mapping similarity of *T. hassleriana* libraries mapped to *A. thaliana* and to its own cds.
(A) Pair-wise Pearson's correlation (*r*) was calculated for all three pairs of biological replicates for each tissue in *T. hassleriana* mapped to *A. thaliana.* **(B)** Pair-wise Pearson's correlation (*r*) between leaf 5, stamen and seed 1 in (n=3) of *T. hassleriana* mapped to its own coding sequence and *A. thaliana*.



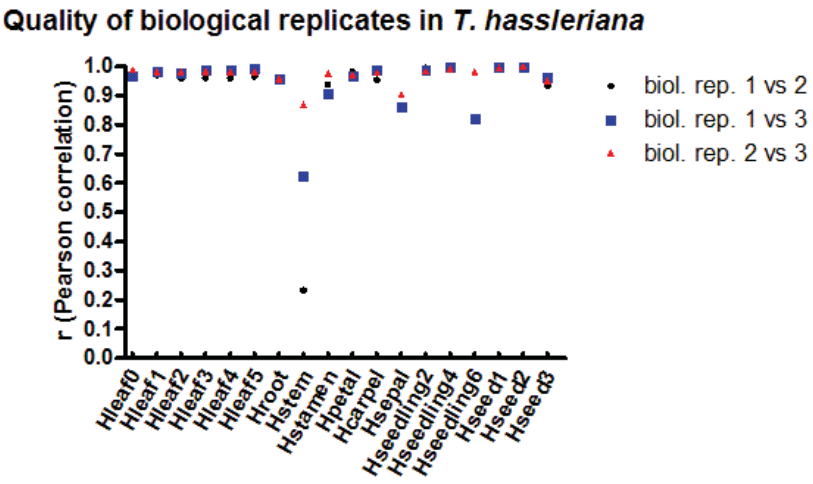
Supplemental Figure 6. Determination of base line gene expression via a histogram of photosystem (PS) I and II transcript abundances reads per mappable million (RPKM) in the *G. gynandra* root.

Y- axis shows frequency and Y- axis depicts RPKM level of PSI and PSII transcript abundance. Red line indicates where threshhold of base line expression was set.

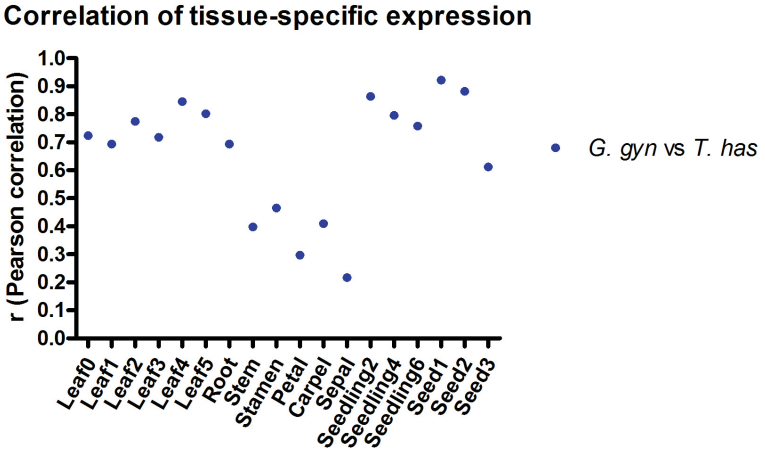
A



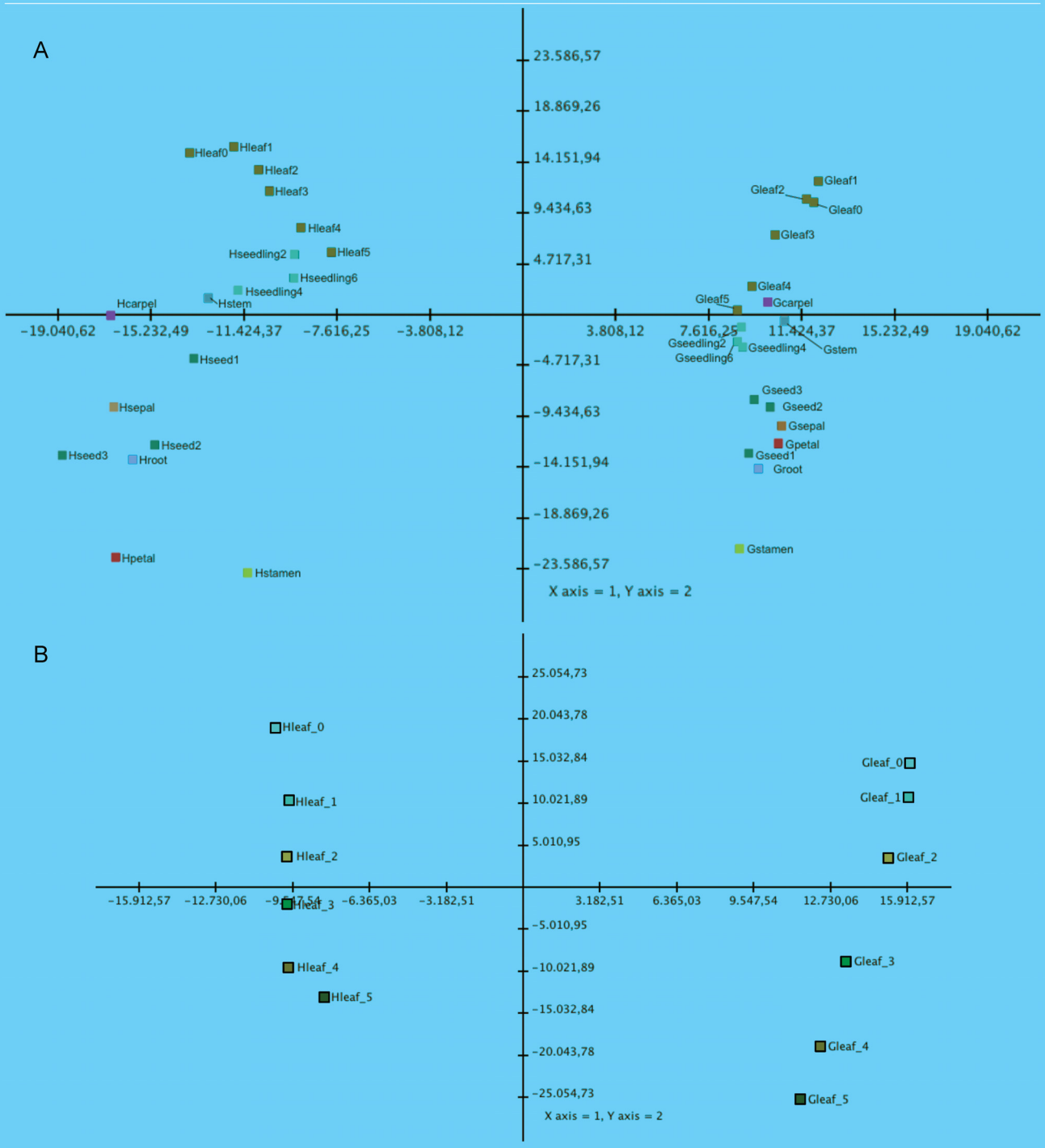
B



C

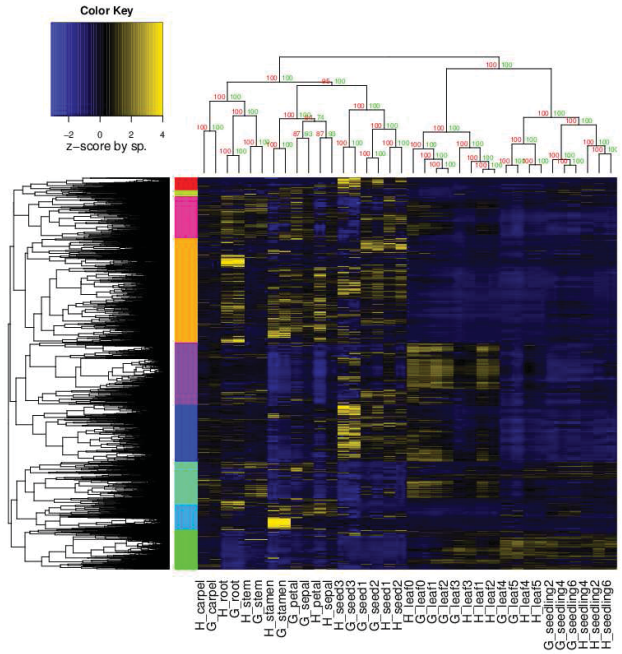


Supplemental Figure 7. Quality assessment of the biological replicates within each species and tissue similarity between *G. gynandra* and *T. hassleriana*. (A) Pair-wise Pearson's correlation (*r*) was calculated for all three pairs of biological replicates for each tissue (n=3) in *G. gynandra*. **(B)** Pair-wise Pearson's correlation (*r*) was calculated for all three pairs of biological replicates for each tissue (n=3) in *T. hassleriana*. **(C)** Pair-wise Pearson‘s correlation between individual tissues of *T. hassleriana* and *G. gynandra*.

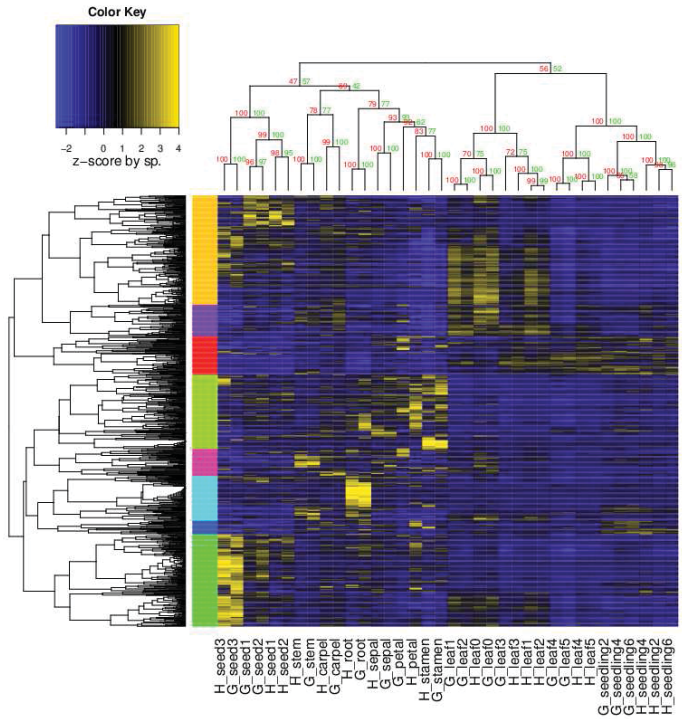


Supplemental Figure 8. Principle component analysis between *G. gynandra* and *T. hassleriana.*
(A) Plot shows all averaged tissues from *G. gynandra* (G) and *T. hassleriana* (H) sequenced (n=3). The first component describes 15% of all data variablility seperating both species. The second component (14%) separates samples by tissue identity within each species. Tissues are indicated by color key (left).
(B) Averaged leaf gradient samples (n=3) from *G. gynandra* (G) and *T. hassleriana* (H) were analysed. First component decribes 44 % and second component describes 29% of variability.

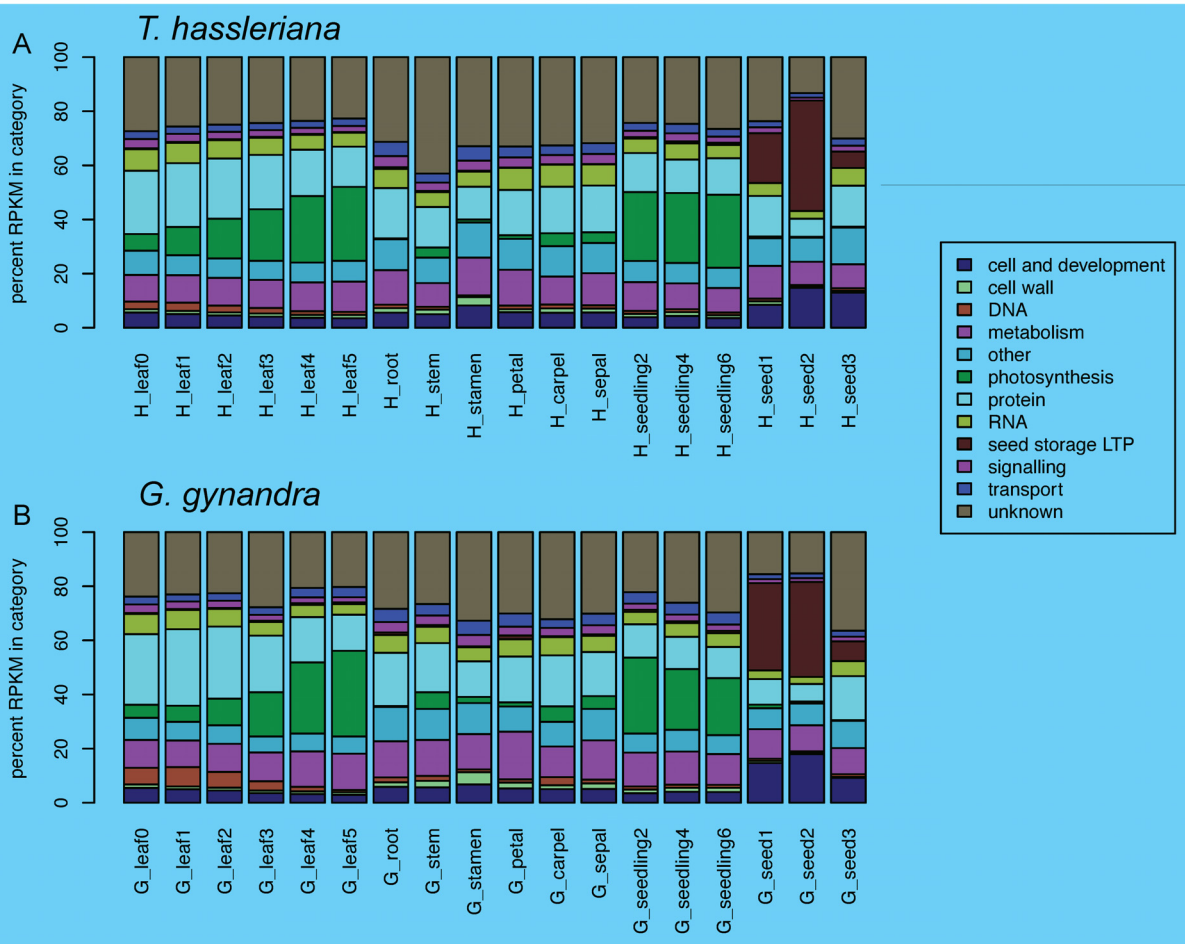
A



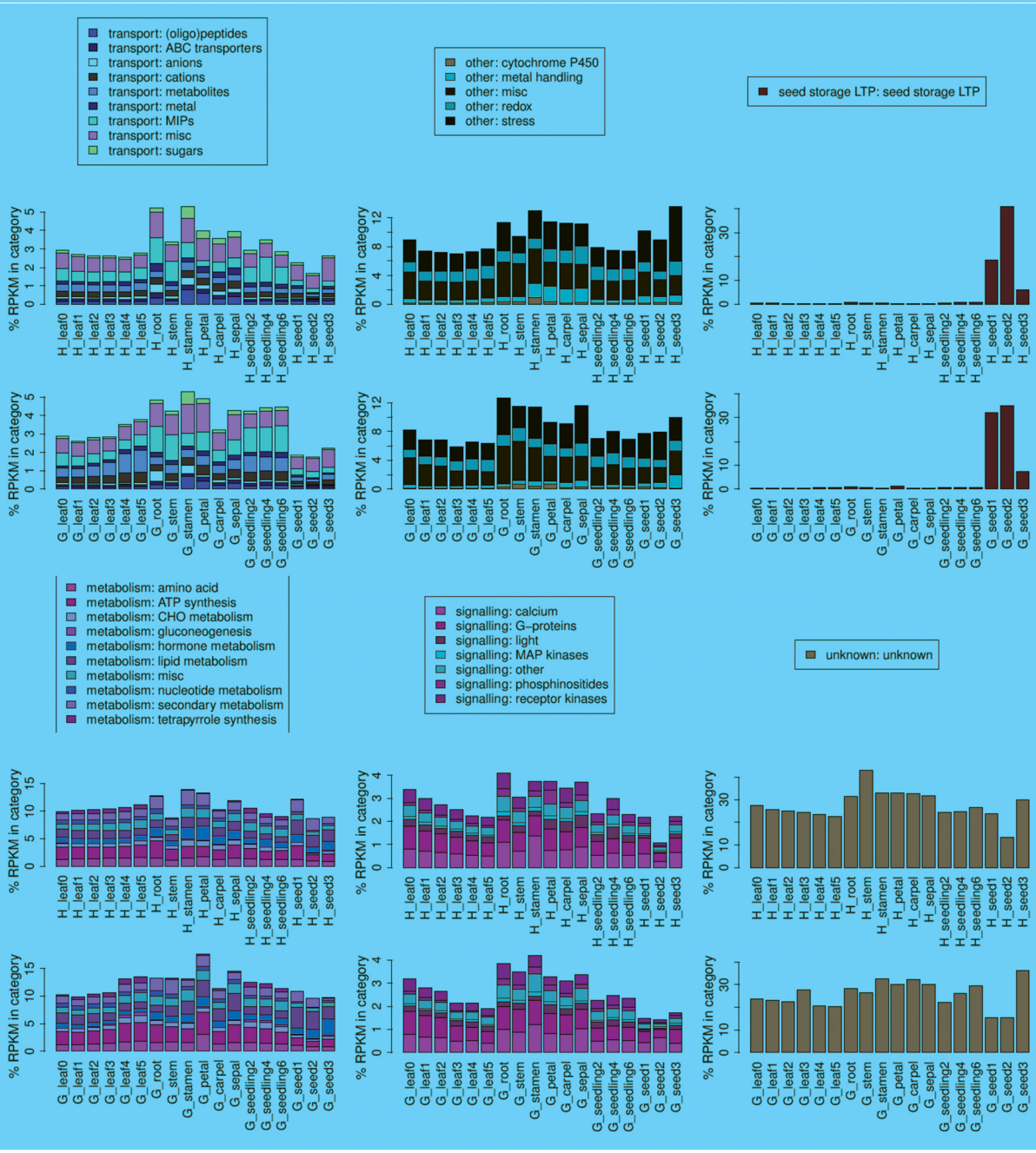
B



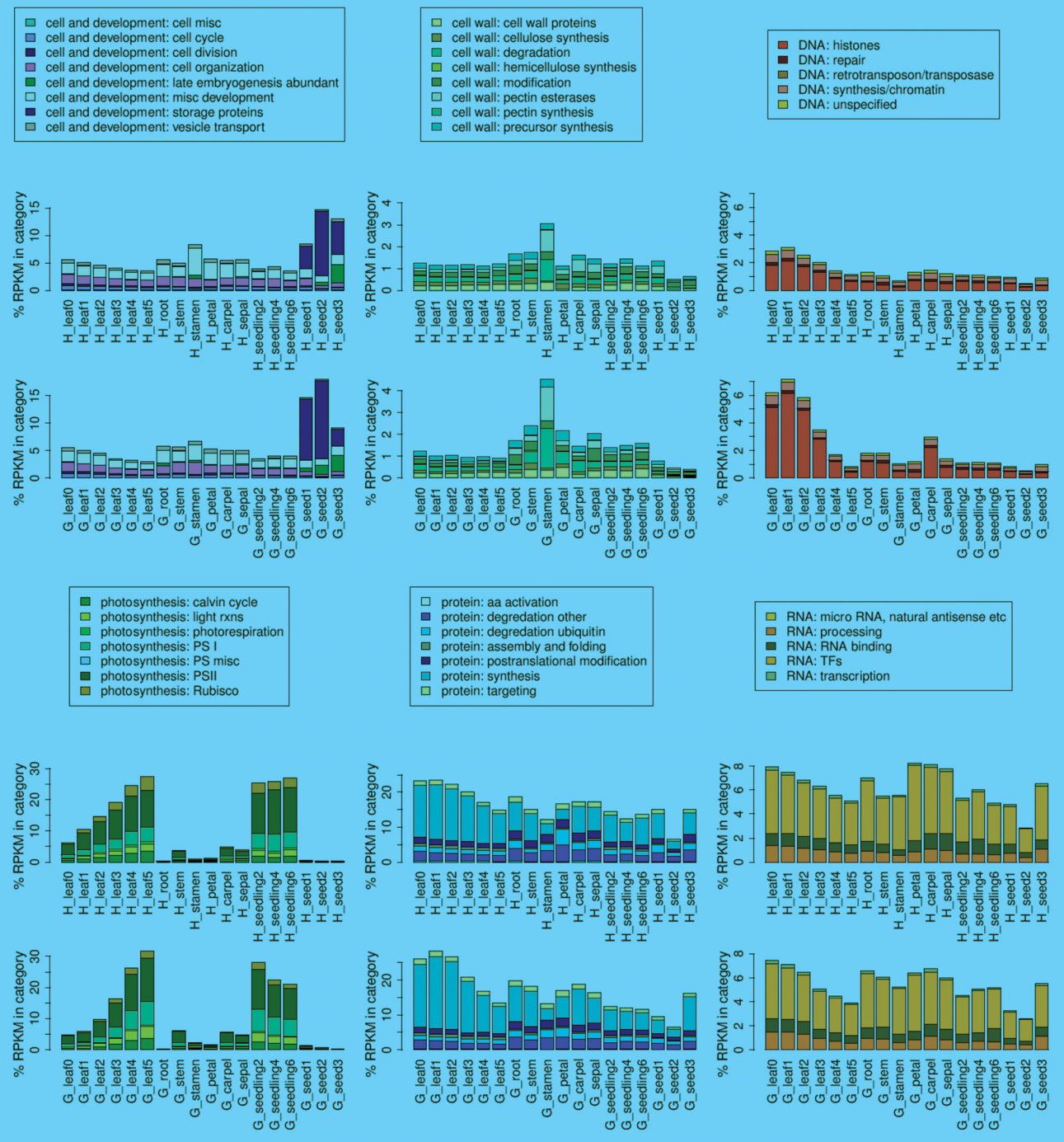
Supplemental Figure 9. Hierarchical cluster analysis with bootstrapped samples of *G. gynandra* and *T. hassleriana*. Numbers above the nodes show the approximately unbiased p-value (red) and the bootstrap probability (green). Blue is lowest expression and yellow highest expression. Left-hand vertical bars denote major clusters in the dendrogram by color. **(A)** Clustering of all over 20 RPKM expressed genes in all averaged samples (n=3). Sample averages were clustered as species scaled Z-scores with Pearson's Correlation. **(B)** Hierarchical Clustering of all transcriptional regulators expressed in all tissues sequenced in *G. gynandra* and *T. hassleriana*. Sample averages (n=3) were clustered as species-scaled Z-scores with Pearson's Correlation.



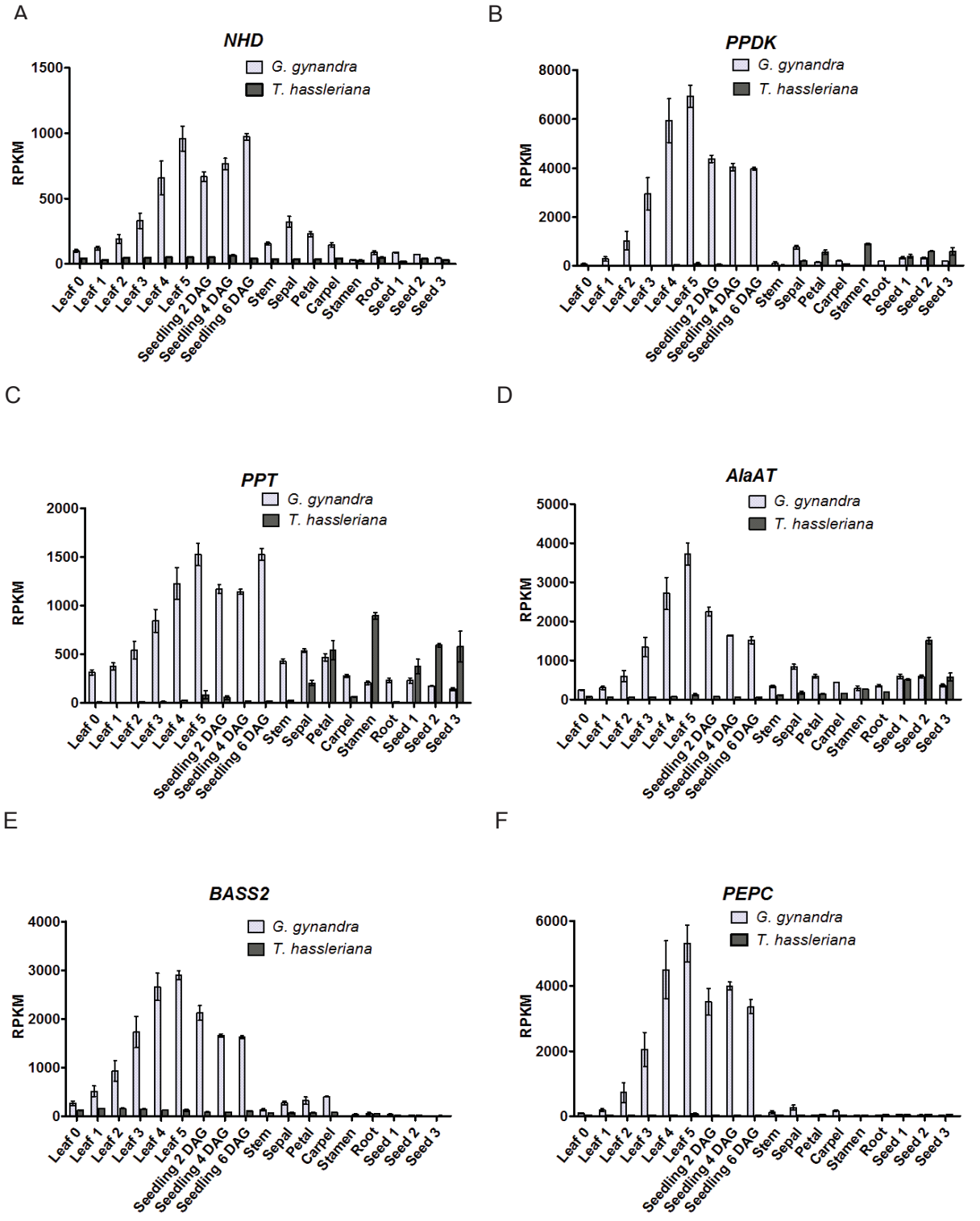
Supplemental Figure 10. Transcriptional investment of each tissue compared in both species. Cumulative average RPKMs in percent for basal Mapman categories for each tissue in *G. gynandra* and *T. hassleriana*.



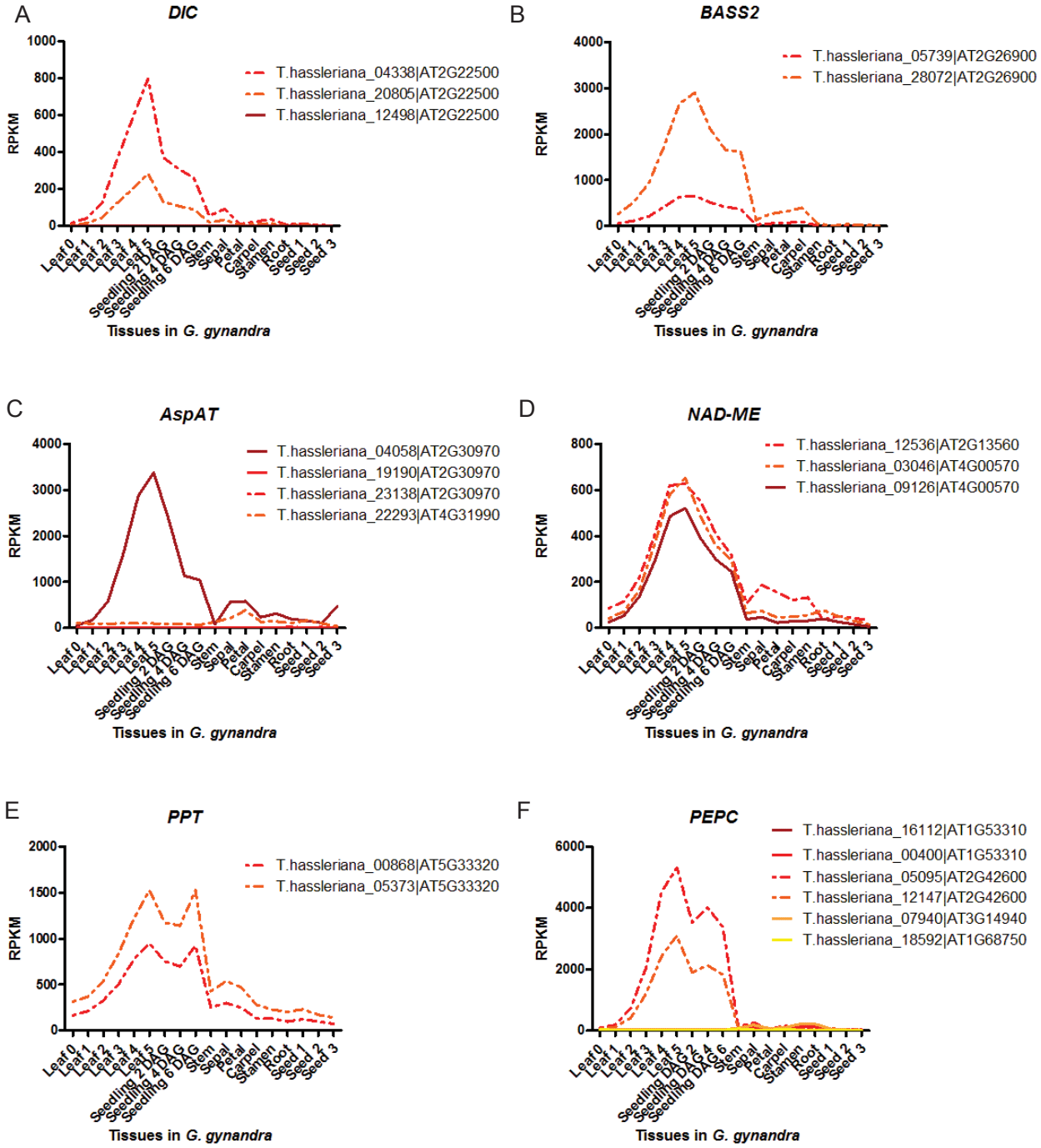
Supplemental Figure 11.1. Transcriptional investment at secondary Mapman category of each tissue compared in both species (Part 1). Distribution of the Mapman categories in each tissue in *G. gynandra* and *T. hassleriana*. Plot shows percent of average RPKMs of the 12 customized secondary Mapman bins for each tissue.



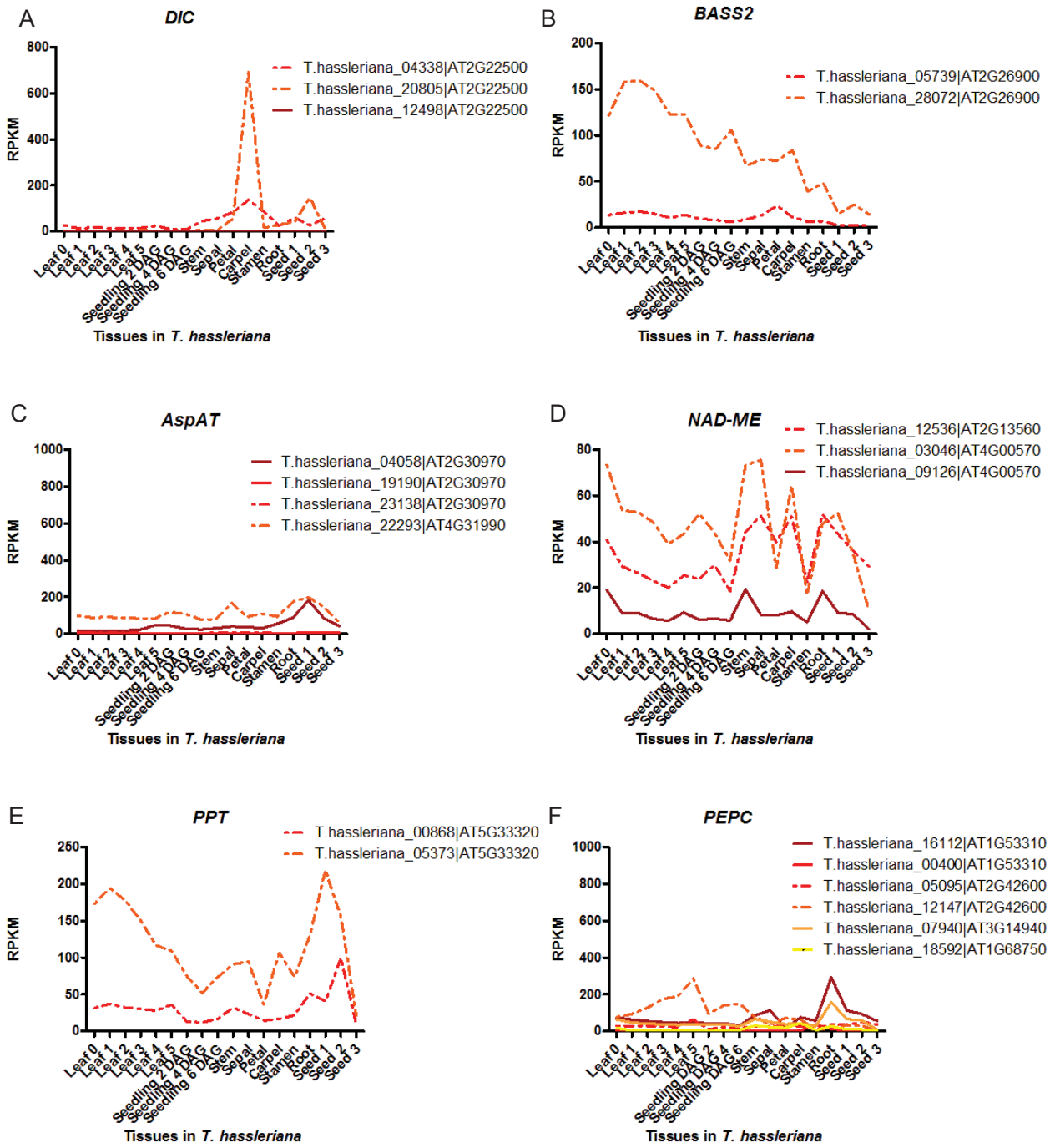
Supplemental Figure 11.2. Transcriptional investment at secondary Mapman category of each tissue compared in both species (Part 2). Distribution of the Mapman categories in each tissue in *G. gynandra* and *T. hassleriana*. Plot shows percent of average RPKMs of the 12 customized secondary Mapman bins for each tissue.



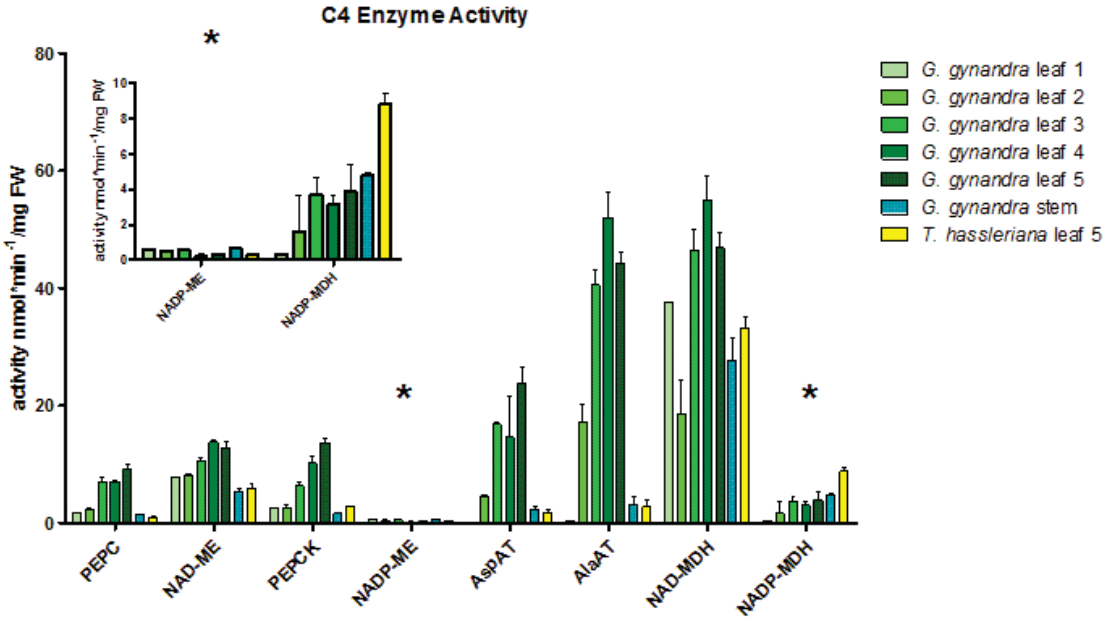
Supplemental Figure 12. Comparison of gene expression dynamics within the leaf gradient of both species.
(A-F) Average expression pattern of highest abundant putative ortholog of C_4 cycle genes (*NHD, PPDK, PPT, AlaAT, BASS2, PEPC*) in photo- and heterotrophic tissues in *G. gynandra* (light grey) and *T. hassleriana* (dark grey); (n=3 ± SE, standard error)



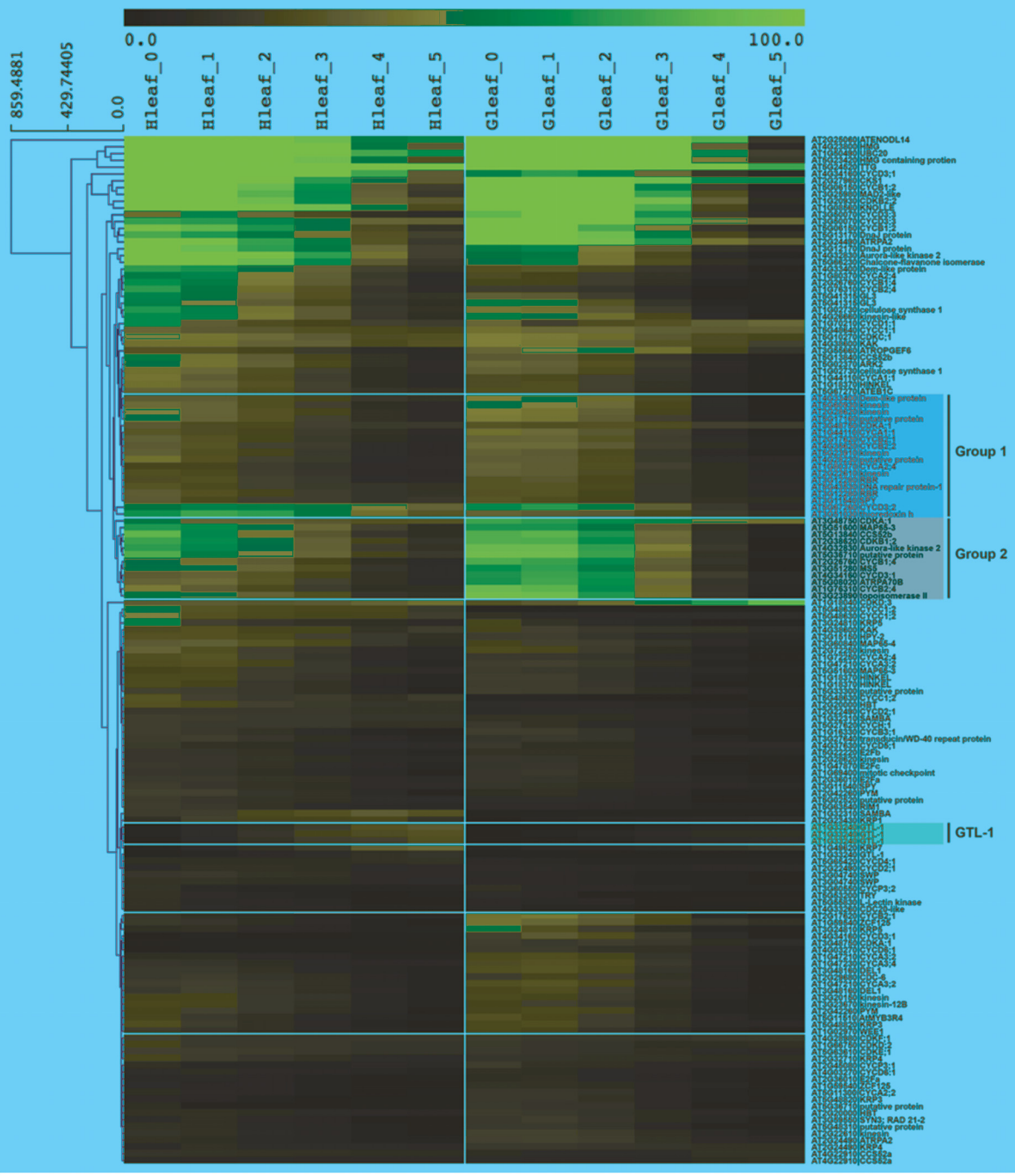
Supplemental Figure online 13. Plot of all C_4 gene putative orthologs expression pattern (RPKM) in *G. gynandra,* that were annotated as C_4 genes with AGI identifier and respective *T. hassleriana* ID. **(A-F)** Average expression pattern of putative ortholog of C_4 cycle genes (*DIC, BASS2, AspAT, NAD-ME, PPT, PEPC*) in photo- and heterotrophic tissues in *G. gynandra* (n=3).



Supplemental Figure online 14. Plot of all C_4 gene putative orthologs expression pattern (RPKM) in *T. hassleriana,* that were annotated as C_4 genes with AGI identifier and respective *T. hassleriana* ID. **(A-F)** Average expression pattern of putative ortholog of C_4 cycle genes (*DIC, BASS2, AspAT, NAD-ME, PPT, PEPC*) in photo- and heterotrophic tissues in *T. hassleriana* (n=3).

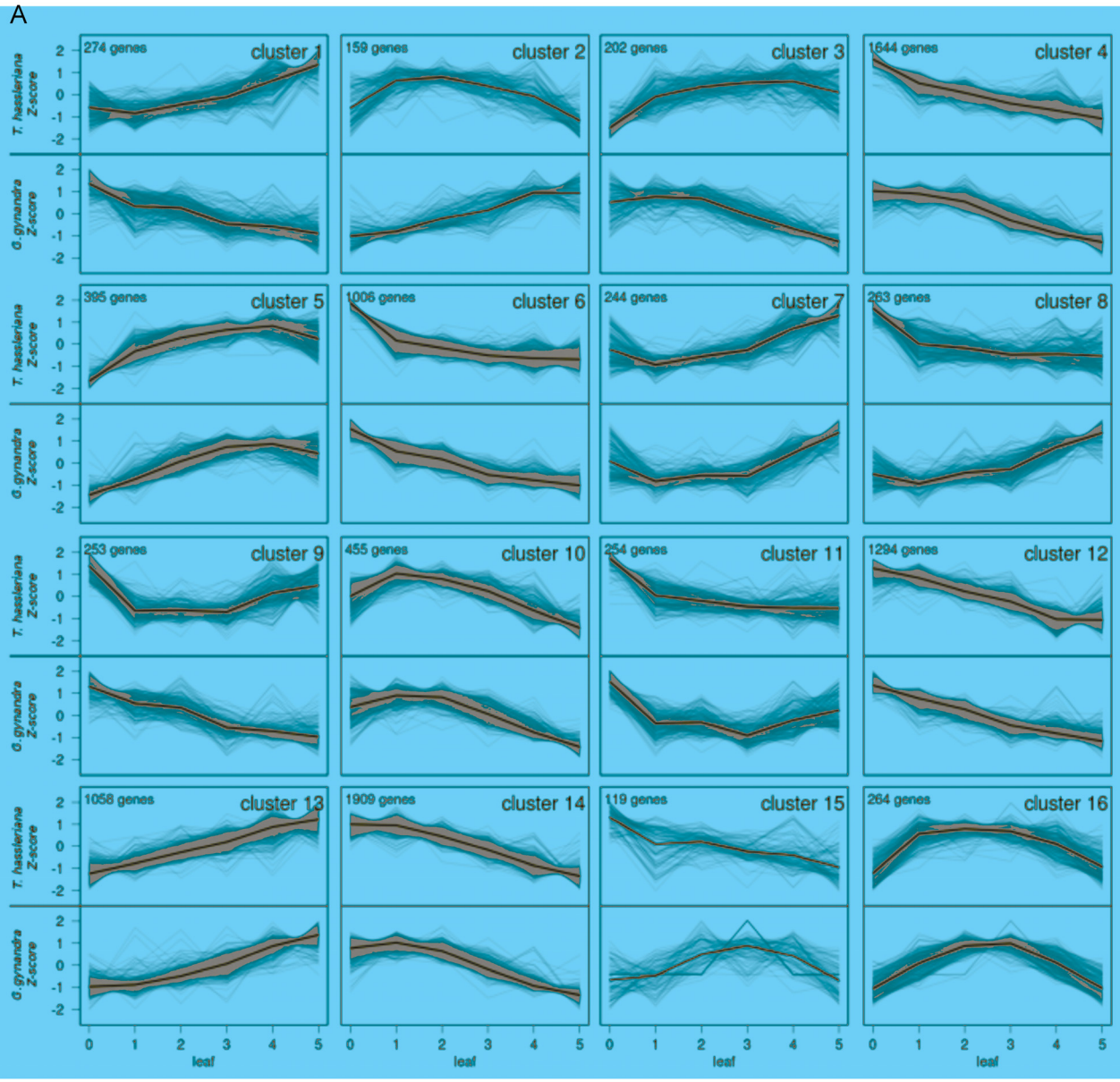


Supplemental Figure 15. Enzyme activity measurement of soluble C_4 cycle enzymes. Enzyme activities of PEPC, NAD-ME, PEPCK, NADP-ME, AspAT, AlaAT, NAD-MDH and NADP-MDH were measured along the developing *G. gynandra* leaf (stage 1-5) with the mature *T. hassleriana* leaf (stage 5) as C_3 control. (FW: fresh weight; n=3 ±SE, standard error; biological replicates with each 3 technical replicates)

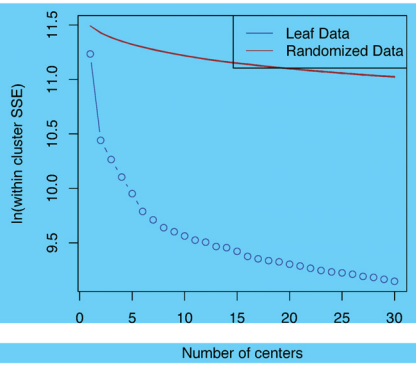


Supplemental Figure 16. Hierarchical clustering of average RPKM with Euclidean distance of core cell cycle genes in *T. hassleriana* and *G. gynandra*. Core cell cycle genes were extracted from (Vandepoele et al., 2002; Beemster et al., 2005). Deregulated cluster of interest are marked with blue and red boxes. *GTL1* cluster is highlighted with green box.

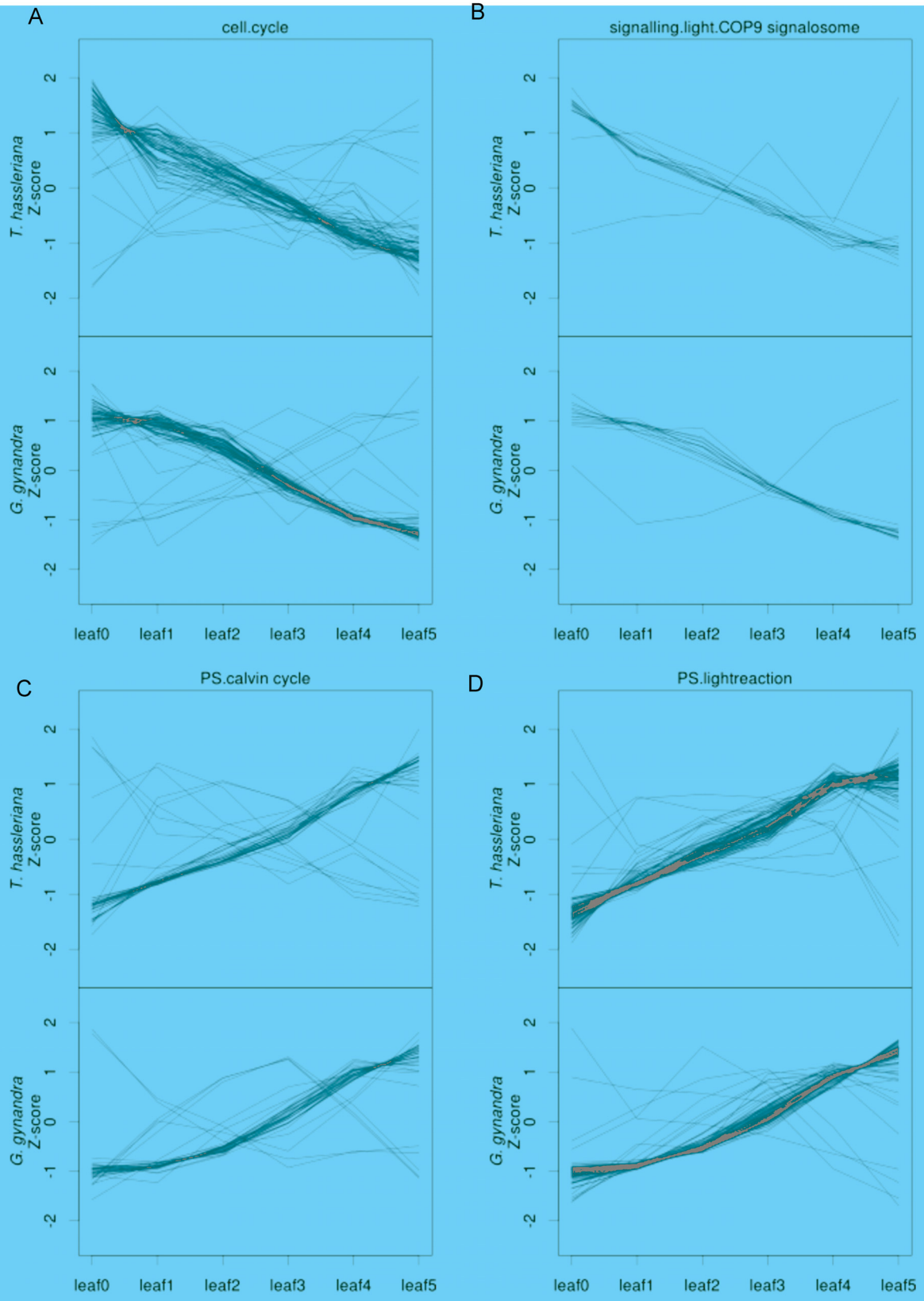
Supplemental Figure 17. Hierarchical clustering with Pearson's correlation of leaf developmental factors. Averaged transcript abundances (RPKM) of leaf gradient sample of transcriptional regulators involved in axial and vasculature fate determination were clustered. Group 1 (orange) and group 2 (red) show genes that are altered between *T. hassleriana* (H) and *G. gynandra* (G).



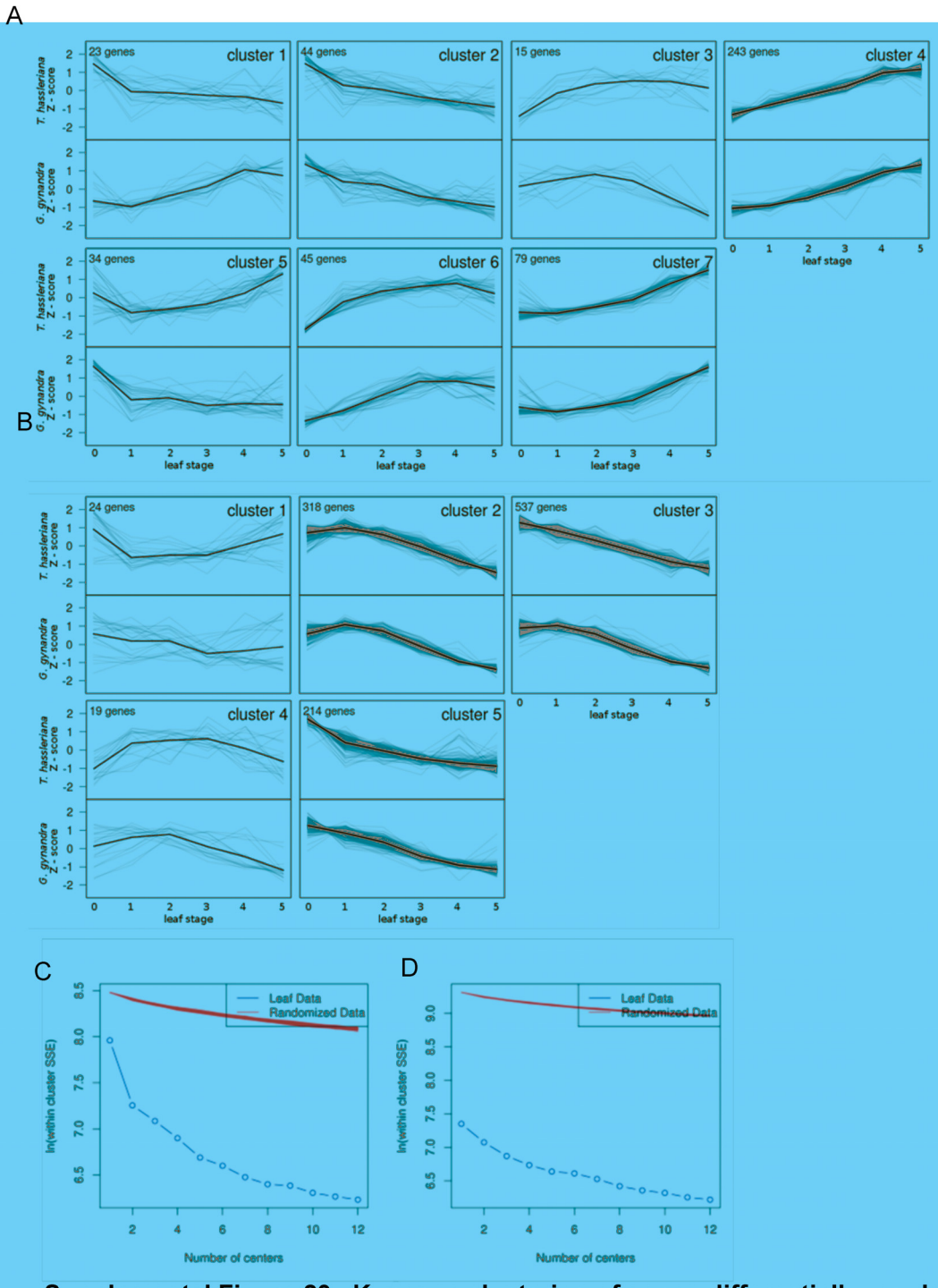
B



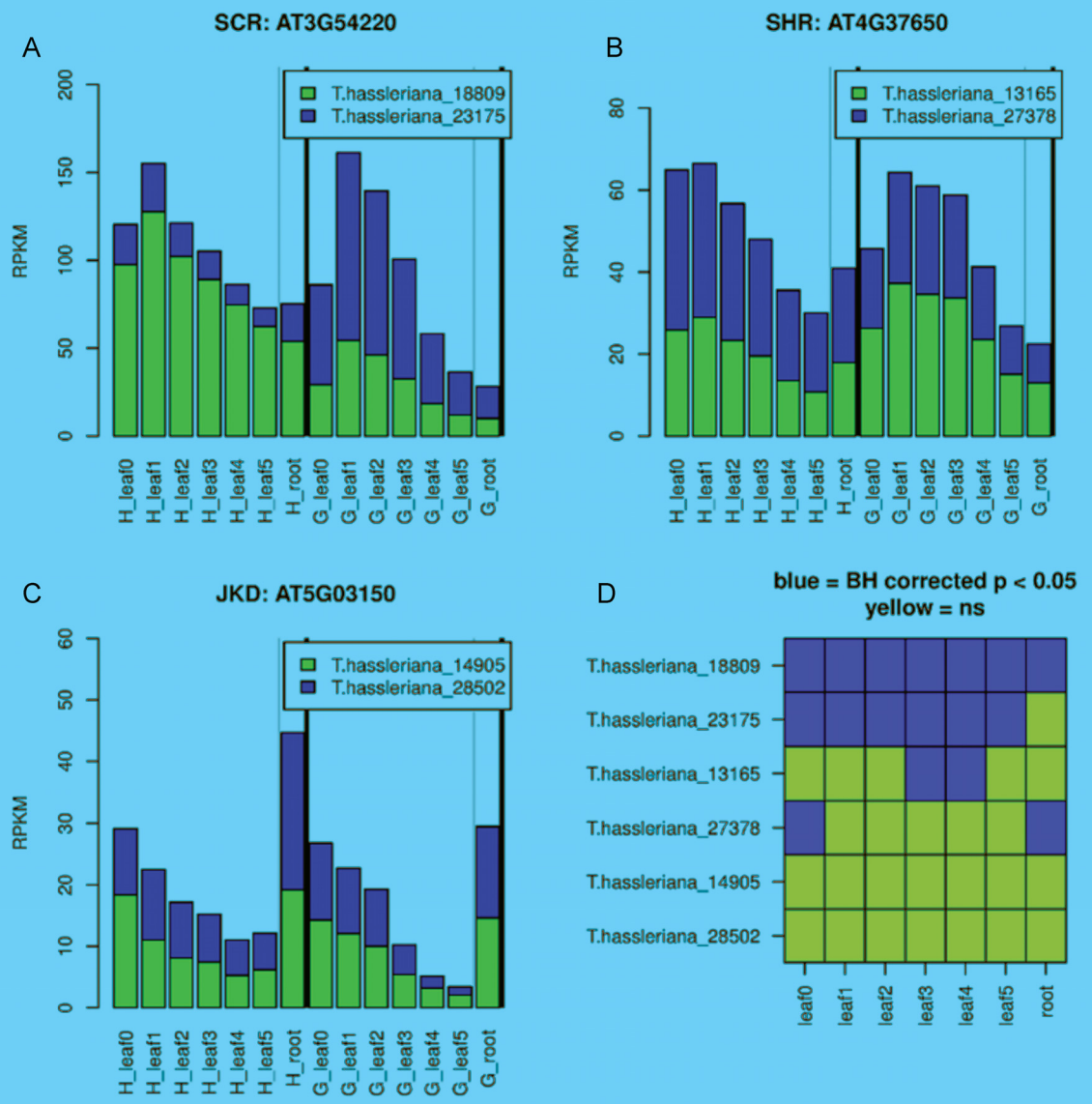
Supplemental Figure 18. *K*-means clustering of leaf gradient expression data and quality assessment. **(A)** *K*-means clustering of transcript abundances (RPKM) of leaf stage averages (*n*=3) between *T. hassleriana* and *G. gynandra* shown as species-scaled Z-scores. Size of each cluster is indicated in each cluster box. **(B)** Ln of the sum of the squared euclidean distance (SSE) between each gene and the center of it's cluster across various numbers of clusters calculated with a *K*-means algorithm for the leaf gradient data (blue) compared to the average of 250 scrambled datasets (red).



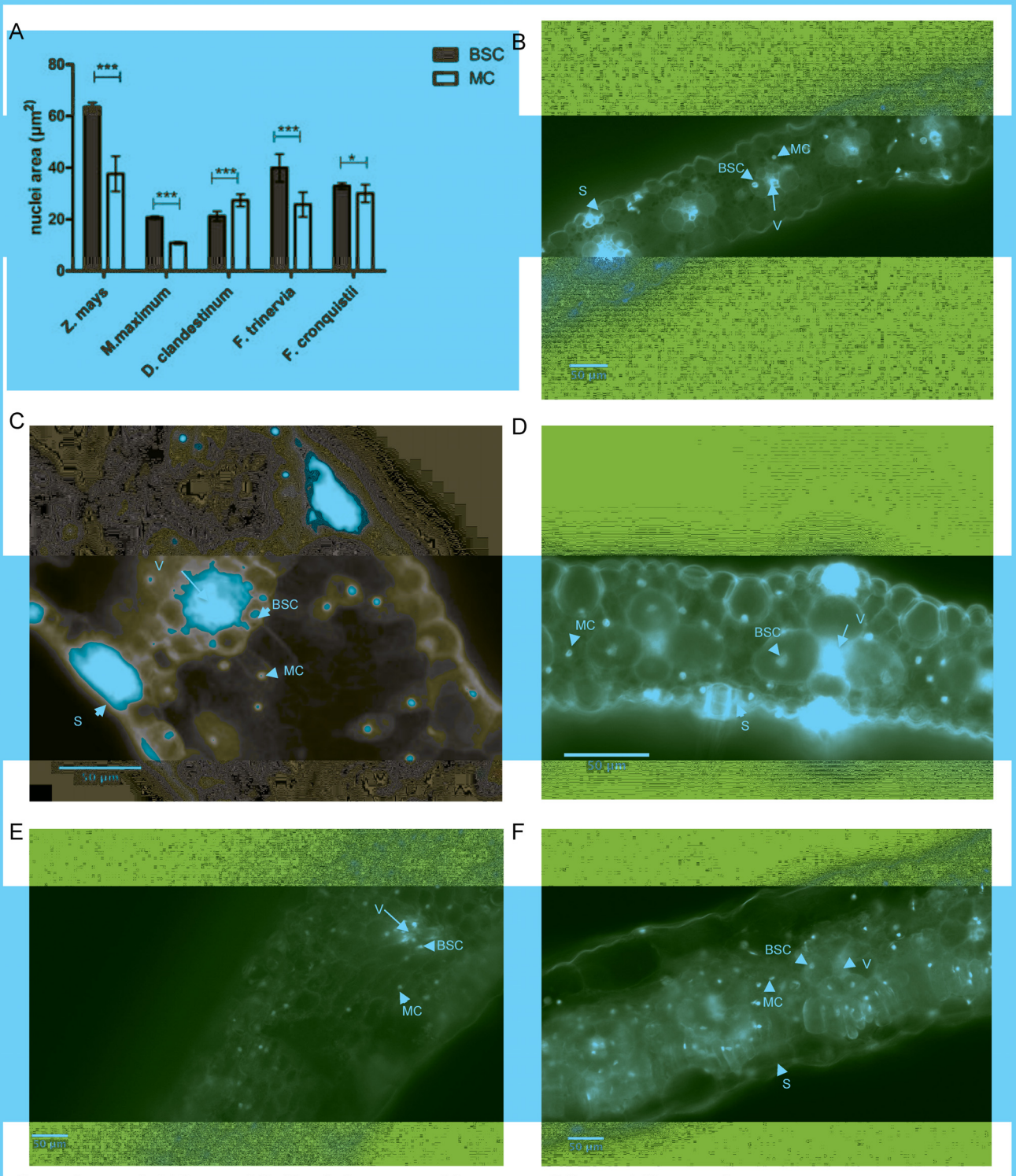
Supplemental Figure 19. Z-score plots of enriched mapman categories in the shifted clusters. Species scaled Z-scores from averaged transcript abundances (RPKM) for each leaf stage per species (n=3). **(A,B)** shifted enriched categories from cluster 4. **(C,D)** shifted enriched categories from cluster 13. Number in brackets are the respective Mapman category bin codes.



Supplemental Figure 20. K-means clustering of genes differentially regulated during the transition from proliferation to enlargement. (A,B) K-means clustering of *T. hassleriana* and *G. gynandra* homologs of gene set that is significantly up-regulated **(A;** p-value<0.05**)** or down-regulated (**B**; p-value<0.05) between day 9 and 10 day in developing *A. thaliana* leaves (Andriankaja et al., 2012). Per species scaled Z-scores from averaged transcript abundances (RPKM) for each leaf stage per species (n=3). **(C,D)** Ln of the sum of the squared Euclidean distance (SSE) between each gene and the center of its clusters across various numbers of clusters calculated with a K-means algorithm for the leaf gradient data (blue) compared to the average of 250 scrambled datasets (red) for **(C)** up- and **(D)** down-regulated.



Supplemental Figure 21. Transcript abundances of *SCARECROW* and *SHORTROOT* homologs in *G. gynandra* (G) and *T. hassleriana* (H) leaf and root.
(A-C) Expression pattern (average RPKM; n=3) of all homologs of *SCARECROW* **(*SCR*; A)**; *SHORTROOT* **(*SHR*; B)** and *JACKDAW* **(*JKD*; C)** in both species. **(D)** Dual color map of significant (blue; FWE corrected p-Value<0.05) or non significant (yellow; n.s) expressed transcripts of *SCR*, *SHR* and *JKD*.



Supplemental Figure 22. Nuclei area and images of C_4 and C_3 species.
(A) Quantification of BSC and MC nuclei area of mature leaves of monocotyledonous (*Zea mays*; *Megathyrsus maximus*; *Dichantelium clandestinum*) and dicotyledonous (*Flaveria trinervia*; *Flaveria cronquistii*) C_4 and C_3 species cross sections (error bars ±SD; n=3). Area of nuclei is given as µm² with at least 100 nuclei analyzed per cell type per species. Asterisks indicate statistically significant differences between BSC and MC (*** p-value<0.001; * p-value<0.05). **(B-F)** Microscopic fluorescence images of propidium iodide stained mature leaf cross sections of *Zea mays*,C_4 (**B**); *Dichantelium clandestinum*;C_3 (**C**); *Megathyrsus maximus*, C_4 (**D**); *Flaveria cronquistii*, C_3 (**E**); *Flaveria trinervia,* C_4 (**F**). Scale bar: 50 µm; closed arrows pointing to nuclei of indicated cell type. BSC: bundle sheath cell; MC: mesophyll cell; V: vein; S: stomata.

Supplemental Table 1 online. Velvet/OASES assembly stats from *G. gynandra* and *T. hassleriana* paired end reads. Backmapping of paired end reads was performed with TopHat standard settings. Annotation via blastp against TAIR10 proteome.

	G. gynandra (C_4)	*T. hassleriana* (C_3)
k-mer	31	31
N50 contig	1916	1996
unigenes	59471	52479
total transcripts	176850	163456
Backmapping %	60	63
Annotation of TAIR10 %	86	87

Supplemental Table 2 online. Cross species mapping results. *T. hassleriana* Leaf 5, Seed 1, Stamen (n=3) was mapped to *A. thaliana* via blat in translated protein (A) mode to assess sensitvity of cross species mapping. Results of mapping were normalized as RPKM and collapsed on 1 AGI per multiple identifier in *T. hassleriana* Pearson's correlation *r* values of collapsed *T. hassleriana* Leaf 5, Seed 1 and Stamen (n=3) mapped to *A. thaliana* (B) and to itself were calculated (C).

A

Species	Sample	Total number of cleaned reads	Total number of mapped reads	Mapping efficiency against A.thaliana reference	Number of genes >20 RPKM	Number of genes >1000 RPKM
T. hassleriana	**Hleaf5_1**	41085063	23502678	57.20492141	5825	151
	Hleaf5_2	26393836	22289304	84.44889936	5675	122
	Hleaf5_3	67907227	43184738	63.59372913	5684	146
	Hstamen_1	46237107	27726175	59.96520284	5923	48
	Hstamen_2	48025041	28220020	58.76105343	5950	47
	Hstamen_3	17855771	14433105	80.83159781	5467	60
	Hseed1_1	38620315	21654259	56.06960741	6253	39
	Hseed1_2	28792149	17462026	60.64856777	6301	48
	Hseed1_3	25372947	14217549	56.03428329	6107	42

B

	collapsed expression by mapping to own cds vs to A. thaliana	1vs1	2vs2	3vs3	average
Hleaf5	r	0.90	0.89	0.91	0.90
	r2	0.81	0.80	0.82	0.81
Hstamen	r	0.79	0.79	0.79	0.79
	r2	0.62	0.62	0.62	0.62
Hseed1	r	0.91	0.86	0.9	0.89
	r2	0.83	0.74	0.81	0.79

C

	T. hassleriana mapped to *A. thaliana*	1vs2	1vs3	2vs3	average
Hleaf5	r	0.98	1.00	0.98	0.99
	r2	0.97	0.99	0.96	0.97
Hstamen	r	0.97	0.96	0.98	0.97
	r2	0.94	0.92	0.96	0.94
Hseed1	r	0.97	0.99	0.98	0.98
	r2	0.94	0.98	0.96	0.96

Supplemental Table 3 online. Pearson's correlation (r) of each individual replicate per tissue in G. gynandra and T. hassleriana respectively (A). Pearson's correlation between G. gynandra and T. hassleriana individual tissues (B).

A

#	Species	Tissue	Pearson correlation *r* between biological replicates 1 vs 2	1 vs 3	2 vs 3
1	*G. gynandra*	**Gleaf0**	0.98	0.99	0.99
2		**Gleaf1**	0.97	0.96	0.98
3		**Gleaf2**	0.95	0.92	0.98
4		**Gleaf3**	0.79	0.92	0.93
5		**Gleaf4**	0.81	0.97	1.00
6		**Gleaf5**	0.99	0.99	0.99
7		**Groot**	0.92	0.93	0.93
8		**Gstem**	0.97	0.94	0.95
9		**Gstamen**	0.61	0.61	0.97
10		**Gpetal**	0.88	0.84	0.84
11		**Gcarpel**	0.99	0.61	0.57
12		**Gsepal**	1.00	0.97	0.97
13		**Gseedling2**	0.99	0.98	0.99
14		**Gseedling4**	0.90	0.92	0.99
15		**Gseedling6**	0.70	0.99	0.75
16		**Gseed1**	0.99	0.99	1.00
17		**Gseed2**	1.00	1.00	1.00
18		**Gseed3**	0.77	0.64	0.94
19	*T. hassleriana*	**Hleaf0**	0.97	0.97	0.99
20		**Hleaf1**	0.97	0.98	0.98
21		**Hleaf2**	0.96	0.98	0.98
22		**Hleaf3**	0.96	0.99	0.98
23		**Hleaf4**	0.96	0.99	0.98
24		**Hleaf5**	0.97	0.99	0.98
25		**Hroot**	0.95	0.96	0.96
26		**Hstem**	0.23	0.62	0.87
27		**Hstamen**	0.94	0.91	0.98
28		**Hpetal**	0.98	0.97	0.97
29		**Hcarpel**	0.95	0.99	0.98
30		**Hsepal**	0.87	0.86	0.90
31		**Hseedling2**	0.99	0.99	0.98
32		**Hseedling4**	0.99	1.00	0.99
33		**Hseedling6**	0.82	0.82	0.98
34		**Hseed1**	0.99	1.00	0.99
35		**Hseed2**	1.00	1.00	1.00
36		**Hseed3**	0.93	0.96	0.95

Supplemental Table 3 online. Pearson's correlation (r) of each individual replicate per tissue in G. gynandra and T. hassleriana respectively (A). Pearson's correlation between G. gynandra and T. hassleriana individual tissues (B).

B

Pearson Correlation *r* between G. *gynandra* and *T. hassleriana*		
#	**Tissue**	**r**
1	**Leaf0**	0.723369664
2	**Leaf1**	0.693967315
3	**Leaf2**	0.774414647
4	**Leaf3**	0.718280077
5	**Leaf4**	0.845767325
6	**Leaf5**	0.801946455
7	**Root**	0.693418487
8	**Stem**	0.397920288
9	**Stamen**	0.465027959
10	**Petal**	0.296842384
11	**Carpel**	0.409336161
12	**Sepal**	0.216833607
13	**Seedling2**	0.864093832
14	**Seedling4**	0.79602302
15	**Seedling6**	0.757896499
16	**Seed1**	0.922002838
17	**Seed2**	0.882400443
18	**Seed3**	0.612106172

Supplemental Table 4 online. Number of significatly up- or downregulated genes in *G. gynandra* compared to *T. hassleriana* within the different tissues. Differential expressed gene p-Values were calculated via EdgeR and Bonferroni-Holms corrected, genes with p<0.05 were classified as differential regulated.

Tissue	UP p< 0.05	UP p< 0.01	UP p< 0.001	DOWN p< 0.05	DOWN p< 0.01	DOWN p< 0.001
leaf0	5435	5061	4539	6076	5696	5237
leaf1	5197	4841	4391	5914	5529	5026
leaf2	4234	3894	3443	5047	4644	4204
leaf3	4646	4283	3833	5484	5070	4576
leaf4	3250	2911	2511	3774	3399	2979
leaf5	3236	2894	2447	4133	3716	3191
root	4343	3973	3511	5151	4755	4254
stem	7835	7497	7123	8462	8129	7698
stamen	4545	4116	3652	5388	4976	4451
petal	4445	4063	3613	5122	4751	4317
carpel	3718	3352	2929	3640	3274	2894
sepal	5650	5276	4780	6422	6023	5539
seedling2	4012	3644	3186	4354	3981	3546
seedling4	4113	3684	3202	4416	4043	3569
seedling6	2874	2534	2180	3542	3154	2714
seed1	4116	3764	3321	4457	4083	3591
seed2	6600	6270	5807	7075	6727	6276
seed3	6108	5725	5307	7088	6674	6190
mean	4686.5	4321.222222	3876.388889	5308.055556	4923.555556	4458.444444
max	7835	7497	7123	8462	8129	7698

Supplemental Table 5 online. List of genes present in root to shoot recruitment module.

T. hassleriana cds ID (Cheng et al., 2013)	Arabidopsis homologue	Coexpressed with TF	TAIR short annotation
T.hassleriana_10164	AT1G70410		beta carbonic anhydrase 4
T.hassleriana_20805	AT2G22500		uncoupling protein 5
T.hassleriana_17885	**AT5G61590**	**ERF**	**Integrase-type DNA-binding superfamily protein**
T.hassleriana_27615	**AT1G04250**	**Aux/IAA**	**AUX/IAA transcriptional regulator family protein**
T.hassleriana_13599	**AT5G13180**	**VND-I2**	**NAC domain containing protein 83**
T.hassleriana_07159	AT4G12730	Aux/IAA	FASCICLIN-like arabinogalactan 2
T.hassleriana_22160	AT5G57560		Xyloglucan endotransglucosylase/hydrolase family protein
T.hassleriana_03276	AT1G11545	Aux/IAA	xyloglucan endotransglucosylase/hydrolase 8
T.hassleriana_11774	AT1G43670		Inositol monophosphatase family protein
T.hassleriana_19959	AT5G19140	ERF	Aluminium induced protein with YGL and LRDR motifs
T.hassleriana_13658	AT1G25230	ERF	Calcineurin-like metallo-phosphoesterase superfamily protein
T.hassleriana_11758	AT3G14690	VND-I2	cytochrome P450, family 72, subfamily A, polypeptide 15
T.hassleriana_00726	AT5G46900		Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily
T.hassleriana_13312	AT3G22120		cell wall-plasma membrane linker protein
T.hassleriana_18867	AT3G54110		plant uncoupling mitochondrial protein 1
T.hassleriana_22110	AT1G14870		PLANT CADMIUM RESISTANCE 2
T.hassleriana_13333	AT5G19190		
T.hassleriana_11698	AT3G13950		
T.hassleriana_01980	AT5G25265		
T.hassleriana_04483	AT5G62900		
T.hassleriana_21987	AT1G13700	ERF	6-phosphogluconolactonase 1
T.hassleriana_15837	AT1G05000		Phosphotyrosine protein phosphatases superfamily protein
T.hassleriana_08797	AT5G23750	Aux/IAA	Remorin family protein
T.hassleriana_08517	AT5G36160		Tyrosine transaminase family protein
T.hassleriana_12936	AT5G25980		glucoside glucohydrolase 2
T.hassleriana_04639	AT2G01660		plasmodesmata-located protein 6
T.hassleriana_22812	AT4G21870	ERF	HSP20-like chaperones superfamily protein
T.hassleriana_10363	AT3G11660	VND-I2	NDR1/HIN1-like 1
T.hassleriana_19882	AT3G04720		pathogenesis-related 4
T.hassleriana_27070	AT2G15220		Plant basic secretory protein (BSP) family protein
T.hassleriana_05312	AT2G37170		plasma membrane intrinsic protein 2
T.hassleriana_05313	AT2G37170		plasma membrane intrinsic protein 2
T.hassleriana_12285	AT2G36830	Aux/IAA	gamma tonoplast intrinsic protein
T.hassleriana_12284	AT2G36830		gamma tonoplast intrinsic protein
T.hassleriana_14369	AT1G11670	Aux/IAA	MATE efflux family protein
T.hassleriana_08980	N.A.		
T.hassleriana_07000	N.A.		

Supplemental Table online 6. List of clustered general leaf developmental and vasculature regulating genes along both leaf gradients.

T. hassleriana **cds ID (Cheng et al., 2013)**	**AGI**	**Annotation based on TAIR10**	**Function in vascular development**
T.hassleriana_16883	AT1G19850	MONOPTEROS (MP)	leaf initiation
T.hassleriana_08823	AT1G19850	MONOPTEROS (MP)	leaf initiation
T.hassleriana_08424	AT1G32240	KANADI 2 (KAN2)	leaf axis formation
T.hassleriana_09176	AT1G32240	KANADI 2 (KAN2)	leaf axis formation
T.hassleriana_20498	AT1G52150	ATHB-15	neg reg of vasc cell diff
T.hassleriana_09793	AT1G52150	ATHB-15	neg reg of vasc cell diff
T.hassleriana_06450	AT1G65620	ASYMMETRIC LEAVES 2 (AS2)	leaf initiation
T.hassleriana_19648	AT1G73590	PIN-FORMED 1 (PIN1)	vein initiation (polar auxin transport)
T.hassleriana_01843	AT1G79430	ALTERED PHLOEM DEVELOPMENT (APL)	vascular cell identity repressed by REV
T.hassleriana_19440	AT1G79430	ALTERED PHLOEM DEVELOPMENT (APL)	vascular cell identity repressed by REV
T.hassleriana_27016	AT2G13820	Bifunctional inhibitor/lipid-transfer protein	vein formation (xylogen)
T.hassleriana_27989	AT2G27230	LONESOME HIGHWAY (LHW)	transcription factor-related
T.hassleriana_09087	AT2G27230	LONESOME HIGHWAY (LHW)	transcription factor-related
T.hassleriana_15265	AT2G27230	LONESOME HIGHWAY (LHW)	transcription factor-related
T.hassleriana_15152	AT2G28510	Dof-type zinc finger DNA-binding family protein	Dof-type zinc finger DNA-binding family protein
T.hassleriana_27908	AT2G28510	Dof-type zinc finger DNA-binding family protein	Dof-type zinc finger DNA-binding family protein
T.hassleriana_06822	AT2G33860	ETTIN (ETT)	leaf axis formation abaxial fate
T.hassleriana_23279	AT2G33860	ETTIN (ETT)	leaf axis formation abaxial fate
T.hassleriana_23086	AT2G37630	ASYMMETRIC LEAVES 1 (AS1)	leaf initiation
T.hassleriana_18733	AT4G08150	KNOTTED-like from Arabidopsis thaliana (KNAT1)	leaf initiation
T.hassleriana_09854	AT4G08150	KNOTTED-like from Arabidopsis thaliana (KNAT1)	leaf initiation
T.hassleriana_25576	AT4G24060	Dof-type zinc finger DNA-binding family protein	Dof-type zinc finger DNA-binding family protein
T.hassleriana_22410	AT4G32880	homeobox gene 8 (HB-8)	vein initiation (post auxin marker of vascular patterning)
T.hassleriana_28697	AT5G16560	KANADI (KAN)	leaf axis formation abaxial; neg reg of PIN1
T.hassleriana_19776	AT5G16560	KANADI (KAN)	leaf axis formation abaxial; neg reg of PIN1
T.hassleriana_18288	AT5G60200	TARGET OF MONOPTEROS 6 (TMO6)	TARGET OF MONOPTEROS 6
T.hassleriana_16642	AT5G60200	TARGET OF MONOPTEROS 6 (TMO6)	TARGET OF MONOPTEROS 6
T.hassleriana_18265	AT5G60690	REVOLUTA (REV)	adaxial leaf axis formation
T.hassleriana_19132	AT5G60690	REVOLUTA (REV)	adaxial leaf axis formation
T.hassleriana_17767	AT5G64080	XYP1	vein formation (xylogen)
T.hassleriana_26861	AT5G64080	XYP1	vein formation (xylogen)

Supplemental Methods

Leaf clearings and safranine staining (Supplemental Figure 1)

For leaf clearings *T. hassleriana* and *G. gynandra* leaves of stage 0 to 5 were destained in 70% EtOH with 1% glycerol added for 24 hrs and cleared in 5% NaOH until they appeared translucent and rinsed with H_2O_{dest}. Leaves were imaged under dark field settings with stereo microcope SMZ1500 (Nikon, Japan). Prior safranine staining, leaves were destained with increasing EtOH series until 100% EtOH and stained for 5 -10 min with 1% safranine (1g per 100ml 96% EtOH). After destaining leaves were analyzed with bright field microscope (Zeiss, Germany). Vein orders were determined by width and position as described by (McKown and Dengler, 2009) for Flaveria species.

Contig assembly and annotation (Supplemental Figure 4, Table 1 and Dataset 3)

Cleaned and filtered paired end (PE) reads were used to create a reference transcriptome for each species. The initial *de novo* assembly was optimized by using 31-kmer using Velvet (v1.2.07) and Oases (v0.2.08) pipeline (Zerbino and Birney, 2008; Schulz et al., 2012). For quality purposes the longest assembled transcript was selected with custom made perl scripts if multiple contigs were present (Schliesky et al., 2012) resulting in 59,471 *G. gynandra* and 52,479 *T. hassleriana* contigs. For quality assessment PE reads were aligned again to the respective contigs for each species via TopHat standard settings with over 60% backmapping efficiency in both species. Assembled longest transcripts were annotated using BLASTX mapping against TAIR10 proteome database (cut-off $1e^{-10}$). The best blastx hits were filtered by the highest bitscore. For quality assessment of contigs, *T. hassleriana* contigs were aligned with BLASTN against *T. hassleriana* predicted cds (Cheng et al., 2013). Multiple matching contigs to one cds identifier were filtered with customized perl script.

Cross species mapping sensitivity assessment (Supplemental Figure 5; Table 2)

All three biological replicates of leaf stage 5, stamen and young seed from *T. hassleriana* were mapped with BLAT V35 in dnax mode (nucleotide sequence of query and reference are translated in six frames to protein) with default parameters to both, the *T. hassleriana* gene models and the *A. thaliana* TAIR10 representative gene models. Subsequently, the BLAT output was filtered for the best match per read based on the highest score. RPKMs were calculated based on mappable reads per million (RPKM). The RPKM expression data was collapsed to single *A. thaliana* AGIs (RPKM were added) to avoid multiple assigned *T. hassleriana*‘s IDs to the same AGI. Pearson’s correlation r was calculated between the mapped *T. hassleriana* replicates mapped on *A. thaliana* gene models among each other. Also Pearson’s correlation r was calculated between cross species mapped *T. hassleriana* leaf5, stamen and seed1 replicates and the replicates of Leaf5 mapped to its own cds in *T. hassleriana*.

Principal component analysis (Supplemental Figure 8)

Principal component analyses (PCA,Yeung and Ruzzo, 2001) was carried out with MULTI EXPERIMENT VIEWER VERSION 4 (MEV4, (Saeed et al., 2003; Saeed et al.,

2006) on gene row SD normalized averaged RPKMs with median centering.

Enzyme Assays (Supplemental Figure 15)

From *G. gynandra* leaf stage 2 to 5, enzymatic activities of known C_4 enzymes were determinedas summarized by Ashton et al. (1990) in three biological replicates.

Comparison of Cleomaceae leaf gradients to *A. thaliana* leaf differentiation (Supplemental Figure 19)

Examination of Cleomaceae expression patterns of genes differentially regulated during the transition from cell proliferation to expansion in *A. thaliana*.
Andriankaja et al. (2012) observed that the transition between cell proliferation and expansion occurred between days 9 and 10. They defined two sets of genes significantly differentially expressed between day 9 and 10, one up-regulated and one down-regulated. The expression of the *T. hassleriana* and *G. gynandra* homologues of these genes were analyzed. The sum of standard error (SSE) was taken as a quality control to determine an appropriate number of clusters. The number of cluster centers chosen was 7 and 5 for up-regulated and down-regulated genes, respectively. The *K*-means clustering was performed the same as before, except that genes were not previously filtered by expression level and genes were only binned once into clusters.

Supplemental References

Andriankaja, M., Dhondt, S., De Bodt, S., Vanhaeren, H., Coppens, F., De Milde, L., Muehlenbock, P., Skirycz, A., Gonzalez, N., Beemster, G.T.S., and Inze, D. (2012). Exit from Proliferation during Leaf Development in Arabidopsis thaliana: A Not-So-Gradual Process. Dev. Cell 22, 64-78.

Ashton A.R., Burnell J.N., Furbank R.T., Jenkins C.L.D., Hatch M.D. (1990). The enzymes in C4 photosynthesis. In Enzymes of Primary Metabolism. Methods in Plant Biochemistry, P.M. Dey and J.B. Harborne, eds (London: Academic Press), pp. 39–72

Cheng, S., van den Bergh, E., Zeng, P., Zhong, X., Xu, J., Liu, X., Hofberger, J., de Bruijn, S., Bhide, A.S., Kuelahoglu, C., Bian, C., Chen, J., Fan, G., Kaufmann, K., Hall, J.C., Becker, A., Braeutigam, A., Weber, A.P.M., Shi, C., Zheng, Z., Li, W., Lv, M., Tao, Y., Wang, J., Zou, H., Quan, Z., Hibberd, J.M., Zhang, G., Zhu, X.-G., Xu, X., and Schranz, M.E. (2013). The T arenaya hassleriana Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers. Plant Cell **25,** 2813-2830.

McKown, A.D., and Dengler, N.G. (2009). Shifts in leaf vein density through accelerated vein formation in C-4 Flaveria (Asteraceae). Annals of Botany **104,** 1085-

1098.

Saeed, A.I., Hagabati, N.K., Braisted, J.C., Liang, W., Sharov, V., Howe, E.A., Li, J., Thiagarajan, M., White, J.A., and Quackenbush, J. (2006). TM4 microarray software suite. In DNA Microarrays, Part B: Databases and Statistics, A. Kimmel and B. Oluver, eds, pp. 134.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J. (2003). TM4: A free, open-source system for microarray data management and analysis. Biotechniques **34,** 374.

Schliesky, S., Gowik, U., Weber, A.P.M., and Brautigam, A. (2012). RNA-Seq Assembly - Are We There Yet? Front Plant Sci **3,** 220-220.

Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA- seq assembly across the dynamic range of expression levels. Bioinformatics **28,** 1086- 1092.

Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C_3 and C_4 Plant Species

Canan Külahoglu, Alisandra K. Denton, Manuel Sommer, Janina Maß, Simon Schliesky, Thomas J. Wrobel, Barbara Berckmans, Elsa Gongora-Castillo, C. Robin Buell, Rüdiger Simon, Lieven De Veylder, Andrea Bräutigam and Andreas P.M. Weber

Plant Cell 2014;26;3243-3260; originally published online August 8, 2014;
DOI 10.1105/tpc.114.123752

Chapter 4

Co-Author Manuscripts

4.1 Manuscript H: Predicting C_4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape

Overview

Title: Predicting C_4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape
Authors: David Heckmann, Stefanie Schulze, Alisandra Denton, Udo Gowik, Peter Westhoff, Andreas P.M. Weber, Martin J. Lercher
Published in Cell, June 2013
Impact factor: 33.116
Co-authorship

Main Findings

This manuscript examines the fitness landscape of the evolution from C_3 to C_4 photosynthesis. The model starts in a anatomically preconditioned C_3 state with environmental conditions appropriate for C_4 evolution. As step wise biochemical changes towards C_4 photosynthesis accumulate there is a constant increase in carbon fixation, a proxy for fitness. Importantly, a fully integrated C_4 photosynthetic trait is evolutionarily accessible from all modeled pointsthere are no reductions in fitness or local maxima. However, in both the evolutionary best path and the Monte Carlo simulations, there is a modular path from C_3 to C_4 with the initial establisment of the photorespiratory pump, followed by the integration of the C_4 cycle including increasing PEPC activity and movement of Rubisco to the BS, fine tuning of the C_4 cycle kinetics, reduction in BS conductance, and

changes in the kinetics of Rubisco. This path was validated by comparison to literature and wet lab measurements for C_3, intermediate, and C_4 species.

Contributions

- Assistance in design and execution of wetlab analyses
- PEPC enzymatic activity assays of Flaveria species
- Edited full manuscript

Cell

Theory

Predicting C_4 Photosynthesis Evolution: Modular, Individually Adaptive Steps on a Mount Fuji Fitness Landscape

David Heckmann,[1] Stefanie Schulze,[2] Alisandra Denton,[3] Udo Gowik,[2] Peter Westhoff,[2,4] Andreas P.M. Weber,[3,4] and Martin J. Lercher[1,4,*]
[1]Institute for Computer Science
[2]Institute for Plant Molecular and Developmental Biology
[3]Institute for Plant Biochemistry
Heinrich Heine University, 40225 Düsseldorf, Germany
[4]Cluster of Excellence on Plant Sciences (CEPLAS)
*Correspondence: lercher@cs.uni-duesseldorf.de
http://dx.doi.org/10.1016/j.cell.2013.04.058

SUMMARY

An ultimate goal of evolutionary biology is the prediction and experimental verification of adaptive trajectories on macroevolutionary timescales. This aim has rarely been achieved for complex biological systems, as models usually lack clear correlates of organismal fitness. Here, we simulate the fitness landscape connecting two carbon fixation systems: C_3 photosynthesis, used by most plant species, and the C_4 system, which is more efficient at ambient CO_2 levels and elevated temperatures and which repeatedly evolved from C_3. Despite extensive sign epistasis, C_4 photosynthesis is evolutionarily accessible through individually adaptive steps from any intermediate state. Simulations show that biochemical subtraits evolve in modules; the order and constitution of modules confirm and extend previous hypotheses based on species comparisons. Plant-species-designated C_3-C_4 intermediates lie on predicted evolutionary trajectories, indicating that they indeed represent transitory states. Contrary to expectations, we find no slowdown of adaptation and no diminishing fitness gains along evolutionary trajectories.

INTRODUCTION

To predict the evolution of biological systems, it is necessary to embed a systems-level model for the calculation of fitness into an evolutionary framework (Papp et al., 2011). However, explicit theories to predict strong correlates of fitness exist for very few complex model systems (Papp et al., 2011; Stern and Orgogozo, 2008). A major example is the stoichiometric metabolic network models of microbial species, which have been used to predict bacterial adaptation to nutrient conditions in laboratory experiments (Fong and Palsson, 2004; Hindré et al., 2012; Ibarra et al., 2002). On a macroevolutionary timescale, related methods have been applied to predict the outcome and temporal order of reductive genome evolution in endosymbiotic bacteria (Pál et al., 2006; Yizhak et al., 2011). These studies on microbial evolution have employed metabolic yield of biomass production as a correlate of fitness, an approach that cannot be transferred directly to multicellular organisms.

However, it is likely that the efficiency with which limiting resources are converted into biomass precursors is under strong selection across all domains of life. For multicellular eukaryotes, this trait may be most easily studied in plants, which use energy provided by solar radiation to build sugars from water and CO_2. To fix carbon from CO_2, plants use the enzyme RuBisCO (ribulose-1,5-bisphosphate carboxylase/oxygenase). RuBisCO has a biologically relevant affinity for O_2, resulting in a toxic product that must be recycled in the energy-consuming metabolic repair pathway known as photorespiration (Maurino and Peterhansel, 2010). The decarboxylation of glycine—a key metabolite within this pathway—by the glycine decarboxylase complex (GDC) releases CO_2. About 30 million years ago, photorespiration increased to critical levels in many terrestrial ecosystems due to the depletion of atmospheric CO_2. To circumvent this problem, C_4 photosynthesis evolved to concentrate CO_2 around RuBisCO in specific cell types (Edwards et al., 2010; Sage et al., 2012).

CO_2 first enters mesophyll (M) cells, where most RuBisCO is located in C_3 plants. In contrast, C_4 plants have shifted RuBisCO to neighboring bundle sheath (BS) cells. In the M of C_4 plants, PEPC (phospho*enol*pyruvate carboxylase, which does not react with oxygen) catalyzes the primary fixation of CO_2 as bicarbonate. The resulting C_4 acids enter the BS and are decarboxylated, releasing CO_2 in proximity to RuBisCO. BS cells are surrounded by thick cell walls, believed to reduce CO_2 leakage (Kiirats et al., 2002). Such an energy-dependent biochemical CO_2-concentrating pump is the defining feature of C_4 plants; species differ in the decarboxylating enzyme employed and in the metabolites shuttled between cell types (Drincovich et al., 2011; Furbank, 2011; Pick et al., 2011).

Despite the complexity of C_4 photosynthesis, this trait constitutes a striking example of convergent evolution: it has evolved

Cell

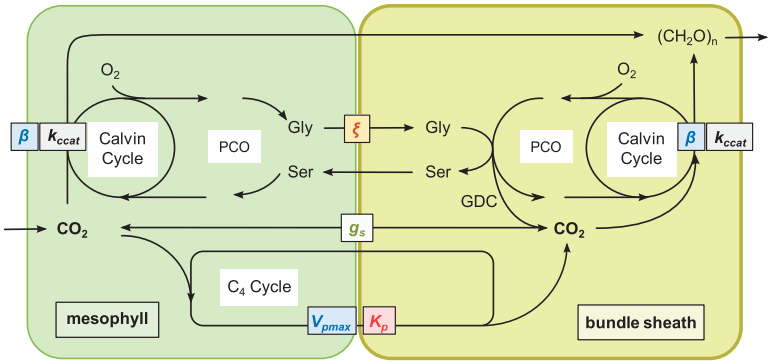


Figure 1. Overview of C_3-C_4 Biochemistry, Modeled as Two Interacting Cell Types

CO_2 enters the M and is either fixed by RuBisCO in the M or shuttled to the BS through the C_4 cycle and fixed by RuBisCO there. The resulting C_3 acids are fed into the Calvin cycle. Deleterious fixation of O_2 by RuBisCO leads to photorespiration (PCO). Model parameters are β, the fraction of RuBisCO active sites in the M; k_{ccat}, the maximal turnover rate of RuBisCO; ξ, the fraction of M derived glycine decarboxylated by GDC in the BS (note that for $\xi < 1$, decarboxylation of glycine also takes place in the M); V_{pmax}, the activity of the C_4 cycle; K_p, the Michaelis-Menten constant of PEPC for bicarbonate; and g_s, the BS conductance for gases. See also Figure S2 and Table S2.

independently in more than 60 angiosperm lineages from the ancestral C_3 photosynthesis (Sage et al., 2011). The leaf anatomy typical for C_4 plants—close vein spacing and prominent BS cells, designated "Kranz" anatomy—is also adaptive for C_3 species in environments associated with C_4 evolution (Brodribb et al., 2010). A rudimentary Kranz anatomy was thus likely already present in the C_3 ancestors of C_4 species (Sage et al., 2012), forming a "potentiating" anatomical state (Christin et al., 2011, 2013). Furthermore, all enzymes required for C_4 photosynthesis have orthologs in C_3 species, where they perform unrelated functions. In the evolution of C_4 biochemistry, these enzymes required concerted changes in their cell-type-specific gene expression as well as adjustment of their kinetic properties (Aubry et al., 2011; Gowik and Westhoff, 2011; Sage, 2004).

Some plant species have biochemistry that is intermediate between C_3 and C_4 (Edwards and Ku, 1987). These species possess a rudimentary Kranz anatomy and divide RuBisCO between M and BS cells. Often, however, photorespiratory glycine decarboxylation by GDC is largely shifted to the BS (see Figure 1), resulting in a moderate increase in the CO_2 concentration in BS cells (Sage et al., 2012).

C_4 plants make up 3% of today's vascular plant species but account for ~25% of terrestrial photosynthesis (Edwards et al., 2010; Sage et al., 2012). How C_4 photosynthesis evolved and why it evolved with such repeatability, are two fundamental questions in plant biology (Sage et al., 2012). Low atmospheric CO_2/O_2 ratio, heat, aridity, and high light are discussed as important factors promoting C_4 evolution, explaining the abundance of C_4 plants in tropical and subtropical environments (Edwards et al., 2010; Ehleringer et al., 1991). However, C_4 metabolism also allows higher biomass production rates in temperate regions (Beale and Long, 1995). The resulting accelerated growth makes engineering of the C_4 trait into major crops a promising route toward meeting the growing demands on food production (Hibberd et al., 2008). Rational strategies to approach this challenge require a detailed understanding of not only the C_4 state but also the fitness landscape connecting it with the ancestral C_3 biochemistry.

Here, we map the biochemical fitness landscape on which evolution from C_3 to C_4 photosynthesis occurs. Inserting the fitness estimates into a population genetic framework, we then explore the probability distribution of evolutionary trajectories leading from C_3 to C_4 systems. We thereby predict biochemical evolution in a multicellular eukaryote on macroevolutionary timescales (Hindré et al., 2012; Papp et al., 2011). Our results show that C_4 evolution is repeatable and predictable in its details. Importantly, experimentally determined parameter sets for C_3-C_4 intermediates fall well within the clustered distribution of predicted evolutionary trajectories. This agreement not only validates the model but also further provides important insights into the evolutionary nature of these species as transitory states in the evolution toward full C_4 photosynthesis.

RESULTS

A Biochemical Model for C_3-C_4 Evolution

RuBisCO is the most abundant protein on earth, responsible for up to 30% of nitrogen investment and 50% of total protein investment in plants (Ellis, 1979). C_4 plants typically contain lower amounts of RuBisCO per leaf area than C_3 plants (Ghannoum et al., 2011), explaining their lower nitrogen requirements (Brown, 1978). Reduced RuBisCO production is facilitated by higher CO_2 assimilation per RuBisCO protein, allowing C_4 plants to channel protein investment into other processes. In addition, C_4 plants do not need to open their stomata as much as C_3 plants to ensure sufficient internal CO_2 partial pressure, and they thus lose less water in hot and arid environments (Ghannoum et al., 2011). We assume that the overall fitness gain associated with C_4 photosynthesis is proportional to the amount of CO_2 that can be fixed using a given quantity of RuBisCO per leaf area (A_c).

To predict the steady-state enzyme-limited net CO_2 assimilation rate, A_c, from phenotypic parameters, we modified a mechanistic biochemical model developed by von Caemmerer (2000) to describe C_3-C_4 intermediates (Figure 1 and Experimental Procedures; see also Peisker, 1986). The underlying von Caemmerer model is itself based on models describing gas exchange in C_3 and in C_4 plants (Berry and Farquhar, 1978; Farquhar et al., 1980; von Caemmerer, 1989, 2000); these models have been used and validated in a variety of contexts (Yin and Struik, 2009). An extensive discussion of the model's generality and the choice of parameters can be found in the von Caemmerer book (2000).

C_3 and C_4 metabolisms represent limiting cases of the model, and representative parameter ranges were derived from C_3 and

Cell

C_4 species (Experimental Procedures). Evolution is modeled via changes in the following parameters: β, the fraction of RuBisCO active sites in the M, which ranges from ~95% in C_3 to 0% in some C_4 plants (where all RuBisCO is shifted to the BS); k_{ccat}, the maximal turnover rate of RuBisCO, which is lower in C_3 plants due to a trade-off with CO_2 specificity (Savir et al., 2010); ξ, the fraction of glycine derived from unwanted fixation of O_2 in M cells that is decarboxylated by GDC in the BS, ranging from 0 in C_3 to 1 in many C_3-C_4 intermediates (i.e., activity of the photorespiratory CO_2 pump); V_{pmax}, quantifying the activity of the C_4 cycle (i.e., the PEPC-dependent CO_2 pump); K_p, the Michaelis-Menten constant of PEPC (the core protein of the C_4 cycle) for bicarbonate; and g_s, the BS gas conductance (which quantifies the combined effects of cell geometry and cell wall properties).

Other kinetic parameters for RuBisCO were shown to be strongly linked to k_{ccat} (Savir et al., 2010) and are modeled accordingly (Extended Experimental Procedures and Figure S1 available online). The model describes the core steps of carbon fixation in communicating M and BS cells (Figure 1). CO_2 and O_2 enter M cells, with diffusion into and out of BS cells (g_s). CO_2 can be fixed in both cell types at rates characterized by the allocation (β) and kinetics (k_{ccat}) of RuBisCO. Alternatively, CO_2 may initially be fixed into a C_4 acid through the action of the C_4 cycle in M cells, characterized by the activity (V_{pmax}) and the kinetics (K_p) of its rate-limiting enzyme, PEPC. The C_4 acids then diffuse into the BS cells, where they are decarboxylated to free CO_2. We assume PEPC to be rate limiting (von Caemmerer, 2000), and thus neither this part of the C_4 cycle nor the recycling of the CO_2 carrier to the M is modeled explicitly. Finally, due to downregulation of GDC in the M, a fraction of the glycine resulting from the fixation of O_2 in the M is decarboxylated by GDC in BS cells (ξ).

The C_3 ancestors of C_4 species likely possessed a potentiating anatomy, characterized by decreased vein spacing and increased BS size (Christin et al., 2011, 2013). These anatomical features enable efficient diffusion of photorespiratory and C_4 cycle metabolites between compartments. C_3 plants that are closely related to C_4 species were further shown to exhibit a specific localization of chloroplasts and mitochondria in the BS cells. This "proto-Kranz" anatomy (Muhaidat et al., 2011) may be necessary for the establishment of a photorespiratory CO_2 pump by allowing the loss of GDC activity in the M to be compensated by the BS (Sage et al., 2012). Accordingly, our model starts from a C_3 state with proto-Kranz anatomy. This morphology can evolve further toward full C_4 Kranz anatomy (McKown and Dengler, 2007) via two main processes: (1) a reduction in the relative number of M cells and (2) an increase of BS cell size. Both processes influence our model exclusively by changing the proportion of RuBisCO allocated to BS cells instead of M cells (i.e., by decreasing β).

All parameters were normalized to total leaf area. At environmental conditions relevant for the evolution of C_4 photosynthesis and the constant RuBisCO concentration assumed in the model, C_3 and C_4 parameterizations lead to A_c values of 15.5 and 83.8 μmol m^{-2} s^{-1}, respectively. These hypothetical A_c values are assumed to reflect fitness gains during C_4 evolution, even if these fitness gains are in fact partially realized by the channeling of resources from RuBisCO production into other processes.

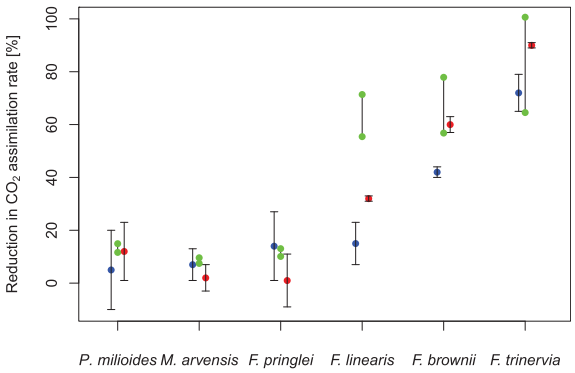


Figure 2. The Model Predicts the Reduction in Carbon Fixation Rate when the C_4 Cycle Is Reduced by Inhibiting PEPC

Blue and red dots show A_c reduction at 1 mM and 4 mM DCDP, respectively, with error bars indicating SD (Brown et al., 1991). Green dots show the range of predicted A_c reduction at 80%–100% inhibition of the C_4 cycle. See Extended Experimental Procedures for details.

C_4 species have been categorized into three subtypes, depending on the predominant decarboxylating enzyme (NAD malic enzyme, NAD-ME; NADP malic enzyme, NADP-ME; or phospho*enol*pyruvate carboxykinase, PEPCK) (Hatch et al., 1975). Our model is compatible with the stoichiometry of all three of these pathways under excess light. This agrees with experimental observations, which show that fitness-relevant traits are independent of C_4 subtype (Ehleringer and Pearcy, 1983; Ghannoum et al., 2001).

One major reason for the generality of our modeling approach is that carbon fixation is largely decoupled from other parts of plant metabolism. When light and nitrogen are available in excess, we thus expect that biomass production is strictly proportional to the carbon fixation rate, A_c. To confirm this, we coupled our C_3/C_4 model to a full plant metabolic network (Dal'Molin et al., 2010). The full model can be modified to reflect the different subtypes of C_4 metabolism (NAD-ME, NADP-ME, PEPCK). We sampled the parameter space of our C_3/C_4 model, using the predicted metabolite fluxes to constrain flux-balance analyses (FBA) of the full model (Oberhardt et al., 2009). For each of the three C_4 subtypes, we demonstrated that biomass production is indeed directly proportional to A_c (Figure S2; Pearson's $R^2 > 0.999$). These results support the robustness of our model to differences in the metabolism of different plant lineages.

As long as RuBisCO is active in both M and BS ($0 < \beta < 1$), our model predicts that CO_2 assimilation increases with decreasing M GDC expression (i.e., decreasing ξ). This prediction is consistent with experimental data from crosses between C_3-C_4 intermediate *Moricandia* and C_3 *Brassica* (Hylton et al., 1988). Furthermore, the model predicts the quantitative influence of experimentally suppressed C_4 cycles in phylogenetically diverse C_3-C_4 intermediates and C_4 plants (Brown et al., 1991) (Figure 2). A discrepancy between model and experiments is observed only for *F. linearis*. In this species, PEPC activity appears to be a suboptimal predictor for C_4 cycle activity, likely because of insufficient activity of PPDK (pyruvate, Pi dikinase) (Ku et al., 1983).

Cell

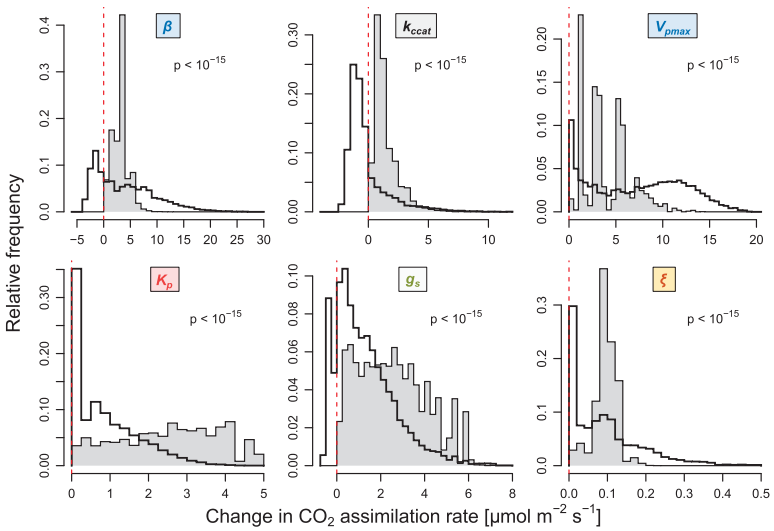


Figure 3. Realized Fitness Gains Are More Narrowly Distributed Than Potential Fitness Gains

White bars show potential fitness gains when one parameter is changed towards the C4 value. Gray bars show fitness gains realized in the evolutionary simulations. Negative values (to the left of the dashed red lines) indicate fitness reductions. Fitness is approximated by CO_2 assimilation rate. Although potential fitness gains vary widely, realized fitness gains are comparable between parameters. The distributions of potential and of realized fitness gains are significantly different ($p < 10^{-15}$ for each parameter, median tests). See also Figure S4.

Changes of the model parameters are ultimately caused by DNA mutations of protein coding or regulatory regions, and hence occur in discrete steps. Although each model parameter is known to show genetic variation, we currently lack a detailed understanding of the genotype-phenotype relationships. We thus divided each parameter range into six equidistant phenotypic states, with C_3 and C_4 states as endpoints. Choosing different discretizations did not change the observed patterns (Figure S3), except for ξ (see Discussion).

Despite Extensive Epistasis, the C_4 State Is Accessible from Every Point in the Fitness Landscape

The phenotypic parameters that distinguish C_3 from C_4 metabolism span a six-dimensional fitness landscape. Due to functional dependencies between the parameters, this landscape shows strong epistasis: fitness effects of changes in one parameter vary widely depending on the values of other parameters (Figure 3). Parameters differ in their potential influence on fitness. Whereas any individual increase in ξ raises A_c by at most 0.5 μmol m^{-2} s^{-1} (and never decreases fitness), a single increase in β can boost A_c by as much as 27 μmol m^{-2} s^{-1} or diminish A_c by as much as 3.7 μmol m^{-2} s^{-1}.

For half of the parameters (β, k_{ccat}, g_s), the same parameter change toward C_4 can both increase and decrease fitness, depending on the background provided by the remaining parameter values. This type of interaction has been termed sign epistasis (Weinreich et al., 2005) and affects 5.5% of the discretized fitness landscape (25,145 out of 486,000 pairwise combinations of parameter changes). Sign epistasis can be further classified as reciprocal if changing either of two parameters modifies fitness in one direction, while subsequently adding the second change modifies fitness in the opposite direction (Poelwijk et al., 2011). Reciprocal sign epistasis is a necessary (though not sufficient) condition for the existence of multiple fitness maxima (Poelwijk et al., 2011). The discrete C_3/C_4 fitness landscape contains only 20 points with reciprocal sign epistasis. All 20 involve an interaction between β and k_{ccat} at intermediate activity of the C_4 cycle (V_{pmax}). At these points, changes toward C_4 of β or k_{ccat} individually increase fitness. However, the C_4 cycle is not sufficiently active to compensate for the associated reduction in M photosynthetic efficiency when both parameters change simultaneously.

Maximal fitness is achieved when all parameters reach their C_4 values. Despite strong and often sign-changing epistasis, there is always at least one parameter change (median four changes) toward the C_4 state that increases fitness (Figure S4). Thus, the global fitness optimum is evolutionary accessible (Weinreich et al., 2005) from every position in the landscape. It immediately follows that there are no local maxima, giving the biochemical fitness landscape an exceedingly simple, smooth, "Mount (Mt.) Fuji-like" structure.

Modular Evolution of a Complex Trait

To evolve from C_3 to C_4 metabolism, our model requires 30 individual mutational changes (five steps in each of the six parameters). Parameters change with unequal probabilities. For example, the mutational target for inactivation of M GDC (increasing ξ) is large (Sage, 2004). Active GDC is a multienzyme system consisting of four distinct subunits, and downregulation of any of these will result in reduced GDC activity (Engel et al., 2007). Furthermore, M expression of each subunit is likely regulated by several transcription factor binding sites, each with several nucleotides important for binding. Random mutations at any of these sites are likely to downregulate M GDC expression. This inactivation is sufficient to establish a photorespiratory CO_2 pump, as we assume a low diffusional distance between M and BS cells, as well as a specific subcellular distribution of organelles in the BS (proto-Kranz anatomy). Due to this photorespiratory pump, any RuBisCO present in the BS will operate under increased CO_2 pressure, thereby increasing organismal fitness. Conversely, reduced GDC activity in BS cells would lead to decreased CO_2 pressure in the BS and hence would reduce organismal fitness. Thus, while random mutations may be equally likely to diminish GDC activity in M and in BS cells, only reductions in M activity are likely to be fixed in a population.

Cell

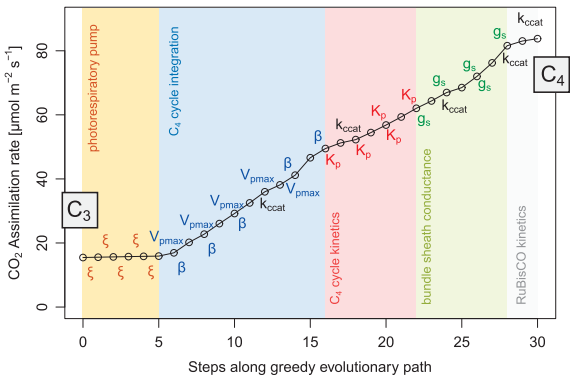


Figure 4. Fitness Changes along the "Greedy" Path through the Fitness Landscape from C_3 to C_4

This trajectory always chooses the most likely parameter change, combining mutation and fixation probabilities. The label centered above or below each edge indicates the mutation connecting two states. Evolution along the greedy path is modular (colored areas), except for the RuBisCO turnover rate k_{ccat}. CO_2 assimilation rate is used as a proxy for fitness. See also Figures S3 and S5.

In contrast to the large mutational target for the reduction of M GDC expression, other parameter changes involve increases in tissue-specific gene expression or changes in enzyme kinetics, which require specific mutations, restricted to only a few potential target nucleotides. Specifically, mutations that increase C_4 cycle activity appear much less likely, as different enzymes need to be upregulated in BS and in M cells, respectively. In the absence of precise estimates, we used plausible relative mutational probabilities for the model parameters (Extended Experimental Procedures). The general evolutionary patterns were found to be robust over a wide range of mutational probabilities and discretizations (Figure S3B).

Once a mutation that changes a model parameter occurs, its probability of fixation in the evolving plant population is determined by the associated change in fitness. Our simulations assume a "strong selection, weak mutation" regime, such that beneficial mutations are fixed in the population before the next mutation occurs (Gillespie, 1983). We estimated the fixation probability using a population genetic model first derived by Kimura (1957), assuming a constant population size of 100,000 individuals.

Each sequence of evolutionary changes linking the C_3 to the C_4 state defines an adaptive trajectory (or path) through the biochemical fitness landscape. The probability of individual steps is estimated as a combination of mutation and fixation probabilities. Figure 4 shows fitness changes associated with a unique "greedy" path, which always realizes the most likely parameter change. Here, changes for all but one of the six parameters are strictly clustered in modules (Figure 4). First, photorespiration is shifted to the BS ($\xi \uparrow$). Next, the C_4 cycle is established ($V_{pmax} \uparrow$), while RuBisCO is simultaneously shifted to the BS ($\beta \downarrow$). Then, the Michaelis-Menten constant of PEPC is adjusted ($K_p \downarrow$). Finally, gas diffusion is reduced ($g_s \downarrow$) in order to avoid leakage of CO_2 from the BS. The only parameter whose changes are not modular in this scenario is the maximal turnover rate of RuBisCO ($k_{ccat} \uparrow$), which is continuously adjusted along the greedy evolutionary trajectory, reflecting a shifting optimum due to the different CO_2 concentrations in M and BS.

Evolution is not deterministic, and the greedy path shown in Figure 4 represents only one of more than 10^{19} possible sequences of changes from C_3 to C_4. To more realistically characterize the evolution of C_4 biochemistry, we thus performed Monte Carlo simulations. At each step, we chose one parameter at random, weighted by the relative mutational probabilities. Using the biochemical model (Figure 1), we calculated the fitness change associated with adjusting the chosen parameter one step toward C_4. The change was accepted with a corresponding probability, derived from the population genetics model.

Despite the strong influence of chance, our Monte Carlo simulations support the same qualitative succession of modular changes in C_4 evolution (Figures S3A and S5). As observed in the greedy path, k_{ccat} is the only parameter that is continuously adjusted along the evolutionary trajectory, whereas ξ, V_{pmax} combined with β, K_p, and g_s tend to cluster with themselves ($p < 10^{-15}$ for dispersion higher than random of k_{ccat} and for modularity of ξ, V_{pmax} combined with β, K_p, and g_s; median tests for the distance between changes in the same parameter compared to random model).

Changes Early and Late in Adaptation Lead to Similar Fitness Increases

Strikingly, the greedy path through the fitness landscape (Figure 4) shows an almost linear fitness increase toward the C_4 state, with each evolutionary step resulting in a similar fitness increase. The only exceptions are the early establishment of a photorespiratory pump (ξ), the initial establishment of the C_4 cycle (V_{pmax}), and the two last adjustments of k_{ccat}. Thus, *realized* fitness gains along the greedy evolutionary path are very similar among the different parameters. This finding is in stark contrast to the broad distribution of *potential* fitness changes across the landscape (Figure 3).

Again, the stochastic evolutionary simulations support the result for the greedy path. Figure 3 shows that the distributions of realized fitness changes are much narrower than those of possible fitness changes. Furthermore, the median of realized fitness gains is similar across parameters, and lies around 2 μmol m^{-2} s^{-1} for all parameters except ξ. Accordingly, the time needed until the next parameter change is fixed in the population remains similar along evolutionary trajectories (Figure S6).

Repeatability of Evolution

The observed modularity and the narrow distributions of realized fitness gains demonstrate that the order of evolutionary changes toward C_4 is not arbitrary. Thus, evolution of this biochemical system is expected to repeat itself qualitatively in different species. Simulated evolutionary trajectories indeed cluster narrowly around a "mean path" ($p < 10^{-15}$; Figures 5 and S7).

Experimental Data from C_3-C_4 Intermediates Validate the Model

Our model of C_4 evolution is based on a number of simplifying assumptions and uses rough estimates of relative mutational

Cell

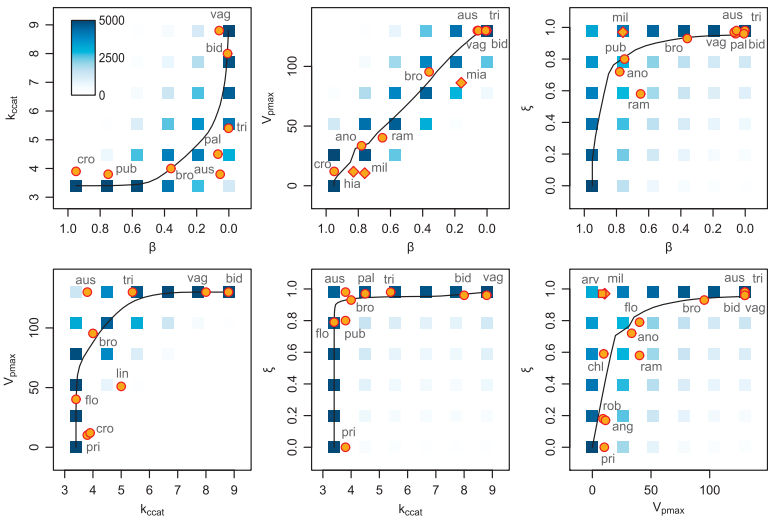


Figure 5. Projections of Trajectories through the Six-Dimensional Fitness Landscape Predicted by the Combined Biochemical and Stochastic Populations Genetics Model

Density of blue dots is proportional to the number of times a given parameter combination was crossed by a simulated trajectory. Black lines show the mean path of the set of trajectories. Orange dots are the *Flaveria* data described in the text, except for V_{pmax}, which was capped at 130 μmol m^{-2} s^{-1}. Abbreviations of species names: ang, *F. angustifolia*; ano, *F. anomala*; aus, *F. australasica*; bid, *F. bidentis*; bro, *F. brownii*; chl, *F. chloraefolia*; cro, *F. cronquistii*; flo, *F. floridana*; lin, *F. linearis*; pal, *F. palmeri*; pri, *F. pringlei*; pub, *F. pubescens*; ram, *F. ramosissima*; rob, *F. robusta*; tri, *F. trinervia*; vag, *F. vaginata*. Diamonds correspond to *Panicum* species: mil, *P. milioides*; hia, *P. hians*; mia, *P. miliaceum*. The square corresponds to *Moricandia arvensis*. See also Figures S6 and S7.

probabilities and population size. To assess its ability to quantitatively describe the evolution of real plants, we compared the model predictions to experimental data from the genera *Flaveria*, *Moricandia*, and *Panicum.* The experimental parameter sets for four plants and one plant correspond to the C_3 and C_4 endpoints, respectively. In addition, our data set included 15 species that have measured biochemical parameters intermediate between C_3 and C_4 (Figure 5); some of these species were previously classified as either C_3 or C_4 based on other criteria (McKown et al., 2005). Each of the intermediate species constitutes a separate point on evolutionary trajectories that started at C_3 biochemistry.

We collected experimental estimates of the biochemical model parameters for each of the 20 species from the literature, and we extended this data set by experimentally determining V_{pmax} and ξ for several *Flaveria* species (Experimental Procedures). With few exceptions, the experimentally determined parameter sets indeed lie very close to the predicted mean path through the fitness landscape (Figure 5). The model predicts experimental parameter combinations much better than a null model assuming a random order of evolutionary changes (Figure 6; $p < 10^{-15}$, median test).

DISCUSSION

The evolution of C_4 photosynthesis represents a rare opportunity to predict the functional evolution of a complex system: a closed six-parameter model calculates a phenotypic variable (A_c) of high relevance to fitness. The comparison to experimental data from diverse C_3-C_4 intermediates confirms the model's ability to quantitatively predict biochemical evolution over a timescale of several million years (Sage et al., 2012). While the majority of the data describe the genus *Flaveria*, the model also correctly predicts data from two phylogenetically distant genera (Figure 5). Comparisons to additional C_3-C_4 intermediates are currently limited by the availability of species-specific protocols for the separation of BS and M cells.

A hypothesis for the evolutionary succession of biochemical and morphological changes in the evolution of the C_4 syndrome was previously derived from phylogenetically informed analyses of C_3-C_4 intermediates (Sage et al., 2012). This hypothesis assumes modular biochemical changes, starting with a shift of photorespiration to the BS, followed by the establishment of a C_4 cycle in conjunction with a shift of RuBisCO to the BS, and finally an optimization stage in which parameters are fine-tuned. Our simulations support this scenario, narrowing it further by indicating that upregulation of the C_4 cycle usually precedes a shift of RuBisCO to the BS (Figure S3) even after previous establishment of a photorespiratory pump.

As expected due to the stochastic nature of evolution, the simulations indicate that modules are not strict and that the order of events may vary between independently evolving species. In particular, we find that the initial establishment of a photorespiratory pump (or C_2 cycle) is typical of evolutionary trajectories toward C_4 photosynthesis but may not be mandatory, as suggested previously (Sage et al., 2012).

Model Assumptions

While our model tracks changes in a phenotypic biochemical space, evolution is ultimately based on genomic mutations. We used qualitative reasoning when choosing relative mutational probabilities and the distribution of discrete steps linking C_3 and C_4 states. The sensitivity analysis (Figure S3B) demonstrates that other parameterizations lead to qualitatively very similar results. The only exception is the early establishment of a photorespiratory pump (ξ), which occurs with high probability only when the large mutational target for deactivation of the M GDC is taken into account.

The full C_4 cycle requires expression shifts in at least four separate enzymes. At each point in evolution, one of the enzymes that constitute the C_4 cycle will be rate limiting, making it the next target for fitness-enhancing upregulation. Distinct implementations of the C_4 cycle were shown to overlap in a

Cell

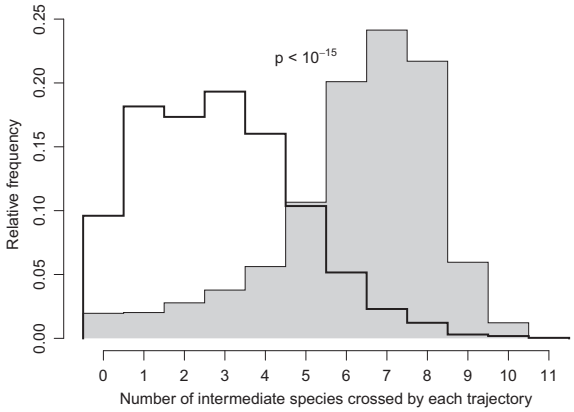


Figure 6. Distribution of the Number of Different C_3-C_4 Intermediate Species Whose Experimental Parameter Combinations Are Crossed by Each Single Predicted Trajectory

The combined biochemical and population genetics model (gray) fits the experimental data much better than a random model that ignores fitness effects (white) ($p < 10^{-15}$, median test). The parameter sets for *F. robusta*, *F. pringlei*, *F. cronquistii*, *F. angustifolia*, *and F. vaginata* are located at the C_3 or C_4 endpoints and hence crossed by every trajectory; they were excluded from this analysis.

single species (Furbank, 2011; Pick et al., 2011), potentially increasing the size of the mutational target. Our model uses the central enzyme PEPC to represent the complete pathway, accounting for the complexity of the C_4 cycle by using a low relative mutational probability.

A Simple, Mt. Fuji-like Biochemical Fitness Landscape

We found that the biochemical fitness landscape is exceedingly smooth: there are no local maxima besides the C_4 endpoint, as there is always at least one parameter change toward the C_4 value that increases the CO_2 fixation rate.

Comparison to experimental data from C_3-C_4 intermediate species indicates that our model indeed captures their evolutionary dynamics. The single-peaked fitness landscape suggests that these species are transitory states rather than evolutionary dead ends, continuously evolving toward the full C_4 syndrome as long as selective environmental conditions persist. The origin of *Flaveria* C_4 traits in the past 5 million years, together with the unusually large number of C_3-C_4 intermediate species in this genus (Sage et al., 2012), is consistent with this notion.

Half of the parameters in our model exhibit sign epistasis (Figure 3). Certain evolutionary trajectories thus involve reductions in fitness and are deemed not accessible (Weinreich et al., 2005); their inaccessibility contributes to the clustering of evolutionary trajectories. The paucity of reciprocal sign epistasis provides a partial explanation for the smooth landscape structure (Poelwijk et al., 2011).

Fitness landscapes resulting from interactions of mutations within the same gene can be rough and multipeaked (Weinreich et al., 2006). However, experimental fitness landscapes spanned by independently encoded functional units are similar in structure to the biochemical fitness landscape observed here: interactions among alleles of different genes rarely exhibit sign epistasis and often lead to simple, single-peaked landscapes (Chou et al., 2011; Khan et al., 2011; but see Kvitek and Sherlock, 2011).

Evolutionary Trajectories

Due to extensive sign epistasis among mutations within the same coding sequence, it was concluded that protein evolution may be largely reproducible and even predictable (Lozovsky et al., 2009; Weinreich et al., 2006). Despite the relatively low incidence of sign epistasis, we find that the same is true for the evolution of a complex biochemical system. Thus, different plants that independently "replay the tape of evolution" toward C_4 photosynthesis tend to follow similar trajectories of phenotypic changes (Figure 5). This resembles the high level of phenotypic and often genotypic parallelism in microbial evolution observed in experiments (Hindré et al., 2012) and predicted based on stoichiometric metabolic modeling (Fong and Palsson, 2004; Ibarra et al., 2002).

To explain the polyphyly of the C_4 syndrome, it has been hypothesized that each evolutionary step comes with a fitness gain (Gowik and Westhoff, 2011; Sage, 2004). We found that reality may be even more extreme: the fitness gain achieved by each individual change remained comparable along evolutionary trajectories (Figure 4). Accordingly, realized fitness advantages were much more similar across parameters than expected for random trajectories (Figure 3). This differs markedly both from theoretical expectations (Fisher, 1930; Orr, 2005) and from experimental observations in some genetic landscapes (Chou et al., 2011; Khan et al., 2011), which find diminishing fitness increases and a slowdown of adaptation along adaptive trajectories.

In the case of C_4 evolution, late-changing parameters (C_4 cycle kinetics, BS conductance) benefit from an already optimized background provided by previous evolution. Because everything else required for C_4 photosynthesis is already in place, their potential to contribute favorably to fitness is increased. Accordingly, we find no clear pattern of decelerated evolution along simulated trajectories, except for the last steps in PEPC kinetics and for late-occurring fixations of the now-superfluous photorespiratory pump (Figure S6). Conversely, the first few steps in C_4 evolution (initial establishment of CO_2 pumps) are only weakly selected, as only little RuBisCO is available in the BS at this time (Figure 4). Their fixation thus takes substantially longer than later changes (Figure S6): the first step is the most difficult one, also in C_4 evolution.

Why do C_3 plants still dominate many habitats, despite the simple, single-peaked fitness landscape and the substantial fitness gains resulting from individual evolutionary changes toward C_4 metabolism? A partial explanation is provided by weak selection on the first mutations. Furthermore, C_4 metabolism is strongly favored by selection only under specific environmental conditions, such as drought, high temperatures, and high light (excluding, for example, plants in woodlands). Finally, the potentiating Kranz-like anatomy in the C_3 ancestors of C_4 lineages (Christin et al., 2011, 2013; Sage et al., 2012) is not present in many other lineages, making the evolution of C_4 metabolism in these species unlikely.

Cell

The evolutionary dynamics uncovered above may shed light onto plans for experimental evolution of C_4 photosynthesis in C_3 plants through the application of increased selection pressure (Sage and Sage, 2007). Our results indicate that this endeavor may be accelerated by genetically engineering the first, slow steps of C_4 evolution. In particular, it may be advisable to pre-establish a photorespiratory CO_2 pump by knocking out M-specific GDC expression.

EXPERIMENTAL PROCEDURES

Biochemical Model and Fitness Landscape

The steady-state enzyme-limited net CO_2 assimilation rate (A_c) was used as a proxy for fitness of C_3, C_4, and intermediate evolutionary phenotypes. To predict A_c from phenotypic parameters, we slightly modified a mechanistic biochemical model for C_3-C_4 intermediates developed by von Caemmerer (2000) (Figure 1).

The CO_2 assimilation rates in the M and in the BS are calculated from the respective rates of carboxylation, oxygenation, and mitochondrial respiration (in addition to photorespiration). We assume constant concentrations of CO_2 (250 μbar) and O_2 (200 mbar) in M cells. Carboxylation and oxygenation are modeled as inhibitory Michaelis-Menten kinetics. RuBisCO kinetic parameters were shown to be subject to trade-offs (Savir et al., 2010); accordingly, we model these parameters as a function of RuBisCO maximal turnover rate (k_{ccat}). Activity of the C_4 cycle is assumed to be limited by PEPC activity and to follow Michaelis-Menten kinetics. The parameterization corresponds to a temperature of 25°C. The resulting set of equations can be solved for A_c in closed form. Equations, parameters, and further details are given in Extended Experimental Procedures and Tables S1 and S2.

For each evolving model parameter, we obtained representative C_3 and C_4 values (see below). The resulting range was subdivided into equidistant steps, leading to a discrete six-dimensional phenotype space. Based on the biochemical model, we calculated A_c for each parameter combination.

Calculation of Evolutionary Trajectories

We simulated a set of 5,000 evolutionary trajectories on the discrete fitness landscape, starting with the C_3 state. At each step, a trait (parameter) to be changed was chosen at random, with relative probabilities derived from current qualitative knowledge about the genetic complexity of the trait (Extended Experimental Procedures). We estimated selection coefficients (s) as the relative difference in A_c between ancestral and derived state, calculated using the biochemical model. We assumed a randomly mating population of diploid hermaphrodites, with incomplete dominance of mutations. The derived state was accepted with its probability of fixation, estimated using a formula first derived by Kimura (1957). We repeated the simulation process until reaching the C_4 parameter set.

To calculate a mean path from the set of 5,000 simulated trajectories, we averaged each parameter at each step (i.e., β at the first step of the mean path is the average of β values across the first steps of all simulated trajectories, etc.). Parameters were normalized to the interval [0,1]. Clustering of trajectories was quantified by calculating for each trajectory the mean of the normalized point-wise Manhattan distances to this mean trajectory. This measure is closely related to the recently introduced mean path divergence (Lobkovsky et al., 2011).

To estimate evolutionary modularity for each parameter, we used a distance measure defined as the number of other fixation events that occurred between two subsequent fixation events of the same parameter. V_{pmax} and β evolve together and were treated as a joint parameter in this context.

To determine a greedy trajectory through the landscape, we changed at each step the parameter that maximized the product of mutational probability and probability of fixation.

Comparison to Experimental Data

Data for the partitioning of RuBisCO between M and BS cells (β) and RuBisCO turnover rates (k_{ccat}), as well as PEPC activities (V_{pmax}) and decarboxylation of M-derived glycine in the BS (ξ) for *Moricandia* and *Panicum*, were obtained from the literature. We assayed PEPC activity in leaf extracts (summarized by Ashton et al., 1990) from 14 *Flaveria* species as a proxy for V_{pmax}. ξ was estimated for 14 *Flaveria* species by comparing the transcript levels of glycine decarboxylase P subunit genes that are expressed specifically in the BS (*gldpA*) to those expressed in all inner leaf tissues (*gldpD*). GldP transcript levels in leaves of 14 *Flaveria* species were determined by RNA sequencing. Data on K_p and g_s in intermediate species were not available. See Extended Experimental Procedures for more details on experimental data. We mapped experimental parameter values to the closest point in the discrete space of the model fitness landscape.

Random Null Model and Statistical Methods

To assess the statistical significance of our findings, we used a random null model to predict evolutionary trajectories. In this model, each trajectory starts with the C_3 state and evolves randomly, i.e., with equal probability for each directed parameter change, until the C_4 state is reached.

All simulations and statistical analyses were performed in the R environment (R Development Core Team, 2010). Statistical significance was assessed using Fisher's exact test and the median test implemented in the coin package (Hothorn et al., 2006).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.04.058.

ACKNOWLEDGMENTS

We thank Veronica Maurino, Itai Yanai, Eugene Koonin, Joachim Krug, and Andrea Bräutigam for helpful discussions. Computational support and infrastructure was provided by the "Center for Information and Media Technology" (ZIM) at the Heinrich Heine University Düsseldorf. This work was supported by the Deutsche Forschungsgemeinschaft (IRTG 1525 to D.H. and A.D.; FOR 1186 to S.S.; EXC 1028 to M.J.L., A.P.M.W., and P.W.; and CRC 680 to M.J.L.).

Received: December 18, 2012
Revised: March 21, 2013
Accepted: April 23, 2013
Published: June 20, 2013

REFERENCES

Ashton, A.R., Burnell, J.N., Furbank, R.T., Jenkins, C.L.D., and Hatch, M.D. (1990). The enzymes in C_4 photosynthesis. In Enzymes of Primary Metabolism, P.M. Dey and J.B. Harboene, eds. (London, UK: Academic Press), pp. 39–72.

Aubry, S., Brown, N.J., and Hibberd, J.M. (2011). The role of proteins in C_3 plants prior to their recruitment into the C_4 pathway. J. Exp. Bot. *62*, 3049–3059.

Beale, C.V., and Long, S.P. (1995). Can perennial C_4 grasses attain high efficiencies of radiant energy-conversion in cool climates. Plant Cell Environ. *18*, 641–650.

Berry, J.A., and Farquhar, G.D. (1978). The CO_2 concentrating function of C_4 photosynthesis: a biochemical model. In Proceedings of the Fourth International Congress on Photosynthesis Biochemical Society, London, pp. 119-131.

Brodribb, T.J., Feild, T.S., and Sack, L. (2010). Viewing leaf structure and evolution from a hydraulic perspective. Funct. Plant Biol. *37*, 488–498.

Brown, R.H. (1978). A difference in N use efficiency in C_3 and C_4 plants and its implications in adaptation and evolution. Crop Sci. *18*, 93–98.

Brown, R.H., Byrd, G.T., and Black, C.C. (1991). Assessing the degree of $c_{(4)}$ photosynthesis in c_3-c_4 species using an inhibitor of phosphoenolpyruvate carboxylase. Plant Physiol. *97*, 985–989.

Cell

Chou, H.H., Chiu, H.C., Delaney, N.F., Segrè, D., and Marx, C.J. (2011). Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science *332*, 1190–1192.

Christin, P.A., Sage, T.L., Edwards, E.J., Ogburn, R.M., Khoshravesh, R., and Sage, R.F. (2011). Complex evolutionary transitions and the significance of c_3-c_4 intermediate forms of photosynthesis in Molluginaceae. Evolution *65*, 643–660.

Christin, P.A., Osborne, C.P., Chatelet, D.S., Columbus, J.T., Besnard, G., Hodkinson, T.R., Garrison, L.M., Vorontsova, M.S., and Edwards, E.J. (2013). Anatomical enablers and the evolution of C_4 photosynthesis in grasses. Proc. Natl. Acad. Sci. USA *110*, 1381–1386.

Dal'Molin, C.G., Quek, L.E., Palfreyman, R.W., Brumbley, S.M., and Nielsen, L.K. (2010). C4GEM, a genome-scale metabolic model to study C_4 plant metabolism. Plant Physiol. *154*, 1871–1885.

Drincovich, M.F., Lara, M.V., Andreo, C.S., and Maurino, V.G. (2011). Evolution of C_4 decarboxylases: Different solutions for the same biochemical problem: provision of CO_2 in Bundle Sheath Cells. In C_4 photosynthesis and related CO_2 concentration mechanisms, A.S. Raghavendra and R.F. Sage, eds. (Dordrecht: Springer), pp. 277–300.

Edwards, G.E., and Ku, M.S.B. (1987). Biochemistry of C_3-C_4 intermediates. In The biochemistry of plants, *Volume 10* (New York: Academic Press, Inc.), pp. 275–325.

Edwards, E.J., Osborne, C.P., Strömberg, C.A.E., Smith, S.A., Bond, W.J., Christin, P.A., Cousins, A.B., Duvall, M.R., Fox, D.L., Freckleton, R.P., et al.; C4 Grasses Consortium. (2010). The origins of C_4 grasslands: integrating evolutionary and ecosystem science. Science *328*, 587–591.

Ehleringer, J., and Pearcy, R.W. (1983). Variation in Quantum Yield for CO_2 Uptake among C_3 and C_4 Plants. Plant Physiol. *73*, 555–559.

Ehleringer, J.R., Sage, R.F., Flanagan, L.B., and Pearcy, R.W. (1991). Climate change and the evolution of C_4 photosynthesis. Trends Ecol. Evol. *6*, 95–99.

Ellis, R.J. (1979). Most abundant protein in the world. Trends Biochem. Sci. *4*, 241–244.

Engel, N., van den Daele, K., Kolukisaoglu, U., Morgenthal, K., Weckwerth, W., Pärnik, T., Keerberg, O., and Bauwe, H. (2007). Deletion of glycine decarboxylase in Arabidopsis is lethal under nonphotorespiratory conditions. Plant Physiol. *144*, 1328–1335.

Farquhar, G.D., Caemmerer, S., and Berry, J.A. (1980). A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species. Planta *149*, 78–90.

Fisher, R.A. (1930). The Genetical Theory of Natural Selection (Oxford: Oxford Univ. Press).

Fong, S.S., and Palsson, B.O. (2004). Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. Nat. Genet. *36*, 1056–1058.

Furbank, R.T. (2011). Evolution of the C_4 photosynthetic mechanism: are there really three C_4 acid decarboxylation types? J. Exp. Bot. *62*, 3103–3108.

Ghannoum, O., von Caemmerer, S., and Conroy, J.P. (2001). Carbon and water economy of Australian NAD-ME and NADP-ME C_4 grasses. Funct. Plant Biol. *28*, 213–223.

Ghannoum, O., Evans, J.R., and von Caemmerer, S. (2011). Nitrogen and water use efficiency of C_4 plants. In C_4 Photosynthesis and Related CO_2 Concentrating Mechanisms, A.S. Raghavendra and R.F. Sage, eds. (Dordrecht, The Netherlands: Springer), pp. 129–146.

Gillespie, J.H. (1983). A simple stochastic gene substitution model. Theor. Popul. Biol. *23*, 202–215.

Gowik, U., and Westhoff, P. (2011). The path from C_3 to C_4 photosynthesis. Plant Physiol. *155*, 56–63.

Hatch, M.D., Kagawa, T., and Craig, S. (1975). Subdivision of C_4-pathway species based on differing C_4 acid decarboxylating systems and ultrastructural features. Funct. Plant Biol. *2*, 111–128.

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

Hindré, T., Knibbe, C., Beslon, G., and Schneider, D. (2012). New insights into bacterial adaptation through in vivo and in silico experimental evolution. Nat. Rev. Microbiol. *10*, 352–365.

Hothorn, T., Hornik, K., van de Wiel, M.A., and Zeileis, A. (2006). A Lego system for conditional inference. Am. Stat. *60*, 257–263.

Hylton, C.M., Rawsthorne, S., Smith, A.M., Jones, D.A., and Woolhouse, H.W. (1988). Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C_3-C_4 intermediate species. Planta *175*, 452–459.

Ibarra, R.U., Edwards, J.S., and Palsson, B.O. (2002). *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. Nature *420*, 186–189.

Khan, A.I., Dinh, D.M., Schneider, D., Lenski, R.E., and Cooper, T.F. (2011). Negative epistasis between beneficial mutations in an evolving bacterial population. Science *332*, 1193–1196.

Kiirats, O., Lea, P.J., Franceschi, V.R., and Edwards, G.E. (2002). Bundle sheath diffusive resistance to CO_2 and effectiveness of C_4 photosynthesis and refixation of photorespired CO_2 in a C_4 cycle mutant and wild-type *Amaranthus edulis*. Plant Physiol. *130*, 964–976.

Kimura, M. (1957). Some problems of stochastic processes in genetics. Ann. Math. Stat. *28*, 882–901.

Ku, M.S.B., Monson, R.K., Littlejohn, R.O., Jr., Nakamoto, H., Fisher, D.B., and Edwards, G.E. (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 Physiol. *71*, 944–948.

Kvitek, D.J., and Sherlock, G. (2011). Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLoS Genet. *7*, e1002056.

Lobkovsky, A.E., Wolf, Y.I., and Koonin, E.V. (2011). Predictability of evolutionary trajectories in fitness landscapes. PLoS Comput. Biol. *7*, e1002302.

Lozovsky, E.R., Chookajorn, T., Brown, K.M., Imwong, M., Shaw, P.J., Kamchonwongpaisan, S., Neafsey, D.E., Weinreich, D.M., and Hartl, D.L. (2009). Stepwise acquisition of pyrimethamine resistance in the malaria parasite. Proc. Natl. Acad. Sci. USA *106*, 12025–12030.

Maurino, V.G., and Peterhansel, C. (2010). Photorespiration: current status and approaches for metabolic engineering. Curr. Opin. Plant Biol. *13*, 249–256.

McKown, A.D., and Dengler, N.G. (2007). Key innovations in the evolution of Kranz anatomy and C_4 vein pattern in *Flaveria* (Asteraceae). Am. J. Bot. *94*, 382–399.

McKown, A.D., Moncalvo, J.-M., and Dengler, N.G. (2005). Phylogeny of *Flaveria* (Asteraceae) and inference of C_4 photosynthesis evolution. Am. J. Bot. *92*, 1911–1928.

Muhaidat, R., Sage, T.L., Frohlich, M.W., Dengler, N.G., and Sage, R.F. (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.

Oberhardt, M.A., Palsson, B.O., and Papin, J.A. (2009). Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. *5*, 320.

Orr, H.A. (2005). The genetic theory of adaptation: a brief history. Nat. Rev. Genet. *6*, 119–127.

Pál, C., Papp, B., Lercher, M.J., Csermely, P., Oliver, S.G., and Hurst, L.D. (2006). Chance and necessity in the evolution of minimal metabolic networks. Nature *440*, 667–670.

Papp, B., Notebaart, R.A., and Pál, C. (2011). Systems-biology approaches for predicting genomic evolution. Nat. Rev. Genet. *12*, 591–602.

Peisker, M. (1986). Models of carbon metabolism in C_3-C_4 intermediate plants as applied to the evolution of C_4 photosynthesis. Plant Cell Environ. *9*, 627–635.

Pick, T.R., Bräutigam, A., Schlüter, U., Denton, A.K., Colmsee, C., Scholz, U., Fahnenstich, H., Pieruschka, R., Rascher, U., Sonnewald, U., and Weber, A.P. (2011). Systems analysis of a maize leaf developmental gradient redefines the

current C_4 model and provides candidates for regulation. Plant Cell *23*, 4208–4220.

Poelwijk, F.J., Tănase-Nicola, S., Kiviet, D.J., and Tans, S.J. (2011). Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. J. Theor. Biol. *272*, 141–144.

R Development Core Team. (2010). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

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

Sage, R.F., and Sage, T.L. (2007). Learning from nature to develop strategies for directed evolution of C_4 rice. In Charting New Pathways to C_4 Rice, J.E. Sheehy, P.L. Mitchell, and B. Hardy, eds. (Hackensack, NJ, USA: World Scientific Publishing), pp. 195–216.

Sage, R.F., Christin, P.A., and Edwards, E.J. (2011). The C_4 plant lineages of planet Earth. J. Exp. Bot. *62*, 3155–3169.

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

Savir, Y., Noor, E., Milo, R., and Tlusty, T. (2010). Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. Proc. Natl. Acad. Sci. USA *107*, 3475–3480.

Stern, D.L., and Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? Evolution *62*, 2155–2177.

von Caemmerer, S. (1989). A model of photosynthetic CO_2 assimilation and carbon-isotope discrimination in leaves of certain C_3-C_4 intermediates. Planta *178*, 463–474.

von Caemmerer, S. (2000). Biochemical Models of Leaf Photosynthesis (Collingwood, Australia: Csiro Publishing).

Weinreich, D.M., Watson, R.A., and Chao, L. (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution *59*, 1165–1174.

Weinreich, D.M., Delaney, N.F., Depristo, M.A., and Hartl, D.L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. Science *312*, 111–114.

Yin, X., and Struik, P.C. (2009). C_3 and C_4 photosynthesis models: an overview from the perspective of crop modelling. NJAS-Wagen. J. Life Sci. *57*, 27–38.

Yizhak, K., Tuller, T., Papp, B., and Ruppin, E. (2011). Metabolic modeling of endosymbiont genome reduction on a temporal scale. Mol. Syst. Biol. *7*, 479.

Chapter 5

Addendum: Manuscripts prepared for publication

5.1 Manuscript K: Plasticity of C_4 photosynthesis in the amphibious sedge *Eleocharis retroflexa*

Overview

Title: Plasticity of C_4 photosynthesis in the amphibious sedge *Eleocharis retroflexa*
Authors: Canan Külahoglu, Simon Schliesky, Manuel Sommer, Alisandra K. Denton, Andreas Hussner, C. Robin Buell, Andrea Bräutigam and Andreas P. M. Weber
Submitted to Plant Cell and Environment, December 2014
Impact factor: 5.906
Co-authorship

Main Findings

This study compared the transcription and anatomy of the photosynthetically flexible sedge Eleocharis retroflexa in different environments. Growing on land E. retroflexia is a classical C_4 plant, with kranz anatomy and high expression of members of the C_4 cycle. In contrast, when grown underwater E. retroflexa shows a relaxed version of kranz anatomy, up regulation of the photorespiratory cycle and other adaptations to an aquatic environment. These aquatic adaptations include an up-regulation of the light harvesting apparatus to compensate for lower light availability under water, a reduced investment in structural genes, and an increase in chloroplasts in mesophyll tissue. Additionally, E. retroflexa showed a sensitive environmental response, with genes differentially expressed between aquatic and terrestrial environments being responsive to minor differences including variation between replicates and a water availability gradient.

Providing insight into the regulation of environmental acclimation, ABA signalling and synthesis was upregulated in terrestrial culms. Additionally, 101 transcription factors, and the epigenetic categories Histone modifications and DNA methyltransferases were differentially expressed between environments.

Contributions

- Discussion and assistance with data analysis
- Enrichment testing
- Differential expression testing
- Editing of full manuscript

1. Title Page

Plasticity of C_4 photosynthesis in the amphibious sedge Eleocharis retroflexa

Canan Külahoglu[a], Simon Schliesky[a], Manuel Sommer[a], Alisandra K. Denton[a], Andreas Hussner[b], C. Robin Buell[c], Andrea Bräutigam[a] and Andreas P. M. Weber[a1]

[a]Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences, Heinrich-Heine-University, 40225 Düsseldorf, Germany

[b]Institute of Plant Biochemistry- Photosynthesis and Stress Physiology of Plants, Cluster of Excellence on Plant Sciences, Heinrich- Heine-University, 40225 Düsseldorf, Germany

[c]Department of Plant Biology, Michigan State University, 48824 East Lansing, MI , USA

[1]Corresponding author; e-mail andreas.weber@uni-duesseldorf.de.

Main Text: 8,930 words excluding references and legends

2. Abstract

It has been hypothesized that evolution of C_4 photosynthesis comes at the cost of reduced phenotypical plasticity, owing to the complex anatomical and biochemical specialization required for operating the C_4 carbon concentrating mechanism. However, C_4 photosynthetic terrestrial wetland species of the genus Eleocharis display remarkable phenotypic plasticity in their mode of photosynthetic carbon assimilation. Eleocharis retroflexa, the most C_4-like species amongst the known Eleocharis C_4 performing sedges, thrives under submerged conditions by reconfiguration of culm anatomy and its photosynthetic mode. The underlying molecular mechanisms permitting adaptation to environmental change through metabolic plasticity are however unknown. To unravel these mechanisms, we employed deep RNA-sequencing of aquatically and terrestrially grown culms and contextualized these molecular data with physiological parameters and enzyme activity measurements. E. retroflexa undergoes structural and metabolic rewiringduring submergence by adapting its culms fully to the new habitat and adjusting its carbon metabolism. While the submerse aquatic E. retroflexa culm transcriptome reflects characteristics of flooding tolerant plants, its carbon metabolism displays a typical C_3- C_4 intermediate signature, featuring high abundance of transcripts encoding proteins involved in photorespiration. At the same time the C_4 cycle is maintained. E. retroflexa represents an interesting model to unravel the molecular mechanisms of adaptation to changing environments by phenotypic plasticity.

Keywords:

RNA-sequencing, Transcriptomics, C_4-Photosynthesis, C_3- C_4 Intermediate, environmental acclimation, phenotypic plasticity, non-model plant species, Eleocharis, Cyperaceae

Introduction

Phenotypic plasticity describes the ability of organisms to accommodate and react to variable environmental conditions by changing their characteristics for better acclimatization (Pigliucci 2001; Sage & McKown 2006). The CO_2 concentrating mechanism of C_4 photosynthesis is considered a specialized adaptation derived from C_3 ancestors. It is a complex trait, which is employed for carbon gain in hot, often arid and high light environments to circumvent high photorespiration rates. The high degree of anatomical and biochemical specialization of C_4 photosynthetic species is thought to reduce their potential for phenotypic plasticity and photosynthetic acclimation to variable environments, as compared to C_3 plants (reviewed by Sage & McKown 2006).

C_4 photosynthesis requires a distinct anatomical and biochemical infrastructure for optimal functionality (Hatch 1987). In general, the C_4 pathway acts as a carbon concentrating mechanism that works on top of the C_3 photosynthetic carbon assimilation by increasing the local CO_2 concentration in the vicinity of ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO; (Bowes et al. 1971; Furbank & Hatch 1987). Typically, with few exceptions, C_4 leaves have two types of photosynthetic cells, with carbon in the form of HCO_3^- initially fixed by the phosphoenolpyruvate carboxylase (PEPC) in the outer mesophyll cells (MCs) and then shuttled in the form of a C_4 carbon compound into the inner bundle sheath cells (BSCs; (Hatch &Slack 1970). In the BSCs, the C_4 carbon compound is decarboxylated, releasing CO_2 at the site of the RuBisCO by either the NAD-dependent malic enzyme (NAD-ME), the NADP-dependent malic enzyme (NADP-ME), or the phosphoenolpyruvate carboxykinase (PEPCK), followed by assimilation into carbohydrates by the Calvin-Benson-Bassham cycle (CBBC; Hatch 1987; Hatch & Slack 1970). The remaining C_3 molecule is transported back to the MCs. The BSCs are situated outside the vascular bundle, encompassing it like a wreath, which is termed “Kranz”-anatomy (Haberlandt 1904). The above-described C_4-specific coordinated modifications of both metabolism and anatomy may have reduced the ability of C_4 performing plants to acclimate their photosynthetic apparatus to

altering environments (reviewed by Sage & McKown 2006).
The sedge family (Cyperaceae) contains more than 20% of the currently known C_4 plant species (Besnard et al. 2009; Sage 2004). Among terrestrial wetland species, members of leafless Eleocharis (Cyperaceae) genus display a remarkable degree of acclimation to varying habitats. These species can grow underwater as well as in air (Ueno 2001; Ueno et al. 1989; Ueno et al. 1988). Among the amphibious Eleocharis species (e.g. E. retroflexa, E. vivipara, E. baldwinii), the photosynthetic modes can be highly variable between an aquatic and terrestrial habitat (Ueno 2004; Ueno & Wakayama 2004). While the culms of the terrestrial form of the three species show a C_4 photosynthesis signature of the NAD-ME subtype and Kranz anatomy, in the aquatic environment, the culms of E. retroflexa, E. vivipara, and E. baldwinii appear more C_4-like, C_3- C_4 intermediate, or C_3, respectively (Ueno, 2004). Upon flooding, E. retroflexa culms undergo acclimatization of the terrestrial culms within days, while new aquatic adapted culms grow (Ueno & Wakayama 2004). However, the submerse aquatic culms, when exposed to air, die away due to rapid drying, while new C_4 terrestrial accustomed culms grow out, as reported for other amphibious Eleocharis species (Ueno 2001; Ueno et al. 1988). Different strategies are known, which terrestrial wetland species are employing to overcome the new challenges of the aquatic environment during submergence. The challenges of the aquatic environment are, for example, physical restrictions on light availability, gas exchange, and nutrient availability (Krause-Jensen & Sand-Jensen 1998; Pedersen et al. 2013). In particular, the gas diffusion rates in water are 104-fold slower than in air. The limitation of CO_2 and O_2 gas flowseverely affects photosynthesis and likely promotes increased rates of photorespiration. Also, light intensity is subdued in turbid flooding water, which decreases photosynthesis efficiency further (Vervuren et al. 2003). The resulting imbalance between carbohydrate assimilation and consumption has lethal consequences for most flooding non-adapted terrestrial plants (Colmer & Voesenek 2009). As an adaptation to flooding, some species grow out of water by shoot elongation to reestablish aerial photosynthesis (Setter & Laureles 1996), whereas others develop specialized “aquatic leaves” specifically accustomed to the wet habitat (Bailey-Serres & Voesenek 2008). Aquatic adapted leaves display reduced gas diffusion resistance as a consequence of reduced cuticle thickness, chloroplast reorientation close towards the epidermis, and reduced leaf thickness (Frost-Christensen et al. 2003a; Mommer et al. 2006; Mommer et al. 2005b; Sand-Jensen & Frost-Christensen 1999).
Previous studies indicated that E. retroflexa might have the capability to change its photosynthetic mode from C_4 to C_4-like, depending on the habitat (Ueno & Wakayama 2004). In this study, we address which physiological and transcriptional programs enable E. retroflexa to thrive during submergence and on land, and how the change of environment affects the photosynthetic modes (C_4-like and C_4 photosynthesis).

5. Material and Methods

Plant material and cultivation
Plants were purchased from an online aquarist-shop (http://www.wasserflora.de; B030PP). E. retroflexa plants were cultivated in terrestrial and aquatic culture for transcriptome profiling by Illumina Sequencing between January and March 2012. Terrestrial E. retroflexa cultures were grown on turf soil in boxes that were partially flooded with tap water under greenhouse conditions (21°C, 12:12h of light/darkness). For the drought stress experiment, E. retroflexa seedlings were transferred to soil and grown for 14 days under the experimental conditions (group 1: control, every day 250 ml water; group 2: every two days 250ml water; group 3: every four days 150ml water; group 4: no water for 14 days). The aquatic culture was set up by transferring viviparous plantlets to aquariums covered with turf soil and a 3 cm upper layer of gravel. Aquariums were filled with tap water (dissolved inorganic carbon, DIC 3 mM) and filamentous algal growth was suppressed by co-cultivation of shrimp. Temperature in the aquarium was constant at approx. 25°C with an approximate pH of 6.8, and fresh-air was constantly supplemented by an aquarium pump. Culms of

aquatic and terrestrial culture were harvested after four weeks of growth. Up to 20 individual plants were pooled for each biological replicate.

Culm anatomy analysis
Fresh culms of E. retroflexa grown submerged and on soil were cut transversally and imaged with the fluorescence microscope Axio Imager M2M (Zeiss, Germany) with light and fluorescence using an UV filter. Images were processed with ZEN10 software (Zeiss, Germany).

Internal transcribed spacer sequence analysis and phylogeny
Internal transcribed spacer (ITS) sequence of E. retroflexa were sub-cloned with ITS1 and ITS4 primer (according to (Inda et al. 2008) and sequenced. Plant identity was confirmed by comparison to the public database Genbank. ITS sequences of other Eleocharis species were manually curated and submitted to http://phylogeny.lirmm.fr/ using the “á la carte” settings. Alignments were performed with Clustal W (Larkin et al.2007) and tree was calculated with PhyML and bootstrapped with 100 repetitions. Tree was drawn with TreeDyn.

RNA extraction, library construction and sequencing
Plant material was extracted with 65°C pre-heated CTAB buffer working solution (1ml buffer per 100mg ground tissue): 50 % (v/v) CTAB buffer stock solution, 2 % (v/v) BME, and 50 % (v/v) acidic phenol. Ground tissue was incubated for 20min at 65°C, followed by two consecutive protein extractions by adding equal volume chloroform-isoamyl alcohol (24:1) to extract and 20 min centrifugation at 10,000xg at 10°C. The aqueous supernatant was transferred to fresh reaction tube and 0.5 volumes of 96% (v/v) ethanol was added. This mixture was loaded onto RNA binding silica columns (Plant RNeasy extraction kit; Qiagen, Hilden, Germany) and further processed as recommend by the manufacturer. RNA was treated twice with RNAse-free DNAse, first on-column and after elution a second time in solution (New England Biolabs, MA, USA). RNA integrity, sequencing library quality, and fragment size were checked on a 2100 Bioanalyzer (Agilent, CA, USA). Libraries were prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina, San Diego, CA) and library quantification was performed with a Qubit 2.0 instrument (Invitrogen, Germany). Single end sequenced samples were multiplexed (6 libraries per lane with approximately 20 million reads per library). All libraries were sequenced on the HISEQ2000 Illumina platform (San Diego, CA). Libraries were sequenced in the single-end mode with read lengths ranging from 80-100 nucleotides.

Transcriptome assembly and annotation
Reads were checked for quality with FASTQC (http://bioinformatics.babraham.ac.uk/projects/fastqc/) and subsequently cleaned and filtered for quality scores greater than 20 and read length greater than 50 nucleotides using the FASTX toolkit (Blankenberg et al. 2010); http://hannonlab.cshl.edu/fastx_toolkit).
Trimmed reads were split into subgroups and were assembled by CAP3 (Huang & Madan 1999). The resulting contigs were merged, split into subgroups again and assembled by CAP3. With this second assembly step, contigs and singlets were merged and assembled in CAP3. The resulting E. retroflexa filtered unigene set (contigs > 200bases length) was annotated using BLASTX searches (cut-off 1e-10) against the S. italica primary transcript database V2.1 (Bennetzen et al. 2012) and Uniref100 (Bairoch et al. 2005). The best blast hit per read was filtered by the highest bitscore. Multiple matching contigs to one S. italica identifier were filtered out with customized Perl script. The unigenes were filtered for contigs that either match an S. italica identifier or a plant accession in Uniref100. This resulted in a unigene database of 27,021 contigs and reduced possible contamination through non-plant contigs.
The final unigene set was uploaded to the KAAS server (http://www.genome.jp/tools/kaas/) to test the representation of KEGG annotated pathways (Moriya et al. 2007). Resulting maps were manually curated for pathways present in plants by comparison to model species A. thaliana and

analyzed for coverage.

Accession Numbers
Sequence data from this article can be found at NCBI Genbank under the accession number SRP ....
Gene expression profiling
Expression abundances were determined by mapping the single-end read libraries (each replicate for each condition) against S. italica primary transcript coding sequences V2.1 (Bennetzen et al. 2012) using BLAT V35 with default parameters (Kent 2002) in dnax mode (nucleotide sequence of query and reference are translated in six frames to protein) followed by parsing out the best mapping hit based on the number of matching bases for each read. Expression was normalized to reads per million mappable reads (RPM). A threshold of 20 RPM per transcript in at least one condition present in at least one replicate was chosen to discriminate against background transcription. Differentially expressed transcripts were determined via EdgeR (Robinson et al. 2010) in R (R Development Core Team, 2009). A significance threshold of 0.05 was applied after P-value was adjusted for the False Discovery Rate (FDR) via Benjamini-Hochberg correction (Benjamini & Hochberg 1995).

Cross species mapping sensitivity assessment
Each E. retroflexa read library was mapped to the unigene database and to the S. italica reference by Blat as described above (see Gene expression profiling). Raw read count files were sorted descending by number of aligned reads per identifier mapped reads and the amount of reads relative to all mapped reads per sample was summed up in R and plotted on a log10 scale. For comparison with other species and cross species mapping against a cognate genome mapping, we mapped (i) T. hassleriana mature leaf reads (Kulahoglu et al. 2014) against the Arabidopsis thaliana TAIR10 representative gene models (Lamesch et al. 2012) and (ii) T. hassleriana mature leaf reads against its own gene models (Cheng et al. 2013).

Data analysis
Data analysis was performed with the R statistical package (R Development Core Team, 2009) and Multi Experiment Viewer 4 (Saeed et al. 2006; Saeed et al. 2003)(MEV4; http://www.tm4.org/mev/) unless stated otherwise. Before Principal Component Analyses (PCA) with median centering, the sample averages were z-score normalized. Hierarchical clustering of samples was performed with MEV4 by normalizing them to z-scores and clustering with average linkage in Euclidean Distance. Sample enrichment was tested for tissue 'signature genes' with expression over 1,000 RPM in each tissue using S. italica V2.1 Mapman categories (from http://mapman.gabipd.org). Significantly differentially expressed transcripts (FDR<0.05) were tested for enrichment by Fisher's Exact Test and p-values were adjusted to FDR via Benjamini-Yekutieli correction (Yekutieli & Benjamini 1999). Mapman fold-change heatmaps were generated using the latest MAPMAN tool V3.6 with the Setaria italica V2.1 as reference (Thimm et al. 2004; Usadel et al. 2005). The Wilcoxon rank test was used for testing significance of fold-changes between the averages of aquatic and terrestrial transcriptomes for specific data subsets, with Benjamini-Yekutieli FDR correction of P-values (Usadel et al. 2005).

Quantitative real-time PCR
Quantitative real time PCR (qRT-PCR) was performed with three biological and three technical replicates per sample using the relative quantification technique by normalizing the gene of interest to a house-keeping gene (UBQ10). SYBR-green (MESA GREEN qPCR MasterMix Plus; Eurogentec) and gene specific primers (Supplemental Table 1) were employed as described by Schmittgen and Livak (2008) (Schmittgen & Livak 2008). Mean normalized expression (MNE) was calculated via the ΔΔCT method after Pfaffl (2001)(Pfaffl 2001).

Enzyme activity and chlorophyll measurements
For the enzyme activity assays under water stress, E. retroflexa terrestrial culms were grown with decreasing amounts of water in four biological replicates for two weeks (see Plant material and cultivation). Enzymatic activities of PEPC, PEPCK, AlaAT, AspAT and NAD-ME were determined as summarized by Ashton et al. (1990)(Ashton et al. 1990) in three biological replicates with three technical replicates per sample. Chlorophyll measurements were performed according to Porra et al (1989) (Porra et al. 1989) with three biological replicates and three technical replicates per sample of aquatically and terrestrially grown culms.

Carbon isotope discrimination
For ^{13}C isotope discrimination leaf powder was freeze-dried and analyzed using the isotope ratio mass spectrometer IsoPrime 100 (ISOTOPE cube; Elementar Analysensysteme). Results were expressed as relative values compared to the international standard (Vienna Pee Dee Belemnite) Element Analysis and calibration for $\delta\ ^{13}C$ measurements followed the two-point method described by Coplen et al. (2006) (Coplen et al. 2006).

6. Results

Sequencing and assembly of aquatic and terrestrial E. retroflexa libraries provides a unigene database covering most of the plant relevant pathways present in KEGG
To provide a reference transcriptome for further studies, three biological replicates of terrestrial culms and two biological replicates of aquatic culms yielding between 29 and 21 million high quality reads, were obtained by Illumina RNA-sequencing (Table 1; Dataset 1). Due to the absence of a reference genome, reads were de novo assembled with CAP3 (Huang & Madan 1999), producing a contig database of 43,817 unigenes (unigene length >200 bases length; Supplemental Table 2; Supplemental Figure 1A). Fifty-eight percent (25,386) of the E. retroflexa contigs were annotated by mapping them to the evolutionary closest available C_4 grass genome of Setaria italica (Bennetzen et al. 2012), matching to 37% (13,204) of the known S. italica genes (Table 1). To estimate the quality of the contigs and contamination due to co-cultivation in the aquaria, the contigs were annotated against the Uniref100 protein database (Bairoch et al. 2005); Supplemental Table 2). Out of 43,817 contigs, 29,512 (67%) found a best match in the Uniref100 database (Figure 1A). From these 21,832 (50%) were annotated to a Viridiplantae identifier (ID) by best hit (Figure 1A). Around 7,112 (16%) contigs fell into the category of non-plant annotated IDs matching to fungi, bacteria or insects (Figure 1A). In the subsequent analyses, the unigene database was limited to 27,021 contigs matching either S. italica identifier, plant identifier of Uniref100 or both with an N50 of 1199 bases (Dataset 2; Supplemental Figure 1B; Supplemental Table 2).
To assess whether core plant metabolism was well represented by the filtered E. retroflexa unigene database (27,021 contigs), it was benchmarked against plant pathways from KEGG (Moriya et al. 2007); Supplemental Dataset 1). The contig database covered all genes involved in light and dark reactions of photosynthesis, as well as starch and sucrose metabolism, tricarboxylic acid cycle (TCA cycle), glycolysis, galactose metabolism, pyruvate metabolism, amino sugar and nucleotide sugar and nucleotide metabolism. Other pathways of carbohydrate metabolism, such as the pentose phosphate pathway, fructose and mannose metabolism, and glyoxylate and dicarboxylate metabolism lacked full coverage by few genes (Supplemental Dataset 1). In general, the metabolism of lipids, amino acids and nucleotides were fully represented. Secondary metabolism involving synthesis of lignin precursors derived from phenylpropanoid, carotenoid, flavonoid, and porphyrinand chlorophyll biosynthesis were completely covered. Terpenoid and anthocyanin synthesis were incomplete.
In summary, the presented E. retroflexa unigene database exhibited good coverage of all core plant metabolic pathways (Supplemental Dataset 1) as well as central cellular processes (DNA repair, DNA transcription, translation and protein; Supplemental Dataset 2), including regulatory networks

and plant hormone signaling (Supplemental Dataset 3).

Mapping of E. retroflexa reads to S. italica improves transcriptome representation relative to mapping the reads to E. retroflexa contigs.
To quantify gene expression, the reads were mapped to a reference sequence. Two options are available for this: mapping of reads (i) to the de novo assembled contigs or (ii) to the genome of a related species. For E. retroflexa transcript quantification using both approaches, mapping to the contigs or a cross-species reference database, were compared and evaluated. At least 70% of the reads mapped to the unigene set (Table 1), however, annotation of the unigene database with S. italica revealed known issues with the de novo assemblies (Schliesky et al. 2012). Around 58% of the annotated E. retroflexa unigenes aligned with a S. italica identifier, which was assigned as the best hit to more than one unigene (Figure 1B; Supplemental Dataset 4). In a more extreme case, 100 contigs matched one S. italica identifier (Si020831m) encoding a protein of unknown function suggesting redundancy of the transcripts in the de novo transcript assembly. C_4 cycle genes, such as the alanine aminotransferase (AlaAT) and the triose-phosphate transporter (TPT) were absent in the E. retroflexa transcript assemblies.

For comparison, E. retroflexa reads were aligned to the S. italica gene models using the Blat algorithm. In total, 21,679 S. italica gene models were aligned through cross-species mapping of E. retroflexa reads, with mapping efficiencies between 28-36% for all mapped samples (Table 1). To compare both mapping approaches, the fraction of reads mapping to higher and lower expressed sequences was visualized for all samples. Mapping the reads to S. italica gene models delivered higher similarity between the individually mapped biological replicates than mapping to the E. retroflexa unigenes (Supplemental Figure 2A). When mapping E. retroflexa reads to its own unigene set, on average, 20% of all reads matched to four of the most highly expressed unigenes (Supplemental Figure 2A), resulting in the high starting point of the curve displayed in Supplementary Figure 2A. On the basis of these results and previous experience (Brautigam et al. 2010; Gowik et al. 2011), we opted to conduct transcript quantification by cross-species mapping of E. retroflexa to S. italica.

The E. retroflexa transcriptomes reflect changes between different habitats and display unexpected variability between replicates
For analyzing the degree of variation and gene expression dynamics of the E. retroflexa transcriptomes, all samples were hierarchically clustered (Figure 2A) and reduced to their main variances by principle component analysis (PCA; Figure 2B). Biological replicates of the aquatic and terrestrial E. retroflexa culm transcriptomes clustered together and were separated by habitat (Figure 2A). The PCA reflected the hierarchical clustering, with the first component separating the samples by habitat, accounting for 36% of total sample variation and the second component with 21% describing the biological variation within the biological replicates (Figure 2B). These results indicated that one third of the gene expression was changed by the culm's habitat. In total, 8% of the whole transcriptome (1,356 genes) was significantly differentially regulated between submersed aquatic (630 up-regulated; BH corrected P-value<0.05) and terrestrial (726 up-regulated; BH corrected P-value<0.05) culms (Supplemental Figure 3). Despite the environmental influence, all samples were qualitatively similar with regard to the Pearson's correlation (Supplemental Table 3), but showed a constant factor of variability between biological replicates. To test whether the observed variability between replicated samples was random or resulted from unintended variation of experimental conditions, we compared the variation between biologicalreplicates of the genes defined as differentially expressed (BH corrected P-value<0.05) and the remaining gene set. If the variation among the samples was random, it would be randomly distributed between the significantly changed transcripts and the remaining transcripts. We found a significant enrichment in a set of genes (Fisher's exact test P-value<0.001) that fulfill both criteria (i) at least two-fold change in transcript abundance between the biological replicates and (ii) genes that were

significantly differentially expressed between habitats (Figure 2C). Therefore, the genes that were related to habitat acclimatization were the ones showing greater variation between replicates than other genes.

Functional changes in the submersed aquatic and terrestrial culm transcriptomes mirror the acclimatization of E. retroflexa to the respective habitats

The terrestrial E. retroflexa culms express more transcripts related to structure

Submersed aquatic E. retroflexa culms grew fast under water, but never lifted themselves beyond the water surface. This is similar to what has been described for the growth habits of aquatic E. vivipara plants (Supplemental Figure 4; (Ueno 2001). Thus, to identify the differences between aquatic and terrestrial culm structure, the respective transcriptomes were analyzed for pathway enrichment of cell wall related categories (Figure 3A). All changes mentioned in the text were statistically significant. The submersed aquatic culms invested much less in transcription of genes related to phenolic compounds, phenylpropanoid and lignin biosynthesis (Figure 3A). Transcripts of these categories were up-regulated and enriched in the terrestrial culms (Fisher's Exact BY corrected P-value 4.2E-4; Figure 3A; Supplemental Figure 5; Supplemental Dataset 6). The higher transcript abundance of structure-related genes in terrestrial culms was reflected by higher fold-changes of the category "cellulose synthesis for cell wall enforcement" (SEC61 BETA, CELLULOSE SYNTHASE LIKE D4; Wilcoxon rank test BY corrected P-value 0.014; Supplemental Dataset 7) as well as enrichment of fold changed transcript levels regarding cell wall modification, e.g. several classes of pectin esterases (PECTIN ESTERASE 11, PECTIN METHYLESTERASE, QUARTET) known to cause cell wall stiffening (Wilcoxon rank test BY corrected P-value 1.69E-4; Supplemental Dataset 7; (Micheli 2001)). There were at least two-fold-changes in expression levels of transcripts involved in cell wall loosening and expansion (Wilcoxon rank test BY corrected P-value 1.79E-10), such as glycosyl hydrolases, beta-xylosidases (BETA-XYLOSIDASE 2 and 3) endotransglucolases (XYLOGLUCAN ENDOTRASNGLYCOSE 3, 4 and 8) and polygalacturonases (POLYGALACTURONASE1 and 3 and QUARTET2), in the terrestrial culms (Figure 3A; Supplemental Dataset 7).

Within the Mapman category of cell wall modification (Wilcoxon rank test BY corrected P-value 0.0016; Supplemental Dataset 7), we could observe up to five-fold-changed expression levels of transcripts annotated as expansins (e.g. EXPANSIN7, EXPANSIN8, EXPANSIN11 and EXPANSIN16; Figure 3A). Further, the terrestrial culms exhibited up-regulation of transcripts belonging to wax synthesis needed for cuticle development (Wilcoxon rank test BY corrected P-value 0.03; Figure 3A; Supplemental Dataset 7). These differences in structure related transcripts were reflected by a 2.7-fold difference in dry weight to fresh weight ratio between the aquatic and terrestrial culms (Figure 3B).

Comparing aquatically and terrestrially grown E. retroflexa culms revealed evident changes in culm anatomy (Figure 4A). The terrestrial culms showed higher auto-fluorescence of lignified tissue (xylem) under UV light than the aquatic culms (Figure 4B). Also, the terrestrial BSCs had weak lignification (Figure 4B). The terrestrial epidermis was regularly interspersed with stomata whose cell walls displayed auto-fluorescence, as well as the auto-fluorescence of cuticle waxes. In the aquatic culms, no stomata could be detected. Thus, the anatomical and structural changes within culm anatomy were consistent with changes in gene expression levels in the transcriptome.

The photosynthesis apparatus is enhanced in aquatically grown culms

One of the challenges for photosynthesis under water is low light availability and the quality of the available light spectrum (Kirk 1994; Pedersen et al. 2013). We assessed the submersed aquatic and terrestrial transcriptomes for consequences of the submerged lifestyle on photosynthesis and light capture. Most of the Mapman annotated transcripts for light reactions includingphotosystem I and II polypeptide subunits, and light harvesting complexes, as well as the cytochrome b6f/c were up-regulated in the aquatic form (Wilcoxon rank test BY corrected P-value 1.09E-13; Figure 4A;

Supplemental Dataset 7). Analysis of cumulative gene expression showed that light reactions occupy 10% of the transcriptional investment (Supplemental Figure 6). Synthesis of glycolipids in general was up-regulated (Fisher's exact test BY corrected P-values <0.001; Supplemental Dataset 6). Transcripts associated with the biosynthesis of thylakoid membrane lipids, such as transcripts of DIGALACTOSYL DIACYLGLYCEROL DEFICIENT 1 and 2 (DGD1 and 2) and SULFOQUINOVOSYLDIACYLGLYCEROL 1 and 2 (SQD1 and 2), were up-regulated in the aquatic culms (Figure 4A, Supplemental Dataset 7). Concordantly with the enrichment of light harvesting complexes in submersed culms, tetrapyrrole biosynthesis was up-regulated (Wilcoxon rank test BY corrected P-value 1.52E-4; Figure 5A; Supplemental Dataset 7). Notably, in the aquatic culms, transcripts associated with chlorophyll and carotenoid biosynthesis were twice as abundant as in the terrestrial culms (Figure 5A; Supplemental Dataset 7).
As indicated by the transcriptomes, photometric chlorophyll determination revealed that total chlorophyll content per dry weight was two-fold higher in the aquatic culms (P-value<0.05; Figure 5B). The transcriptome changes associated with light capture and chloroplasts were reflected in culm anatomy. In terrestrial culms, the outer BSCs were enlarged and accumulated high numbers of chloroplasts as seen in most C_4 plants. The MCs appeared much smaller (Figure 4A). Culms grown under submerged conditions featured less enlarged BSCs, while MC size and chloroplast number increased, compared to terrestrial MC culm anatomy (Figure 4A).

The C_4 cycle signature is stronger in terrestrial culms, while aquatic culms display enhanced expression of the Calvin-Benson-Bassham cycle and photorespiration

In addition to the detected changes in culm structure and light capture, the central carbon metabolism was adapted to the aquatic and terrestrial habitats (Figure 5A). Ueno and colleagues previously analyzed the localization of C_4 cycle enzymes (PEPC, NAD-ME, PPDK and RuBisCO Large Subunit) and classified E. retroflexa as NAD-ME subtype C_4 plant (Ueno & Wakayama 2004). The transcriptomes of the terrestrial and aquatic culms now enabled a detailedanalysis of E. retroflexa's full C_4 cycle and carbon concentrating mechanism under different growth conditions. The C_4 cycle showed clear differences between the terrestrial and aquatic habitats regarding transcript levels of C_4 cycle genes typical for the NAD-ME/PEPCK subtype (Figure 6A). The mentioned C_4 cycle genes were expressed between 593-25,621 RPM in terrestrial culms and between 234-9,972 RPM in submersed aquatic culms (Supplemental Table 4). In the submersed aquatic culms, abundance of C_4 cycle genes was between 31 to 83% of the terrestrial expression (Figure 6A; Supplemental Table 4). The three-transporter system BILE ACID:SODIUM SYMPORTER FAMILY PROTEIN2/SODIUM:HYDROGEN ANTIPORTER/ PHOSPHOENOLPYRUVATE TRANSLOCATOR (BASS2/NHD/PPT), importing substrates (phosphate and pyruvate) for pyruvate, phosphate dikinase (PPDK) activity and exporting phosphoenolpyruvate (PEP) in the MC was highly abundant in the terrestrial culms (Figure 6A). The enzymes needed for providing the substrate for HCO_3^- fixation, PPDK (regenerating PEP from pyruvate), a cytosolic CARBONIC ANHYDRASE (CA2; converting CO_2 to HCO_3^-), and PEPC (converting HCO_3^- and PEP to oxaloacetate; OAA) were highly abundant in the terrestrial compared to the aquatic culms. The transcripts encoding the main decarboxylating enzymes of this C_4 cycle subtype, NAD-ME and also PEPCK, were relatively higher in the terrestrial culms (Figure 6A). Also, the transcript levels of ALANINE AMINOTRANSFERASE (AlaAT) and ASPARTATE AMINO-TRANSFERASE (AspAT) and a mitochondrial MALATE DEHYDROGENASE (MDH), which are essential for the conversion of transfer acids, were more abundant in the terrestrial culms (Figure 6A). In accordance with this C_4 photosynthesis profile, the carbon isotope ratio ($\delta^{13}C$) with -15.76 ‰ was comparable to that of other C_4 plants (Figure 6B; Cernusak et al. 2013) (Cernusak et al. 2013).
The aquatically grown culms had decreased C_4 cycle enzyme expression, though the C_4 photosynthesis signature was still higher compared to typical C_3 plants. In the aquatic culms, Calvin-Bassham-Benson cycle (CBBC) related transcripts were up-regulated (Fisher's Exact BY corrected P-value 1.09E-14; Figure 5A; Supplem ental Dataset 6). Furthermore, the small

RUBISCO subunit(Fisher's exact test BY corrected P-value 3.80E-06; Supplemental Dataset 6) and the RUBISCO ACTIVASE (RCA; Fisher's exact test BY corrected P-value 0.026; Figure 5A; Supplemental Dataset 6) were significantly enriched. At the same time, transcripts related to photorespiration were strongly up-regulated (Fisher's exact test BY corrected P-value 7.10E-08; Figure 5A; Supplemental Dataset 6). Especially, SERINE HYDROXYLMETHYLTRANSFERASE (SHM) and GLYCINE DECARBOXYLASE COMPLEX (GDC) subunits were significantly up-regulated in the aquatic culms (Figure 7; Supplemental Dataset 6). Also, the GLYCOLATE OXIDASE (GOX) was up-regulated in the aquatic transcriptome (BH corrected P-value 0.001; Dataset 1). Complementing the strong photorespiratory signature, transcripts related to refixation of photorespiratory ammonia via glutamine/glutamate synthesis were up-regulated in the aquatic as compared to the terrestrial culms (Figure 7; Supplemental Dataset 6 and 7). Reflecting this less pronounced C_4 signature, the carbon isotope ratio of the aquatic culms with -19 ‰ ($\delta\ ^{13}C$ value) showed stronger discrimination against ^{13}C, which is closer to the range of C_3 plants (Figure 6B; (Cernusak et al. 2013)).

C_4 cycle transcripts in the terrestrial culms showed enhanced plasticity depending on water availability and drought stress

During its life cycle, Eleocharis plants can be subjected to great changes in its habitat of fresh water streams and ponds, including episodes of flooding and drying (Ueno 2001). Consequently, these sedges have evolved the ability to adjust their phenotype quickly to changing environments. Signals of adjustments and transcriptomic plasticity were detected by comparative transcriptomic analysis. Genes that were involved in habitat acclimatization, showed a high degree of variation between the replicates of the same growth condition (Figure 2C).

To independently corroborate this finding, we tested the variability of C_4 cycle genes under different degrees of drought. To this end, we grew E. retroflexa plants on soil and provided them with decreasing amounts of water per group for two weeks (Group 1: control every day 250ml water; Group 2: every two days 250ml water; Group 3: every four days 150ml water; Group 4: no water for 14 days). From these plants,transcripts of core C_4 cycle enzymes (NAD-ME, PPDK, PEPC) were measured via qRT-PCR (Figure 8A-C). In addition, PEPC, NAD-ME, PEPCK, AspAT and AlaAT enzyme activities were determined by coupled photometric assays (Figure 8D). Based on transcriptional activity, reducing the water amount from group 1 to group 2 had no significant effect on gene expression (Figure 8A-C). However, limiting the water availability to watering every fourth day (group 3) caused a significant increase in the expression of NAD-ME (3-fold), PPDK (7-fold) and PEPC (13-fold) between group 1 and 3 (P-value<0.001; Figure 8A-C). Enzyme assays showed a trend towards increasing PEPC and NAD-ME activity during drought, however, the magnitude of change was much lower and the changes were not significant between group 1 and group 3 (Figure 8D). Thus, enzyme activity remains more stable, whereas E. retroflexa reacts strongly on transcriptional level to environmental stimuli.

To investigate what might be controlling this drought response, we took a closer look at the comparative transcriptomes. Abscisic acid (ABA) metabolism (BY corrected P-value 0.0041) and synthesis (BY corrected P-value 0.0038) were enriched in the terrestrial culms (Supplemental Dataset 6). In a related species, E. vivipara, changes in C_4 cycle enzymes are tied to changes in ABA concentration (Agarie et al. 2002; Ueno 2001; Ueno et al. 1988) Transcripts related to the biosynthesis or degradation of other phytohormones were not detected as differentially expressed (Supplemental Dataset 6; Supplemental Figure 7). On the level of transcriptional regulators, 101 transcription factors (TFs) were differentially transcribed between terrestrial (26 TFs up-regulated) and aquatic culms (75 TFs up-regulated; Dataset 1). Interestingly, we detected significant changes related to histone modification (BY corrected P-value 0.00059), DNA methyltransferases (DNMT) and (DMT7; DNMT2; MET1; DRM1; CMT1; Wilcoxon rank test BY corrected P-value 0.0137). Furthermore, ALIFIN-LIKE 1 (AL1) transcriptional regulators (AL1; AL3; AL5; AL6; AL7; Wilcoxon rank test BY corrected P-Value 0.012) were enriched in the aquatic culms (Supplemental Figure 7).

7. Discussion

E. retroflexa has been described as NAD-ME C_4 photosynthesis performing sedge under terrestrial conditions and as a C_4–like plant when submerged in water (Ueno & Wakayama 2004). In general, C_4 plants have been proposed to display less plasticity in the range of growth habitats and phenotypic plasticity (Sage & McKown 2006). To determine, which transcriptional programs enable E. retroflexa to acclimatize to terrestrial and aquatic lifestyle rapidly while still performing C_4/C_4-like photosynthesis, we generated comparative transcriptomes of aquatically and terrestrially grown E. retroflexa culms. This data was contextualized with physiological and anatomical parameters, and experiments assessing the adaptability of the C_4 cycle under drought stress. A phylogenetic analysis based on ITS sequences revealed that E. retroflexa is much more closely related to E. baldwinii than E. vivipara. Amongst these previously described C_4 species, E. retroflexa and E. baldwinii have a stronger C_4 signature compared to E. vivipara, which has been suggested to have evolved C_4 photosynthetic traits more recently (Supplemental Figure 8; Ueno 2001). It is unclear, whether E. retroflexa is still evolving towards full C_4-ness or if the display of C_3- C_4 intermediate traits is a reversion from C_4 photosynthesis for better adaption under water.

The E. retroflexa unigene set provides a base for further molecular studies

To date, no Cyperaceae genome is available and transcriptomes have only been published for Eleocharis baldwinii recently (Chen et al. 2014). In this study, we provide a reference database of 27,021 unigenes for E. retroflexa that covers most of all core plant pathways and most cellular processes and regulatory pathways, as represented by the KEGG database (Moriya et al., 2007; Supplemental Dataset 1, 2 and 3). With an N50 of 1,199 bases and average contig length of 789 bases our filtered unigene databse is in the range of other de novo assembled reference transcriptomes, such as radish (Raphanus sativus; (Wang et al. 2013)), scarlet sage (Salvia splendens; (Ge et al. 2014)), Megathrysus maximus and Dichantelium clandestinum (Brautigam et al. 2014). The assembled contigs provide the basis for designing primer-sets for qRT-PCR, indicating that the database represents the E. retroflexa transcripts to a replicable degree (Figure 8A-C).

For differential transcriptome analysis, we opted for cross-species mapping of reads rather than mapping the reads to the unigene database as de novo contig assemblies of short RNA-seq reads still suffer from several shortcomings. A major issue is the occurrence of redundant contigs representing one gene locus (Papanicolaou et al. 2009), which we also observed in our assembly (Figure 1B). This artificial inflation of contig numbers occurs particularly frequent for highly expressed genes, as well as genes that contain highly conserved sequence motifs of large gene families, such as transcriptional regulators (Figure 1B). These redundant contigs display slight differences between each other, due to alternative splicing, sequencing errors, or single nucleotide polymorphisms (SNPs) between alleles (Papanicolaou et al. 2009). Another disadvantage arises from lowly expressed genes with subsequent low read coverage or assembly errors leading to either fragmented contig or absence of transcripts (Martin & Wang 2011; Schliesky et al. 2012). Mapping to related reference genomes can pose challenges for subsequent data analysis. For example, species-specific genes cannot be mapped because they are not represented in the reference. In addition, different genes might display different rates of sequence divergence due to different evolutionary rates, which might lead to a bias in quantifying gene expression levels. However, the advantages of cross-species mapping outweigh the possible shortcomings of cross-reference mapping and this method has been successfullyused for transcriptome analyses of other non-model species with no available reference genome (Bräutigam et al. 2010; Bräutigam et al. 2014; Gowik et al. 2011; Külahoglu et al. 2014).

Transcriptional changes between the terrestrial and the aquatic culms are closely related to

the habitat switch and photosynthetic mode

To place the changes observed between the aquatic and terrestrial culms (mean r=0.86; Supplemental Table 3) in context, the correlation between the two transcriptomes were compared with published expression data from two closely related C_4 (Gynandropsis gynandra) and C_3 (Tarenaya hassleriana) Cleomaceae species (Külahoglu et al. 2014). The comparison of aquatic and terrestrial E. retroflexa culms shows a slightly higher similarity based on Pearson's correlation than between the mature leaves of T. hassleriana (C_3) and G. gynandra (C_4; mean r=0.80; Supplemental Table 3).

At the same time, comparison of the transcriptomes of a leaf derived tissue, such as petals against mature leaf of T. hassleriana (mean r= 0.11), or root against mature leaf (mean r=0.09; Supplemental Table 3), reveals that the adaptation of E. retroflexa culms to different habitats is changing the transcriptome's general dynamic only to a minor degree. Thus, we conclude, that the changes monitored between the terrestrial and the aquatic culms should reflect the habitat switch and photosynthetic mode of the culm tissues.

There is a high similarity between the aquatic and the terrestrial culm transcriptomes based on PCA and HCL with around 8% (1,356 genes) of all transcripts beingdetected as significantly different (Figure 2A-B; Supplemental Figure 3). The number of altered transcripts between the two habitats is comparable to the number of transcripts responding to systemic responses, pathogen or pest attack (9%, (De Vos et al. 2005)). Interestingly, it is higher than the number of transcripts detected as significantly changed between closely related C_4 and C_3 species, even though overall gene expression patterns correlate more closely (~4% in Cleome and 3.4 % in Flaveria; Bräutigam et al. 2010; Gowik et al. 2011).

The E. retroflexa transcriptome shows high plasticity and variability in transcripts linked to the habitat acclimatization

Biological variation of transcript abundances between the replicates of the same habitat displayed a clearly distinct variation, as judged on the basis of PCA and Pearson's correlation values (Figure 2A-B; Supplemental Table 3). The transcripts displaying the strongest fluctuation between biological replicates are identical with those that are significantly differentially regulated between the aquatic and terrestrial habitat (Figure 2C). This indicates that transcriptional variation between the replicates is not arbitrary or due to experimental error, but it is rather associated with transcriptional fine-tuning of culm acclimatization.

Transcript abundance of enzymes associated with C_4 photosynthesis (PPDK, PEPC, NAD-ME) increases under water deprivation (Figure 8A-C), possibly induced through the hormone abscisic acid (ABA). These results indicate that micro-environmental cues can significantly affect the transcriptional program of E. retroflexa. For E. vivipara, a related Eleocharis species, it has been shown that ABA is able to induce C_4-ness under submerged conditions (Agarie et al. 2002; Ueno et al. 1988). Similarly, ABA signaling has been reported to induce CAM photosynthesis in some facultative CAM plants (Chu et al. 1990; McElwain et al. 1992). The hormones ABA and ethylene have been connected to aquatic leaf formation and its regulation in heterophyllous amphibious plant species (Kuwabara et al. 2001; Minorsky 2003). In the terrestrial transcriptome ABA metabolism is significantly up-regulated (Supplemental Dataset 6), with no trace of other plant hormone circuits being significantly altered under aquatic or terrestrial conditions (Supplemental Figure 7). For E. vivipara it has been reported that application of exogenous gibberellic acid to terrestrial culms can trigger submerged (C_3) culm anatomy featuring small BSCs and the absence of stomata (Ueno 2001). However, no enhanced gibberellic acid signaling could be detected in the aquatic transcriptome (Supplemental Figure 7), indicating that on transcriptional level different regulatory mechanisms may play a role in E. retroflexa.

Under aquatic conditions plants can suffer from hypoxia and impeded gas exchange leading to increased ethylene concentrations within the submerged plants (Bailey-Serres & Voesenek 2008; Jackson 2008). In the submersed aquatic E. retroflexa transcriptome, no significant alterations related to hypoxia-induced signaling pathways were detected (Lee et al. 2011; Mustroph et al.

2009), implying that E. retroflexa culms were well acclimatized to the aquatic habitat at the time of their harvest.
Intriguingly, we found significant changes related to histone modification and DNA methyltransferases (DNMT; Supplemental Figure 7; Supplemental Dataset 7). Five ALIFIN-LIKE transcriptional regulators were also enriched in the aquatic culms. Among those, AL1, AL5, AL6 and AL7 are known to bind to di- and trimethylated histone H3 at lysine 4 (H3K4me3/2), which are markers of transcriptionally active chromatin (Lee et al. 2009). Enrichments of trimethylation of histone H3 Lys4 (H3K4me3) and acetylation of histone H3 Lys9 (H3K9ac), often used as a positive marker of histone modifications, are associated with transcriptional activity and correlate with gene activation in response to drought stress (reviewed by (Kim et al. 2010). In rice, modification levels of acetylation of histone H3, dimethylation of histone H3 Lys4 (H3K4me2) and trimethylation of histone H3 Lys4 (H3K4me3) are altered on submergence-inducible genes during the process from submergence to re-aeration (Tsuji et al. 2006). There, the submergence treatments resulted in the decrease of H3K4me2 levels and increase of H3K4me3 levels on the 5′- and 3′-coding regions of submergence inducible genes alcohol dehydrogenase1 (ADH1) and pyruvate decarboxylase1 (PDC1) genes (Tsuji et al. 2006).
In summary, up-regulation of transcripts encoding histone-modifying enzymes could play a role in altering the overall expression profile in submersed culms. Interestingly, submersion of E. retroflexa culms does not leave traces of hypoxia-stress induced signaling in the surveyed transcriptome.

E. retroflexa culms reflect acclimatization to the habitat by changes in culm structure and photosynthesis
Earlier studies by Ueno and colleagues showed that E. retroflexa plants develop new adapted photosynthetic culms under water, which rapidly dry out when the plants are transferred to soil (Ueno 2001; Ueno & Wakayama 2004). When grown under water, E. retroflexa culms grow fast; however, they never lift beyond the water surface (Supplemental Figure 4B), which is similar to E. vivipara's growth habit (Ueno 2001). This could be connected with a decreased investment in genes related to phenolic compounds, phenylpropanoid and lignin biosynthesis in aquatic culms (Figure 3A). Characteristically, submerged leaves have two main strategies to survive in the wet habitat, by either growing out of the water by stem elongation or the development of aquatically accustomed leaves (Mommer & Visser 2005). Clearly, the latter is true for E. retroflexa. Comparison of cross sections of aquatic and terrestrial culms, revealed decreased auto-fluorescence of phenolic compounds indicating the presence of lignin (Figure 4B). Reduced investment and thus reduced cost in fixed carbon in vascular bundles and lignin are a characteristic for aquatic plants (Sculthorpe 1967). Aquatic plants apparently need less cell wall reinforcement in the water, since a cell wall constrains the rate and direction of turgor-driven cell growth (Bailey-Serres & Voesenek 2008). Also evident in the transcriptome is the significantly higher fold-change in transcript levels related to cell wall modification and enlargement, such as various expansins in the terrestrial culms (Figure 3A, Supplemental Dataset 7). This seems to be concomitant with the observed enlarged BSCs in the terrestrial culms (Figure 4; Ueno & Wakayama 2004). Besides the decreased need for leaf rigidity under water, the aquatic environment poses unique challenges for photosynthesis by low light availability and a shift in light spectrum (Kirk 1994; Pedersen et al. 2013). We find that E. retroflexa overcomes these challenges by a 10% higher investment in transcripts related to light reactions and photosystems (Supplemental Figure 6; Figure 5A). The aquatically adapted plant Rumex palustris shows lower PSII abundance after acclimatization to submergence compared to terrestrially acclimation (Mommer et al. 2005b). In submerged E. retroflexa transcriptomes both photosystems are increased (Figure 5A). When submerged, E. retroflexa culms have a higher demand of light absorption, which is structurally supported by higher abundance of transcripts related to galactolipid biosynthesis and transcripts needed for thylakoid membrane assembly (Figure 5A). The higher transcript abundances portioned into both photosystem polypeptide subunits, light harvesting complexes and in chlorophyll and carotenoid biosynthesis, indicate that aquatic culms have to adjust their photosynthetic apparatus as an adaption to lower light intensities

and a change in light spectrum under water (Holmes & Klein 1987; Sand-Jensen 1989). Interestingly, during shade avoidance plants display higher chlorophyll per dry weight (Bailey et al. 2004), similar to what is observed in submerged E. retroflexa for chlorophyll content (Figure 5B). Hence, the adaption of the light harvesting machinery of E. retroflexa might derive from a shade avoidance response (SAR) as it has been suggested for other submerged growing plant species (Boeger & Poulson 2003; Frost-Christensen & Sand-Jensen 1995; Mommer et al. 2005a).
On the regulatory level, the GOLDEN-2-LIKE 1 (GLK1) transcription factor is significantly (three-fold) up-regulated in the aquatic culms (150/50 RPM). Prior studies in the C_4 plant Zea mays showed that the GOLDEN-2 gene is exclusively expressed in the BSCs, whereas GLK1 is only expressed in MCs (Langdale & Kidner 1994; Rossini et al. 2001). In both cell types, these transcription factors are important for chloroplast biogenesis (Rossini et al. 2001). The results presented here for aquatic culms are consistent with the described function of GLK1 acting as nuclear regulator of photosynthetic capacity (Waters et al. 2009), especially under submerged conditions when MCs are being enlarged and accumulate more chloroplasts (Figure 4A).

E. retroflexa culms display a change in its C_4 photosynthesis profile depending on environmental cues

E. retroflexa has been described as a NAD-ME subtype C_4-like photosynthesis performing species under aquatic and terrestrial conditions, though the C_4 cycle enzymes (PPDK, PEPC and NAD-ME) show slightly lower protein abundance in the aquatic form (Ueno & Wakayama 2004); Figure 6A). In both transcriptomes and in the enzyme activity assays we found evidence for the presence of two decarboxylating enzymes –NAD-ME and PEPCK (Figure 8D), as it has been described for G. gynandra and M. maximum (Brautigam et al. 2014; Sommer et al. 2012).

During submergence, photosynthesis rates drop in non-adapted terrestrial species, due to the increased gas diffusion resistance, restricted access to light and biochemical limitations (Centritto et al. 2003; Long & Bernacchi 2003). Aquatic acclimated culms are thinner and have reduced cuticles to decrease the internal diffusion path for CO_2 to the chloroplasts (Maberly & Madsen 2002; Madsen & Sandjensen 1991). Aquatic E. retroflexa culms visually appear to be thinner, have less dry matter (Figure 3B) and their transcriptional investment in cuticle waxes is down-regulated (Figure 3A). The reduction of cuticle thickness has been reported to lead to a reduction of the gas diffusion resistance in aquatic plants (Frost-Christensen et al. 2003b). Typical C_4 architecture is dissolved in aquatic culms: the MCs adjacent to epidermis are massively enlarged with high chloroplast content in aquatic culms (Figure 4A), as it has been reported for other aquatically adapted plants (Mommer & Visser 2005). For reducing the diffusion path length, the chloroplasts are present in all epidermal and sub-epidermal cells and positioned towards the exterior of the cells (Mommer et al. 2005b). All these monitored acclimations of submersed aquatic E. retroflexa culms support the hypothesis that CO_2 directly enters the MCs of the aquatic leaves via diffusion through the epidermis and not via stomata (Mommer et al. 2005b).

Another mechanism of aquatic plants for reducing gas diffusion resistance is the conversion of CO_2 to HCO_3^- for highersolubility catalyzed by carbonic anhydrases (CA; reviewed by Pedersen et al. 2013). In E. retroflexa, CARBONIC ANHYDRASE (CA2) appears to be recruited to the C_4 cycle by its up-regulation in the terrestrial culms, whereas in the aquatic culms the BETA CARBONIC ANHYDRASE 5 is 1.5 fold higher accumulated (Dataset 1; BH corrected P-value 0.0003).

Per definition, C_4-like species display higher C_4 cycle activities than in a C_3- C_4 intermediate, but lack complete BSC compartmentalization of RuBisCO (Edwards & Ku 1987). It has been postulated earlier, that E. retroflexa culms maintain a C_4-like profile when submerged based on C_4 cycle enzymes and RuBisCO protein immuno-localizations (Ueno & Wakayama 2004). Transcriptome analysis revealed, that the terrestrial culms have a stronger C_4 cycle signature than the aquatic culms (Figure 6A). The small subunit of the RUBISCO is two-fold more highly expressed in the aquatic culms (Dataset 1) and the large RuBisCO subunit is present in BSCs as well as MCs (Ueno & Wakayama 2004). However, the aquatic culms show an atypically high

expression of transcripts related to photorespiration for either aquatically adapted (Mommer et al. 2006) or C_4-like plants (Mallmann et al. 2014); Figure 5A). Especially, SHM and GDC subunits are significantly up-regulated in the aquatic culms (Figure 7). Typically, underwater photosynthesis in non-acclimated terrestrial plants is characterized by high photorespiratory rates, as reduced gas diffusion rates under water will lead to relatively low internal CO_2 concentrations compared with the internal oxygen concentrations in the presence of light (Jahnke et al. 1991; Maberly & Spence 1989). Similar conditions can occur for aquatic adapted plants, when they grow a dense canopy, which leads CO_2 depletion and O_2 super-saturation of the water during the day time (Keeley 1999). In the genus Flaveria, C_3- C_4 intermediate species have been reported to display a similar photorespiratory signature as C_3 species (Mallmann et al. 2014), while maintaining the C_4 cycle in parallel. In evolutionary terms, the establishment of a photorespiratory CO_2 pump–also termed as C_2 photosynthesis- is thought to be a necessary step towards C_4-ness (Gowik et al. 2011; Heckmann et al. 2013; Mallmann et al. 2014; Sage 2004; Schulze et al. 2013).
In aquatic E. retroflexa culms this high photorespiratory signature might stem from two factors: (i) abolishment of the strict RuBisCO compartmentalization to the BSCs in the aquatic culms as seen in the terrestrial culms (Ueno & Wakayama 2004) and (ii) dense vegetation of E. retroflexa plants leading to CO_2 depletion under water. A possible explanation for the photorespiratory signature not being associated with an apparent fitness penalty in aquatic E. retroflexa culms could be that the photorespiratory cycle is used for an efficient CO_2 re-fixation and balancing O_2 and CO_2 availability at the site of RuBisCO, which arises from higher-diffusion resistance for CO_2 uptake and continuous photosynthesis action. Thus, these plants might complement thereby the C_4 cycle possibly by using photorespiration for cycling each intercellular CO_2 until it is fixed in form of carbon compounds, as it is known from C_3- C_4 intermediate species.
Besides the genus Eleocharis, the monocotyledonous Orcuttiea family has amphibious C_4 species (Keeley 1998). When grown on soil, these species perform C_4 photosynthesis and when submerged they switch to C_4-like photosynthesis without classic Kranz anatomy (Keeley 1998). Single-cell C_4 photosynthesis has been also found in facultative aquatic species, e.g. Hydrilla and Egeria, under limited CO_2 availability and warm water temperatures (Bowes et al. 2002; Casati et al. 2000; Reiskind et al. 1997). To date, no examples of classic two-cell BSC/MC C_4 photosynthesis have been discovered for an aquatic plant species. With the variety of mechanisms evolvedto circumvent the gas diffusion resistance and optimize CO_2 fixation, one may wonder whether performing two-cell C_4 photosynthesis is actually feasible under water.

Conclusions

In this study, we present an in depth analysis of E. retroflexa transcriptional acclimatization to terrestrial and aquatic habitats. The assembly of the transcriptome provides a unigene set for further molecular studies. The transcriptomes of the terrestrial and aquatic culms enabled a detailed analysis of E. retroflexa's full C_4 cycle, carbon concentrating mechanism and metabolism under different growth condition. The aquatic E. retroflexa transcriptome reflects many traits known for other heterophyllous aquatic plant species. E. retroflexa is surprisingly flexible in its usage of the C_4 cycle and reacts fast to micro-environmental changes, such as water deprivation. While classic Kranz anatomy is lost under water, E retroflexa possibly uses a C_2-like photorespiratory cycle to supplement the C_4 cycle as seen in C_3- C_4 intermediate plant species.

8. Acknowledgements

We thank Katrin L. Weber and Elizabeth Klemp for excellent technical support for the $\delta\ ^{13}C$ analysis. Work in the authors' laboratory was supported by grants of the Deutsche Forschungsgemeinschaft (EXC 1028, IRTG 1525, WE 2231/8-2, and WE 2231/9-2 to APMW). We are grateful to the HHU Biomedical Research Center (BMFZ) for support with RNA-Seq analysis

and to the MSU High Performance ComputingCluster (HPCC) for support with computational analysis of RNA-Seq data.

Author Contributions

C.K. performed experimental work, analyzed data and wrote the paper; S.S. set up growth conditions for plants and cultivated plants, took photographic images of plants and performed CAP3 assembly; M.S. performed analysis of transcriptome variability; A.K.D. performed relative cumulative expression and Edge R analyses; A.H. assisted with set up of growth conditions in aquaria; C.R.B assisted in data analysis; A.B. co-wrote the paper, assisted in data analysis and experimental design; A.P.M.W. designed study and co-wrote the paper.

9. References

Agarie S., Kai M., Takatsuji H. & Ueno O. (2002) Environmental and hormonal regulation of gene expression of C_4 photosynthetic enzymes in the amphibious sedge Eleocharis vivipara. Plant Science, 163, 571-580.

Ashton A., Burnell J., Furbank R., Jenkins C. & Hatch M. (1990) Enzymes of C_4 photosynthesis. Methods in Plant Biochemistry, Volume 3,.(Ed. PJ Lea) pp. 39-72. Academic Press: San Diego, CA.

Bailey-Serres J. & Voesenek L.A. (2008) Flooding stress: acclimations and genetic diversity. Annual Review of Plant Biology, 59, 313-339.

Bailey S., Horton P. & Walters R.G. (2004) Acclimation of Arabidopsis thaliana to the light environment: the relationship between photosynthetic function and chloroplast composition. Planta, 218, 793-802.

Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H.Z., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi N. & Yeh L.S.L. (2005) The universal protein resource (UniProt). Nucleic Acids Research, 33, D154-D159.

Benjamini Y. & Hochberg Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological, 57, 289-300.

Bennetzen J.L., Schmutz J., Wang H., Percifield R., Hawkins J., Pontaroli A.C., Estep M., Feng L., Vaughn J.N., Grimwood J., Jenkins J., Barry K., Lindquist E., Hellsten U., Deshpande S., Wang X., Wu X., Mitros T., Triplett J., Yang X., Ye C.-Y., Mauro-Herrera M., Wang L., Li P., Sharma M., Sharma R., Ronald P.C., Panaud O., Kellogg E.A., Brutnell T.P., Doust A.N., Tuskan G.A., Rokhsar D. & Devos K.M. (2012) Reference genome sequence of the model plant Setaria. Nat Biotech, 30, 555-561.

Besnard G., Muasya A.M., Russier F., Roalson E.H., Salamin N. & Christin P.-A. (2009) Phylogenomics of C_4 Photosynthesis in Sedges (Cyperaceae): Multiple Appearances and Genetic Convergence. Molecular Biology and Evolution, 26, 1909-1919.

Blankenberg D., Gordon A., Von Kuster G., Coraor N., Taylor J., Nekrutenko A. & Galaxy T. (2010) Manipulation of FASTQ data with Galaxy. Bioinformatics, 26, 1783-1785.

Boeger M.R.T. & Poulson M.E. (2003) Morphological adaptations and photosynthetic rates of amphibious Veronica anagallis-aquatica L. (Scrophulariaceae) under different flow regimes. Aquatic Botany, 75, 123-135.

Bowes G., Ogren W.L. & Hageman R.H. (1971) Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase. Biochemical and Biophysical Research Communications, 45, 716-722.

Bowes G., Rao S.K., Estavillo G.M. & Reiskind J.B. (2002) C_4 mechanisms in aquatic angiosperms: comparisons with terrestrial C_4 systems. Functional Plant Biology, 29, 379-392.

Bräutigam A., Kajala K., Wullenweber J., Sommer M., Gagneul D., Weber K.L., Carr K.M., Gowik U., Mass J., Lercher M.J., Westhoff P., Hibberd J.M. & Weber A.P.M. (2010) An mRNA Blueprint for C_4 Photosynthesis Derived from Comparative Transcriptomics of Closely Related C_3 and C_4 Species. Plant Physiology, 155, 142-156.

Bräutigam A., Schliesky S., Külahoglu C., Osborne C.P. & Weber A.P. (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.

Casati P., Lara M.V. & Andreo C.S. (2000) Induction of a C_4-like mechanism of CO_2 fixation in Egeria densa, a submersed aquatic species. Plant Physiology, 123, 1611-1621.

Centritto M., Loreto F. & Chartzoulakis K. (2003) The use of low CO_2 to estimate diffusional and non-diffusional limitations of photosynthetic capacity of salt-stressed olive saplings. Plant Cell and Environment, 26, 585-594.

Cernusak L.A., Ubierna N., Winter K., Holtum J.A., Marshall J.D. & Farquhar G.D. (2013) Environmental and physiological determinants of carbon isotope discrimination in terrestrial plants. The New phytologist, 200, 950-965.

Chen T., Zhu X.-G. & Lin Y. (2014) Major alterations in transcript profiles between C_3- C_4 and C_4 photosynthesis of an amphibious species Eleocharis baldwinii. Plant Molecular Biology, 86, 93-110.

Cheng S., van den Bergh E., Zeng P., Zhong X., Xu J., Liu X., Hofberger J., de Bruijn S., Bhide A.S., Kuelahoglu C., Bian C., Chen J., Fan G., Kaufmann K., Hall J.C., Becker A., Braeutigam A., Weber A.P.M., Shi C., Zheng Z., Li W., Lv M., Tao Y., Wang J., Zou H., Quan Z., Hibberd J.M., Zhang G., Zhu X.-G., Xu X. & Schranz M.E. (2013) The Tarenaya hassleriana Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers. Plant Cell, 25, 2813-2830.

Chu C., Dai Z., Ku M.S. & Edwards G.E. (1990) Induction of Crassulacean Acid Metabolism in the Facultative Halophyte Mesembryanthemum crystallinum by Abscisic Acid. Plant Physiology, 93, 1253-1260.

Colmer T.D. & Voesenek L.A.C.J. (2009) Flooding tolerance: suites of plant traits in variable environments. Functional Plant Biology, 36, 665-681.

Coplen T.B., Brand W.A., Gehre M., Groning M., Meijer H.A.J., Toman B. & Verkouteren R.M. (2006) New guidelines for delta C-13 measurements. Analytical Chemistry, 78, 2439-2441.

De Vos M., Van Oosten V.R., Van Poecke R.M., Van Pelt J.A.,Pozo M.J., Mueller M.J., Buchala A.J., Metraux J.P., Van Loon L.C., Dicke M. & Pieterse C.M. (2005) Signal signature and transcriptome changes of Arabidopsis during pathogen and insect attack. Molecular plant-microbe interactions, 18, 923-937.

Edwards G.E. & Ku M.S. (1987) Biochemistry of C_3- C_4 intermediates. New York: Academic Press, In M.D. Hatch & N.K. Boardmann (Eds.), 275-325.

Frost-Christensen H., JØrgensen L.B. & Floto F. (2003) Species specificity of resistance to oxygen diffusion in thin cuticular membranes from amphibious plants. Plant, Cell & Environment, 26, 561-569.

Frost-Christensen H. & Sand-Jensen K. (1995) Comparative kinetics of photosynthesis in floating and submerged Potamogeton leaves. Aquatic Botany, 51, 121-134.

Furbank R.T. & Hatch M.D. (1987) Mechanism of C_4 photosynthesis: the size and composition of the inorganic carbon pool in bundle sheath cells. Plant Physiology, 85, 958-964.

Ge X., Chen H., Wang H., Shi A. & Liu K. (2014) De novo assembly and annotation of Salvia splendenstranscriptome using the Illumina platform. PLoS ONE, 9, e87693.

Gowik U., Bräutigam A., Weber K.L., Weber A.P. & Westhoff P. (2011) Evolution of C_4 photosynthesis in the genus Flaveria: how many and which genes does it take to make C_4? The Plant Cell, 23, 2087-2105.

Haberlandt G. (1904) Physiologische Pflanzenanatomie. W. Engelmann.

Hatch M.D. (1987) C_4 photosynthesis – a unique blend of modified biochemistry, anatomy and

ultrastructure. Biochimica Et Biophysica Acta, 895, 81-106.
Hatch M.D. & Slack C.R. (1970) Photosynthetic CO_2-Fixation Pathways. Annual Review of Plant Physiology, 21, 141-161.
Heckmann D., Schulze S., Denton A., Gowik U., Westhoff P., Weber A.P. & Lercher M.J. (2013) Predicting C_4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. Cell, 153, 1579-1588.
Holmes M. & Klein W. (1987) The light and temperature environments. SPEC. PUBL. BR. ECOL. SOC. 1987.
Huang X.Q. & Madan A. (1999) CAP3: A DNA sequence assembly program. Genome Research, 9, 868-877.
Inda L.A., Torrecilla P., Catalán P. & Ruiz-Zapata T. (2008) Phylogeny of Cleome L. and its close relatives Podandrogyne Ducke and Polanisia Raf. (Cleomoideae, Cleomaceae) based on analysis of nuclear ITS sequences and morphology. Plant Systematics and Evolution, 274, 111-126.
Jackson M.B. (2008) Ethylene-promoted elongation: An adaptation to submergence stress. Annals of Botany, 101, 229-248.
Jahnke L.S., Eighmy T.T. & Fagerberg W.R. (1991) Studies of Elodea nuttallii grown under photorespiratory conditions. 1. Photosynthetic characteristics. Plant Cell and Environment, 14, 147-156.
Keeley J.E. (1998) C_4 photosynthetic modifications in the evolutionary transition from land to water in aquatic grasses. Oecologia, 116, 85-97.
Keeley J.E. (1999) Photosynthetic pathway diversity in a seasonal pool community. Functional Ecology, 13, 106-118.
Kent W.J. (2002) BLAT--the BLAST-like alignment tool. Genome Research, 12, 656-664.
Kim J.M., To T.K., Nishioka T. & Seki M. (2010) Chromatin regulation functions in plant abiotic stress responses. Plant, Cell & Environment, 33, 604-611.
Kirk J.T.O. (1994) Light and photosynthesis in aquatic ecosystems. Cambridge university press.
Krause-Jensen D. & Sand-Jensen K. (1998) Light attenuation and photosynthesis of aquatic plant communities. Limnology and Oceanography, 43, 396-407.
Külahoglu C., Denton A.K., Sommer M., Mass J., Schliesky S., Wrobel T.J., Berckmans B., Gongora-Castillo E., Buell C.R., Simon R., De Veylder L., Bräutigam A. & Weber A.P. (2014) Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C_3 and C_4 Plant Species. The Plant Cell, 26, 3243-3260.
Kuwabara A., Tsukaya H. & Nagata T. (2001) Identification of factors that cause heterophylly in Ludwigia arcuata Walt. (Onagraceae). Plant Biology, 3, 670-670.
Lamesch P., Berardini T.Z., Li D., Swarbreck D., Wilks C., Sasidharan R., Muller R., Dreher K., Alexander D.L., Garcia-Hernandez M., Karthikeyan A.S., Lee C.H., Nelson W.D., Ploetz L., Singh S., Wensel A., Huala E. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Research, 40, D1202-1210.
Langdale J.A. & Kidner C.A. (1994) Bundle-sheath defective, a mutation that disrupts cellular differentiation in maize leaves. Development, 120, 673-681.
Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. & Higgins D.G. (2007) Clustal W and clustal X version 2.0. Bioinformatics, 23, 2947-2948.
Lee S.C., Mustroph A., Sasidharan R., Vashisht D., PedersenO., Oosumi T., Voesenek L.A. & Bailey-Serres J. (2011) Molecular characterization of the submergence response of the Arabidopsis thaliana ecotype Columbia. The New phytologist, 190, 457-471.
Lee W.Y., Lee D., Chung W.-I. & Kwon C.S. (2009) Arabidopsis ING and Alfin1-like protein families localize to the nucleus and bind to H3K4me3/2 via plant homeodomain fingers. Plant Journal, 58, 511-524.
Long S.P. & Bernacchi C.J. (2003) Gas exchange measurements, what can they tell us about the underlying limitations to photosynthesis? Procedures and sources of error. Journal of

Experimental Botany, 54, 2393-2401.
Maberly S.C. & Madsen T.V. (2002) Freshwater angiosperm carbon concentrating mechanisms: processes and patterns. Functional Plant Biology, 29, 393-405.
Maberly S.C. & Spence D.H.N. (1989) Photosynthesis and photorespiration in fresh-water organisms – amphibious plants. Aquatic Botany, 34, 267-286.
Madsen T.V. & Sand-Jensen K. (1991) Photosynthetic carbon assimilation in aquatic macrophytes. Aquatic Botany,41, 5-40.
Mallmann J., Heckmann D., Bräutigam A., Lercher M.J., Weber A.P., Westhoff P. & Gowik U. (2014) The role of photorespiration during the evolution of C_4 photosynthesis in the genus Flaveria. eLife, e02478.
Martin J.A. & Wang Z. (2011) Next-generation transcriptome assembly. Nature Reviews Genetics, 12, 671-682.
McElwain E.F., Bohnert H.J. & Thomas J.C. (1992) Light moderates the induction of phosphoenolpyruvate carboxylase by NaCl and abscisic-acid in Mesembryanthemum crystallinum. Plant Physiology, 99, 1261-1264.
Micheli F. (2001) Pectin methylesterases: cell wall enzymes with important roles in plant physiology. Trends in Plant Science, 6, 414-419.
Minorsky P.V. (2003) The hot and the classic. Plant Physiology, 133, 1671-1672.
Mommer L., de Kroon H., Pierik R., Bogemann G.M. & Visser E.J.W. (2005a) A functional comparison of acclimation to shade and submergence in two terrestrial plant species. New Phytologist, 167, 197-206.
Mommer L., Pons T.L. & Visser E.J. (2006) Photosynthetic consequences of phenotypic plasticity in response to submergence: Rumex palustris as a case study. Journal of Experimental Botany, 57, 283-290.
Mommer L., Pons T.L., Wolters-Arts M., Venema J.H. & Visser E.J. (2005b) Submergence-induced morphological, anatomical, and biochemical responses in a terrestrial species affect gas diffusion resistance and photosynthetic performance. Plant Physiology, 139, 497-508.
Mommer L. & Visser E.J. (2005) Underwater photosynthesis in flooded terrestrial plants: a matter of leaf plasticity. Annals of Botany, 96, 581-589.
Moriya Y., Itoh M., Okuda S., Yoshizawa A.C. & Kanehisa M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research, 35, W182-W185.
Mustroph A., Zanetti M.E., Jang C.J., Holtan H.E., Repetti P.P., Galbraith D.W., Girke T. & Bailey-Serres J. (2009) Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America, 106, 18843-18848.
Papanicolaou A., Stierli R., ffrench-Constant R.H. & Heckel D.G. (2009) Next-generation transcriptomes for next-generation genomes using est2assembly. BMC Bioinformatics, 10, 447.
Pedersen O., Colmer T.D. & Sand-Jensen K. (2013) Underwater photosynthesis of submerged plants - recent advances and methods. Frontiers in plant science, 4, 140.
Pfaffl M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Research, 29.
Pigliucci M. (2001) Phenotypic plasticity: beyond nature and nurture. Syntheses in ecology and evolution.
Porra R.J., Thompson W.A. & Kriedemann P.E. (1989) Determination of accurate extinction coefficients and simultaneous-equations for assaying chlorophyll-a and chlorophyll-b extracted with 4 different solvents – verification of the concentration of chlorophyll standards by atomic-absorption spectroscopy. Biochimica Et Biophysica Acta, 975, 384-394.
R Development Core Team (2009). R: A Language and Environment for Statistical Computing. (Vienna, Austria: R Foundation for Statistical Computing).
Reiskind J.B., Madsen T.V., Van Ginkel L.C. & Bowes G. (1997) Evidence that inducible C_4-type photosynthesis is a chloroplastic CO_2-concentrating mechanism in Hydrilla, a submersed

monocot. Plant, Cell & Environment, 20, 211-220.
Robinson M.D., McCarthy D.J. & Smyth G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139-140.
Rossini L., Cribb L., Martin D.J. & Langdale J.A. (2001) The maize Golden2 gene defines a novel class of transcriptional regulators in plants. Plant Cell, 13, 1231-1244.
Saeed A.I., Hagabati N.K., Braisted J.C., Liang W., Sharov V., Howe E.A., Li J., Thiagarajan M., White J.A. & Quackenbush J. (2006) TM4 microarray software suite. In: DNA Microarrays, Part B: Databases and Statistics (eds A. Kimmel & B. Oluver), pp. 134-9.
Saeed A.I., Sharov V., White J., Li J., Liang W., Bhagabati N., Braisted J., Klapa M., Currier T., Thiagarajan M., Sturn A., Snuffin M.,Rezantsev A., Popov D., Ryltsov A., Kostukovich E., Borisovsky I., Liu Z., Vinsavich A., Trush V. & Quackenbush J. (2003) TM4: A free, open-source system for microarray data management and analysis. Biotechniques, 34, 374-8.
Sage R.F. (2004) The evolution of C-4 photosynthesis. New Phytologist, 161, 341-370.
Sage R.F. & McKown A.D. (2006) Is C_4 photosynthesis less phenotypically plastic than C_3 photosynthesis? Journal of Experimental Botany, 57, 303-317.
Sage R.F., Sage T.L. & Kocacinar F. (2012) Photorespiration and the Evolution of C-4 Photosynthesis. In: Annual Review of Plant Biology, Vol 63 (ed S.S. Merchant), pp. 19-47.
Sand-Jensen K. (1989) Environmental variables and their effect on photosynthesis of aquatic plant communities. Aquatic Botany, 34, 5-25.
Sand-Jensen K. & Frost-Christensen H. (1999) Plant growth and photosynthesis in the transition zone between land and stream. Aquatic Botany, 63, 23-35.
Schliesky S., Gowik U., Weber A.P. & Brautigam A. (2012) RNA-Seq Assembly - Are We There Yet? Frontiers in plant science, 3, 220.
Schmittgen T.D. & Livak K.J. (2008) Analyzing real-time PCR data by the comparative C-T method. Nature protocols, 3, 1101-1108.
Schulze S., Mallmann J., Burscheidt J., Koczor M., Streubel M., Bauwe H., Gowik U. & Westhoff P. (2013) Evolution of C-4 Photosynthesis in the Genus Flaveria: Establishment of a Photorespiratory CO_2 Pump. Plant Cell, 25, 2522-2535.
Sculthorpe C. (1967) The biology of vascular plants. Edward Arnold, London, 610.
Setter T.L. & Laureles E.V. (1996) The beneficial effect of reduced elongation growth on submergence tolerance of rice. Journal of Experimental Botany, 47, 1551-1559.
Sommer M., Brautigam A. & Weber A.P. (2012) The dicotyledonous NAD malic enzyme C_4 plant Cleome gynandra displays age-dependent plasticity of C_4 decarboxylation biochemistry. Plant Biology, 14, 621-629.
Thimm O., Blasing O., Gibon Y., Nagel A., Meyer S., Kruger P., Selbig J., Muller L.A., Rhee S.Y. & Stitt M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant Journal, 37, 914-939.
Tsuji H., Saika H., Tsutsumi N., Hirai A. & Nakazono M. (2006) Dynamic and reversible changes in histone H3-Lys4 methylation and H3 acetylation occurring at submergence-inducible genes in rice. Plant and Cell Physiology, 47, 995-1003.
Ueno O. (2001) Environmental Regulation of C_3 and C_4 Differentiation in the Amphibious Sedge Eleocharis vivipara. Plant Physiology, 127, 1524-1532.
Ueno O. (2004) Environmental regulation of photosynthetic metabolism in the amphibious sedge Eleocharis baldwinii and comparisons with related species. Plant Cell and Environment, 27, 627-639.
Ueno O., Samejima M. & Koyama T. (1989) DISTRIBUTION AND EVOLUTION OF C-4 SYNDROME IN ELEOCHARIS, A SEDGE GROUP INHABITING WET AND AQUATIC ENVIRONMENTS, BASED ON CULM ANATOMY AND CARBON ISOTOPE RATIOS. Annals of Botany, 64, 425-438.
Ueno O., Samejima M., Muto S. & Miyachi S. (1988) Photosynthetic characteristics of an amphibious plant, Eleocharis vivipara – Expression of C_4 and C_3 modes in constrasting environments. Proceedings of the National Academy of Sciences of the United States of

America, 85, 6733-6737.
Ueno O. & Wakayama M. (2004) Cellular expression of C_3 and C_4 photosynthetic enzymes in the amphibious sedge Eleocharis retroflexa ssp. chaetaria. Journal of plant research, 117, 433-441.
Usadel B., Nagel A., Thimm O., Redestig H., Blaesing O.E., Palacios-Rojas N., Selbig J., Hannemann J., Piques M.C., Steinhauser D., Scheible W.R., Gibon Y., Morcuende R., Weicht D., Meyer S. & Stitt M. (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of coresponding genes, and comparison with known responses. Plant Physiology, 138, 1195-1204.
Vervuren P.J.A., Blom C. & de Kroon H. (2003) Extreme flooding events on the Rhine and the survival and distribution of riparian plant species. Journal of Ecology, 91, 135-146.
Wang Y., Pan Y., Liu Z., Zhu X., Zhai L., Xu L., Yu R., Gong Y. & Liu L. (2013) De novo transcriptome sequencing of radish (Raphanus sativus L.) and analysis of major genes involved inglucosinolate metabolism. BMC Genomics, 14, 836.
Waters M.T., Wang P., Korkaric M., Capper R.G., Saunders N.J. & Langdale J.A. (2009) GLK transcription factors coordinate expression of the photosynthetic apparatus in Arabidopsis. The Plant Cell, 21, 1109-1128.
Yekutieli D. & Benjamini Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference, 82, 171-196.

10. Tables

Table 1. Sequencing and mapping statistics and transcriptome dynamics of E. retroflexa read samples aligned to the S. italica genome and the E. retroflexa transcriptome assembly.

Eleocharis samples	Aquatic 1	Aquatic 2	Terrestrial 1	Terrestrial 2	Terrestrial 3
Raw reads	30469686	34239945	22011612	30043954	25010427
Cleaned reads	28203253	33838855	21729249	29391792	24629693
Mapped reads to S. italica	7768905	12337799	7648795	8182178	7585432
Mapped reads to unigene database (>200 bases)	7593124	23674037	18779815	24787798	17606148
Mapping efficiency to S. italica	28	36	35	28	31
Mapping efficiency to unigenes	27	70	86	84	71
Number of S. italica IDs >20 RPM	4882	5263	5826	5807	5495
Number of S. italica IDs > 1,000 RPM	132	143	135	136	136
Number of S. italica ID matching	19298	20248	19893	20041	19814
Number of unigenes matching	34971	38548	38324	38352	38489
S.italica IDs covered by reads (%)	54.4	57.1	56.1	56.5	55.9
E. retroflexa unigenes covered by reads (%)	79.8	88	87.5	87.5	87.8
Transcript number >0 RPM aligned to S. italica	19298	20248	19893	20041	19814
Transcript number >1 RPM aligned to S. italica	14875	15378	16001	15910	15654
Transcript number >20 RPM aligned to S. italica	4882	5263	5826	5807	5495
Transcript number >1,000 RPM aligned to S. italica	132	143	135	136	136

11. Figure legends

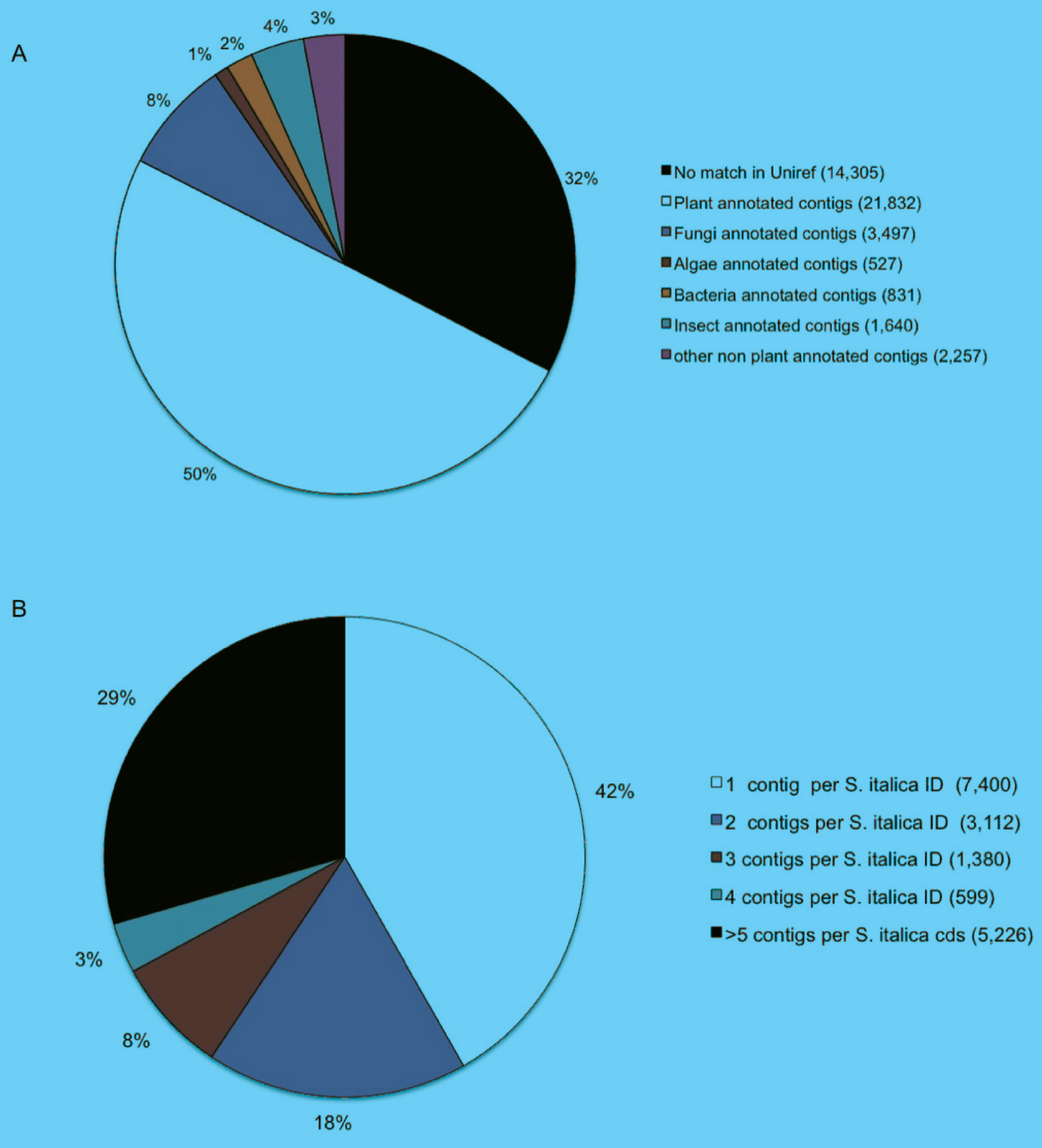


Figure 1. Annotation of E. retroflexa contigs.
(A) Annotation of contigs against Uniref 100. Distribution of contigs annotated by Uniref100 based on major taxonomic categories of plant, bacteria, algae, fungi and other non-plant annotated contigs. Total contig number is indicated in parentheses. **(B)** Annotation of contigs using alignments to the S. italica predicted proteome. The percentage of contigs per unique S. italica protein, showing redundancy in assembled contigs is displayed. The number of best matching contigs per predicted S. italica identifier is indicated in parentheses.

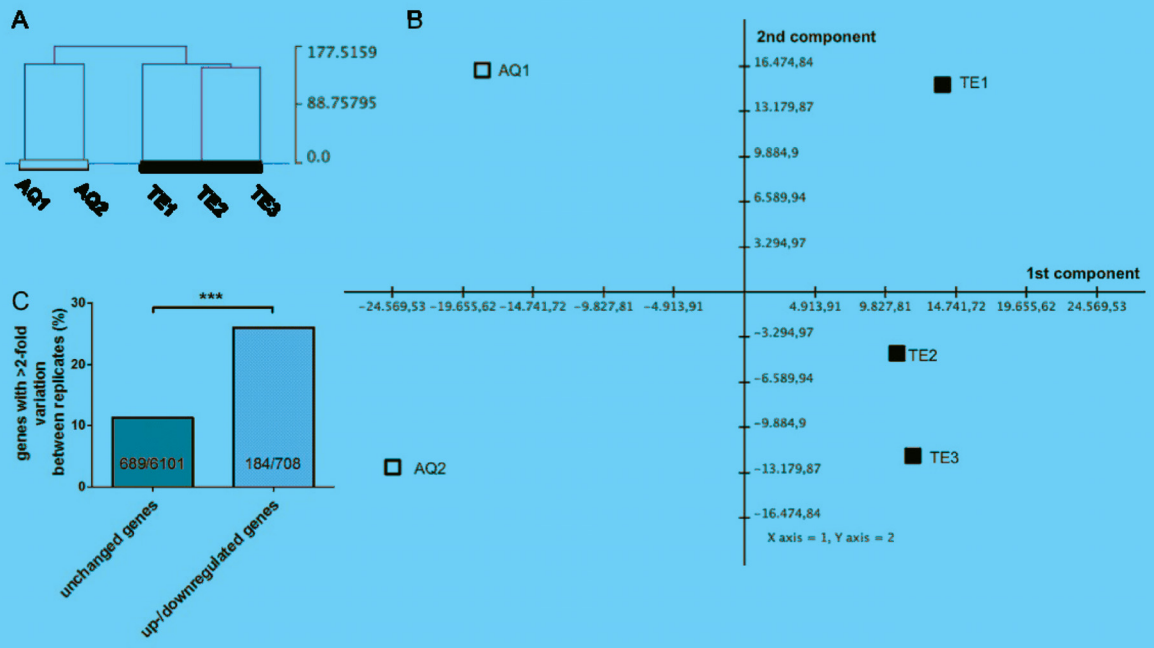


Figure 2. Transcriptome dynamics and variability between samples and habitat.
(A) Hierarchical clustering of all sequenced E. retroflexa samples. E. retroflexa transcriptomes (>1 RPM filtered) were clustered, after normalizing to z-scores per row,with Euclidean distance and average linkage. **(B)** Principle component analysis (PCA) between submersed aquatic and terrestrial E. retroflexa transcriptomes. Plot shows all sequenced samples from aquatic (white) and terrestrial (black) E. retroflexa (n=3; RPM). First component (x-axis) separates samples by habitat (36%) of all data variability, and second component (y-axis) describes biological sample variability (21%) within each growth condition. **(C)** Variance plot between biological replicates and significant changed transcripts.

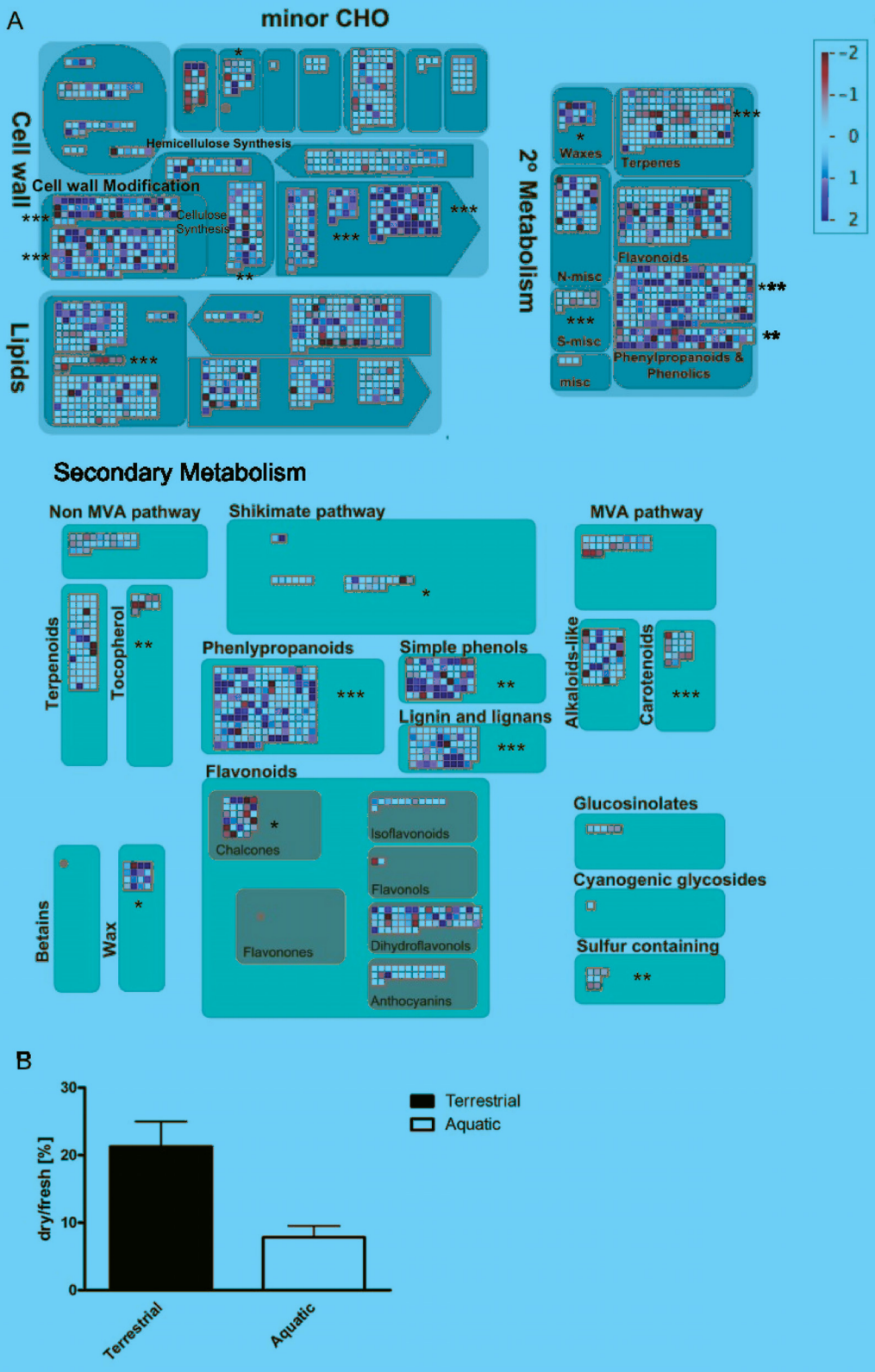

A
minor CHO
Cell wall
Hemicellulose Synthesis
Cell wall Modification
Cellulose Synthesis
Lipids
2° Metabolism
Waxes
Terpenes
N-misc
Flavonoids
S-misc
Phenylpropanoids & Phenolics
misc
~2
~1
0
1
2
Secondary Metabolism
Non MVA pathway
Shikimate pathway
MVA pathway
Terpenoids
Tocopherol
Phenlypropanoids
Simple phenols
Lignin and lignans
Alkaloids-like
Carotenoids
Flavonoids
Chalcones
Isoflavonoids
Flavonols
Flavonones
Dihydroflavonols
Anthocyanins
Glucosinolates
Cyanogenic glycosides
Sulfur containing
Betains
Wax
B
dry/fresh [%]
30
20
10
0
Terrestrial
Aquatic
Terrestrial
Aquatic

Figure 3. Structural differences between submersed aquatic and terrestrial E. retroflexa culms.
(A) Overview of secondary and carbon metabolism transcript levels in E. retroflexa culms. Heatmaps depict log2 fold-changes of submersed aquatic versus terrestrial transcript levels in RPM. Red (ratio<0) represents a decrease of transcript levels in submersed aquatic culms. Blue indicates (ratio>0) an increase of transcript levels in terrestrial culms. Asterisks indicate significant fold-changes calculated by Wilcoxon Rank sum test. Benjamini-Yekutieli (BY) FDR corrected P-values (* P-value<0.05; ** P-value<0.01; *** P-value<0.001). Heatmaps were generated with Mapman tool (Usadel et al. 2005). **(B)** Comparison of water content and biomass between terrestrial (black) and aquatic (white culms) as ratio dry weight (DW) against fresh weight (FW) in percent. n= 3 biological replicates; error bars ± SE, standard error.

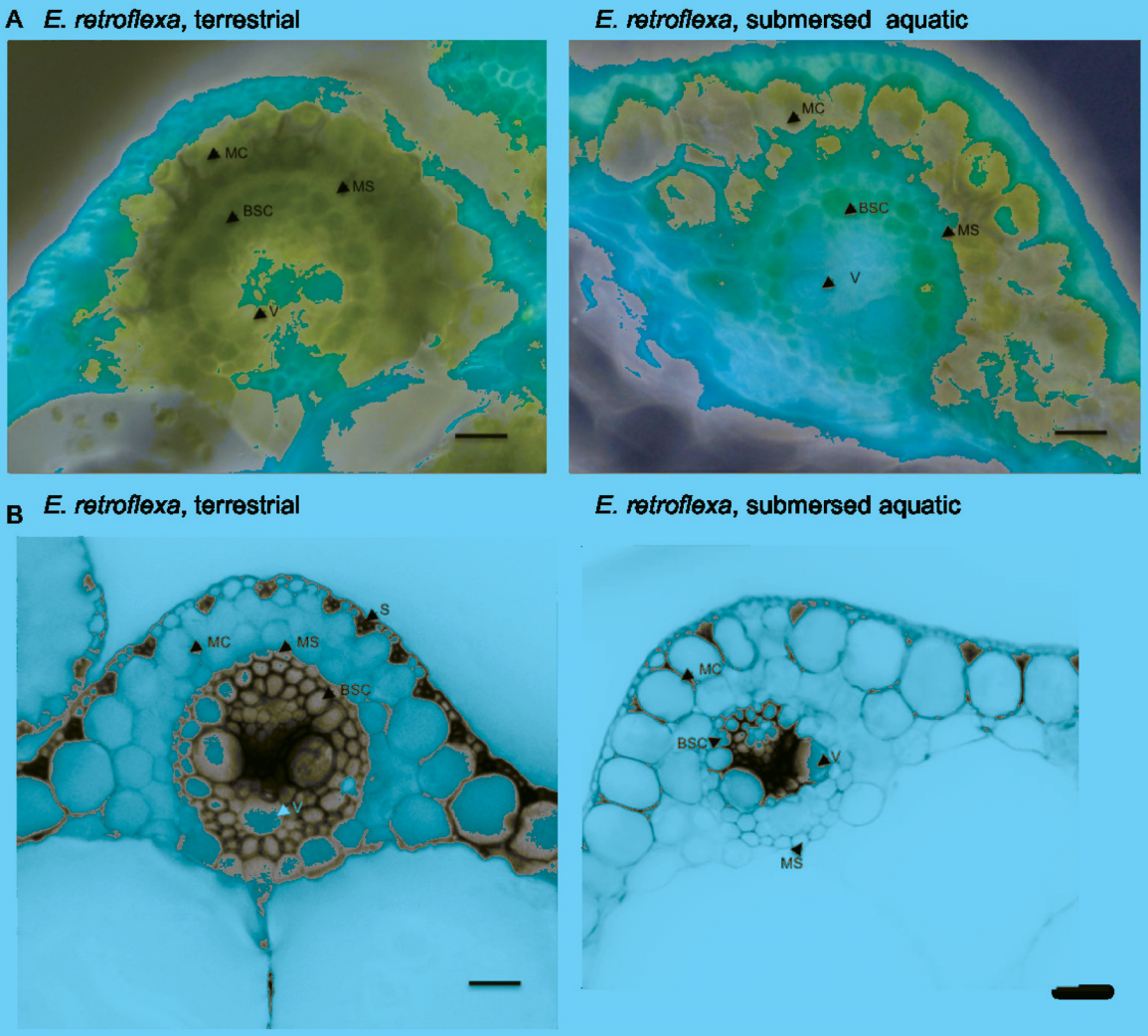


Figure 4. Culm anatomy of mature E. retroflexa plants grown under terrestrial (left) and aquatic (right) conditions.
(A) Microscopic images of cross-sectioned E. retroflexa culms grown either on soil or submersed aquatic conditions. **(B)** Auto-fluorescence microscopic images of E. retroflexa cross-sectioned culms grown either on soil or under water. Scale bar: 20µm. Cell types are indicated by closed arrows. BSC: bundle sheath cell; MC: mesophyll cell; MS: mestome sheath; V: vein.

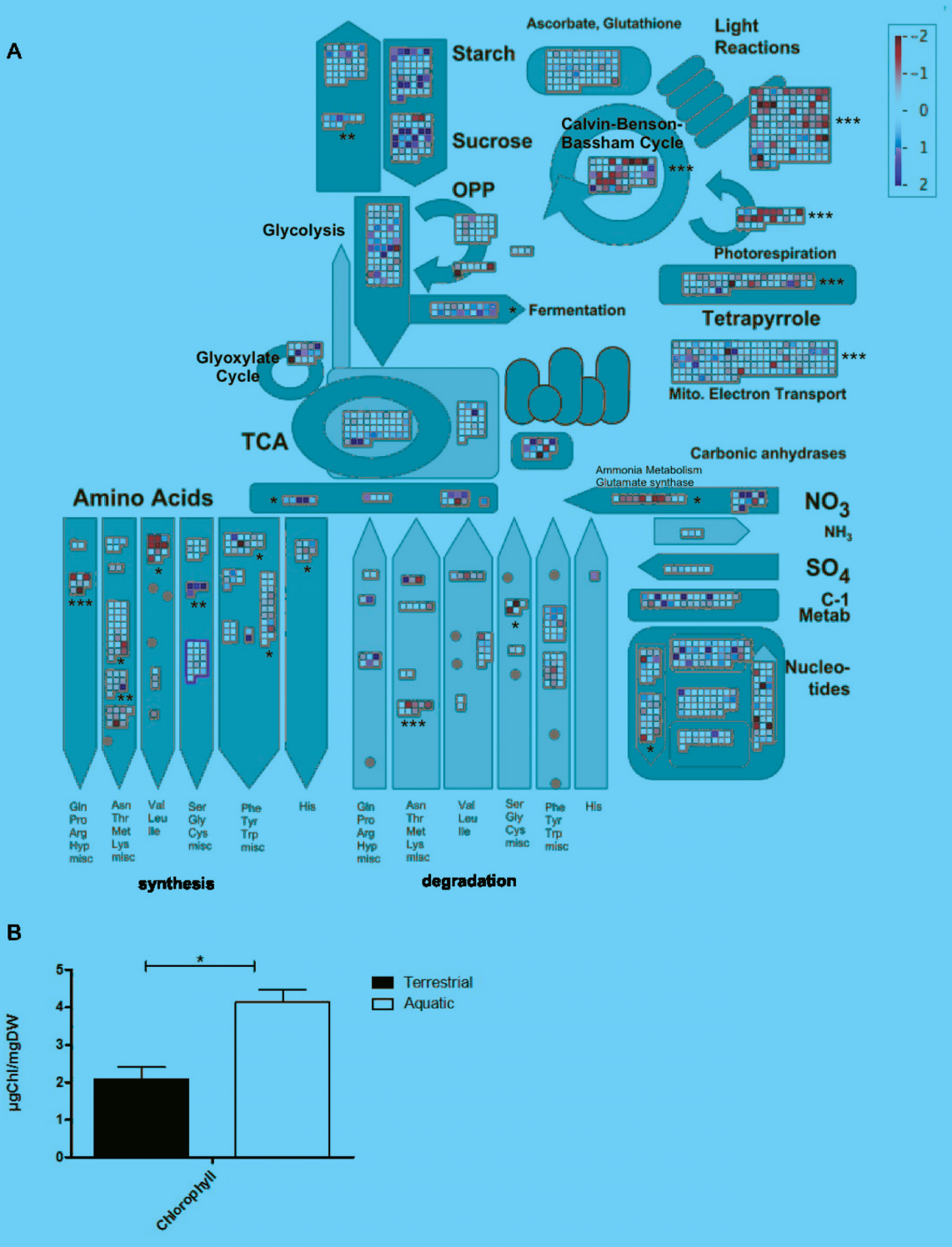


Figure 5. Transcriptional and physiological differences between submerse aquatic and terrestrial E. retroflexa culms.
(A) Overview of central metabolism transcript levels in E. retroflexa. Heatmaps show log2 fold-changes of submerse aquatic versus terrestrial transcript levels in RPM. Red (ratio<0) represents an increase of transcript levels in submersed aquatic and blue (ratio>0) an increase of transcript levels in terrestrial culms. Asterisks indicate significant fold-changes calculated with Wilcoxon Rank Test.

P-values (* P-value<0.05; ** P-value<0.01; *** P-value<0.001) were Benjamini-Yekutieli FDR corrected. Heatmaps were generated with Mapman tool (Usadel et al. 2005). **(B)** Total chlorophyll content in terrestrial (black) and submersed aquatic (white) culms as µg per mg dry weight (DW). n= 4 biological replicates; error bars ± SE, standard error. Asterisks indicate statistically significant differences between terrestrial and submerse aquatic samples (* P-value<0.05).

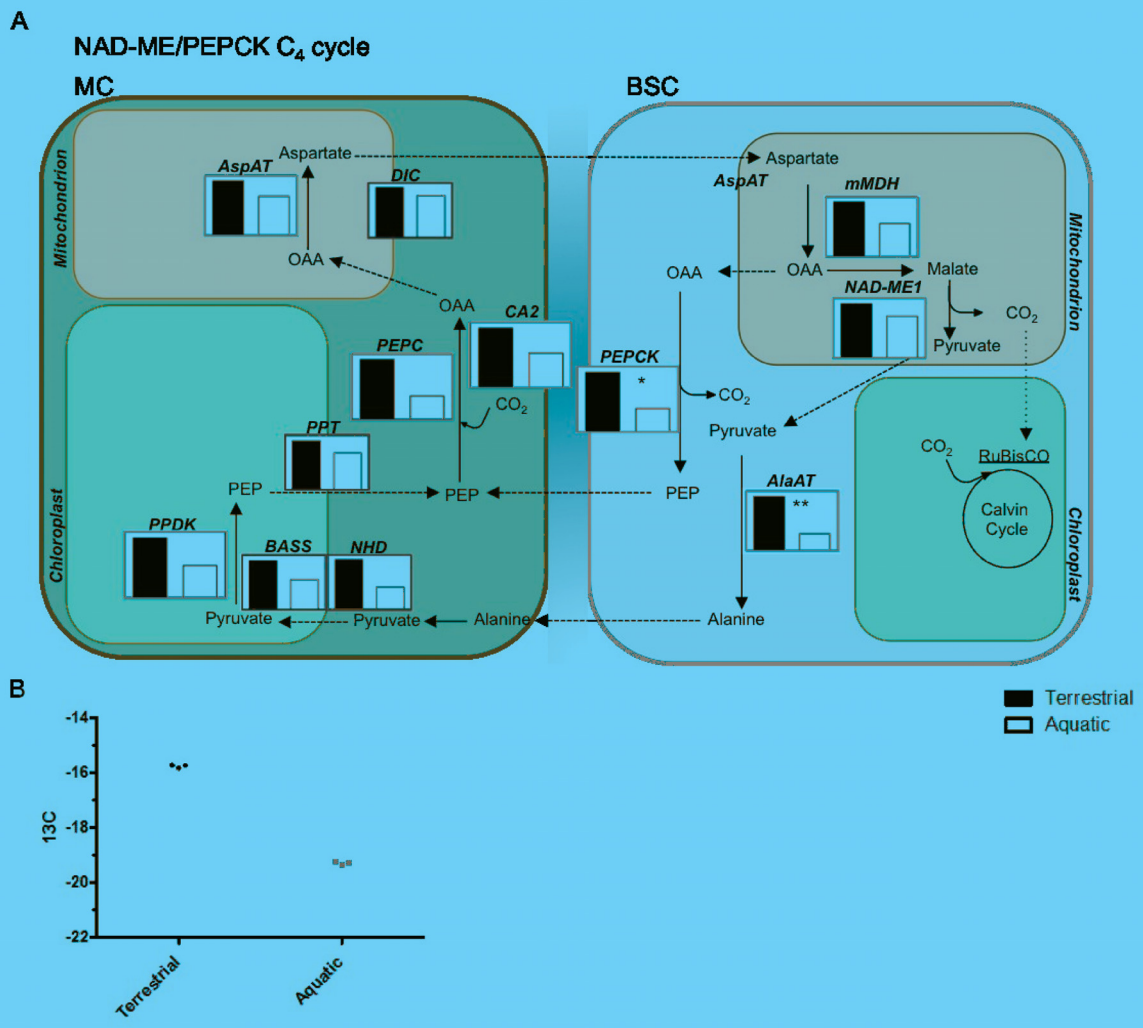


Figure 6. Genes encoding for C_4 photosynthesis are altered between terrestrial and submersed aquatic E. retroflexa culms.
(A) Schematic overview of the NAD-ME/PEPCK C_4 cycle known for C_4 plants (adapted from Sommer et al 2012). Relative transcript abundances between terrestrial (black) and submersed aquatic (white) transcriptomes are shown in small insets and were normalized by setting the highest expressed condition to 1 for each gene. Asterisks denote significant expression changes between aquatic and terrestrial samples (Edge R; FDR Benjamini-Hochberg corrected P-values). ** P-value<0.01; * P-value<0.05. Localization of C_4 enzymes in E. retroflexa is assumed from literature (Ueno et al., 2004). Red boxes indicate relevant C_4 cycle transporter and blue boxes soluble C_4 cycle enzymes. **PEPC:** PHOSPHOENOLPYRUVATE CARBOXYLASE; **CA2**: CARBONIC ANHYDRASE2; **DIC**: DICARBOXYLATE CARRIER; **AspAT**: ASPARTATE AMINOTRANSFERASE; **mMDH**: mitochondrial MALATE DEHYDROGENASE; **NAD-ME1**: NAD-dependent MALIC ENZYME1; **AlaAT**: ALANINE AMINOTRANSFERASE; **PEPCK**: PHOSPHOENOLPYRUVATE CARBOXYKINASE; **BASS**: BILE ACID:SODIUM SYMPORTER; **NHD**: SODIUM:HYDROGEN ANTIPORTER; **PPDK**: PYRUVATE ORTHOPHOSPHATE DIKINASE; **PPT**: PHOSPHATE/PHOSPHOENOLPYRUVATE TRANSLOCATOR **(B)** $^{13}C/^{12}C$ isotope ratio of terrestrial (black) and aquatic (grey) culms; n=3.

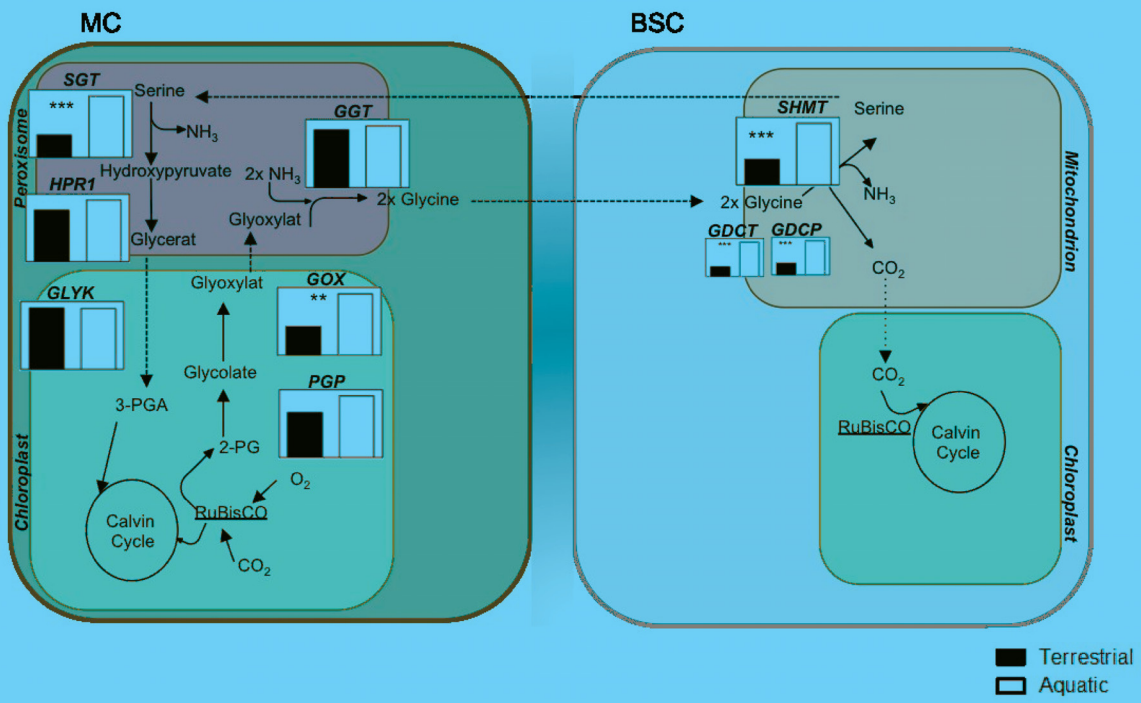


Figure 7. Genes encoding for photorespiration are enhanced in the submersed aquatically grown culms.
Schematic overview of the photorespiratory pathway known for C_3- C_4 photosynthesis intermediate plants (adapted from Gowik et al 2012). Relative transcript abundances between terrestrial (black) and submersed aquatic (white) transcriptomes are shown in small insets and were normalized by setting the highest expressed condition to 1 for each gene. Asterisks denote significant expression changes between aquatic and terrestrial samples (Edge R; Benjamini-Hochberg FDR corrected P-value). ** P-value<0.01; *** P-value<0.001. Localization of photorespiratory enzymes is assumed from literature (reviewed by Sage et al. 2012). **PGP**: 2-PHOSPHOGLYCOLATE PHOSPHATASE; **GOX**: GLYCOLATE OXIDASE; **GGT**: GLUTAMATE:GLYOXYLATE OXIDASE; **GDC**: GLYCINE DECARBOXYLASE; **SHM**: SERINE HYDROXYLMETHYL TRANSFERASE; **SGT**: SERINE:GLYOXYLATE AMINOTRANSFERASE; **HPR1**: NADH-dependent HYDROXYPYRUVATE REDUCTASE; **GLYK**: GLYCERATE KINASE

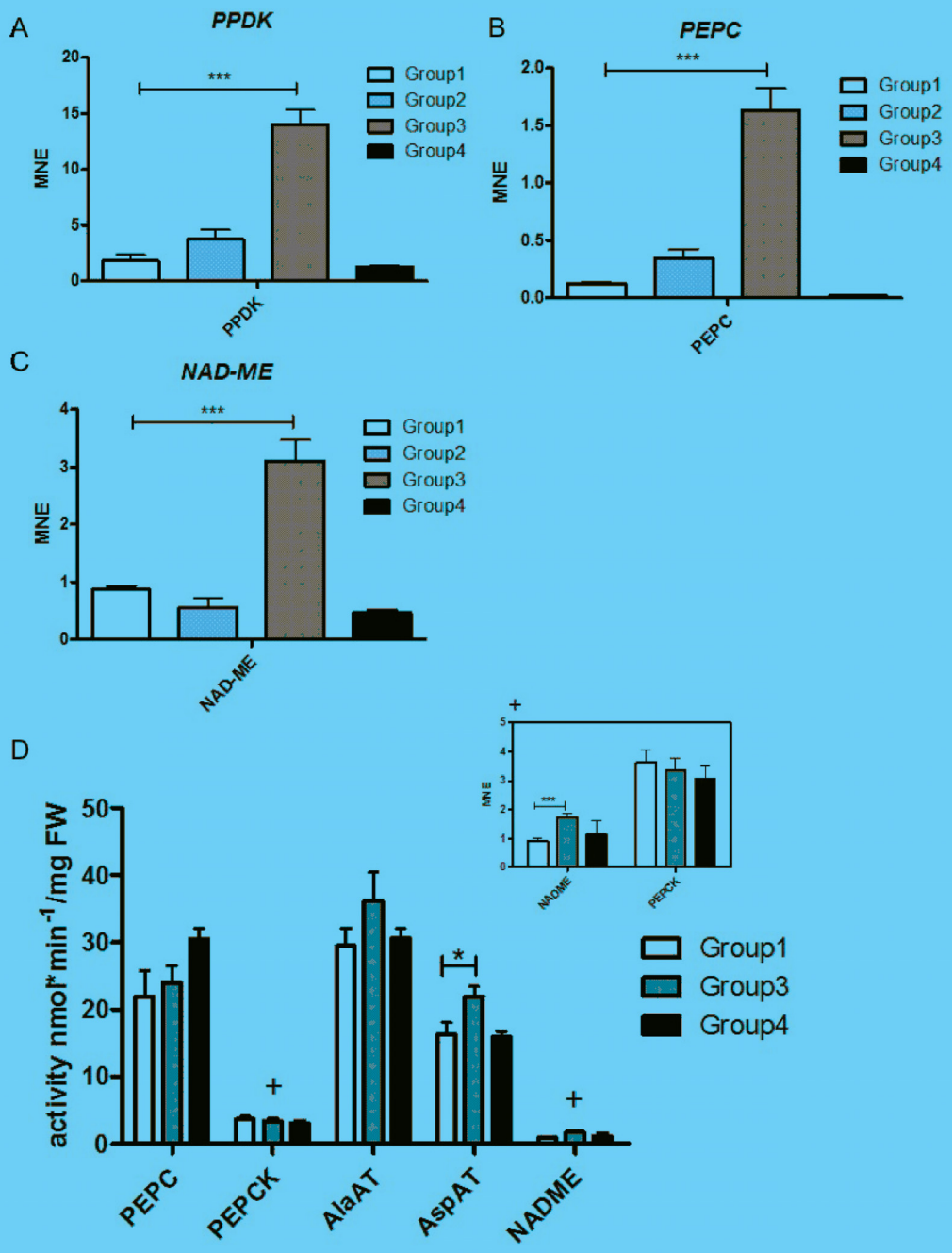


Figure 8. Analysis of metabolic plasticity of E. retroflexa terrestrial culms under increasing drought stress.
(A-C) Transcriptional pattern of select C_4 and photorespiratory genes in water-deprived terrestrially grown E. retroflexa culms. Quantitative real-time PCR was performed on samples from 12 week old terrestrial E. retroflexa culms subjected to drought stress for 14 days (group 1: control, every day 250 ml water (white); group 2: every two days 250ml water (light grey); group 3: every four days 150ml water (dark grey); group 4: no water for 14 days (black)). PPDK **(A)**, PEPC **(B)**, NAD-ME **(C)** were normalized with UBQ10 as housekeeping control. MNE: Mean Normalized Expression; n=3 ± SE, Standard Error. Asterisks indicate statistically significant differences between control and group 3 (***P-Value<0.001). **(D) Enzyme activity measurement of**

soluble C_4 cycle enzymes in water deprived terrestrially grown E. retroflexa culms. Enzyme activities of PEPC, NAD-ME, PEPCK, AspAT, AlaAT were measured from 12 week old terrestrial E. retroflexa culms subjected to drought stress for 14 days (Group 1: control, every day 250ml water (white); every four days 150ml water (dark grey); Group 4: no water for 14 days (black)). FW: fresh weight; error bars ±SE; 3 biological replicates each with 3 technical replicates. Asterisks indicate statistically significant differences between control and group 3 (* P-values<0.01; *** P-Value<0.001). Cross indicates inset.

Datasets

Dataset 1. Annotated transcriptome expression data (RPM) of E. retroflexa submersed aquatic and terrestrial culms.
Dataset 2. CAP3 assembled filtered E. retroflexa unigene database.

13. Supporting Information

Supplemental Datasets

Supplemental Dataset 1. Metabolic pathways covered by E.retroflexa unigene database.
Supplemental Data Set 2. Cellular processes covered by E. retroflexa unigene database.
Supplemental Dataset 3. Regulatory processes covered by E. retroflexa unigene database.
Supplemental Dataset 4. E retroflexa contigs annotated with Uniref100 via tblastx based on highest bitscore.
Supplemental Dataset 5. E. retroflexa contigs annotated against S. italica V2.1 primary transcripts.
Supplemental Dataset 6. Mapman category enrichment analysis (Fisher's Exact test).
Supplemental Dataset 7. Wilcoxon rank sum test of Mapman categories.

Supplemental Figures

Supplemental Figure 1. Histogram of E. retroflexa unigene database.
Supplemental Figure 2. Cumulative relative expression plots of reads mappedagainst various references by BLAT.
Supplemental Figure 3. Comparison of submersed aquatic and terrestrial E. retroflexa transcriptomes.
Supplemental Figure 4. Photographic images of E. retroflexa cultivation.
Supplemental Figure 5. Quantitative transcript abundance patterns between submersed aquatic and terrestrial E. retroflexa culms.
Supplemental Figure 6. Transcriptional investment of submersed aquatic and terrestrial E. retroflexa culms.
Supplemental Figure 7. Transcriptional regulation and plant hormone expression patterns of submersed aquatic or terrestrial E. retroflexa culms.
Supplemental Figure 8. Phylogeny of E. retroflexa based on internal transcribed spacer (ITS) sequences.

Supplemental Tables

Supplemental Table 1. Overview of E. retroflexa qRT-PCR primer.
Supplemental Table 2. Overview of E. retroflexa CAP3 assembly statistics and annotation of contigs.
Supplemental Table 3. Pearson's correlation (r) of transcriptome data.
Supplemental Table 4. Averaged C_4 cycle transcripts (RPMs) of E. retroflexa submersed aquatic and terrestrial transcriptomes

5.2 Manuscript DM: Expression divergence following gene duplication contributes to the evolution of the complex trait C_4 photosynthesis.

Overview

Title: Expression divergence following gene duplication contributes to the evolution of the complex trait C_4 photosynthesis.
Authors: Alisandra K. Denton, Janina Maß, Canan Külahoglu, Martin Lercher, Shin-Han Shiu, Andrea Bräutigam and Andreas P.M. Weber
Submission-ready
Co-first authorship

Main Findings

This manuscript integrates transcriptional data from multiple species to investigate how changes in gene expression following gene duplication have contributed to the evolution of C_4 photosynthesis. The enzymes, transporters, and direct regulators of the core-C_4 cycle show dramatic changes in expression from their nearest paralogs; this is true even in duplicates originating after C_4 evolution. Further, the classical features of C_4 expression (high expression, tissue-specific expression, and photosynthetic-like expression pattern) are not shared with the nearest non-C_4 orthologs. Looking beyond the core C_4 cycle, we find several functions enriched in duplicates that are split between various M, BS, mature or immature expression patterns. Some of these functions are consistent for areas of specialization in C_4 photosynthesis, including ATP-consuming members of photosynthetic cycles, auxin response and various subcategories of cell wall. Enlarged gene families showed a flexibility in evolving gene expression patterns relevant to C_4-photosynthesis. A genome wide correlation of gene family size with expression divergence included frequent gain/loss of photosynthetic expression pattern. Finally, in maize there was a correlation between gene family size and tissue specificity, which was not the case in rice, indicating gene duplication helps precondition development of tissue specificity important for C_4 photosynthesis.

Contributions

- Experimental design
- Optimization of tissue separation method
- Wet lab work: RNAseq, Enzyme assays, metabolite extraction

- Mapping and analysis of transcriptional data
- Data integration, analysis and interpretation
- Writing manuscript

Expression divergence following gene duplication contributes to the evolution of the complex trait C_4 photosynthesis.

Alisandra K Denton[1,4], Janina Maß[2,4], Canan Külahoglu[1], Martin Lercher[2], Shin-Han Shiu[3], Andrea Bräutigam[1] & Andreas P.M. Weber[1]

[1]*Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences, iGRAD-plant program, Heinrich-Heine-University, 40225 Düsseldorf, Germany.*
[2]*Institute of Informatics, Cluster of Excellence on Plant Sciences, iGRAD-plant program, Heinrich-Heine University, 40225 Düsseldorf, Germany.*
[3]*Programs in Genetics and Quantitative Biology, Michigan State University, East Lansing, Michigan 48824.*
[4]*These authors contributed equally to this manuscript.*

Whole genome duplications at the base of the vertebrate and angiosperm lineages are hypothesized to have promoted the rapid evolution, radiation, and success of these lineages. However, the evidence is limited to the timing of radiation shortly after whole genome duplications and examples from individual gene families. Here we test how expression divergence following gene duplication is exploited by the complex trait C_4 photosynthesis on a genome-wide scale. Known C_4 genes and functional categories related to C_4 anatomy and energy balance show expression divergence between tissues where additional specialization is required in C_4 plants. Higher levels of gene duplication were associated with higher specificity in the key C_4 tissues: mesophyll and bundle sheath. This held in ancient duplicates, providing evidence that whole genome duplications precondition the evolution of C_4 photosynthesis.

Whole genome duplications (WGD) are proposed to have facilitated the evolution of morphological diversity in lineages such as vertebrates [1–3] and flowering plants (Angiosperms) [4, 5]. The Angiosperm lineage underwent a WGD prior to the radiation and diversification of the two major clades: monocots and dicots [4]. Similarly early vertebrates underwent two WGDs, shortly before and after the branching off of the hagfish, and before the radiation of the main vertebrate lineages [6]. However, timing of ancient WGDs should not be over interpreted as there are many known WGDs, which are not apparently associated with major morphological specialization or radiation events [7, 8]. There is anecdotal evidence for the contribution of ancient WGDs to lineage specific features, such as the expression of neural regulatory paralogs in the neural crest, a tissue type only found in vertebrates [3]. There is a limited amount of experimental evidence linking gene duplication to trait evolution on a broader scale; for instance, genes with stress-responsive expression are enriched in tandem duplicates in *Arabidopsis thaliana* [9]. However, most genome wide studies have focused on the ramifications of gene-duplication on the divergence of gene sequence and expression.

Expression divergence is a promising target to unravel the functional importance and consequences of gene duplication on a genome-wide scale. Analyses of the transcriptomes of fungi, animals and plants have elucidated basic trends in expression divergence [10]. These trends include greater divergence in duplicates from small scale duplications than WGDs [11, 12] and a correlation of synonymous and–in young duplicates–non synonymous substitutions rate (dS and dN) with expression divergence [13–15]. Importantly, young duplicates show relaxed purifying selection on both their sequence and expression patterns [16, 17]. To elucidate the contribution of gen(ome) duplication to the evolution of novel traits, we examine how expression divergence has contributed to the evolution of the complex trait C_4 photosynthesis on a genome wide scale.

The C_4 trait is an evolutionary "patch" to the ancestral C_3 photosynthetic type. This "patch" helps plants thrive in hot and arid environments. The core of C_4 photosynthesis is a biochemical cycle that pumps CO_2 from the outer mesophyll (M) tissue to the interior bundle sheath (BS) tissue. Evolution of C_4 involves extensive anatomical changes and change in expression of hundreds or perhaps thousands of genes [18–22]. Despite this complexity, C_4 has evolved in more than 66 different lineages [23]. This surprising degree of convergent evolution is possible, in part, because of the pre-existence of all necessary enzymes [24]. Gene duplication is thought to precondition the evolution of C_4 photosynthesis by allowing maintenance of ancestral gene function and recruitment to C_4 photosynthesis. However, this necessity has been recently questioned [25, 26] and largely investigated only for the core genes of the C_4 cycle [27].

We investigate how expression divergence following gene duplication is utilized by the C_4 trait in *Zea mays*. *Z. mays* is a highly duplicated C_4 species that underwent two readily traceable WGDs: the pan-grass duplication ($\sim$70 mya) and the *Z. mays* specific tetraploidy event (5-12 mya)[28]. *Z. mays* belongs to the Andropogoneae, which evolved C_4 photosynthesis $\sim$20 mya [29]. Integrating transcriptomics [30–36] and phylogenetics, we characterize how the expression divergence of known C_4 genes and functions contribute to the specialized biochemistry, energy balance, and anatomy of the trait. We find that large gene families have more divergent expression, and more BS and M tissue specificity genome wide, even in duplications that occurred long before the evolution of C_4 photosynthesis.

Results

Measurement and compilation of expression data in grasses to cover tissues of interest to C_4-photosynthesis. To evaluate how gene duplication could contribute to the evolution of C_4 photosynthesis in terms of expression divergence, we first needed a transcriptome dataset covering tissues where C_4 photosynthesis requires specialization. The specialized anatomy found in C_4 species is set up during leaf development [18, 37], and leaf gradients provide a powerful tool to understand C_4 photosynthesis [30, 38–40]. Coordination of the C_4 cycle requires extensive anatomical and metabolic specialization between M and BS tissue [41]. Therefore, we sampled, enriched and analyzed M and BS tissues across a developing *Z. mays* leaf (Supplementary Fig. 1).

We harvested a developmental gradient from a *Z. mays* leaf; enriched BS and M tissues by a method modified from [42]; and measured levels of metabolites, transcripts, and enzyme activity (Supplementary Dataset 1, 2). This mechanical tissue enrichment method provides high quality RNA and metabolites because tissues are snap frozen at harvest and not thawed until their extraction for down-stream analyses. To estimate the original distributions in M and BS tissues from the partial enrichment data, we "deconvoluted" the data based on marker enzymes or transcripts (see methods; Supplementary Fig. 2). The deconvolution included a test for whether a target transcript, enzyme activity, or metabolite was significantly closer in distribution to either the M or BS marker. The deconvoluted data for mature tissues was consistent with previous studies with M and BS specific transcriptomes [39, 43, 44] (Supplementary Fig. 3, 4, 5, Supplementary Table 1), indicating that the separation method was effective.

To categorize the data and analyze the developmental processes covered by it, we performed functional enrichment analysis for clusters and genes significantly enriched in BS or M. Of the eight clusters, six showed simple patterns, and were high in the M, BS, or both in the leaf tip or base, respectively; while the remaining two clusters were less distinct and termed "mixed-base" and "mixed-M" (Supplementary Fig. 6). Enrichments in mature tissue were consistent with previous separation studies (Supplementary Fig. 4, 5, supplementary note). In developing tissue, the BS cluster enrichments (Fisher's exact, fdr <0.05) included categories of cell and cell wall, as well as lignin biosynthesis. The developing M cluster was enriched in categories including lipid biosynthesis and tetrapyrrole biosynthesis (Supplementary Dataset 3).

To provide perspective for evolutionary questions, published transcriptome data was collected from *Z. mays*, *Sorghum bicolor*, *Setaria italica*, *Brachypodium distachion*, and *Oryza sativa* [30–36] and remapped to the latest respective genome release. This data included photosynthetic and non-photosynthetic tissues; and, where possible, also included developing leaf or separation of BS and M tissue. In each grass, the total collected data was sufficient to show a clear pattern for photosynthetic genes (Supplementary Fig. 7, 8); for *Z. mays*, *S. italica*, and *O. sativa*, the compiled data included separation of BS and M tissue.

Known C_4 genes show high expression divergence, photosynthetic expression patterns and tissue specificity. To investigate how gene duplication and expression divergence relate to the evolution of C_4 photosynthesis, we first examined how much and what sort of expression divergence occurs in the known genes of the core-C_4 cycle. To this end, we qualitatively analyzed and quantitatively compared the expression divergence of core-C_4 genes to the genome-wide background.

The core-C_4 genes showed high, photosynthetic-like, and tissue-specific expression patterns (Fig. 1), which were not shared with their nearest paralogs and orthologs (Fig. 2). Many genes of the C_4 cycle are known to be tissue specific to orchestrate the pumping of CO_2 from M to BS [41], which was consistent with our data (Fig. 1, Supplementary Fig. 9, 10, 11, 12, supplementary

note) Similarly, core-C_4 genes required in both tissue types, were expressed fairly evenly between mature M and BS tissue, with the exception of Aspartate Aminotransferase. Interestingly, four C_4 genes had a paralog that clustered with the opposite tissue specificity of the C_4-paralog. All C_4 paralogs had expression patterns peaking in mature leaf tissue (Fig. 1), highly correlated with other photosynthesis genes (defined from MapMan categories [45]; Supplementary Fig. 7), and the enzymes and transporters were expressed very highly (>300 FPKM; Fig. 2).

To understand when divergence in core-C_4 expression occurs, we examined expression patterns in a phylogenetic context. For each C_4 gene tree, we selected the (non-C_4) homologs in each species that were phylogenetically closest to the *Z. mays* C_4 gene and compared their expression to the remaining homologs (Fig. 2). No significant differences were found between the nearest and remaining homologs in expression level, correlation to photosynthetic pattern, or tissue specificity (Fig. 2; Supplementary Table 2). Even in young, syntenic paralogs, there are large changes in expression pattern and level (Supplementary Fig. 13, 14, 15). Quantitatively, the C_4 genes were significantly more divergent in pattern ($p = 0.016$) and level ($p = 1.01 * 10^{-5}$) in BS & M gradient and other photosynthetic, but not heterotrophic tissues (Fig. 2; Supplementary Table 3). Further, C_4 genes were sufficiently divergent to significantly improve (pattern $p = 6.88 * 10^{-6}$; level $p = 0.0014$) a multiple regression model for correlations to expression divergence (synteny, dS, dN, # *Z. mays* paralogs; Supplementary Fig. 16; Supplementary Table 4, 5).

As the protein sequences of many core C_4 genes are known to evolve under positive selection [27, 46], and the core C_4 genes were highly divergent in expression; we checked for a general relationship between positive selection and expression pattern divergence. While no significant relation between pairwise dN/dS and expression divergence was observed (pattern $p = 0.38$; level $p = 0.73$), there was a significant relation between dN and expression divergence in duplicates originating in the *Z. mays* tetraploidy, which are all of the same age ($r^2 = 0.036$, p=$6.57 * 10^{-13}$; Supplementary Fig. 17). Positive selection can be more readily and reliably identified when more sequence information is included [47]; therefore, we compared positive selection to divergence in a test set of 64 whole gene families. While there was a negative correlation (Spearman's R $(r_s) = -0.08$) between the p-value for significance of positive selection at a branch, and the sum of pairs for expression divergence between this branch and its sister branch, this was not significant ($p = 0.11$). With only one of the three measures for sequence level selective pressure (dN/dS, dN of duplicates from WGDs, significance of dN/dS at tree branch), significantly related to expression divergence, but all trending in the same direction of more divergence with more selective pressure, the relation between sequence level selective pressure and expression divergence remains inconclusive. However, the C_4 genes are clear outliers.

Divergent expression between paralogs relates to specialization in C_4 anatomy and energy balance. An integrated C_4 trait requires modifications to metabolism and anatomy that go far beyond the establishment of the core C_4 cycle, which is reflected in the high number of genes– and functional groups there of–that are differentially regulated between closely related C_3 and C_4 species [18–22]. If WGDs, and not just gene duplications, are important for the evolution of C_4

photosynthesis, we expect gene duplication to contribute to the greater complexity of the C_4 trait.

To determine whether expression divergence contributed to the greater complexity of C_4 photosynthesis, we asked whether any gene functions (MapMan categories) showed a tendency towards particular patterns of expression divergence. We used a graph theory approach to categorize the patterns of expression divergence, with the k-means expression clusters as nodes, and paralog pairs as edges (Fig. 3). For example, a pair of paralogs expressed in clusters 1 and 2, respectively, were assigned to the edge 1_2, while a pair of paralogs that were both expressed in cluster 1, were assigned to the edge connecting cluster 1 to itself, that is the "loop", 1_1. To reduce noise, we excluded paralogs in different clusters that were more similar in expression to each other than to their cluster centers. Then we tested all edges for functional enrichments (MapMan categories). Most significant enrichments (fdr $<$0.05) were found in loops, and these were very similar to the enrichment of all genes in the cluster (Supplementary Dataset 3, 4). Among the 76 enrichments in non-loop edges, the photosynthesis category was enriched in the edge 3_5 ("M-tip" to "BS-tip"). Specifically, edge 3_5 included subunits of photosystem I, and ATP-consuming enzymes from the Calvin-Benson-Bassham (CBB) and photorespiratory cycle (Supplementary Table 6). Other enrichments included several categories of cell wall and auxin response in edge 2_7 ("BS-base" to "even-base") and miscellaneous gluco- galacto- and mannosidases in edge 4_7 ("M-base" to "even-base"; Supplementary Dataset 4).

Evidence for a preconditioning effect of gene duplication in changes in photosynthetic expression pattern and tissue-specific expression patterns. C_4 photosynthesis involves extensive anatomical and metabolic changes to leaf tissue, in particular specialization of functions between M and BS tissue. If gene duplication contributes to C_4 photosynthesis, we expect it contribute to modifications in which genes are expressed in photosynthetic tissues and tissue specificity.

For the observed correlation between expression divergence and gene family size (Supplementary Fig. 16; Supplementary Table 5) to facilitate the evolution of C_4 photosynthesis, we further expect expression divergence to include recruitment of genes to a photosynthetic pattern. To test this, we first classified the expression pattern of every gene in every species as photosynthetic or not. This classification was based on the bi-modal distribution of r_p between each gene's expression pattern and the expression pattern of the photosynthesis genes (MapMan category; Supplementary Fig. 18). This was used to test whether duplication level (# *Z. mays* paralogs) related to how frequently photosynthetic expression patterns were shared between species or were species specific. We found that higher levels of gene duplication were associated with species specificity in both presence and absence of photosynthetic pattern compared to conserved photosynthetic pattern across all species (Fig. 4; Supplementary Fig. 19). This indicates the general correlation between gene duplication level and expression divergence includes both recruitment to, and loss of, photosynthetic pattern.

To test whether gene duplication preconditions the tissue specificity that is characteristic

of C_4 photosynthesis, we tested for a correlation between gene family size and tissue specificity. The average p-value for tissue specificity along the developmental leaf gradient was negatively correlated (r_s=-0.071; p<0.001) with gene family size (Fig. 4). To determine if this was related to C_4 photosynthesis, we compared the p-value for tissue specificity in rice to gene family size and found the opposite pattern (Fig. 4; r_s=0.029; p<0.05). Thus, in the *Z. mays* data, but not in the rice data, larger gene families show higher tissue specificity.

For the C_4 genes, we saw divergence in expression both before and after C_4 evolution. Therefore, we asked if the observed correlations between change in photosynthetic pattern and increased tissue-specificity of expression occurred before or after C_4-evolution. We identified "ancient" orthogroups, which had not further expanded after the time of the pan-grass genome duplication (min pairwise dS >1), and "young" orthogroups, which have expanded entirely since the time of the *Z. mays* tetraploidy (max pairwise dS <0.3). The association between gene family size and increased change in photosynthetic expression pattern, largely held in both ancient and young gene families (Supplementary Fig. 20, 21). The increase in tissue-specificity with gene family size was found in ancient gene families; however, the opposite correlation was found in young gene families (Supplementary Fig. 22). This may relate to silencing of younger duplicates. In summary, ancient gene duplications are associated with increased changes in gene expression patterns relevant to the evolution of C_4, in particular increased tissue specificity.

Discussion

Ancient whole genome duplications are thought to have promoted the evolution of the morphological diversity observed in vertebrates and angiosperms today. However, few studies link gene duplication to evolutionary traits on a genome wide scale. Here, we have tested how gene duplication (including the WGDs at ∼70mya and 5-12mya) [28] and the following expression divergence have contributed to the evolution and integration of the complex trait C_4 photosynthesis at ∼20mya [29].

We find high expression divergence in the core-C_4 genes, in particular recruitment of the C_4 paralog to a high amplitude, photosynthetic-like, tissue-specific expression pattern. This divergence was significantly more than expected, especially when accounting for the sequence features and *Z. mays* gene family size. Similarly, there is a striking co-occurrence of positive selection on the amino acid sequence [27, 46] and expression divergence in core-C_4 genes, but on a wider scale we found only tentative evidence for a relationship between gene sequence positive selection and expression divergence.

Functional categories key to C_4 photosynthesis show specialized expression divergence. Paralogs with functions related to C_4 anatomy are enriched in particular patterns of divergence in immature tissue. The modifications in vascular patterning required for C_4 photosynthesis are thought to require changes in auxin perception [48], and both auxin response transcription factors and their

downstream targets are enriched in the edge between the BS-base and even-base clusters (Supplementary Dataset 4). This edge is further enriched in cell wall categories, which could support the specialized anatomy observed in the BS cell wall [49]. Genes classified under miscellaneous gluco-galacto- and mannosidases were enriched in the edge between the M-base and even-base clusters (Supplementary Dataset 4). These genes included 1,3 beta-galactosidases and various cellulases, which are often associated with loosening or modification of the cell wall [50, 51].

Paralogs with functions relating to energy balance are enriched between M-tip and BS-tip clusters. Maize has a complex energy balance between cell types, with photosystem II restricted to M cells, and several cycles shuttling reducing equivalents into the BS. Further, the use of two decarboxylation enzymes is proposed to add robustness to the energy balance between subtypes in fluctuating light conditions [52, 53]. Similarly, the ATP-consuming enzymes with paralogs in the edge between BS-tip and M-tip clusters could add robustness or fine-regulation to energy balance. Alternatively, two of these enzymes have their highest expression in the M, and the secondary BS specific paralog could provide an overflow mechanism if and when diffusion were to become limiting to the photorespiratory or CBB cycles.

Duplication promotes changes or specialization in expression related to C_4 photosynthesis. Paralogs show relaxed purifying selection after duplication in both sequence and expression pattern [16, 17]. This is consistent with the correlation between gene family size and expression divergence found here (Supplementary Fig. 16) considering that older duplicates show only partial redundancy [54] and the initial relaxation in purifying selection decreases over time [55].

One particular change in expression associated with the evolution of C_4 photosynthesis is the recruitment of genes to a photosynthetic-like pattern with expression peaking in mature leaf tissue. This has been reported [18] and was observed here for the core C_4 genes. Further recruitment to the leaf is thought to relate to other aspects of the C_4-leaf, for instance BS tissue characteristics are thought to derive, to some degree, from endodermis tissue of the root by recruitment of the scarecrow regulatory module [30, 56, 57]. This recruitment, as well as the changes in expression of hundreds to thousands of genes observed between mature leaves of closely related C_4 species [18–22], could be facilitated by gene duplication as larger gene families showed more frequent gain and loss of photosynthetic-like expression pattern.

Evolution of C_4 involves a massive functional change for BS tissue, which is ancestrally a "smart pipe" regulating access to the vasculature [58, 59], but takes on a major photosynthetic function. Similarly, M tissue undergoes metabolic specialization as many functions are divided between M and BS tissues [35, 37, 39, 43, 44, 60]. Here we showed a correlation between gene-family size and tissue specificity in *Z. mays*, which could support recruitment of the BS to photosynthesis or support any of the extensive specialization seen between M and BS cells. This correlation held in ancient gene-families, which have not expanded since the pan-grass duplication roughly 50 million years before C_4 evolution [28, 29]. Thus, ancient genome duplications precondition the evolution

of expression patterns important for C_4 photosynthesis.

An interesting question is how duplication mechanistically facilitates changes in expression. Small scale duplications show more divergence in expression, which has been attributed to duplication without the regulatory sequence. While for genes duplicated with *cis*-regulatory region, it could be due to more rapid accumulation of single nucleotide polymorphisms associated with reduced purifying selection. Alternatively, in a species like *Z. mays*, which is still undergoing diploidization and massive genome arrangement after the tetraploidy event [61, 62], expression divergence may result from the rearrangement or loss of neighboring genes causing a switch in *cis*-regulatory region. Notably, the core-C_4 duplicates with high divergence despite their youth and colinearity (PEPCK, PPDK, and PPDK-RP) all show rearrangement of genes in the immediate upstream region (Supplementary Fig. 23) [63]. We speculate that this may have contributed to the high divergence of C_4 genes, which could not be accounted for based on their sequence characteristics (dS, dN, dN/dS) nor gene family size. Finally, where transcriptional regulators show expression divergence (e.g. Auxin Response Factors diverging between "BS-base" and "even-base" clusters; Supplementary Dataset 4), their down-stream targets can be affected, which could have an effect, for instance, if a duplicated transcriptional enhancer showed a conserved DNA binding, but degenerate protein-protein interactions, and thereby became antagonistic to the other copy. A change of this type would be one way to achieve the observed tissue specificity of some C_3 *cis*-regulatory regions when heterogously expressed in C_4 species [25].

Overall this study connects genome-wide changes in expression to the evolution of a complex trait, showing both that duplication is beneficial and elucidating its advantages. It builds a bridge between the numerous single-gene family studies and large-scale correlations to improve our understanding of evolutionary processes. To further understand the contribution of WGDs to complex trait evolution, it will be important to perform further large-scale, yet function-oriented studies. In particular, examining species with relatively recent WGDs and evolution of trait of interest would increase specificity and potentially allow for a more mechanistic understanding of divergence after WGD.

Figures

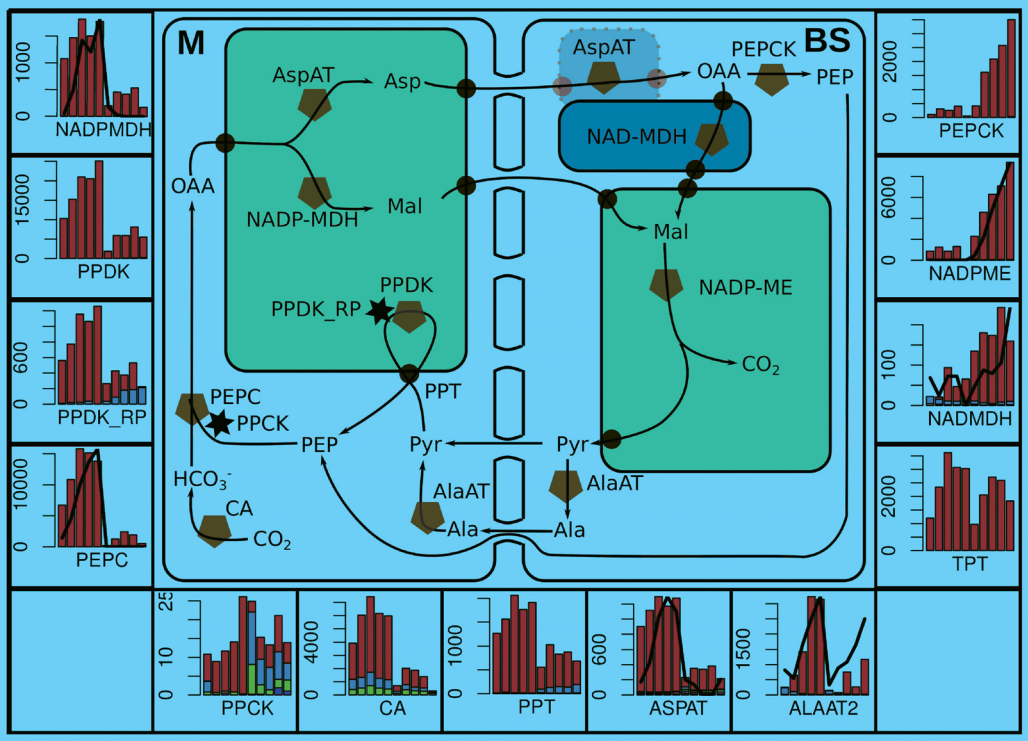


Fig. 1: The core C_4 cycle: abundance and distribution of transcripts and enzyme activities. Ordering of all bars from left to right is immature to mature M, followed by immature to mature BS. Bars show transcript abundance in FPKM with colors denoting different paralogs. Black lines represent relative enzyme activity. Inside: schematic summary of the core C_4 cycle with enzymes as pentagons, transporters as circles, and regulatory proteins as stars. Chloroplastic enzymes are in green compartments and the putative mitochondrial reaction in the purple compartment.

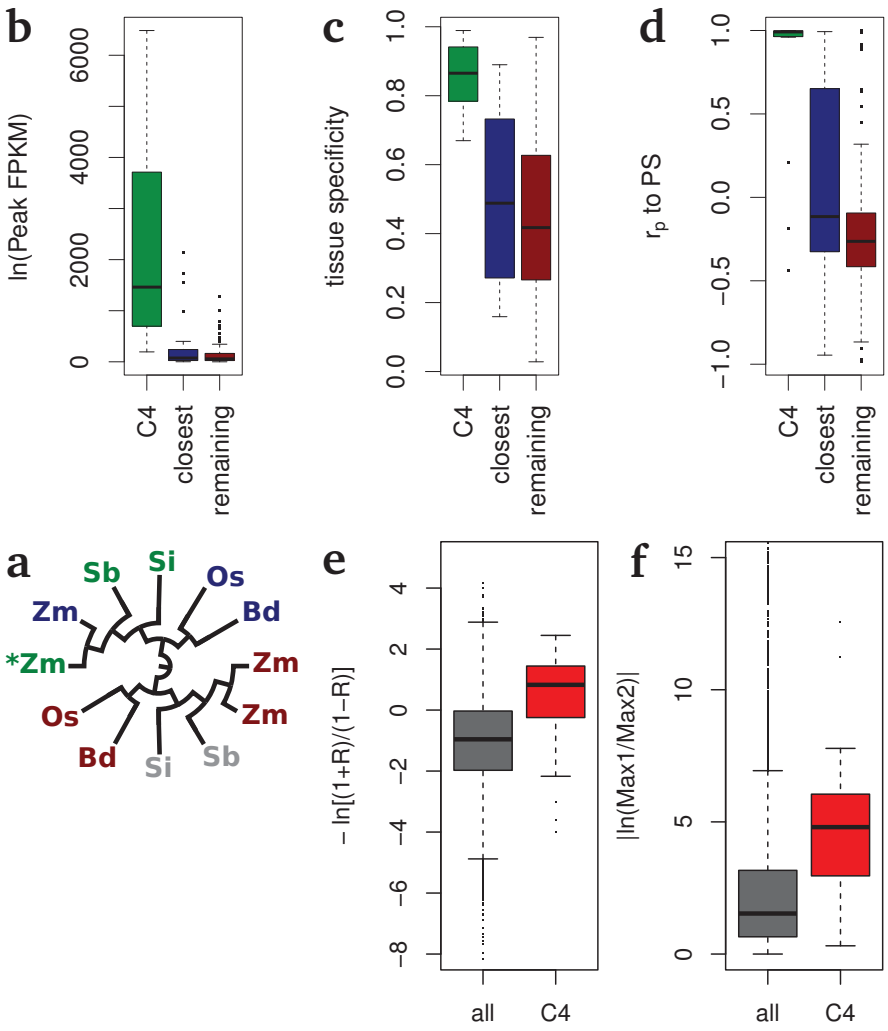


Fig. 2: Expression characteristics and divergence of the core C_4 genes. The similarity (r_p) in expression to the PS (photosynthesis) MapMan category (a), the absolute expression level (excluding regulatory genes)(b), and the BS or M tissue specificity (for C_4 genes tissue specific in *Z. mays* only)(c) between C_4 genes (green), their phylogenetically closest homologs in each species(blue), and the remaining homologs (red). Example classification of homologs on a perfect, no-loss gene tree (d). Where there was only one non-C_4 homolog in any species (grey), these homologs could not be classified as closest nor remaining and were excluded. Quantification of the divergence in expression pattern (e) and level (f) between the C_4 genes and their paralogs vs between all other paralogs.

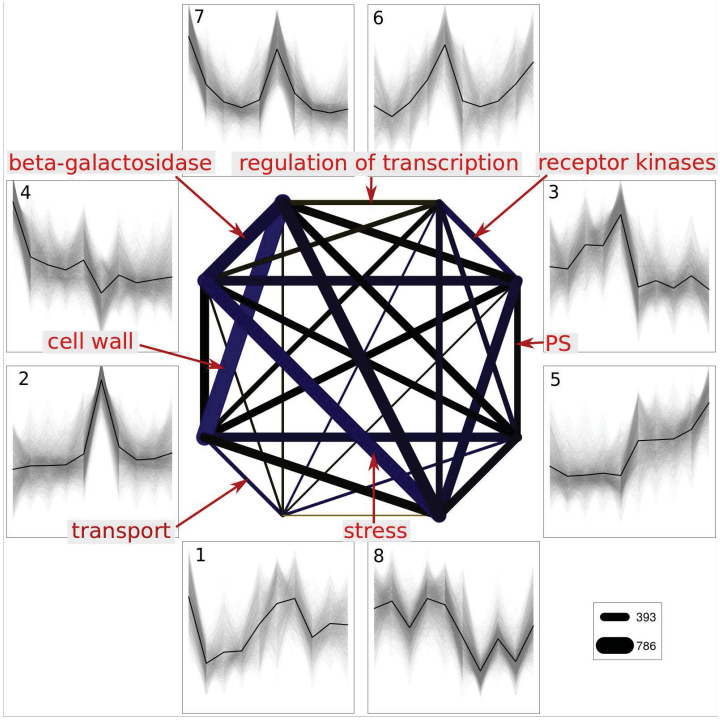


Fig. 3: Paralogs in the edges between expression clusters and their functional enrichments. Example's of significantly enriched functions (fdr <0.05) are shown with an arrow to edge between clusters in which they are enriched (all significant enrichments included in Supplementary Dataset 4). Clusters plotted as z-scores with M base to tip followed by BS base to tip from left to right. The width of the lines connecting clusters is relative to the number of pairs in the edge connecting the respective clusters, while the color indicates whether the edges are larger (blue) or smaller (yellow) than expected based on the total number of pairs in non-loop edges of the connected clusters.

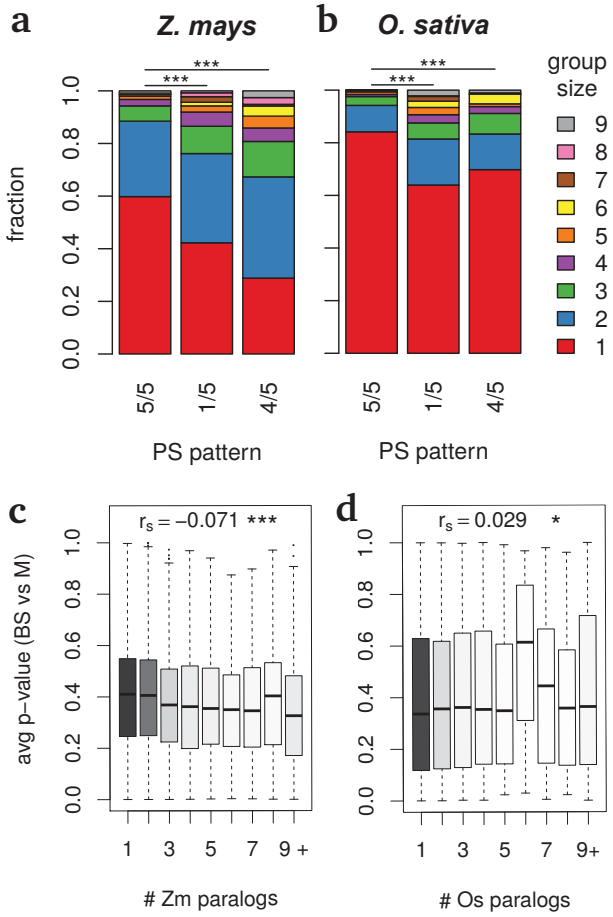


Fig. 4: The relationship between paralog number and expression characteristics related to C_4 evolution. The relation between photosynthetic pattern evolution and group size (# paralogs in orthogroup in respective species)(a-b). Cases where all 5 species show a photosynthetic-like expression pattern (see Supplementary Fig. 18) are considered conserved, while cases where 4 of 5 or 1 of 5 species show a photosynthetic-like expression pattern are considered gain or loss, respectively. The odd species out is *Z. mays* in (a) or *O. sativa* in (b). The significance of tissue specificity (average p-value) vs the group size in (c) *Z. mays* and (d) *O. sativa*.

Accession Numbers The reads related to this article have been deposited in the Sequence Read Archives under the accession number XXXXXX.

Acknowledgements We acknowledge Dr. Katrin Weber for her assistance with the GC/MS analysis; Simon Schliesky for support with data management; the Michigan State University High Performance Computing Cluster team for a good computing experience; and the German Research Foundation for financial support (IRTG 1525 supporting J.M., C.K. and A.K.D.; XXXXXXXXXXX).

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to A.P.M. Weber (email: Andreas.Weber@uni-duesseldorf.de).

Author Contributions A.K.D., A.B. and A.P.M.W. designed the study. A.K.D. performed wetlab measurements, performed all computational and statistical data analysis not related to phylogenetics, performed general data analysis and wrote manuscript. J.M. performed all computational and statistical data analysis related to phylogenetics, performed general data analysis, and assisted in wetlab measurements and writing manuscript. C.K., M.L., S.H.S., A.B., and A.P.M.W. analyzed data and assited in writing manuscript. All authors read and approved the final manuscript.

References

1. Holland, P. W., Garcia-Fernàndez, J., Williams, N. A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Development* **1994,** 125–133 (1994).
2. Holland, L. Z. Evolution of new characters after whole genome duplications: insights from amphioxus. *Seminars in cell & developmental biology* **24,** 101–9 (Feb. 2013).
3. Van Otterloo, E., Cornell, R. A., Medeiros, D. M. & Garnett, A. T. Gene regulatory evolution and the origin of macroevolutionary novelties: insights from the neural crest. *Genesis (New York, N.Y. : 2000)* **51,** 457–70 (July 2013).
4. Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473,** 97–100 (May 2011).
5. De Bodt, S., Maere, S. & Van de Peer, Y. Genome duplication and the origin of angiosperms. *Trends in ecology & evolution* **20,** 591–7 (Nov. 2005).
6. Sidow, A. Gen (om) e duplications in the evolution of early vertebrates. *Current opinion in genetics & development* **6,** 715–722 (1996).
7. Wood, T. E. *et al.* The frequency of polyploid speciation in vascular plants. *Proceedings of the national Academy of sciences* **106,** 13875–13879 (2009).
8. Cannon, S. B. *et al.* Polyploidy did not predate the evolution of nodulation in all legumes. *PLoS One* **5,** e11630 (2010).

9. Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K. & Shiu, S.-H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant physiology* **148,** 993–1003 (2008).

10. Maere, S. & de Peer, Y. V. in *Evolution after gene duplication* (eds Dittmar, K. & Liberles, D.) 31–56 (Wiley-Blackwell, Hoboken, New Jersey, 2010). doi:10.1002/9780470619902.ch3.

11. Casneuf, T., De Bodt, S., Raes, J., Maere, S. & Van de Peer, Y. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. *Genome biology* **7,** R13 (Jan. 2006).

12. Kim, J., Shiu, S.-H., Thoma, S., Li, W.-H. & Patterson, S. E. Patterns of expansion and expression divergence in the plant polygalacturonase gene family. *Genome biology* **7,** R87 (Jan. 2006).

13. Chung, W.-Y., Albert, R., Albert, I., Nekrutenko, A. & Makova, K. D. Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. *BMC bioinformatics* **7,** 46 (Jan. 2006).

14. Gu, Z., Nicolae, D., Lu, H. H.-S. & Li, W.-H. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics* **18,** 609–613 (Dec. 2002).

15. Makova, K. D. & Li, W.-H. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome research* **13,** 1638–1645 (2003).

16. Chain, F. J. J., Ilieva, D. & Evans, B. J. Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evolutionary Biology* **8,** 43 (Jan. 2008).

17. Hellsten, U. *et al.* Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. *BMC Biology* **5,** 31 (Jan. 2007).

18. Külahoglu, C. *et al.* Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C_3 and C_4 Plant Species. *The Plant Cell* **26,** 3243–3260 (Aug. 2014).

19. Bräutigam, A. *et al.* An mRNA blueprint for C_4 photosynthesis derived from comparative transcriptomics of closely related C_3 and C_4 species. *Plant Physiology* **155,** 142–56 (Jan. 2011).

20. Gowik, U., Bräutigam, A., Weber, K. L., Weber, A. P. M. & Westhoff, P. Evolution of C_4 photosynthesis in the genus Flaveria: how many and which genes does it take to make C_4? *The Plant Cell* **23,** 2087–105 (June 2011).

21. Bräutigam, A., Schliesky, S., Külahoglu, C., Osborne, C. P. & Weber, A. P. M. 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,** 3579–93 (July 2014).

22. Mallmann, J. *et al.* The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. *Elife* **3,** e02478 (2014).

23. Sage, R. F., Sage, T. L. & Kocacinar, F. Photorespiration and the evolution of C_4 photosynthesis. *Annual Review of Plant Biology* **63,** 19–47 (June 2012).
24. Aubry, S., Brown, N. J. & Hibberd, J. M. The role of proteins in C(3) plants prior to their recruitment into the C(4) pathway. *Journal of experimental botany* **62,** 3049–59 (May 2011).
25. Williams, B. P., Aubry, S. & Hibberd, J. M. Molecular evolution of genes recruited into C_4 photosynthesis. *Trends in Plant Science* **17,** 213–20 (Apr. 2012).
26. Van den Bergh, E. *et al.* Gene and genome duplications and the origin of C 4 photosynthesis: Birth of a trait in the Cleomaceae. *Current Plant Biology* **1,** 2–9 (2014).
27. Wang, X. *et al.* Comparative genomic analysis of C_4 photosynthetic pathway evolution in grasses. *Genome biology* **10,** R68 (Jan. 2009).
28. Schnable, J. C., Freeling, M. & Lyons, E. Genome-wide analysis of syntenic gene deletion in the grasses. *Genome biology and evolution* **4,** 265–77 (Jan. 2012).
29. Christin, P.-A., Salamin, N., Kellogg, E. a., Vicentini, A. & Besnard, G. Integrating phylogeny into studies of C_4 variation in the grasses. *Plant physiology* **149,** 82–7 (Jan. 2009).
30. Wang, P., Kelly, S., Fouracre, J. P. & Langdale, J. a. Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C_4 Kranz anatomy. *The Plant Journal : For Cell and Molecular Biology* **75,** 656–70 (Aug. 2013).
31. Sekhon, R. S. *et al.* Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. *PloS one* **8,** e61005 (Jan. 2013).
32. Davidson, R. M. *et al.* Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *The Plant journal : for cell and molecular biology* **71,** 492–502 (Aug. 2012).
33. Bennetzen, J. L. *et al.* Reference genome sequence of the model plant Setaria. *Nature biotechnology* **30,** 555–61 (June 2012).
34. Zhang, G. *et al.* Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nature biotechnology* **30,** 549–54 (June 2012).
35. John, C. R., Smith-Unna, R. D., Woodfield, H., Covshoff, S. & Hibberd, J. M. Evolutionary convergence of cell-specific gene expression in independent lineages of C_4 grasses. *Plant physiology* **165,** 62–75 (May 2014).
36. Jiao, Y. *et al.* A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nature genetics* **41,** 258–63 (Feb. 2009).
37. Majeran, W. *et al.* Structural and metabolic transitions of C_4 leaf development and differentiation defined by microscopy and quantitative proteomics in maize. *The Plant cell* **22,** 3509–42 (Nov. 2010).

38. Pick, T. R. *et al.* Systems analysis of a maize leaf developmental gradient redefines the current C_4 model and provides candidates for regulation. *The Plant Cell* **23,** 4208–20 (Dec. 2011).

39. Li, P. *et al.* The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* **42,** 1060–1067 (Oct. 2010).

40. Wang, L. *et al.* Comparative analyses of C_4 and C_3 photosynthesis in developing leaves of maize and rice. *Nature biotechnology* **32,** 1158–1165 (2014).

41. Sheen, J. C_4 Gene Expression. *Annual Review of Plant Physiology and Plant Molecular Biology* **50,** 187–217 (1999).

42. Stitt, M. & Heldt, H. W. Control of photosynthetic sucrose synthesis by fructose-2, 6-bisphosphate: Intercellular metabolite distribution and properties of the cytosolic fructose-bisphosphatase in leaves of Zea mays L. *Planta* **164,** 179–188 (1985).

43. Tausta, S. L. *et al.* Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C_4-related processes. *Journal of Experimental Botany* **65,** 3543–55 (July 2014).

44. Chang, Y.-M. *et al.* Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. *Plant physiology* **160,** 165–177 (2012).

45. Lohse, M. *et al.* Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, cell & environment* **37,** 1250–1258 (2014).

46. Christin, P.-A. *et al.* Oligocene CO2 decline promoted C_4 photosynthesis in grasses. *Current Biology : CB* **18,** 37–43 (Jan. 2008).

47. Anisimova, M., Bielawski, J. P. & Yang, Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular biology and evolution* **18,** 1585–1592 (2001).

48. McKown, A. D. & Dengler, N. G. Vein patterning and evolution in C_4 plants. *Botany* **88,** 775–786 (Sept. 2010).

49. Eastman, P. A. K., Dengler, N. G. & Peterson, C. A. Suberized Bundle Sheaths in Grasses (Poaceae) of Different Photosynthetic Types I. Anatomy, Ultrastructure and Histochemistry. *Protoplasma* **142,** 92–111 (1988).

50. Goulao, L. F. & Oliveira, C. M. Cell wall modifications during fruit ripening: when a fruit is not the fruit. *Trends in Food Science & Technology* **19,** 4–25 (2008).

51. Levy, I., Shani, Z. & Shoseyov, O. Modification of polysaccharides and plant cell wall by endo-1, 4-β-glucanase and cellulose-binding domains. *Biomolecular engineering* **19,** 17–30 (2002).

52. Bellasio, C. & Griffiths, H. The Operation of Two Decarboxylases, Transamination, and Partitioning of C_4 Metabolic Processes between Mesophyll and Bundle Sheath Cells Allows Light Capture To Be Balanced for the Maize C_4 Pathway. *Plant Physiology* **164,** 466–80 (Jan. 2014).

53. Wang, Y., Bräutigam, A., Weber, A. P. M. & Zhu, X.-G. Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis. *Journal of Experimental Botany* **65,** 3567–78 (July 2014).

54. Ihmels, J., Collins, S. R., Schuldiner, M., Krogan, N. J. & Weissman, J. S. Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. *Molecular systems biology* **3,** 86 (Jan. 2007).

55. Pegueroles, C., Laurie, S. & Albà, M. Accelerated Evolution after Gene Duplication: A Time-Dependent Process Affecting Just One Copy. *Molecular biology and evolution* **30,** 1830–1842 (2013).

56. Slewinski, T. L., Anderson, A. a., Zhang, C. & Turgeon, R. Scarecrow plays a role in establishing Kranz anatomy in maize leaves. *Plant & cell physiology* **53,** 2030–7 (Dec. 2012).

57. Slewinski, T. L. Using evolution as a guide to engineer kranz-type c4 photosynthesis. *Frontiers in plant science* **4,** 212 (Jan. 2013).

58. Shatil-Cohen, A. & Moshelion, M. Smart pipes: the bundle sheath role as xylem-mesophyll barrier. *Plant signaling & behavior* **7,** 1088–1091 (2012).

59. Griffiths, H., Weller, G., Toy, L. F. & Dennis, R. J. You're So Vein: Bundle Sheath Physiology, Phylogeny and Evolution in C_3 and C_4 Plants. *Plant, Cell & Environment* **36,** 249–261 (2013).

60. Friso, G., Majeran, W., Huang, M., Sun, Q. & van Wijk, K. J. Reconstruction of metabolic pathways, protein expression, and homeostasis machineries across maize bundle sheath and mesophyll chloroplasts: large-scale quantitative proteomics using the first maize genome assembly. *Plant physiology* **152,** 1219–50 (Mar. 2010).

61. Lai, J. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature genetics* **42,** 1027–30 (Nov. 2010).

62. Springer, N. M. *et al.* Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS genetics* **5,** e1000734 (Nov. 2009).

63. Proost, S. *et al.* PLAZA 3.0: an access point for plant comparative genomics. *Nucleic acids research* **43,** D974–D981 (2015).

Methods

Statistical notes. Unless otherwise noted, all statistical analysis was performed in the R statistical environment. Whenever a test was performed more than 20 times, the false discovery rate [64] was calculated from the resulting p-value.

Obtaining and processing plant Genome Data. Genome and gene-model data was downloaded for 12 grasses with available genomes and for banana as an outgroup from Phytozome 10.0 (*Z. mays*, *S. bicolor*, *S. italica*, *O. sativa*, *B. distachyon*, *Panicum halli*, *Panicum virgatum*; [65]) or Gramene V40 (*Oryza brachyantha*, *Oryza glaberrima*, *Triticum aestivum*, *Triticum urartu*, *Hordeum vulgare*, *Musa acuminata*; [66]). In cases with multiple gene models, the longest protein sequence was used for further analysis.

Defining homology. Three methods were used to define homologous genes as appropriate for the context. First, BLAST [67] was used to define pairs of homolgous genes by reciprocal best hits as well as the 'best' ortholog for a *Z. mays* gene by one-directional best blast hit. Second, OrthoMCL [68] was used to more inclusively define whole orthogroups/gene families. Third, we used paralogs which were previously found to have originated from the pan-grass WGD, the *Z. mays* specific tetraploidy, or from tandem duplications [28].

Mapping between species and genome annotations. Combining data for this study required confident mapping of gene identity between different genome releases. As not all genes with the same identifier show any homology, we used a combination of BLAST and provided mappings (i.e. matching IDs, ftp://ftp.gramene.org/pub/gramene/maizesequence.org/release-5a/working-set/4a_discontinued_ids.txt) to obtain confident mappings. Mappings were given a score of 0 for a provided mapping and a reciprocal best BLAST hit, 2 for only a reciprocal best blast hit, 3 for a provided mapping and best BLAST hit from *Z. mays* 6a to the other genome, and 5 for only a best BLAST hit from 6a to the other genome. Ties were broken randomly. The same scoring was used for interspecies mappings, but without provided mappings. Finally, before using the annotated duplicate origins [28] we filtered pairs that didn't pass a final quality check to see if the mapped WGD derived duplicates showed collinearity using McscanX [69] and if the mapped tandem duplicates occurred within 40 genes of each other.

Phylogenetic analysis. Multiple sequence alignment for orthologous groups was performed with prank [70], and in the case of pairwise *Z. mays* sequences with MAFFT [71]. The ungapped alignment area of the resulting multiple sequence alignment was maximized by filtering poorly aligned and gap-causing sequences with seqSieve (https://pypi.python.org/pypi/seqSieve/0.9.1). Resulting protein alignments were translated to codons with pal2nal [72]. Phylogentic trees were constructed with RaxML [73]. Plots were produced using the ete2 python package [74]. For display only, we manually corrected the PPDK tree so that the paralogs originating from the *Z. mays* tetraploidy were sister to each other. Pairwise estimates for the synonymous and non-synonymous substitution rate (dS and dN) were calculated using codeml from the PAML package [75]. In a test set (described at end of methods) the signature of positive selection (dN/dS >1) was tested using

the branch site model, and significance calculated with a likelihood ratio test [76]. This test was performed at all *Z. mays* genes, their parental branches, and the parental branches there of.

Plant Growth conditions and harvest. *Z. mays* B73 were grown in the summer of 2012 in the same green house and conditions as previously described [38]. The 3rd leaf was harvested when it measured 18 cm from the 2nd ligule to the leaf tip. Two different harvesting methods were performed. In the first, a leaf gradient consisting of 5 sequential developmental slices (4 cm each) were harvested simultaneously using the "leaf guillotine" [38]. This method required 10s to extract the 3rd leaf and properly align it, which does not allow for reliable estimates of the metabolite distributions for high-turnover photosynthetic metabolites. Therefore, a second harvesting method was performed, in which the plants were positioned above two liquid nitrogen containers and two 8 cm slices were cut with connected scissors (Supplementary Fig. 1). With this method there was a delay of less than 1s between slicing and quenching. The full, five slice gradient was used for RNA sequencing, and the faster two slice gradient was used for metabolite extraction.

Tissue enrichment. Mesohpyll and bundle sheath tissues were mechanically enriched by serial filtration on liquid nitrogen using a method modified from [42]. Ground material was filtered through 250, 80, and $41\mu M$ meshes on liquid nitrogen. Three fractions were selected for further analysis because they showed the most enrichment of bundle sheath tissue (did not pass through $80\mu M$ mesh), most enrichment in mesophyll tissue (passed through $41\mu M$ mesh) or intermediate, but consistent proportions of tissues (did not pass through $41\mu M$ mesh).

Extraction and abundance measurements metabolites/enzymes. Enzymes were extracted and desalted as described in [21] from the three enrichment fractions, and the enzyme activity was measured through chlorometric assays as described in [77, 78]. Metabolites were extracted and quantified via gas chromatography/electron-impact time-of-flight mass spectrometry as described in [79]. To consistently exclude data where the peak was hard to distinguish from the background, low-signal metabolites were excluded. Further individual replicates with a raw % abundance in BS of more than 3 standard deviations from the mean were excluded. The integrated peaks were divided by the area of the ribitol (internal standard) peak and the fresh weight, and to further reduce noise and compensate for FW/DW differences between the cell types by the mean abundance for the replicate. Therefore, normalized differences between metabolites represent not absolute distribution, but distribution relative to the other metabolites, particularly sucrose and the other highly-abundant metabolites.

Sequencing and estimating transcriptional abundances. RNA was extracted with QIAGEN RNeasy Plant kits, according to the manufactures instructions except for the addition of an extra wash step in 80% EtOH. Libraries were prepped from RNA with a RNA integrity number >8 and sequenced with the Illumina HiSeq 2000 platform. All additional reads were downloaded from the Sequence Read Archives [80]. Illumina adaptors were trimmed using cutadapt [81] and trimmed for quality using FASTX (Hannon Lab). Trimmed reads were mapped to the 6a release of the *Z. mays* B73 genome with Tophat2 [82] and transcripts abundance calculated with Cufflinks [83]. However, one study [39] used for minor comparisons was mapped only to the 5a genome. For the

one microarray study included [36] data was downloaded from Gene Expression Ombibus [84], and the expression and significance calculated with GEO2R, which uses the Limma R package [85]. Non-default parameters used for all bioinformatics programs are provided (Supplementary Table 7).

Estimation of initial tissue specificity by "deconvolution". The abundance of metabolites, enzymes and transcripts was compared to abundance of BS and M markers to estimate the original tissue specificity by a method modified from [42]. First, to allow for comparison of data with different absolute expression levels, all data was converted into fraction of total transcript in developmental slice. Second, marker transcript (or marker enzyme) levels were used as proxies for the amount of M and BS tissue in each enrichment fraction. The natural log of the BS marker/M marker was plotted against the natural log of a target unknown/M marker across all samples, and the slope of a regression line between these two log ratios estimated the fraction of target gene transcripts that are localized to the BS Supplementary Fig. 2. To determine if target unknowns were more related to either of the tissue markers, we tested whether the slope of this line was significantly different from 0.5 (corresponding to a null enrichment of 50% M, 50% BS). This was automated with a linear regression in R and calculated for every non-marker enzyme, metabolite, and every gene that had a minimum FPKM >0. Tissue specificity was estimated independently in each developmental slice. We assumed the average abundance between the raw values of all enrichment fractions was equal to the average abundance between M and BS. Therefore, to estimate the "pure" abundance values the estimated fraction in BS and M (1 - fraction BS) were multiplied by 2 x the average FPKM value for the developmental slice. For enzyme and metabolite data, the enzyme activity of PEPC and NADPME were used as markers for M and BS respectively. For RNA sequencing data, Lipoxygenase 2 (GRMZM2G015419) and the sum of Ribulose-phosphate 3-epimerase (GRMZM2G026807) and Phenylalanine ammonia-lyase 1 (GRMZM2G074604) were used as M and BS markers, because these markers showed similar enrichment to-, but more steady enrichment than- PEPC and NADPME throughout development.

K-means clustering. K-means clustering was performed to get an overview of the data and allow qualitative categorization of divergence between paralogs. K-means clustering was performed on all genes where the initial tissue specificity could be estimated in every developmental slice (minimum raw FPKM >0). To choose the number of clusters, the sum of standard error (SSE) of clusters with the original data was compared to the SSE of clusters with scrambled data [86]. We proceeded with 8 clusters as this provided a fairly low SSE for the original data and a large difference in SSE between original and scrambled data Supplementary Fig. 24. Clustering was repeated 10,000 times and the solution with the lowest SSE was selected. Each cluster was tested for functional enrichment in all distinct MapMan [45] categories with a Fisher's Exact test.

Defining divergence. We employ two quantitative methods and one qualitative method to estimate divergence. First, we use transformed Pearson correlation between expression patterns as an interval scaled variable for the amount of divergence in expression pattern. The transformation is performed to provide an unbounded and more normally distributed value. The transformation of Pearson's r (r_p) is equal to $ln(\frac{1+r_p}{1-r_p})$. Second, to measure divergence in expression level, we

recorded the absolute value of the natural log for the ratio between peak expression of the paralogs ($|ln(\frac{peakFPKM1}{peakFPKM2})|$). Finally, to evaluate divergence in a qualitative fashion, we developed a clustering based method to track particular patterns of divergence. Using graph theory, we considered the k-means clusters nodes, and pairs of paralogs formed edges either between them or, when both paralogs occurred in the same cluster, loops. To avoid assigning pairs of paralogs to a divergent (non-loop) edge, if they had a conserved expression pattern that was intermediate between the clusters, we excluded "boundary" pairs from further analyses. Boundary pairs were defined as pairs in a non-loop edge where the r_p between the expression pattern of the two paralogs was higher than the the r_p of either pair to its cluster center.

Regression analysis. Multiple linear regression was used to compare the expression divergence (for both pattern and level) of two paralogs to their other characteristics (dN, dS, dN/dS, number of *Z. mays* paralogs in orthogroup, whether either paralog was a C_4 gene or not). The calculated p-value represents the chance of seeing the observed improvement in model fit of adding the factor in question to a model already containing all the other factors if the null hypothesis (no relation) is true. When comparing values that did not approach a normal distribution (e.g. p-values, FPKM, % abundance in BS or M) we performed Spearman rank correlation.

Controlling pairwise counting bias. Some analyses could be sensitive to a bias resulting from counting the pairwise combinations of different sized orthogroups. For instance, there are three pairwise combinations of the group "a", "b", "c" ("a-b", "a-c", and "b-c") and every group member is counted twice; however, add 'd' to the group and there are six pairwise combinations ("a-b", "a-c", "a-d", "b-c", "b-d", "c-d") and every group member is counted three times. To control for this, without introducing other bias by sub setting the data (e.g. taking reciprocal best blast hits selects for young paralogs from small gene families), we scrambled the data to get an empirical p-value accounting for this bias. Specifically we scrambled expression information, but held gene family information constant and counted the number of instances where the result was as, or more, extreme than the original to obtain an empirical p-value.

We expected this bias to be most problematic for two analyses: the correlation between number of *Z. mays* genes in orthogroup and the expression divergence and the functional enrichment in edges between clusters. To test the significance of the correlation between the number of *Z. mays* genes in the orthogroup and expression divergence we scrambled the expression patterns of genes and re-calculated rp between the afore mentioned values 3200 times. To test for an enrichment (or depletion) in edges between clusters and MapMan functional categories, we scrambled the cluster assignment of 'divergent' and 'conserved' pairs 63999 times, and counted the number of cases where each functional category was more or less enriched than the original in each cluster pair using the Python Language [87].

Calculating divergence on phylogenetic tree We used a mean of pairs method to calculate the divergence for nodes of a phylogenetic trees. Pairs consisted of any genes originating from the same species and occurring on different daughter branches of the node. The mean divergence across all pairs was taken as the divergence at the node. The test set where this was calculated

consisted of 64 orthogroups of 60 genes or less with at least one divergent pair of *Z. mays* genes and one conserved pair of *Z. mays* genes. The orthogroups were sorted by the expression of lowest paralog contributing to the conserved or divergent pair, and the 64 most highly expressed were chosen.

Methods References

64. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 289–300 (1995).
65. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40,** D1178–D1186 (2012).
66. Monaco, M. K. *et al.* Gramene 2013: comparative plant genomics resources. *Nucleic acids research* **42,** D1193–D1199 (2014).
67. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25,** 3389–3402 (1997).
68. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13,** 2178–2189 (2003).
69. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research* **40,** e49–e49 (2012).
70. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 10557–10562 (2005).
71. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics* **9,** 286–298 (2008).
72. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research* **34,** W609–W612 (2006).
73. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
74. Huerta-Cepas, J., Dopazo, J. & Gabaldón, T. ETE: a python Environment for Tree Exploration. *BMC bioinformatics* **11,** 24 (2010).
75. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24,** 1586–91 (Aug. 2007).
76. Yang, Z. & Dos Reis, M. Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution* **28,** 1217–1228 (2011).

77. Walker, R. P., Trevanion, S. J. & Leegood, R. C. Phosphoenolpyruvate carboxykinase from higher plants: purification from cucumber and evidence of rapid proteolytic cleavage in extracts from a range of plant tissues. *Planta* **196,** 58–63 (1995).

78. Hatch, M. & Mau, S. Properties of phosphoenolpyruvate carboxykinase operative in C_4 pathway photosynthesis. *Functional Plant Biology* **4,** 207–216 (1977).

79. Rudolf, M. *et al.* In vivo function of Tic22, a protein import component of the intermembrane space of chloroplasts. *Molecular plant,* sss114 (2012).

80. Kodama, Y., Shumway, M. & Leinonen, R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic acids research* **40,** D54–D56 (2012).

81. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17,** pp–10 (2011).

82. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14,** R36 (2013).

83. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7,** 562–578 (2012).

84. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research* **41,** D991–D995 (2013).

85. Smyth, G. K. in *Bioinformatics and computational biology solutions using R and Bioconductor* 397–420 (Springer, 2005).

86. Peeples, M. *R Script for K-Means Cluster Analysis* 2011. <http://www.mattpeeples.net/kmeans.html>.

87. Van Rossum, G. & Drake Jr, F. L. *Python reference manual* (1995).

Supplementary Information

Supplemental Note

C_4 **cycle.** The key feature of C_4 photosynthesis is the biochemical pump which concentrates CO_2 at the site of Rubisco and suppress the costly process of photorespiration. This can result in a 50% increase in photosynthetic efficiency [88]. To achieve this, C_4 plants use the non-oxygen sensitive enzyme, Phospho*enol*pyruvate Carboxylase (PEPC), to fix CO_2 onto Phospho*enol*pyruvate (PEP) in the M. The resulting 4-carbon acid must diffuse to the BS, and be decarboxylated, releasing CO_2. In *Z. mays*, the primary decarboxylating enzyme is NADP Malic Enzyme (NADPME); however, around 15% of the carbon appears to flow through the secondary decarboxylating enzyme PEP Carboxykinase (PEPCK) [38, 89, 90]. The resulting 3-carbon acid diffuses back to the M and is regenerated to PEP, as necessary, completing the cycle. In addition to the carbon shuttle, a small part of the Calvin Benson Bassham cycle is localized to the M and the rest to the BS, resulting in a triphosphate based redox shuttle transporting reducing equivalents to the BS. Both the C_4 cycle and redox shuttle require upregulation of metabolite transporters to support the high flux of metabolites in and out of subcellular compartments. However, only two transporters have been fully characterized in *Z. mays*. Here-after, when we refer to the core-C_4 cycle, we are referring to the enzymes of the primary and secondary C_4 cycle, the known transporters Phosphoenolpyruvate/Phosphate Translocator (PPT) and Triose Phosphate Transporter (TPT), and the two established regulatory proteins PEPC Kinase (PPCK) and Pyruvate Phosphate Dikinase - Regulatory Protein (PPDK-RP).

The elements of the key C_4 cycle are well distributed in our data. Transcripts, and where available enzyme activity, for the enzymes responsible for regenerating PEP (Pyruvate Phosphate Dikinase, PPDK), converting (Carbonic Anhydrase; CA) and fixing the CO_2 (PEPC), and converting the resulting oxaloacetate (OAA) to the transfer acid Mal (NADP Malate Dehydrogenase; NADPMDH) are higher in the M as expected ($p < 0.05$, enzymes; fdr < 0.05, transcripts in Slice 3 - 1; except NADPMDH in Slice 2 where fdr = 0.058; Supplementary Fig. 10, 11. The decarboxylation enzymes are both higher in the BS (fdr < 0.05, transcripts in Slice 4 - 1; Supplementary Fig. 10, 11), and the enzyme which can convert OAA that was transported as aspartate to malate (NAD Malate Dehydrogenase; NADMDH) showed a preference for the BS ($p < 0.05$, enzyme in Slice 1; fdr < 0.05, transcripts in Slice 3 - 2);Supplementary Fig. 10). Several activities in the cycle are expected to be balanced between tissue types, including the TPT transporter, and the Aspartate- and Alanine-Amino Transferase (AspAT and AlaAT). TPT expression is quite even between tissues Supplementary Fig. 12, while AlaAT is enriched in the M at the level of transcripts but not enzyme activity Supplementary Fig. 10. In contrast, for AspAT both enzyme activity and transcripts are strongly enriched in the M. However, we find a M specific paralog with high expression level, and a BS specific paralog with a low expression level, which is very consistent with the previous studies [39, 44], and even with *S. italica* [35].

Metabolites. The metabolic data is hard to interpret as separation was not sufficient to produce significant results after multiple hypothesis correction. However, as there is very little data available for the separation of metabolites between BS and M cells, we want to describe the data anyways to provide information that may help in the design or analysis of future studies.

This study shows the care that will be required to confidently measure the values of photosynthetically active metabolites. The major advantage of the employed technique, is the immediate shock freezing, and frozen processing of tissue, which allows very little time for changes in leaf metabolome. Unfortunately, the employed thecnique allows for only modest enrichment of tissues, and in contrast to enzymes and transcripts there are no known internal metabolite controls that are close-to-perfectly tissue specific, and as small molecular weight metabolites can readily diffuse across the plasmodesmata, there are unlikely to be any fully tissue specific and cytoplasmic metabolites. Therefore, enzyme activity was used for normalization.

Although nothing was significant, we will try to briefly summarize the trends in the data. Metabolites in the core-C_4 cycle all behaved similarly in our data, with a tendency towards BS enrichment in slice 3_4 and a tendency towards M enrichment in slice 1_2 (Supplementary Fig. 25). The mature tendency towards M matches expectation for aspartate and malate, which need to diffuse from the M to the BS. Two previous studies [42, 91] also estimated that concentrations of malate were higher in the M than the BS (Supplementary Fig. 30). Glutamate and α-ketoglutarate are not expected to show a net flux between tissues, and the tendency towards M in mature tissue is therefore unexpected; however, the estimated % M is surprisingly consistent with that reported by [91]. In contrast α-alanine showed a tendency opposite to that of the expected concentration gradient, and incosistent with the previously reported even distribution [91] (Supplementary Fig. 30). Notably, there were also major differences between the fast 2-slice harvest and slower 5-slice harvest (e.g. α-ketoglutarate; Supplementary Fig. 30). The lack of statistically significant enrichments, differences between the developmental stages, differences in slow harvest vs fast harvest, and inconsistency with previous data (Supplementary Fig. 30), point to, if nothing else, the lability of metabolites. The same lability that makes metabolites hard to measure between experiments and sensitive for instance to shading or cooling, means the plant must be able to tolerate a non-continuous distribution of metabolites between tissues.

All the measured photorespiratory metabolites had a tendency towards BS enrichment, as is expected with the BS specific localization of the photorespiratory cycle (Supplementary Fig. 26). Other categories of sugars (Supplementary Fig. 27), amino- (Supplementary Fig. 28) and other organic acids (Supplementary Fig. 29) showed a variety of distributions, with frequent change both in level and tendency towards tissue specificity between the two slices. Indeed the metabolites appeared to show more frequent changes in tissue preference than the enzymes or transcripts. While this may simply reflect the generally high error and low-significance, it may also, in part, reflect how dynamic the metabolome is.

Transcription factors of interest There is strong interest in engineering the C_4 trait into C_3 crop species to increase photosynthetic efficiency and ultimately growth and yield. However, the complexity of the C_4 trait goes well beyond the capabilities of even the most successful current engineering methods. However, the highly convergent nature of C_4-evolution provides hope that extreme changes may be facilitated by comparatively simple changes in regulatory architecture. Therefore, we used the compiled expression data to highlight some top-candiate transcription factors of interest to understanding C_4 photosynthesis and its evolution.

Individual studies targeting transcription factors of interest to the C_4 trait in *Z. mays*, have provided candidate lists from X-Y members [30, 38, 39, 43, 44]. While this remains a very ambitious number for individual characterization, taking the intersection of various studies is an extremely strict measure, that results in 0 remaining candidates [43]. Therefore, we take a more permissive and inclusive approach to find transcription factors that are of interest in understanding C_4 photosynthesis supported by four or more of the following six criteria relating to C_4 photosynthesis, its evolution, and kranz anatomy. The criteria were: 1) significantly associated with either the M or BS marker in all 5 slices in this study; 2) consistent direction of enrichment across all samples and studies (all BS >M or all M >BS; 3) the FPKM in *Z. mays* leaf ("V5_Tip_s-2_Leaf", [31] was at least twice that of both *B. distachyon* and *O. sativa* leaves [32]; 4) The peak expression in floral primordia was at least 1.5 times that of husk primodia [30]; 5) expressed at least 20 FPKM in floral primordia; And 6) show a correlation to the PS expression pattern (r_p) higher that 0.4 in *Z. mays*, but not in *B. distachyon* or *O. sativa*. In total, 19 transcription factors met these criteria (Supplementary Dataset 5).

Among the identified transcription factors are ones with particularly interesting orthologs in *A. thaliana*. Three DOF transcription factors were selected (GRMZM2G114998, AC233935.1_FG005, and GRMZM2G179069), all of which had higher FPKM in maize and the other C_4 species than either C_3 species, a photosynthetic-like expression pattern in *Z. mays* but not in either C_3 species, and were more highly expressed in floral than husk primordia. Further, in concordance with the enrichment of the whole DOF family among BS specific genes, all three selected DOF genes were higher in the BS of every comparison, and significantly higher in the BS in every slice of our leaf gradient. The *A. thaliana* ortholog of GRMZM2G114998, AT4G24060 or DOF4.6, is expressed at the sites of early vein formation [92], making DOF4.6 an interesting candidate in understanding the narrower vein spacing in C_4 species. The other two DOF family transcription factoers, GRMZM2G179069 and AC233935.1_FG005, share their closest *A. thaliana* homolog, AT3G55370 or OBP3, which is a mediator of phytochrome signaling [93]. Phytochrome signaling is a major regulator of photomorphogenesis or how a plant develops in response to light [94]. Another mediator of phytochome signaling, the COP9 signalosome, has been putatively linked to the differences in leaf development seen between C_3 and C_4 sister species [18].

Two auxin response regulators were identified, both of which were higher in *Z. mays* than either C_4 species, and higher in floral than husk primordia. Further ARF3 was expressed highly in the floral primordia, and consistently higher in M than BS; while AXR2 had a photosynthetic pat-

tern in *Z. mays* and *S. bicolor* that was not shared with the C_3 species, and was consistently higher in BS than M. In *A. thaliana*, ARF3 (AT2G33860) helps mediate the specification of abaxial and adaxial fate [95, 96]. In a study in grasses, the C_4 leaves showed more asymmetry, and modified M/BS ratios between abaxial and adaxial regions, while the C_3 leaves did not [97]. AXR2 (AT3G23050) is involved in the interplay between ABA and auxin response [98]. Auxin is a major hormone for specifying vascular cell fate [99], and modifications in auxin signaling, through modifications in synthesis, transport, perception and timing, are thought to be related to the specialized vein pattering in C_4 species [48]. Finally, in relation to the enhanced secondary cell walls in BS, MYB52 (GRMZM2G455869) is an exceptionally interesting candidate. MYB52 showed over twice the FPKM in the C_4 species compared to the C_3 species, showed a photosynthetic-like expression pattern specifically in the C_4 species, was expressed more highly in the BS in every comparison, and significantly so across the leaf gradient. A. thaliana over expressing MYB52 (AT1G17950) show hypersensitivity to ABA and increased drought tolerance [100]. MYB52 was further identified in a "post-genomic" screen for secondary cell wall related proteins, and it's mutant showed hyper-lignification [101]. In summary, the transcription factors discussed here and the rest from (Supplementary Dataset 5) are highly interesting candidates, which warrant further investigation to see if their promising expression patterns and annotations might help drive any of the features of BS or M tissue specificity in C_4 species.

Advancements in understanding the differences between BS and M cells. To determine if this separation method was consistent with previous studies at a functional level, we tested sets of genes significantly co-regulated with M and BS markers and our k-means clusters for enrichments in MapMan functional categories. To facilitate the comparison of the various M and BS separation studies, we re-ran enrichment testing for all provided [39, 43] or described [44] gene sets that were considered differentially regulated between BS and M cells. For comparability, each set was compared to a background of the 6a genome release. Enrichments in genes specific to the BS were quite consistent between studies (Supplementary Fig. 5), with a handful of categories shared between all samples and studies. Many of these categories are well understood (e.g. the Calvin Benson Bassham cycle) or have hypothesized benefits (e.g. S- assimilation, the DOF transcription factor family; [39, 44, 60, 102]. However, one previously un-examined category, misc.myrosinases-lectin-jacalin, was consistently enriched in the BS. An *A. thaliana* homolog (AT4G19840) of the *Z. mays* myrosinases-lectin-jacalins is a phloem sap protein with a putative role in defense [103, 104], indicating that this category may relate to conserved BS functions and not C_4 photosynthesis. In addition to functions that were consistent across all tissues, many sub categories of protein synthesis were enriched in BS specific genes specifically in three comparable younger tissues (slice 4, slice 3, and section -1 from [43]).

Enrichments in M specific genes showed greater variability between studies (Supplementary Fig. 4). While no categories were enriched in every sample, there was still a strong bias for particular categories. For instance, 20 categories were enriched in seven or more of the ten samples. These included several subcategories of the photosynthesis light reactions, particularly photosystem II; lipid metabolism and lipid transfer proteins; isoprenoid/carotenoid synthesis; and light signaling.

Interestingly, transport was consistently enriched in both M and BS specific genes, indicating it is a category generally undergoing specialization between tissues.

The above analyses compared genes differentially expressed in each developmental slice individually, and to integrate gradient and tissue specificity patterns we performed a functional enrichment analysis on k-means clusters. As clustering was performed only on genes expressed sufficiently to be "deconvoluted" (min FPKM >0), but compared, as above, to the unfiltered 6a genome; some major categories such as RNA, protein, and signaling were enriched in most to all clusters, and not assigned.unknown was frequently depleted. Therefore, we focus on the smaller categories and more specific enrichments.

Clusters 2 and 5 consisted of genes with high expression in the BS base and tip, respectively, and showed a distinct set of enrichments. In cluster 2 many developmental and structural categories were enriched; including cell and cell organization; cell wall proteins and precursor synthesis; lignin synthesis; and categories likely related to the cell wall such as β- 1,3 glucan hydrolases. Additionally several regulation related categories were enriched, such as hormone metabolism with jasmonate and auxin response, and a few transcription factor families. In contrast, in cluster 5, with expression high in the BS tip, the enrichments were dominated by major energy and metabolism categories. In relation to energy production, cluster 5 was enriched in the photosynthetic categories of Calvin Benson Bassham cycle, and photorespiration, as well as mitochondrial electron transport and the TCA cycle. Related to metabolism, cluster 5 was enriched in major and minor carbohydrate metabolism, sulfur metabolism, nucleotide metabolism, secondary nitrogen metabolism, polyamine metabolism, and the oxidative pentose phosphate pathway. Finally, cluster 5 was enriched in a set of regulatory categories distinct from that of cluster 2, including ethylene metabolism and response, and six transcription factor families, of which, only basic Helix-Loop-Helix was shared with cluster 2. In summary, while genes and categories significantly up-regulated in the BS were fairly constant across the leaf (Supplementary Fig. 5) [43], strong differences could be seen in functions of clusters peaking in the BS base or tip, with the base more specialized in development, cell wall and lignification, while the tip was more specialized in photosynthesis and metabolism. Both BS base and tip were enriched in regulatory genes, but largely distinct subcategories of hormone metabolism/response, and transcription factor families.

Similarly, in the M we observed distinct enrichments in the M base cluster (4) and the M tip cluster (3). In the M base cluster 4, enrichments included lipid metabolism and some development, protein and signaling categories, as well as tetrapyrolle synthesis. While in the M tip cluster 3, there were strong enrichments in photosynthesis including both photosystem I and photosystem II, and a concomitant enrichment in light stress. In addition, cluster 3 was enriched in isoprenoid and flavenoid biosynthesis, and the often down-stream-of-photoreceptors family, CONSTANS.

Clusters expressed highly in both tip tissues (6) or both base tissues (7) showed enrichments distinct from the individual tissue types. Most striking in cluster 6, were not the few enrichments,

such as heat stress, that were specific to this cluster; but the lack of an enrichment in the PS categories that was so characteristic of the tissue specific tip clusters 3 and 5. The even base cluster 7, shared cell wall enrichments with the BS base cluster 2, showed distinct auxin related enrichments (auxin response factor (ARF) and Aux/IAA family instead of the auxin.induced-regulated-responsive-activated in cluster 2), and was the only cluster enriched in brasinosteroid metabolism.

The "mixed" categories 1 and 8 appear to contain biological information despite their weird appearance. The deconvolution method is such that it can induce a small pattern in a fairly evenly expressed gene. Double checking the raw data for these clusters, we see that cluster 1 can be described as expressed evenly high in the base, and otherwise slightly higher in the BS than M, while cluster 8 can be described as expressed highest in tip and base, and shows mild M enrichment in some slices (Supplementary Fig. 31). Cluster 1 shared enrichments with other more basal clusters, like cell wall and protein synthesis, as well as histones. Cluster 8 was enriched in cytoskeleton, lipid degradation, minor CHO metabolism, protein degradation and targeting, various regulation of transcription and signaling pathways, and various stress categories.

Bundle fraction contains not only BS but also tracheary elements. Consistent with expectations for the enrichment method, the transcriptome reflects co-enrichment of the vascular tissue with the BS tissue. Ethylene response is enriched in the BS in every slice, which has been implicated in triggering cambial cell division and xylem growth in populus and Zinna cell cultures [105, 106]. We observed many positive regulators of tracheary elements with peak expression in the basal BS slice (BS5). Both vascular cells and BS cells have highly developed and lignified secondary cell walls, which would be difficult to tease apart from each other in the enrichments in cell wall and lignin biosynthesis in basal BS up-regulated genes. However, LAC17 is necessary for lignification of the protoxylem elements in *A. thaliana* [107], and three of it's homologs in *Z. mays* are expressed (21-309 FPKM), and BS specific (fdr <0.05) in the basal slice. In the vasculature, programmed cell death is induced after secondary cell wall deposition [108]. Among genes associated with programmed cell death we find XYLEM CYSTEIN PROTEASE (XCP) 1 (GRMZM2G066326) and 2 (GRMZM2G367701) highly (664 and 398 FPKM) and specifically (fdr <0.01) expressed in BS5.

Supplemental Datasets

Supplementary Dataset 1: Spreadsheet with transcriptional, annotation, and mapping information for *Z. mays* genes

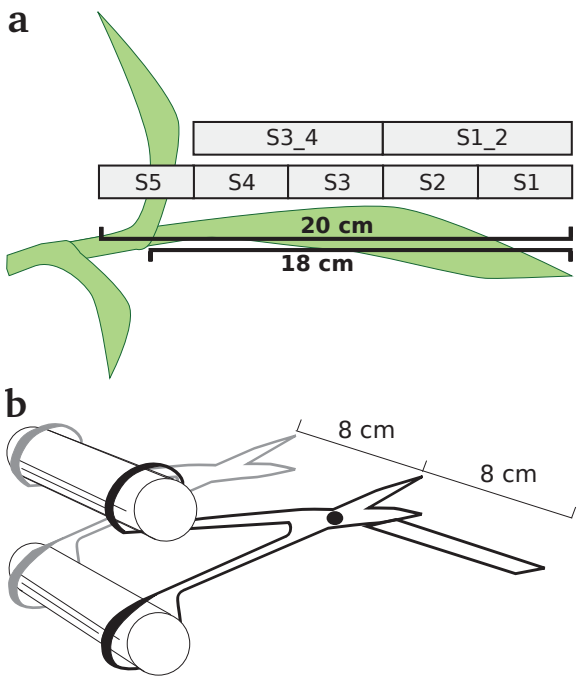
Supplementary Dataset 2: Spreadsheet with metabolic and enzyme activity data

Supplementary Dataset 3: Spreadsheet with significant enrichments for tissue specific genes in each slice and for k-means clusters

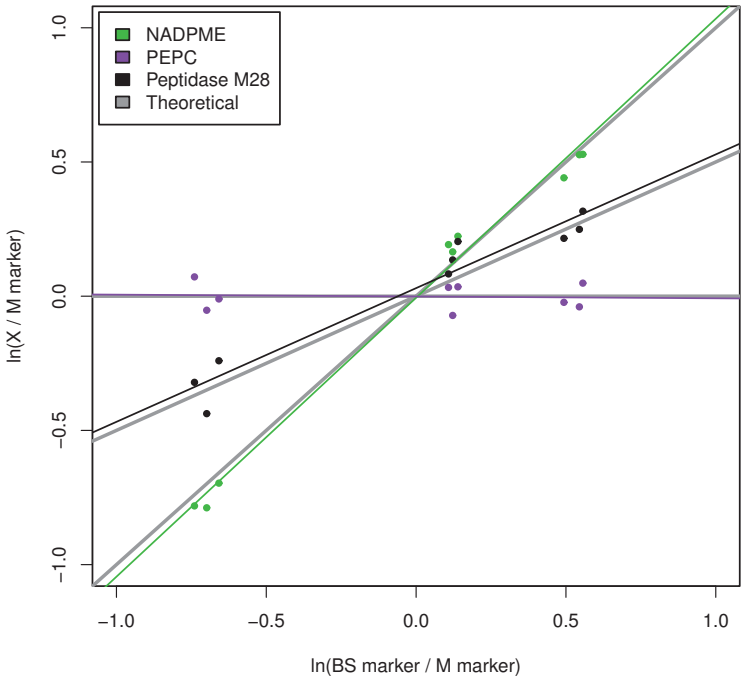
Supplementary Dataset 4: Spreadsheet with significant enrichments for edges between clusters

Supplementary Dataset 5: Spreadsheet with transcription factors meeting the criteria of interest

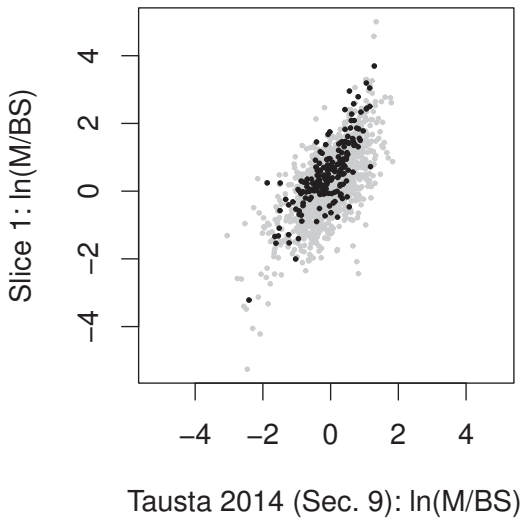
Supplemental figures



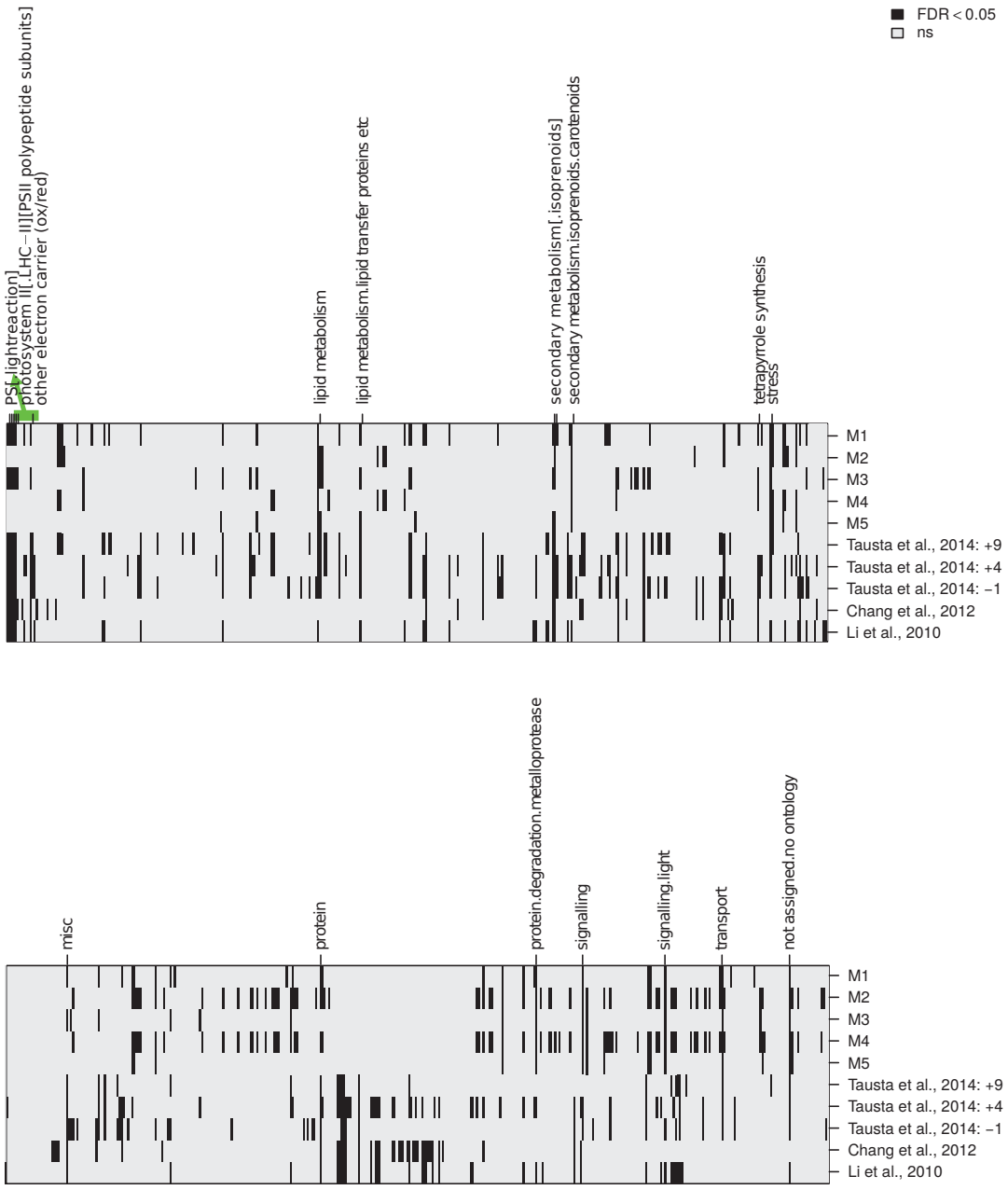
Supplementary Fig. 1: Visual summary of tissues (a) and harvest method (b). The five 4 cm slices (a) were harvested for transcriptome analysis using the leaf "guilotine" [38], while the two 8 cm (a) slices were harvested for metabolite analysis using two pairs of attached scissors (b).



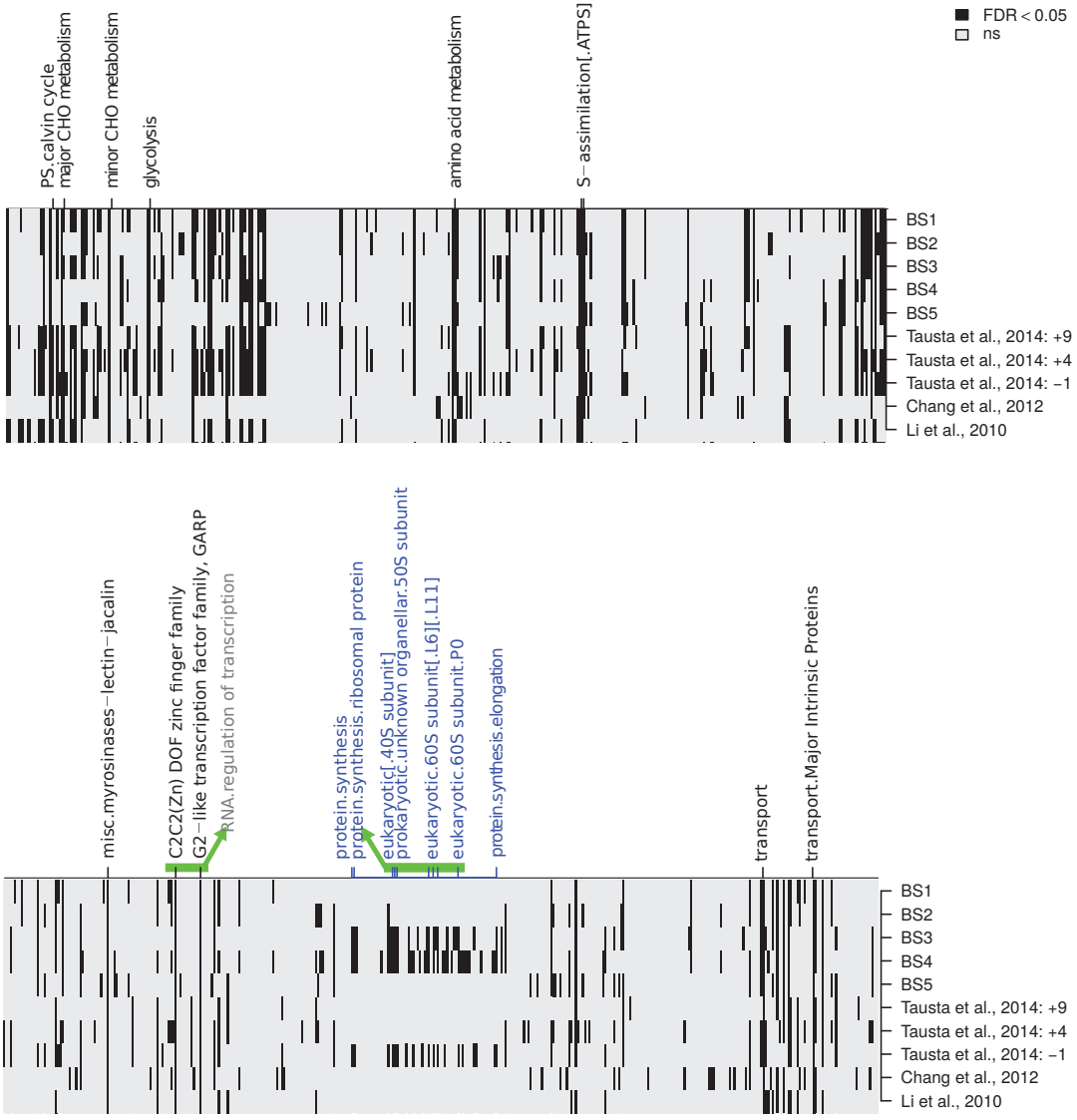
Supplementary Fig. 2: Example comparision between target genes and markers used to "deconvolute" data, that is, estimate the original distribution of target abundance between BS and M cells [42]. PEPC (GRMZM2G083841) as an example of a M specific target, NADPME (GRMZM2G085019) as an example of a BS specific target, and a peptidase M28 (GRMZM2G159171) as an example of a non-enriched target. The slope of the linear regression line yields the estimated fraction of abundance in BS.



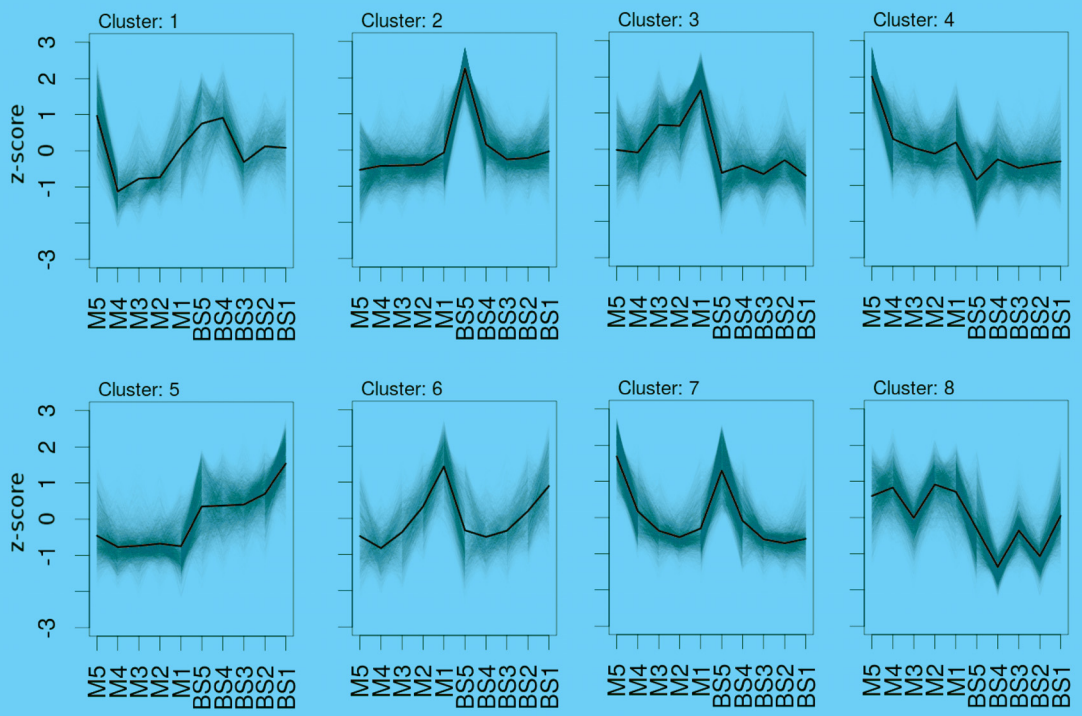
Supplementary Fig. 3: Comparison between transcript BS/M ratios in mature leaf in "deconvoluted" data (Slice 1) and a previous study using laser micro disection (Section +9) [43].



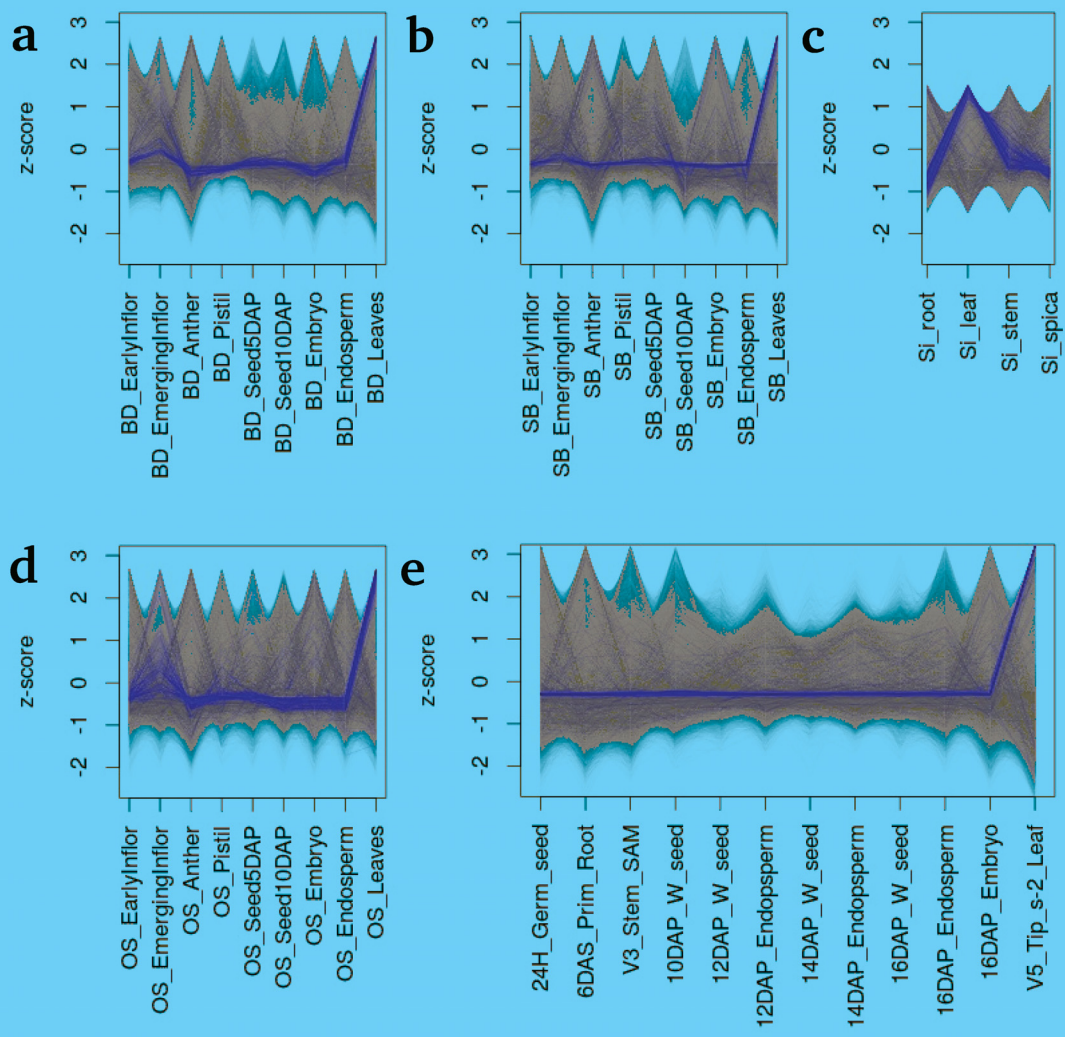
Supplementary Fig. 4: Comparison of tissue enrichments between studies. For comparability all enrichments were calculated with the significantly tissue-specific genes (as defined in each study) in the foreground and the remainder of the unfiltered 6a genome in the background for a Fisher's exact test. The most-consistent enrichments (those enriched in at least 7 samples) are labeled. For the sake of compact display, green bars indicate highlighted categories are subcategories of the more basal category indicated with an arrow, and square bracets indicate enrichment of both basal category and [sub category] 34



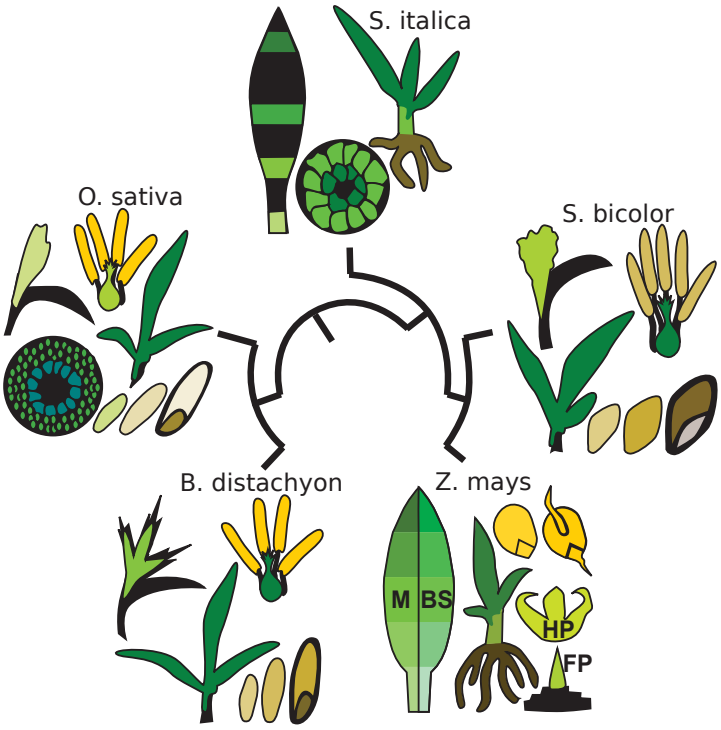
Supplementary Fig. 5: Comparison of tissue enrichments between studies. For comparability all enrichments were calculated with the significantly tissue-specific genes (as defined in each study) in the foreground and the remainder of the unfiltered 6a genome in the background for a Fisher's exact test. The most-consistent enrichments are labeled. Categories enriched in every sample are labeled in black, and those only enriched in the most similar immature tissues (Slice 4 and 3 here, and section -1 from [43]) are labeled in blue. For the sake of compact display, green bars indicate highlighted categories are subcategories of the more basal category indicated with an arrow, and square bracets indicate enrichment of both basal category and [sub category].



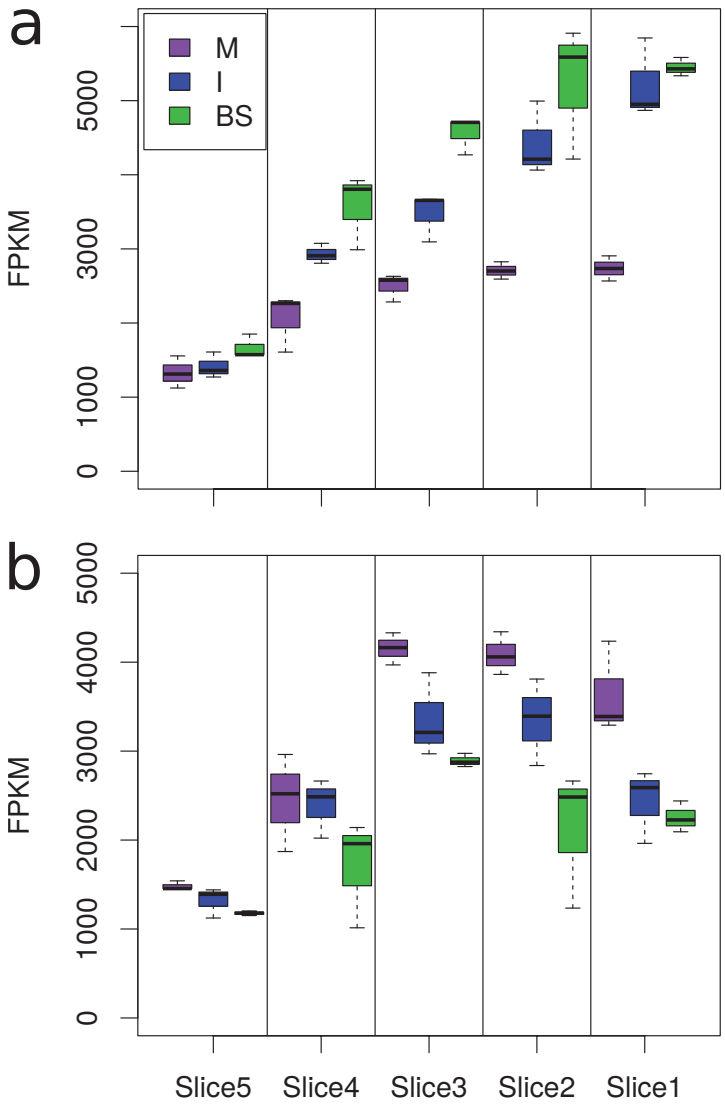
Supplementary Fig. 6: K-means clustering of BS & M gradient, individual genes in grey and centers in black.



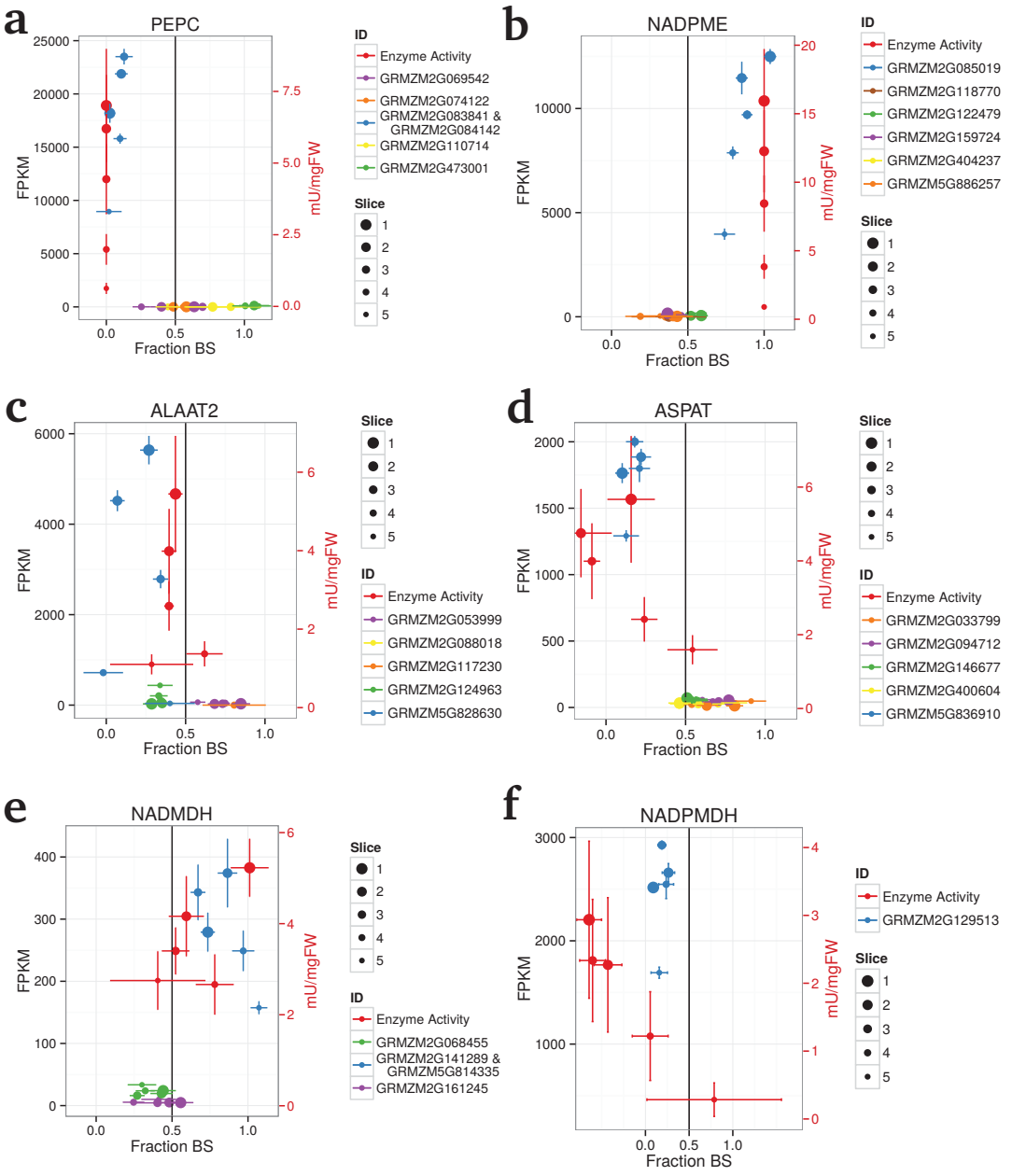
Supplementary Fig. 7: Photosynthetic expression pattern (blue) in tissues used to define it in (a) *B. distachyon*, (b) *S. bicolor*, (c) *S. italica*, (d) *O. sativa*, and (e) *Z. mays*.



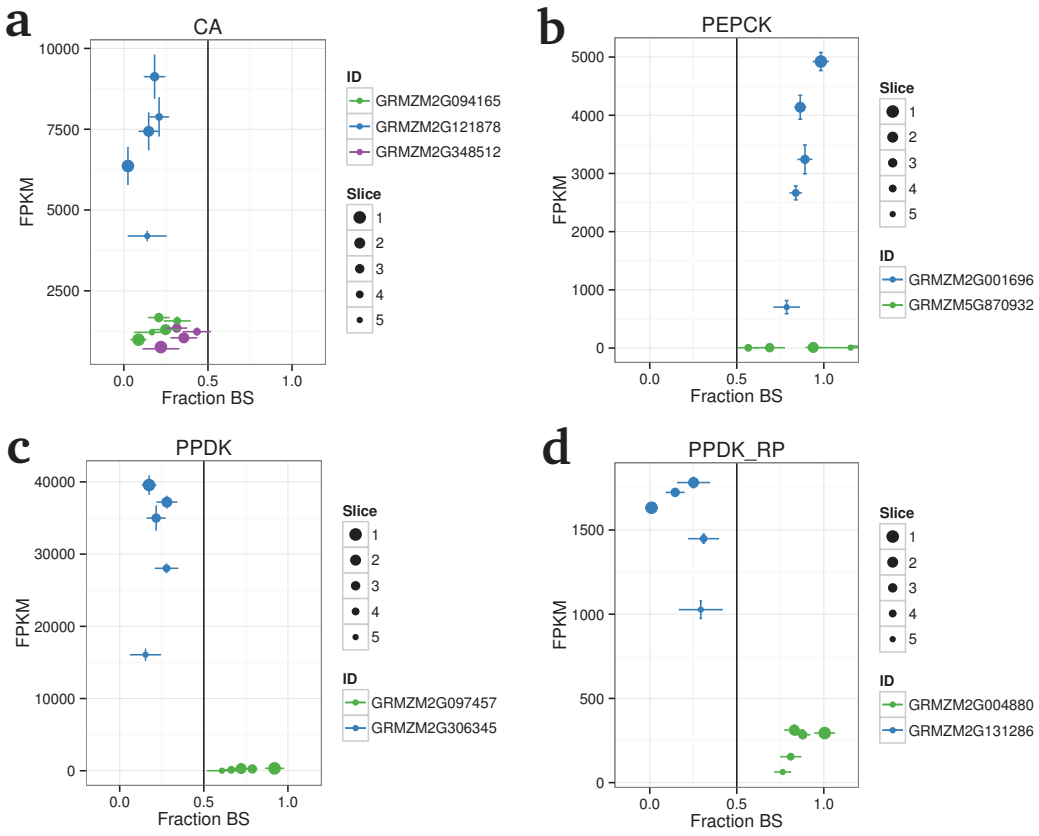
Supplementary Fig. 8: Cartoon summary of the types of tissues coverd by the expression data in each species.



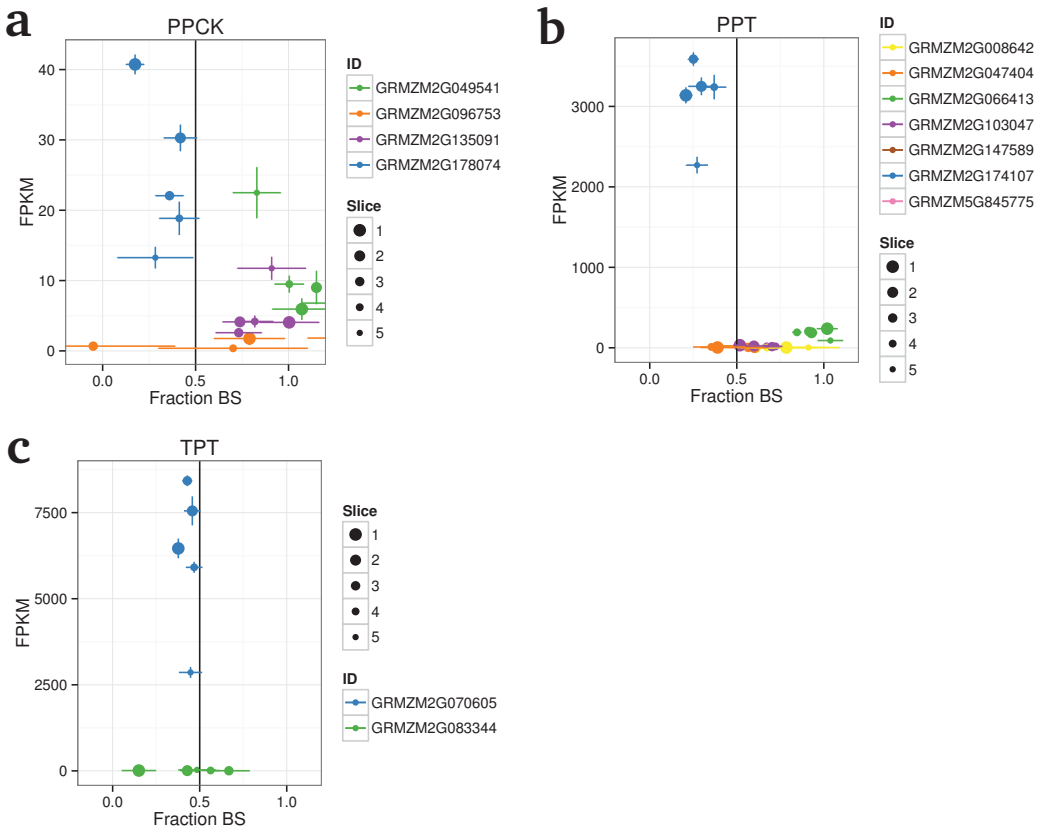
Supplementary Fig. 9: Distribution of transcript abundance of genes with known tissue specificity in raw data. In the BS (a), NADPME (GRMZM2G085019), and in the M (b) PEPC (GRMZM2G083841).



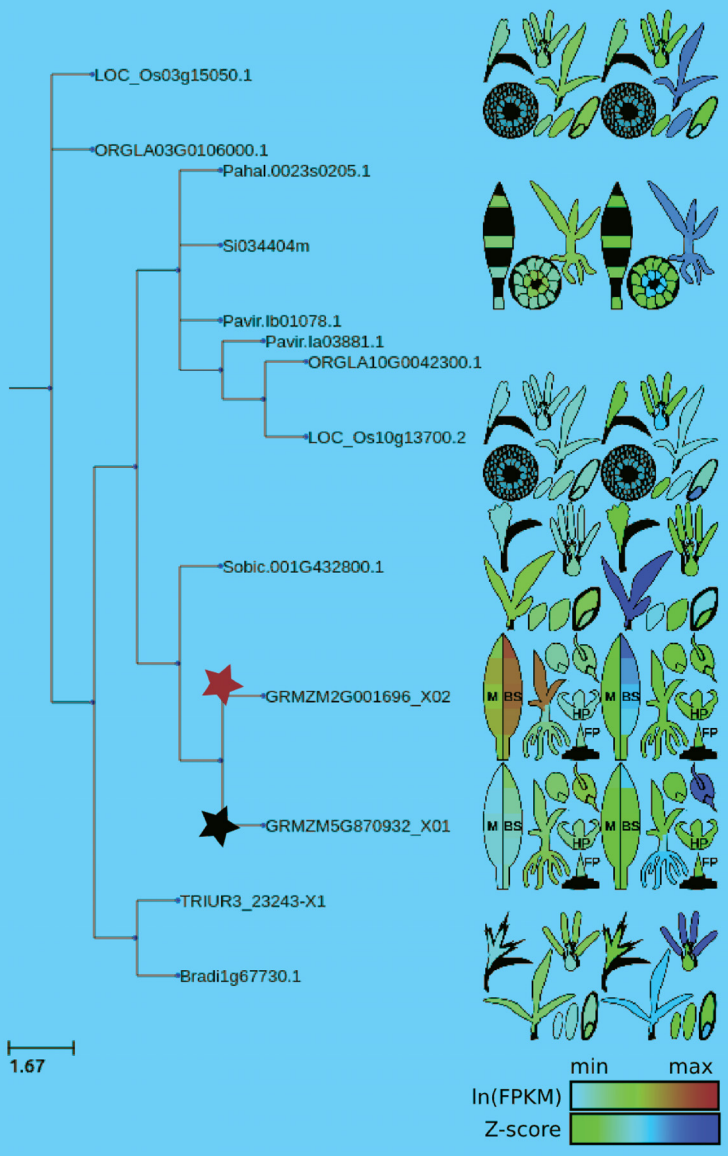
Supplementary Fig. 10: Abundance and specificity of enzyme activity (red) and trancripts (non-red colors represent different paralogs) in core C_4 gene families. Error bars show standard error. In PEPC (a) and NADPME (b) the enzyme activities were used as the markers, and therefore are defined at 0 and 1 fraction in BS, respectively. The remaining families are AlaAT2 (c), AspAT (d), NADMDH (e), NADPME (f). Two identifiers per color indicate the sum of the genes annotated on positive and negative strand at same loci is used.



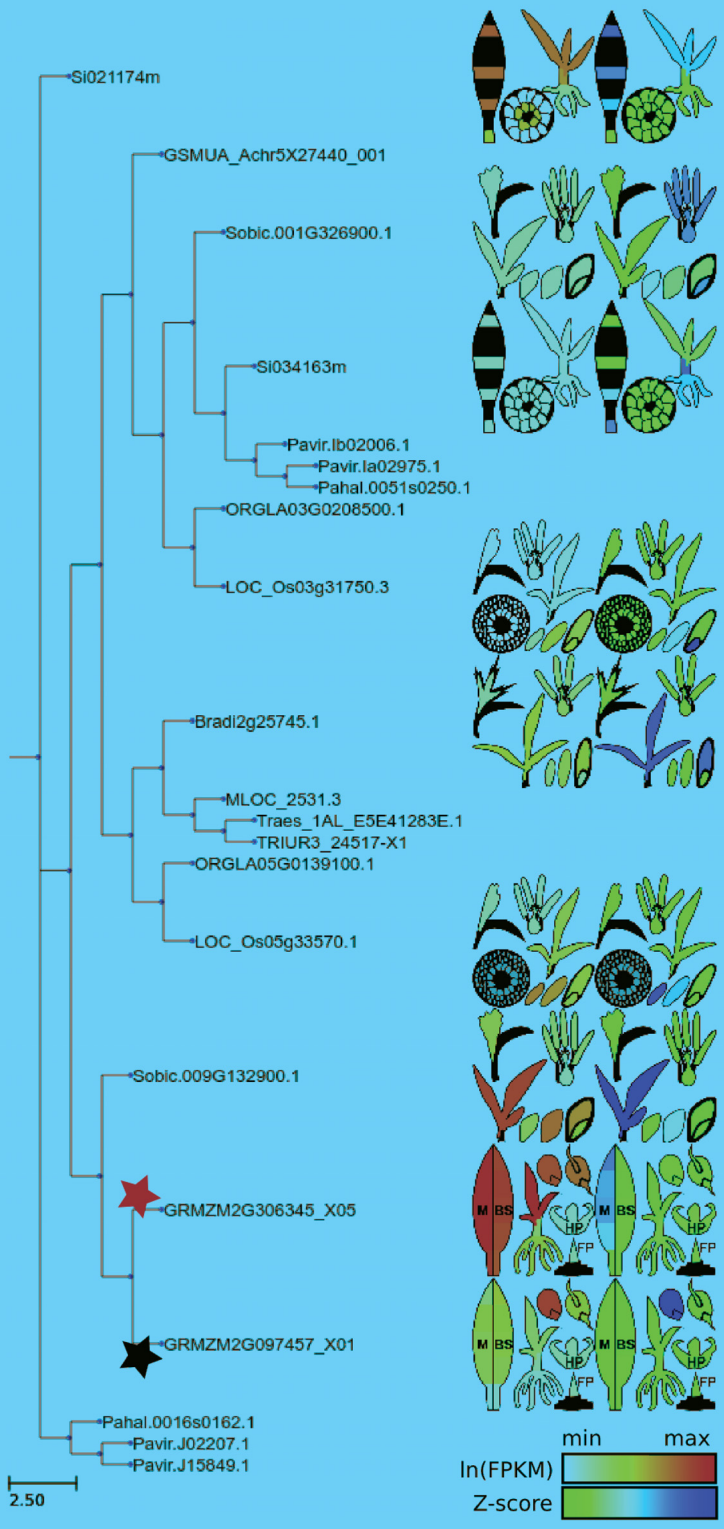
Supplementary Fig. 11: Abundance and specificity of trancripts (colors represent different paralogs) in core C_4 gene families. Error bars show standard error. Families are CA (a), PEPCK (b), PPDK (c) and PPDK-RP (d).



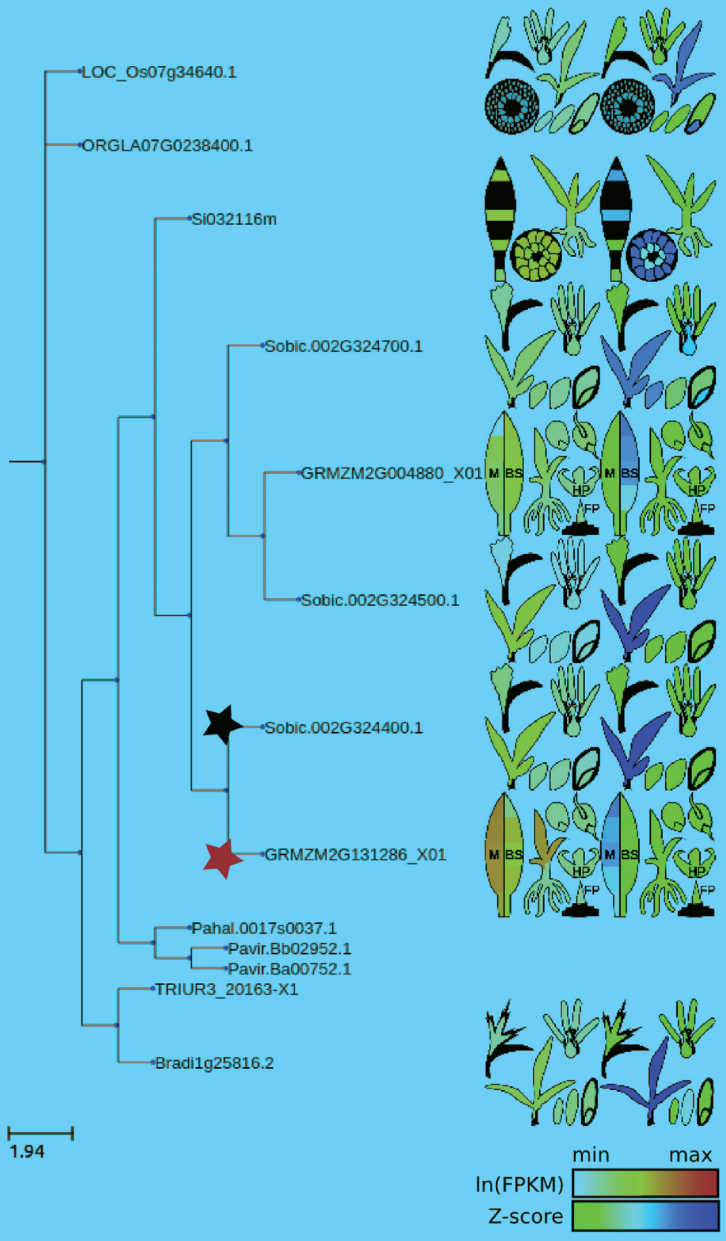
Supplementary Fig. 12: Abundance and specificity of trancripts (colors represent different paralogs) in core C_4 gene families. Error bars show standard error. Families are PPCK (a), PPT (b), and TPT (c).



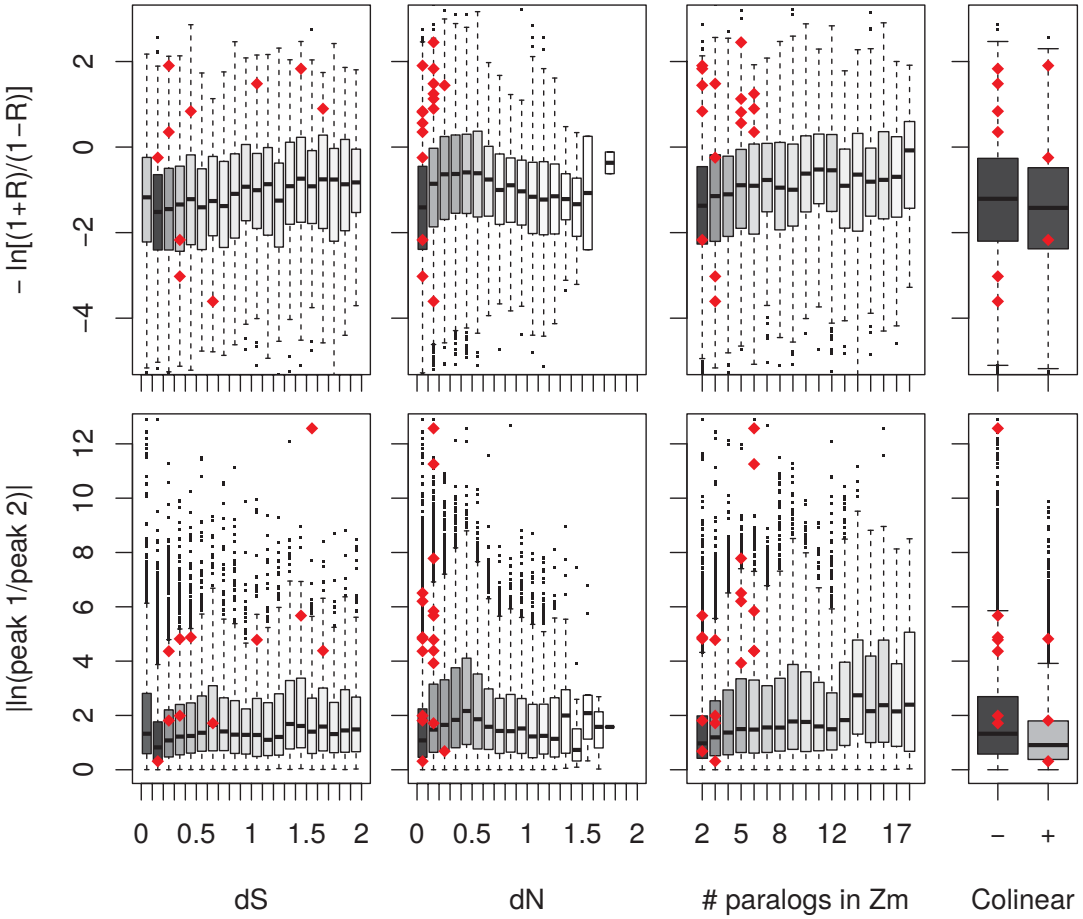
Supplementary Fig. 13: Expression pattern of PEPCK gene family on phylogeny.



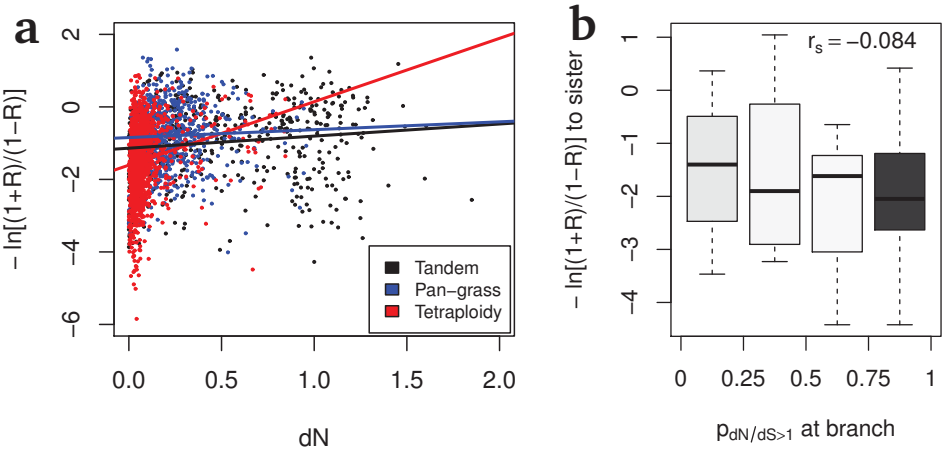
Supplementary Fig. 14: Expression pattern of PPDK gene family on phylogeny.



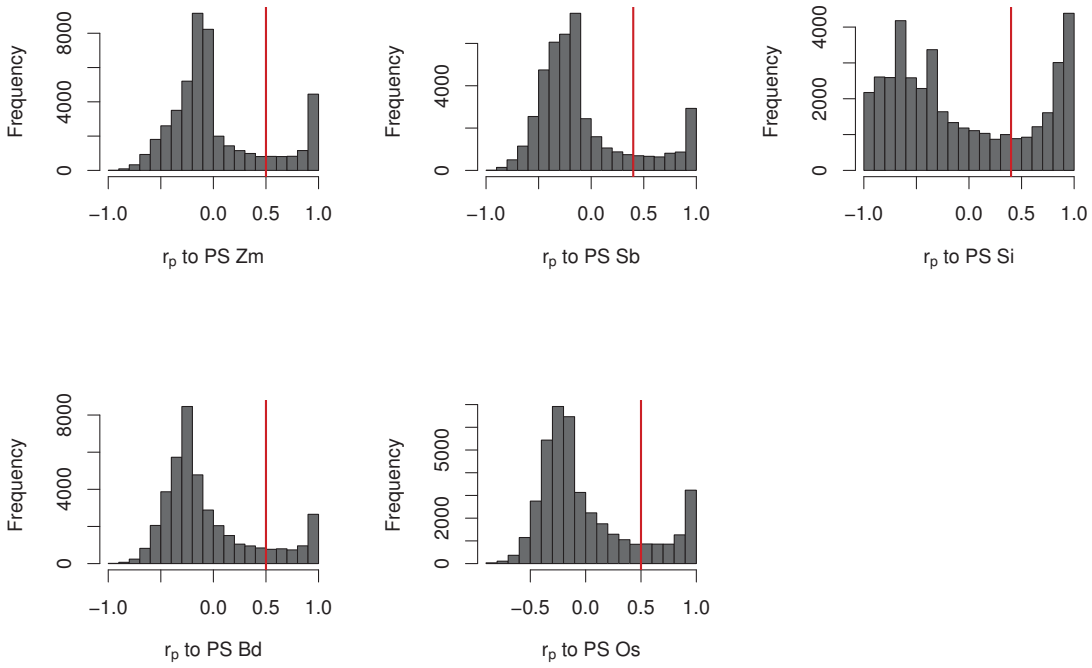
Supplementary Fig. 15: Expression pattern of PPDK-RP gene family on phylogeny.



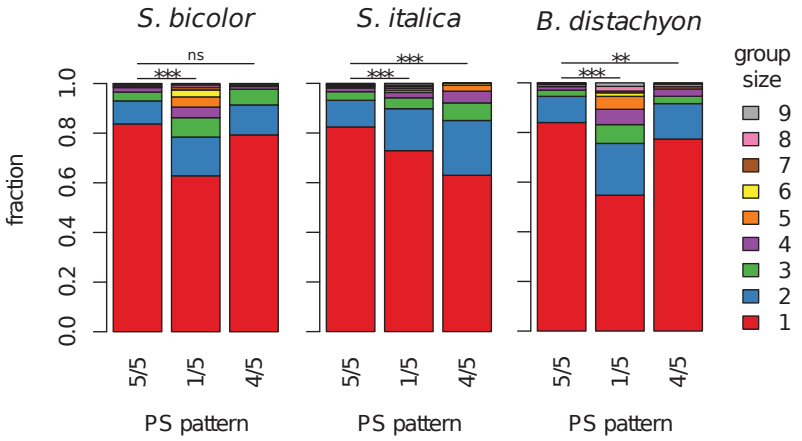
Supplementary Fig. 16: Comparison of sequence and gene family factors to divergence in pattern (transformed pearson correlation; above) and level divergence ($|ln(peakFPKM1/peakFPKM2)|$; below). Shading indicates number of pairs in bin, relative to the largest bin. The red diamonds indicate pairs including core C_4 genes. To control age prior to plotting colinearity, pairs were filterd to those with dS <1.6, after which the mean dS of colinear (0.397) and non-colinear (0.402) pairs did not significantly differ (t-test $p = 0.44$).



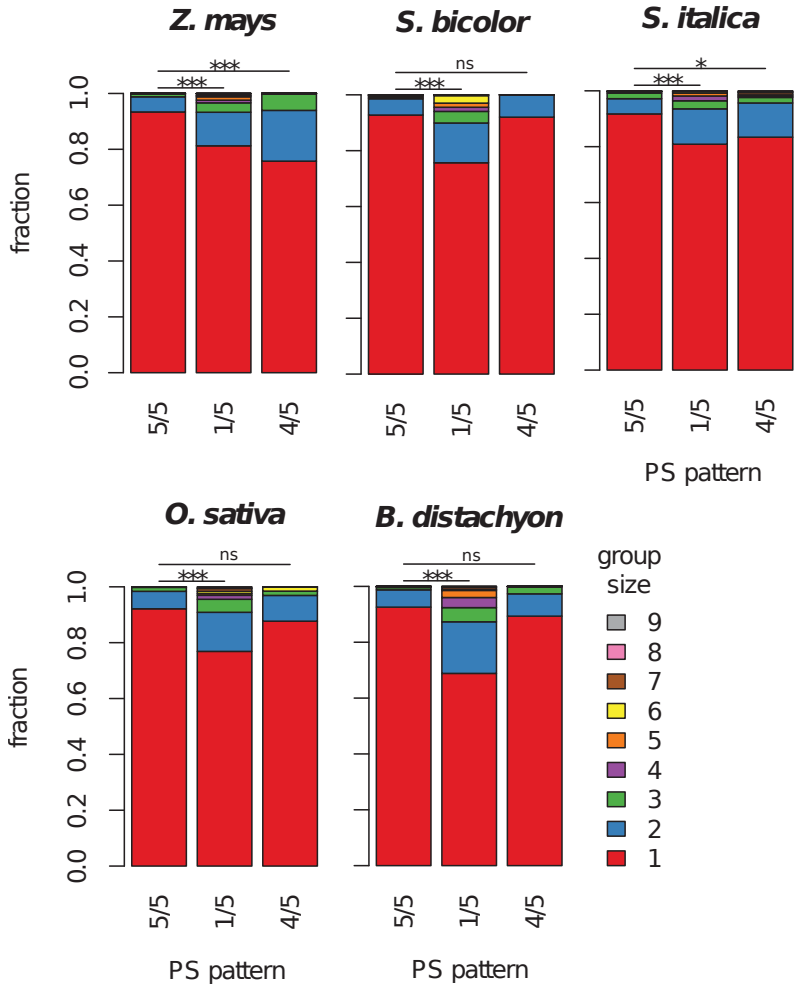
Supplementary Fig. 17: The relationship between expression pattern divergence and possible indicators of selection pressure. In (a) divergence is compared to the dN of genes with different annotated origins [28] including those where all duplicates are of the same age (the *Z. mays* tetraploidy and pan-grass WGDs). In (b) divergence at a branch on a phylogenetic tree is compared to the significance of positive selection at the same branch of the tree.



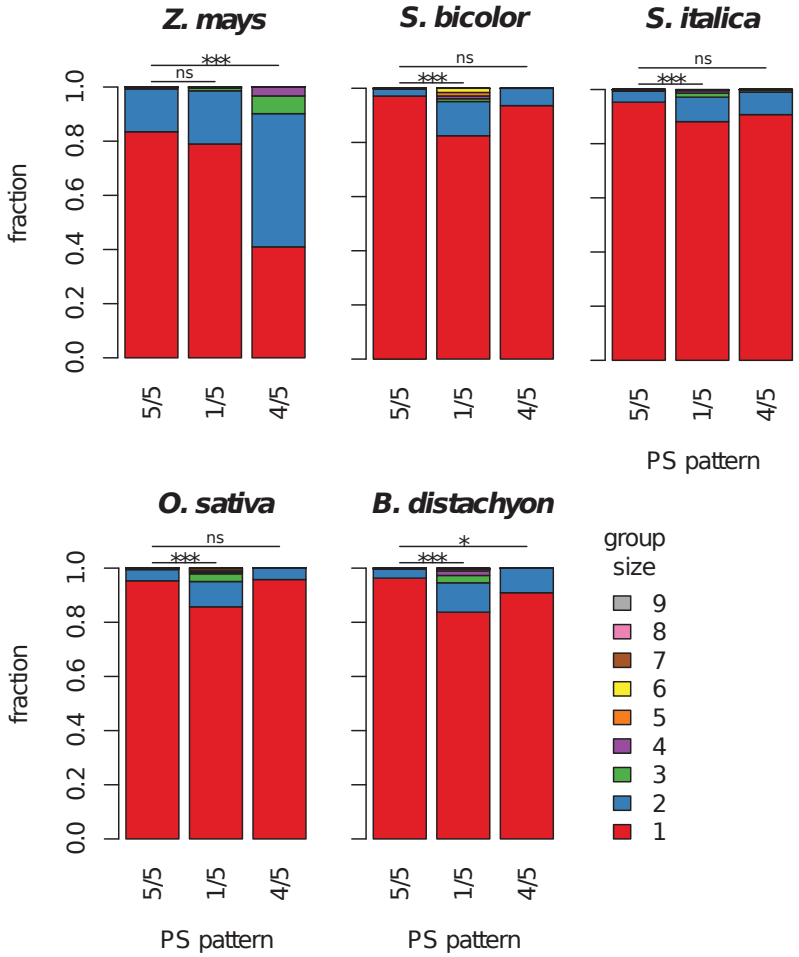
Supplementary Fig. 18: The categorization of expression patterns into photosynthesis-like or not. For the tissues in (Supplementary Fig. 7; Supplementary Table 8) the r_p between the expression pattern of each gene and the mean z-score of the genes in PS (photosynthesis) MapMan [45] category. Threshholds (red) were set to divide the resulting bi-modal distributions.



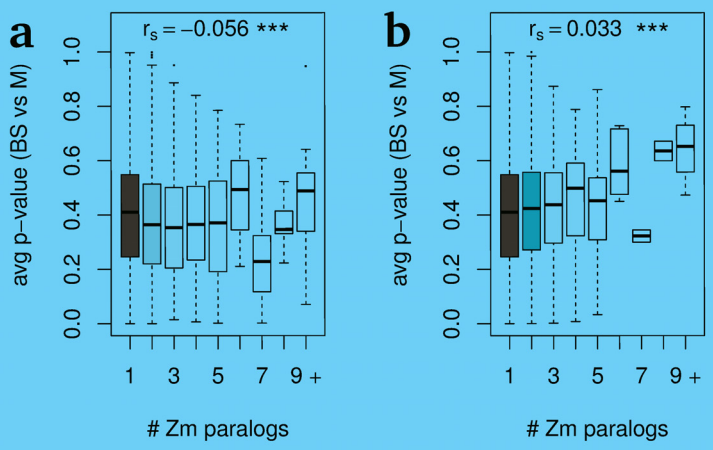
Supplementary Fig. 19: The relationship between group size (# paralogs in orthogroup in respective species) and photosynthetic pattern evolution. Cases where all 5 species show a photosynthetic-like expression pattern (see Supplementary Fig. 18) are considerd conserved, while cases where 4 of 5 or 1 of 5 species show a photosynthetic-like expression pattern are considered gain or loss, respectively. The odd species out is *S. bicolor* in (a), *S. italica* in (b) and *B. distachtyon* in (c).



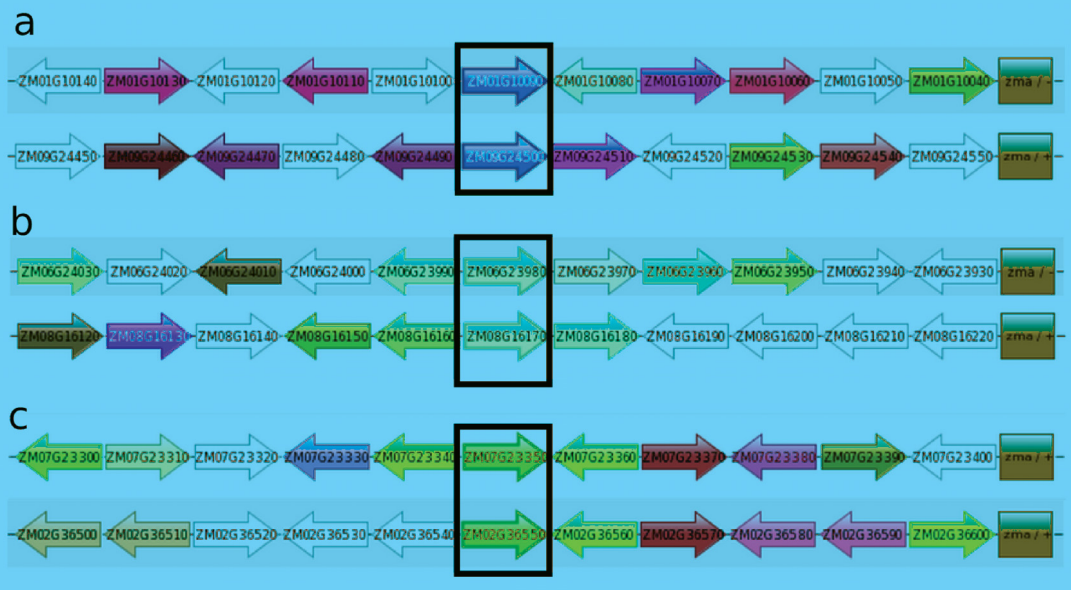
Supplementary Fig. 20: the relationship between group size (# paralogs in orthogroup in respective species) and photosynthetic pattern evolution in "ancient orthogroups" (min dS >1). cases where all 5 species show a photosynthetic-like expression pattern (see Supplementary Fig. 18) are considerd conserved, while cases where 4 of 5 or 1 of 5 species show a photosynthetic-like expression pattern are considered gain or loss, respectively. The odd species out is *Z. mays* in (a), *s. bicolor* in (b), *S. italica* in (c), *o. sativa* in (d), and *B. distachtyon* in (e).



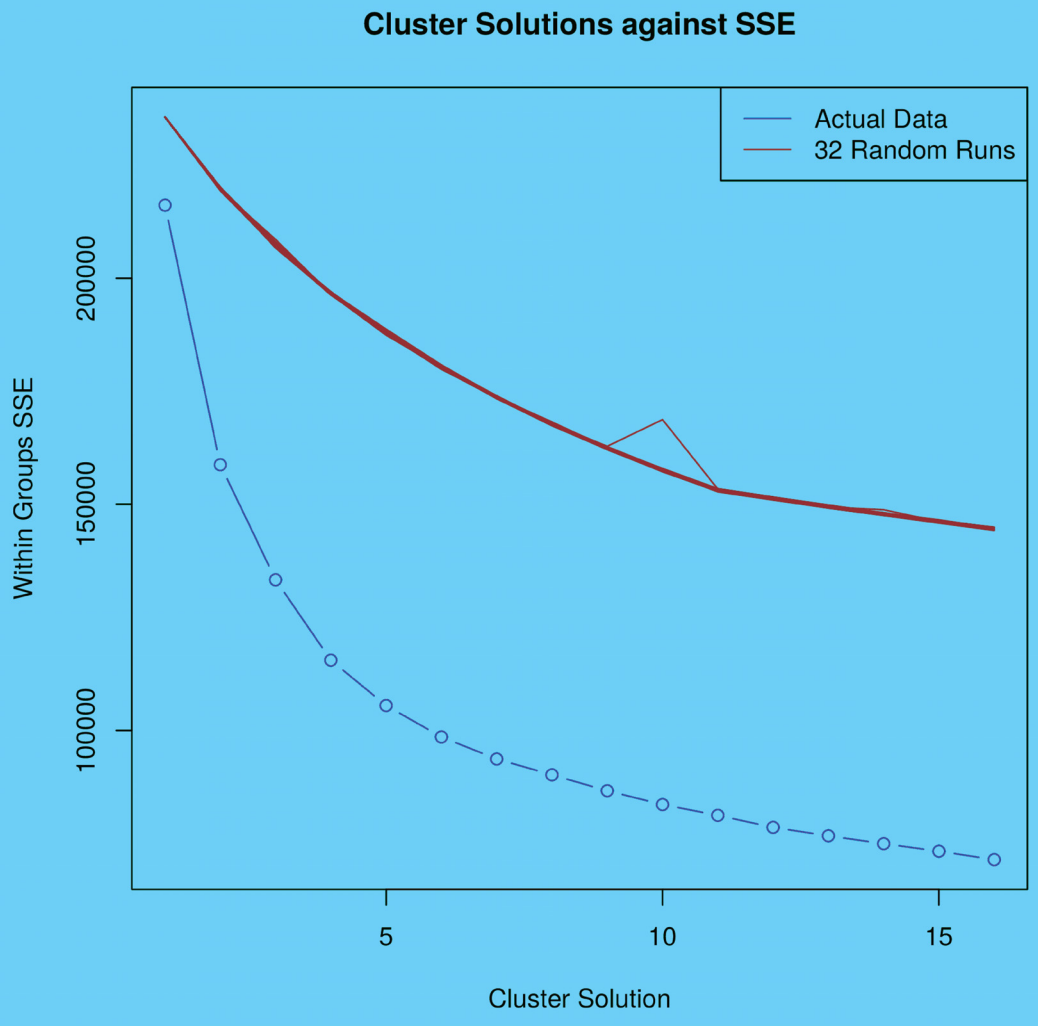
Supplementary Fig. 21: The relationship between group size (# paralogs in orthogroup in respective species) and photosynthetic pattern evolution in "young orthogroups" (min dS <0.3). Cases where all 5 species show a photosynthetic-like expression pattern (see Supplementary Fig. 18) are considerd conserved, while cases where 4 of 5 or 1 of 5 species show a photosynthetic-like expression pattern are considered gain or loss, respectively. The odd species out is *Z. mays* in (a), *S. bicolor* in (b), *S. italica* in (c), *O. sativa* in (d), and *B. distachtyon* in (e).



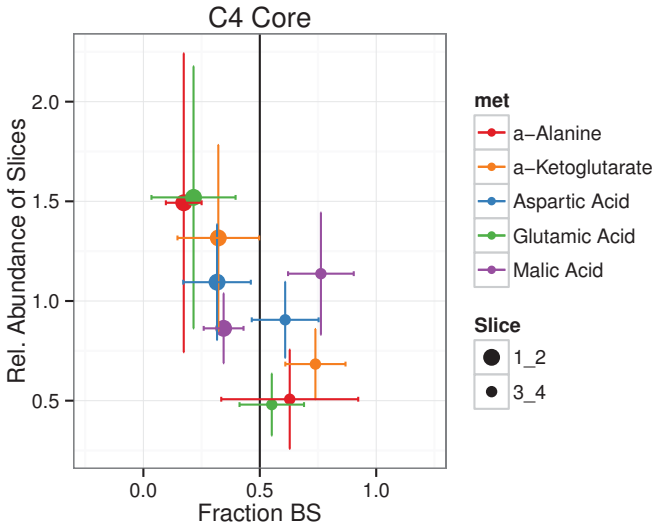
Supplementary Fig. 22: The relationship between group size (# paralogs in orthogroup in respective species) and significance of tissue specificity (average p-value) in (a) "ancient orthogroups" (min dS >1) and (b) "young orthogroups" (min dS <0.3)



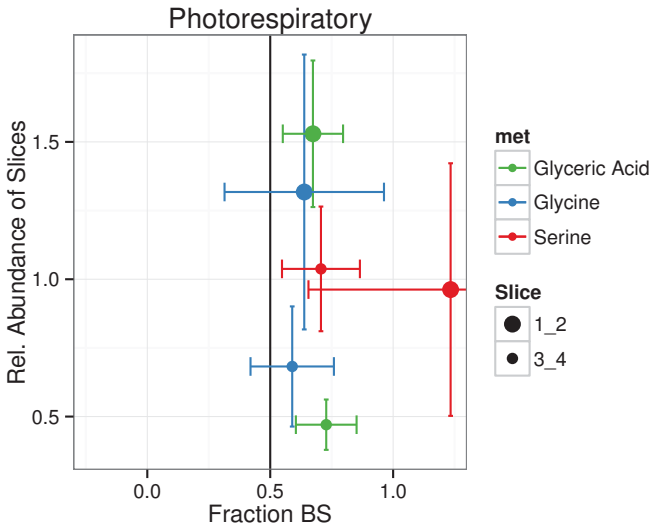
Supplementary Fig. 23: Local gene organization of syntenic, young duplicates with high divergence (in boxes). Namely, (a) PEPCK, (b) PPDK, and (c) PPDK-RP. Data and visualization from Plaza 3.0 [63]. Different colors or shades denote different homologous groups.



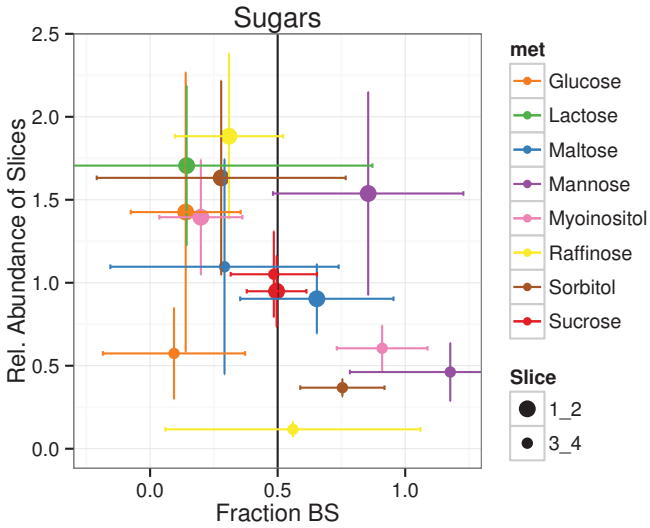
Supplementary Fig. 24: Sum of standard error compared between original and random data with a different number of cluster centers.



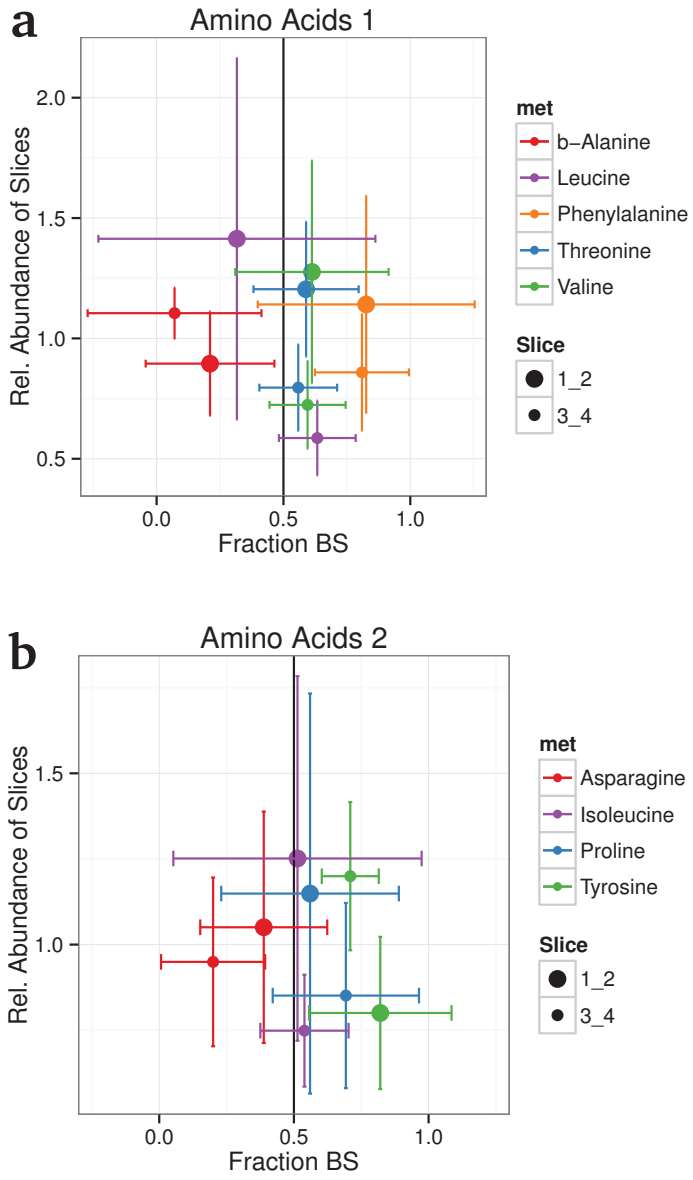
Supplementary Fig. 25: Distribution of measured core C4 metabolites between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute).



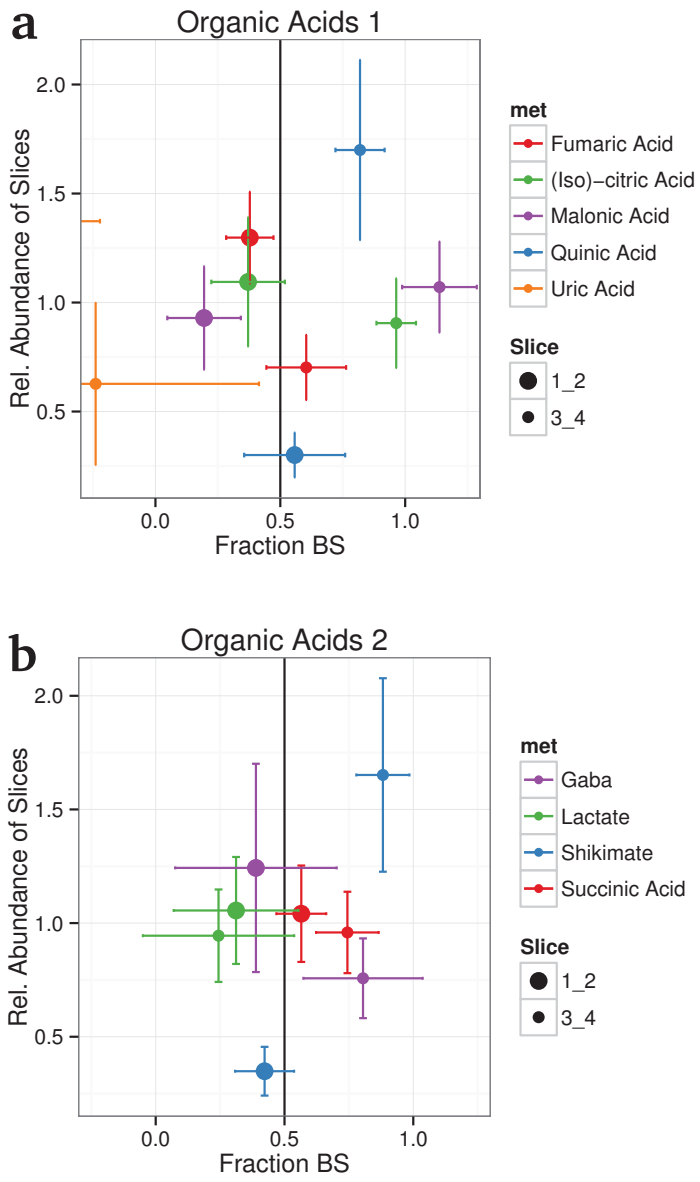
Supplementary Fig. 26: Distribution of measured photorespiratory metabolites between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute).



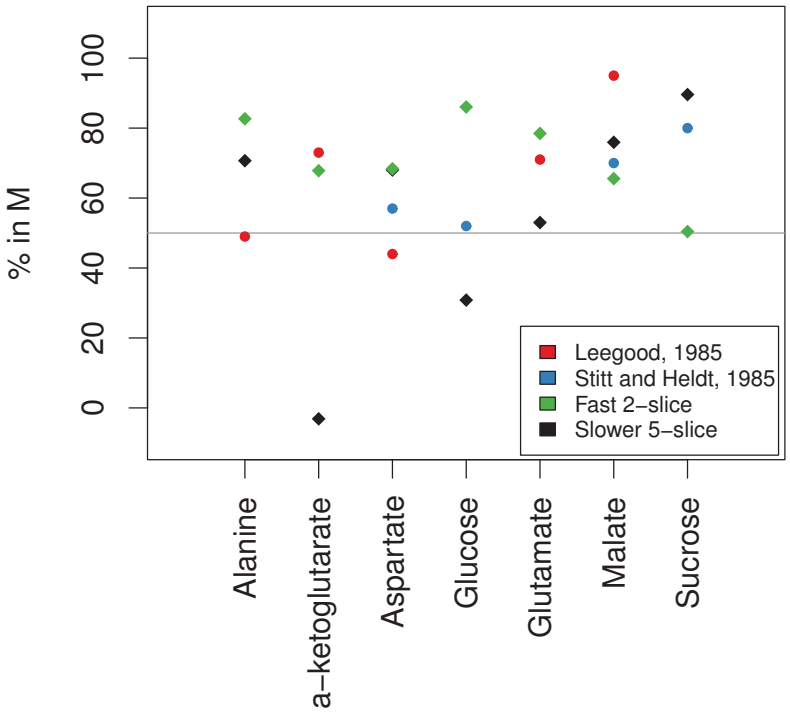
Supplementary Fig. 27: Distribution of measured sugars between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute).



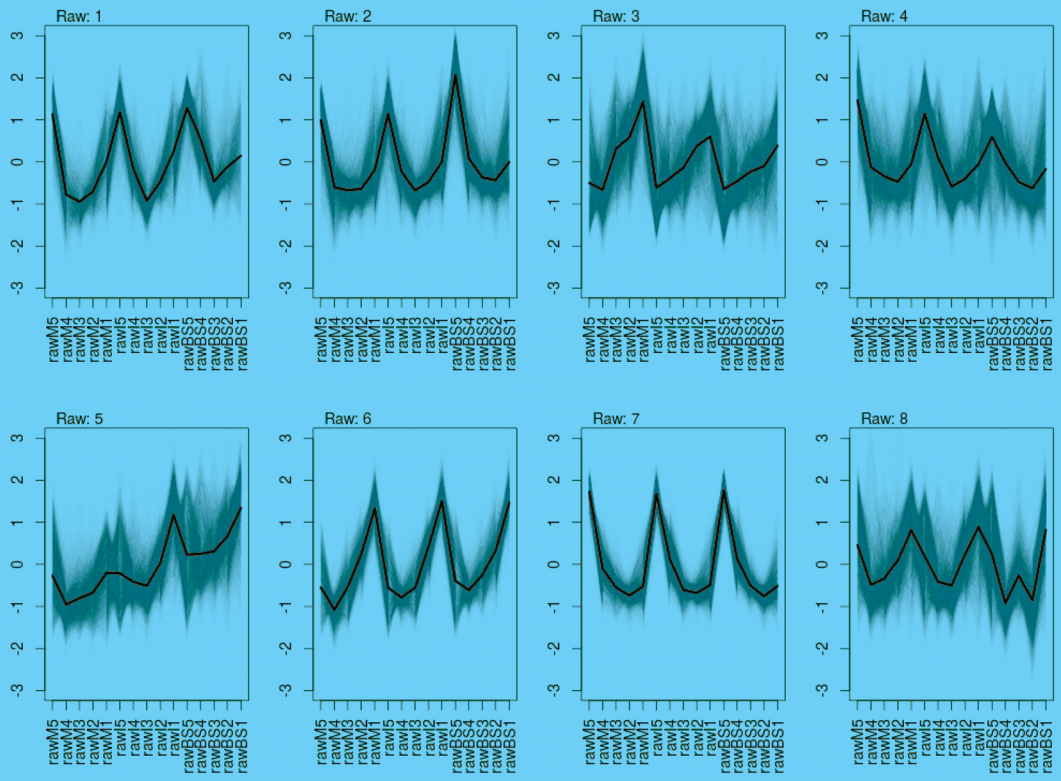
Supplementary Fig. 28: Distribution of measured non-C4 core nor photorespiratory amino acids between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute). Metabolites split arbitrarily into (a) and (b) for plotting clarity.



Supplementary Fig. 29: Distribution of measured non-C4 core nor photorespiratory organic acids between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute). Metabolites split arbitrarily into (a) and (b) for plotting clarity.



Supplementary Fig. 30: Comparison between the cell specificity of metabolites that have been measured in previous studies [42, 91]; the fast, 2-slice metabolite havest; and the slower, 5-slice gradient harvest.



Supplementary Fig. 31: K-means clustering of BS & M gradient, plotted with the z-score of the raw (not deconvoluted) expression data. Individual genes in grey and centers in black.

Supplemental Tables

Supplementary Table 1: Comparison of BS/M values between studies by linear regression. The study using enzymatic and mechanical separation [44] reported the purest tissues, while the studies using laser micro dissection [39, 43] are more able to separate BS from vascular bundle. In mature tissue Slice 1 and Section + 9 [43] were compared; while in immature tissue Slice 4, 5 and Section -1 [43] were compared.

			Min > 10 FPKM			Min > 100 FPKM		
	x.study	y.study	slope	p.value	r^2	slope	p.value	r^2
Mature	[44]	[39]	0.32	1.21×10^{-91}	0.35	0.44	1.32×10^{-27}	0.64
	[44]	[43]	0.33	3.45×10^{-161}	0.36	0.39	8.71×10^{-31}	0.51
	[44]	S1	0.40	4.91×10^{-252}	0.27	0.71	9.22×10^{-54}	0.63
	[39]	[43]	0.96	~ 0	0.94	0.99	8.72×10^{-186}	0.97
	[39]	S1	0.94	2.22×10^{-115}	0.43	1.13	2.54×10^{-25}	0.64
	[43]	S1	0.94	9.65×10^{-220}	0.42	1.16	4.79×10^{-40}	0.59
Immature	[43]	S5	0.61	7.21×10^{-102}	0.24	0.68	1.99×10^{-30}	0.42
	[43]	S4	0.70	1.39×10^{-157}	0.37	0.68	2.92×10^{-30}	0.45

Supplementary Table 2: P-values from wilcox-rank test for differences between the C_4 genes, their closest homologs, and their remaining homologs.

	C_4 vs closest	C_4 vs remaining	remaining vs closest	notes on data
r_p to photosynthesis	1.21×10^{-5}	1.84×10^{-7}	0.15	see Table S.Tissues
peak FPKM	4.77×10^{-6}	7.84×10^{-9}	0.37	no PPDK_RP, PPCK
tissue specificity	9.32×10^{-3}	3.22×10^{-4}	0.61	$C_4 > 0.7$ only

Supplementary Table 3: P-values for a t-test of the divergence between the C4 core genes and all their paralogs vs all other paralog pairs in genome.

	Pattern divergence $ln(\frac{1+r_p}{1-r_p})$	**Level divergence** $\lvert ln(\frac{peakFPKM1}{peakFPKM2})\rvert$
M & BS gradient	0.016	1.01×10^{-5}
primordial leaf/husk gradient [30]	0.638	0.085
Atlas including leaves [31]	0.048	0.130
All non-leaf/husk tissues from above	0.683	0.417
All leaf/husk tissues from above	4.72×10^{-4}	1.40×10^{-6}

Supplementary Table 4: Multiple regression of expression level divergence vs sequence and gene family features.

	features	**estimate**	**standard error**	**p-value**	**r-squared**
deconvoluted BS & M gradient	dS	0.029	0.0099	3.28×10^{-03}	0.0045
	dN	-0.017	0.0102	9.76×10^{-02}	
	Colinearity	-0.050	0.0097	2.90×10^{-07}	
	# paralogs *Z. mays*	0.007	0.0094	4.48×10^{-01}	
	dN/dS	-0.003	0.0087	7.37×10^{-01}	
	C4 or not	0.027	0.0087	1.67×10^{-03}	
non-photosynthetic atlas tissues[31]	dS	0.053	0.0054	1.02×10^{-22}	0.00461
	dN	-0.012	0.0056	3.81×10^{-02}	
	Colinearity	-0.051	0.0055	7.35×10^{-21}	
	# paralogs *Z. mays*	-0.026	0.0055	1.77×10^{-06}	
	dN/dS	0.007	0.0052	2.11×10^{-01}	
	C4 or not	0.003	0.0052	5.41×10^{-01}	
all (developing) photosynthetic tissues (deconvoluted, primordial[30] and atlas[31]	dS	0.040	0.0099	4.89×10^{-05}	0.0061
	dN	-0.020	0.0102	5.33×10^{-02}	
	Colinearity	-0.051	0.0097	1.76×10^{-07}	
	# paralogs *Z. mays*	0.020	0.0094	3.78×10^{-02}	
	dN/dS	0.001	0.0087	8.80×10^{-01}	
	C4 or not	0.025	0.0087	3.46×10^{-03}	

Supplementary Table 5: Multiple regression of expression pattern divergence vs sequence and gene family features.

	features	estimate	standard error	p-value	r-squared
deconvoluted BS & M gradient	dS	0.060	0.0098	5.77×10^{-10}	0.034
	dN	0.029	0.0100	4.30×10^{-03}	
	Colinearity	-0.096	0.0095	8.56×10^{-24}	
	# paralogs Z. mays	0.068	0.0093	2.71×10^{-13}	
	dN/dS	-0.007	0.0086	3.88×10^{-01}	
	C4 or not	0.039	0.0086	6.52×10^{-06}	
non-photosynthetic atlas tissues[31]	dS	0.029	0.0054	9.58×10^{-08}	0.025
	dN	0.027	0.0055	9.52×10^{-07}	
	Colinearity	-0.065	0.0054	2.29×10^{-33}	
	# paralogs Z. mays	0.108	0.0054	3.76×10^{-87}	
	dN/dS	-0.009	0.0052	9.69×10^{-02}	
	C4 or not	0.005	0.0051	3.51×10^{-01}	
all (developing) photosynthetic tissues (deconvoluted, primordial[30], and atlas[31])	dS	0.108	0.0095	6.97×10^{-30}	0.085
	dN	-0.029	0.0097	2.69×10^{-03}	
	Colinearity	-0.148	0.0093	1.78×10^{-56}	
	# paralogs Z. mays	0.160	0.0090	2.54×10^{-69}	
	dN/dS	-0.009	0.0084	2.68×10^{-01}	
	C4 or not	0.041	0.0084	1.18×10^{-06}	

Supplementary Table 6: The paralog pairs of the MapMan PS category, which occurred non-ambiguously in the edge of Clusters 3 “M-tip” and 5 “BS-tip”, and whether they consume ATP

Paralogs	MapMan bincode	MapMan subcategory of PS	Uses ATP	Clusters
		calvin cycle		
GRMZM2G089136, GRMZM2G382914	1.3.3	phosphoglycerate kinase	Y	3, 5
GRMZM2G026024, GRMZM2G463280	1.3.12	PRK	Y	5, 3
GRMZM2G162529, GRMZM2G463280	1.3.12	PRK	Y	5, 3
		photorespiration		
GRMZM2G018786, GRMZM2G054663	1.2.7	glycerate kinase	Y	3, 5
GRMZM2G076239, GRMZM2G129246	1.2.2	glycolate oxydase	N	3, 5
		lightreaction		
GRMZM2G010555, GRMZM2G102349	1.1.40	cyclic electron flow-chlororespiration	N	5, 3
GRMZM5G885392, GRMZM5G896082	1.1.40	cyclic electron flow-chlororespiration	N	3, 5
GRMZM2G048313, GRMZM2G122337	1.1.5.2	other electron carrier (ox/red).ferredoxin	N	5, 3
GRMZM2G329047, GRMZM2G377855	1.1.2.2	photosystem I.PSI polypeptide subunits	N	5, 3

Supplementary Table 7: Non-default parameters used for bioinformatics programs.

tophat2	For studies with reads shorter than 50 bases **--segment-length=N** (N = read length/2) was set so that reads were mapped in at least 2 segments **--b2-very-sensitive** and **--read-realign-edit-dist=0** were set to increase sensitivity **-G** <**file.gtf**> was used to guide mappings to annotated transcriptome
cufflinks2	**-u** was set to improve distribution of reads mapping to more than one position **-G** <**file.gtf**> was used to guide assembly to annotated transcriptome
cutadapt	**-e0.1** was used to set the maximum fraction of errors for a match **-O5** was used to require an adaptor match to be at least 5 bases long
fastq-quality-trimmer	**-Q33** indicates the quality encoding **-l25** was used to discard trimmed reads shorter than 25 bases **-t28** was set for the quality score threshhold
blastall	**-p blastn** was used for BLAST searches in nucleotide space between *Z. mays* genome releases, while **-p blastp** was used for BLAST searches in protein space between species **-m8** was set for a tabular output **-FF** was set to turn off quality filtering, and thereby allow avoid excluding perfect matches between different *Z. mays* genome releases **-e1e-1** was set to skip any matches of a quality where 0.1 or more would be expected by chance based on database size
McscanX	**-w1**, **-k300**, **-m50**, and **-g-0.5** were set to err on the sensitive side while detecting colinearity
Prank	**+F** was set as recommended for sequences with many insertions or deletions
Mafft	**--auto** was used
RaxML	**-m PROTGAMMAIJTT** was set to employ the JTT amino acid substitution matrix with optimized substitution rates, and a gamma model of rate heterogenity including invariant sites. **-k** was used to print branch lengths **-NautoMR** was used to stop bootstrapping after convergence **-b 123** is used to set a seed for random numbers while bootstrapping **-p 12345** was used to set a seed for random numbers in parsimony inference
codeml	**runmode = -2**, **model = 0**, and **Nssites = 0 F3x4 model** were used to estimate pairwise dN and dS **runmode = 0**, **seqtype = 1**, **CodonFreq = 2**, **model = 2**, **Nssites = 2**, **fix_kappa = 0**, and **kapa = 2** were used for both null and alternative branch site models The negative log likelihood of the null model with parameters **fix_omega = 1** and **omega = 1** in the null model was compared to the alternative model with parameters **fix_omega = 0** and **omega = 1.5** to determine significance of dN/dS signature >1 for the branch site models

Supplementary Table 8: Tissues used for differnet analyses

Z. mays	r_p PS	Leaf & husk	mature tissue specificity
6DAS_Prim_Root [31]	Y		
24H_Germ_seed [31]	Y		
16DAP_Embryo [31]	Y		
V3_Stem_SAM [31]	Y		
12DAP_W_seed [31]	Y		
10DAP_W_seed [31]	Y		
16DAP_W_seed [31]	Y		
14DAP_W_seed [31]	Y		
14DAP_Endopsperm [31]	Y		
12DAP_Endopsperm [31]	Y		
16DAP_Endosperm [31]	Y		
V9_13th_Leaf [31]		Y	
V9_11th_Leaf [31]		Y	
V9_Immature_Leaves [31]		Y	
R2_13th Leaf [31]		Y	
VT_13th Leaf [31]		Y	
V9_8th_Leaf [31]		Y	
V5_Tip_s-2_Leaf [31]	Y	Y	
All primordia samples [30]		Y	
M5		Y	
M4		Y	
M3		Y	
M2		Y	Y
M1		Y	Y
BS5		Y	
BS4		Y	
BS3		Y	
BS2		Y	Y
BS1		Y	Y
S. italica	r_p PS		
M [35]			
BS [35]			
leaf ligule 4 + 1 [33]			
leaf ligule 3 - 1 [33]			
leaf ligule 3 + 2 [33]			
leaf tip - 1 [33]			
root [34]	Y		
stem [34]	Y		
leaf [34]	Y		
spica [34]	Y		

Supplemental References

88. Amthor, J. S. From sunlight to phytomass: on the potential efficiency of converting solar radiation to phyto-energy. *New Phytologist* **188,** 939–959 (2010).

89. Furbank, R. T. Evolution of the C(4) photosynthetic mechanism: are there really three C(4) acid decarboxylation types? *Journal of Experimental Botany* **62,** 3103–8 (May 2011).

90. Hatch, M. D. The C_4 -pathway of photosynthesis. Evidence for an intermediate pool of carbon dioxide and the identity of the donor C_4-dicarboxylic acid. *The Biochemical journal* **125,** 425–32 (Nov. 1971).

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

92. Gardiner, J., Sherr, I. & Scarpella, E. Expression of DOF genes identifies early stages of vascular development in Arabidopsis leaves. *International Journal of Developmental Biology* **54,** 1389 (2010).

93. Ward, J. M., Cufr, C. A., Denzel, M. A. & Neff, M. M. The Dof transcription factor OBP3 modulates phytochrome and cryptochrome signaling in Arabidopsis. *The Plant Cell Online* **17,** 475–485 (2005).

94. Nemhauser, J. & Chory, J. Photomorphogenesis. *The Arabidopsis book/American Society of Plant Biologists* **1** (2002).

95. Kelley, D. R., Arreola, A., Gallagher, T. L. & Gasser, C. S. ETTIN (ARF3) physically interacts with KANADI proteins to form a functional complex essential for integument development and polarity determination in Arabidopsis. *Development* **139,** 1105–1109 (2012).

96. Iwasaki, M. *et al.* Dual regulation of ETTIN (ARF3) gene expression by AS1-AS2, which maintains the DNA methylation level, is involved in stabilization of leaf adaxial-abaxial partitioning in Arabidopsis. *Development* **140,** 1958–1969 (2013).

97. Soares-Cordeiro, A. S. *et al.* Variations in the dorso-ventral organization of leaf structure and Kranz anatomy coordinate the control of photosynthesis and associated signalling at the whole leaf level in monocotyledonous species. *Plant, cell & environment* **32,** 1833–1844 (2009).

98. Belin, C., Megies, C., Hauserová, E. & Lopez-Molina, L. Abscisic acid represses growth of the Arabidopsis embryonic axis after germination by enhancing auxin signaling. *The Plant Cell Online* **21,** 2253–2268 (2009).

99. Teale, W. D., Paponov, I. A. & Palme, K. Auxin in action: signalling, transport and the control of plant growth and development. *Nature Reviews Molecular Cell Biology* **7,** 847–859 (2006).

100. Park, M. Y., Kang, J.-y. & Kim, S. Y. Overexpression of AtMYB52 confers ABA hypersensitivity and drought tolerance. *Molecules and cells* **31,** 447–454 (2011).

101. Cassan-Wang, H. *et al.* Identification of novel transcription factors regulating secondary cell wall formation in Arabidopsis. *Frontiers in plant science* **4** (2013).

102. Weckopp, S. C. & Kopriva, S. Are changes in sulfate assimilation pathway needed for evolution of C_4 photosynthesis? *Frontiers in Plant Science* **5,** 773 (2015).

103. Lee, J. R., Boltz, K. A. & Lee, S. Y. Molecular chaperone function of Arabidopsis thaliana phloem protein 2-A1, encodes a protein similar to phloem lectin. *Biochemical and biophysical research communications* **443,** 18–21 (2014).

104. Zhang, C. *et al.* Harpin-induced expression and transgenic overexpression of the phloem protein gene AtPP2-A1 in Arabidopsis repress phloem feeding of the green peach aphid Myzus persicae. *BMC plant biology* **11,** 11 (2011).

105. Love, J. *et al.* Ethylene is an endogenous stimulator of cell division in the cambial meristem of Populus. *Proceedings of the National Academy of Sciences* **106,** 5984–5989 (2009).

106. Pesquet, E. & Tuominen, H. Ethylene stimulates tracheary element differentiation in Zinnia elegans cell cultures. *New Phytologist* **190,** 138–149 (2011).

107. Schuetz, M. *et al.* Laccases direct lignification in the discrete secondary cell wall domains of protoxylem. *Plant physiology* **166,** 798–807 (2014).

108. Groover, A. & Jones, A. M. Tracheary element differentiation uses a novel mechanism coordinating programmed cell death and secondary cell wall synthesis. *Plant Physiology* **119,** 375–384 (1999).