The Transcriptomic Blueprint of C_4 Photosynthesis

Kumulative Dissertation

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

vorgelegt von

Canan Külahoglu

aus Köln

Düsseldorf, November 2014

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. Rüdiger Simon

Tag der mündlichen Prüfung: 15.01.2015

Erklärung

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.

Köln, den 18. November 2014

Canan Külahoglu

„Au milieu de l'hiver
j'ai découvert en moi
un invincible été "

<Albert Camus>

Für meine Familie

TABLE OF CONTENTS

GLOSSARY

2-PG	2-phosphoglycolate
3-PGA	3-phosphoglycerate
ADP	adenine diphosphate
AlaAT	alanine aminotransferase
AspAT	aspartate aminotransferase
ATP	adenine triphosphate
BS	bundle sheath
C_2, C_3, C_4	two, three, four-carbon molecule
CA	carbonic anhydrase
CO_2	carbon dioxide
DNA	deoxyribonucleic acid
e.g.	*exempli gratia*
E. retroflexa	*Eleocharis retroflexa*
g2	golden-2
glk1	golden-2-like1
GDC	glycine decarboxylase complex
G. gynandra	*Gynandropsis gynandra*
i.e.	*id est*
M	mesophyll
MDH	malate dehydrogenase
NAD	nicotinamide adenine dinucleotide
NADH	reduced nicotinamide adenine dinucleotide
NAD-ME	NAD-dependent malic enzyme
NADP	nicotinamide adenine dinucleotidephosphate
NADPH	reduced nicotinamide adenine dinucleotidephosphate
NADP-ME	NADP-dependent malic enzyme
NGS	next-generation sequencing
O_2	oxygen
OAA	oxaloacetate
PACMAD	*Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae, Danthonioideae*
PEP	phospho*enol*pyruvate
PEPC	phospho*enol*pyruvate carboxylase
PEPCK	phospho*enol*pyruvate carboxykinase
PPDK	pyruvate orthophosphate dikinase
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
RBP	ribulose-1,5-bisphosphate
RuBisCO	ribulose 1,5-bisphophate carboxylase/oxygenase
scr	scarecrow
SHM	serine hydroxymethyltransferase
shr	short root
SNP	single nucleotide polymorphism
T. hassleriana	*Tarenaya hassleriana*
Z. mays	*Zea mays*

I. PREFACE

This thesis is divided into two independent chapters. In the first chapter (**Section III-VI**) three first author manuscripts are presented. *Manuscript 1* features a method review explaining and presenting RNA-sequencing technology, experimental design, laboratory practice and data analysis (**V.1 FIRST-AUTHORED MANUSCRIPTS**; Külahoglu and Bräutigam, 2014; Methods Mol. Biol., Vol. 1158, pp. 71-91). The second research article (*Manuscript 2*; **V.2 FIRST-AUTHORED MANUSCRIPTS**) features two comparative transcriptomic atlases between two closely related C_4 and C_3 photosynthesis performing species of the Cleomaceae (Külahoglu et al., 2014; Plant Cell, Vol. 26, No.8, pp. 3243-60). The third research article is added as addendum to **Chapter 1** (**VI. ADDENDUM** *Manuscript 3 Prepared for Publication*) and presents the transcriptional plasticity of the amphibious sedge *Eleocharis retroflexa* with regard to its structure and photosynthetic mode in dynamic environments. A conclusion and outlook of the transcriptional dynamics of C_4 photosynthesis regarding the presented results concludes **Chapter 1** (**VII. CONCLUSION AND OUTLOOK**).

In **Chapter 2** (**VIII. CO-AUTHORED MANUSCRIPTS**) co-authored publications are presented. These publications share C_4 photosynthesis, next-generation sequencing, or both as a common topic. In *Manuscript 4*, the genome of *Tarenaya hassleriana* is introduced (Chen et al., 2013; Plant Cell; Vol. 25, No. 8, pp. 2813-30). The subsequently presented *Manuscript 5* is shedding light on the context of genome duplication and C_4 gene evolution in the Cleomaceae (van den Bergh et al., 2014; Curr. Plant Biol., http://dx.doi.org/10.1016/j.cpb.2014.08.001). *Manuscript 6* features an integrative analysis of all C_4 photosynthesis subtypes and the transcriptomes of the C_4 species *Megathyrsus maximum* and the C_3 species *Dichantelium clandestinum* (Bräutigam et al, 2014; J. Exp. Bot., Vol. 65, No. 13, pp. 3579–3593). *Manuscript 7* is a methodological article, that contains the transcriptome of *Azolla filiculoides* and cultivation methods to advance its domestication for fern biomass based economy (Brouwer et al., 2014; New Phytologist Vol. 202, pp. 1069-1082). This thesis is concluded with a method-based review (*Manuscript 8*) explaining the benefits of co-expression analysis for shedding light on photorespiration (Bordych et al., 2013; Plant Biology, Vol. 15, No. 4, pp. 686-93. doi: 10.1111/plb.12027).

II.1 SUMMARY

By using light as energy source, plants are able to convert inorganic carbon (carbon dioxide) into organic biomass through photosynthesis. The action of photosynthesis increases carbon sequestration and storage of ecosystems. The main enzyme catalyzing the carbon fixation is ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO).

Under permissive conditions in hot, high light and arid environments, C_4 photosynthesis is highly advantageous. Thus, C_4 photosynthesis has evolved more than 65-times convergently in the Angiosperms from ancestral C_3 species (Sage et al., 2011). In comparison to C_3 species, C_4 performing plants have a higher photosynthetic capacity and need less water and nitrogen (Black, 1973; Ruan et al., 2012). C_4 photosynthesis can be described as a biochemical pump on top of conventional C_3 photosynthesis that supercharges photosynthetic carbon assimilation via the Calvin-Benson-Bassham cycle, by concentrating CO_2 at the site of the RuBisCO (Andrews and Lorimer, 1987; Furbank and Hatch, 1987). Even though there are relatively few C_4 species, compared to C_3 species, plants harboring the C_4 photosynthetic trait account for a quarter of all primary fixed carbon on our planet (Still et al., 2003; Edwards et al., 2010; Sage and Zhu, 2011). Thus, transferring the C_4 trait into C_3 crop species is a major effort in plant breeding and is hoped for to trigger a second “green revolution” (von Caemmerer et al., 2012).

One approach to investigate the C_4 trait is to generate transcriptomic and genomic blueprints of different C_4 species and closely related C_3 species. The recent development of next-generation sequencing technologies opened up the path for large-scale transcriptome analyses and comparisons.

This PhD thesis aims at expanding our knowledge about the transcriptional fingerprint of C_4 photosynthesis, identifying C_4 specific gene regulation, and understanding how the C_4 trait is established with special focus on dicotyledonous leaf ontology. The studies presented in this PhD thesis used a systems biology approach for detailed transcriptome comparisons in the Cleomaceae and Cyperaceae families.

The methods applied for generating the presented transcriptome data sets are summarized as workflow in *Manuscript 1*, a published method review on RNA-sequencing (Külahoglu and Bräutigam, 2014).

Two comparative transcriptome atlases featuring the closely related *Gynandropsis gynandra* (C_4) and *Tarenaya hassleriana* (C_3) shed light on dicotyledonous C_4 leaf ontology. We connected two key features of Kranz anatomy to transcriptional changes by integration of expression and anatomical data during leaf ontology (Külahoglu et al., 2014). Larger bundle sheath cell size, which is a key feature of C_4 anatomy, could be connected with higher ploidy levels in these cells. Furthermore, we could link the C_4 specific higher vein density to a delay in photosynthetic gene expression. This delay of mesophyll differentiation is possibly

contributing to the denser venation through prolonged initiation of vein orders in *G. gynandra* as seen in *A. thaliana* (Scarpella et al., 2004). By comparing the C_4 cycle gene expression patterns between the two atlases it became apparent that the C_4 genes had different ancestral expression domains in the C_3 species leading to the assumption that a master regulator of C_4 is unlikely.

The complex adaptive trait of C_4 photosynthesis allows the adaption to hot and arid environments. Since C_4 plants are highly specialized by their complex leaf anatomy and biochemistry, which are needed to run the C_4 cycle, the evolution of C_4 photosynthesis has been hypothesized to come at the cost of reduced phenotypic plasticity (Sage and McKown, 2006). One exception is the terrestrial wetland species *Eleocharis retroflexa*, which displays a remarkable phenotypic plasticity in its mode of photosynthetic carbon assimilation (*Manuscript 3*). By comparing the *E. retroflexa* transcriptomes of culms grown under water and on soil, we found, that *E. retroflexa* is surprisingly flexible in the usage of the C_4 cycle on transcriptional level. Submerged *E. retroflexa* loses classical Kranz anatomy and exhibits features known for aquatic plants, while up-regulating transcripts related to photorespiration and maintaining the C_4 cycle as seen in C_3-C_4 intermediate plant species (Sage et al., 2012). Thus, *E. retroflexa* is an example of a highly plastic and adaptive C_4 species, which contrasts the broad observation that C_4 species display less plasticity in phenotype and metabolism, likely due to their high level of specialization.

In summary, the presented comparative transcriptome studies showed new connections between anatomical specialization and transcriptional dynamics of the highly complex and adaptive C_4 photosynthesis trait.

References

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.

Black, C.C. (1973). Photosynthetic carbon fixation in relation to net CO_2 uptake. Annual Review of Plant Physiology and Plant Molecular Biology **24,** 253-286.

Edwards, E.J., Osborne, C.P., Stroemberg, C.A.E., Smith, S.A., Bond, W.J., Christin, P.-A., Cousins, A.B., Duvall, M.R., Fox, D.L., Freckleton, R.P., Ghannoum, O., Hartwell, J., Huang, Y., Janis, C.M., Keeley, J.E., Kellogg, E.A., Knapp, A.K., Leakey, A.D.B., Nelson, D.M., Saarela, J.M., Sage, R.F., Sala, O.E., Salamin, N., Still, C.J., Tipple, B., and Consortium, C.G. (2010). The Origins of C_4 Grasslands: Integrating Evolutionary and Ecosystem Science. Science **328,** 587-591.

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 Physiology **85,** 958-964.

Külahoglu, C., and Bräutigam, A. (2014). Quantitative transcriptome analysis using RNA-seq. Methods Mol Biol **1158,** 71-91.

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. (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.

Ruan, C.-J., Shao, H.-B., and da Silva, J.A.T. (2012). A critical review on the improvement of photosynthetic carbon assimilation in C_3 plants using genetic engineering. Critical Reviews in Biotechnology **32,** 1-21.

Sage, R.F., and 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., and Zhu, X.-G. (2011). Exploiting the engine of C_4 photosynthesis. Journal of Experimental Botany **62,** 2989-3000.

Sage, R.F., Christin, P.A., and Edwards, E.J. (2011). The C_4 plant lineages of planet Earth. Journal of Experimental Botany **62,** 3155-3169.

Sage, R.F., Sage, T.L., and Kocacinar, F. (2012). Photorespiration and the Evolution of C_4 Photosynthesis. In Annual Review of Plant Biology, Vol 63, S.S. Merchant, ed, pp. 19-47.

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.

Still, C.J., Berry, J.A., Collatz, G.J., and DeFries, R.S. (2003). Global distribution of C_3 and C_4 vegetation: Carbon cycle implications. Global Biogeochemical Cycles **17**.

von Caemmerer, S., Quick, W.P., and Furbank, R.T. (2012). The Development of C_4 Rice: Current Progress and Future Challenges. Science **336,** 1671-1672.

II.2 ZUSAMMENFASSUNG

Pflanzen sind in der Lage, die Energie des Sonnenlichtes mittels Photosynthese zu nutzen und Kohlenstoffdioxid (CO_2) in Biomasse zu verwandeln. Hauptsächlich wird die Kohlenstofffixierung von der Ribulose 1,5-Bisphosphate Carboxylase/Oxygenase (RuBisCO) katalysiert. In warmen, sonnigen, wie auch trockenen Klimaregionen ist die C_4 Photosynthese, eine Abwandlung der C_3 Photosynthese, sehr vorteilhaft. Die Vorteile der C_4 Photosynthese führten zu einer wiederholt auftretenden konvergenten Evolution von C_4 Spezies unter geeigneten Umweltbedingungen (Sage et al., 2011). Vergleicht man C_3 und C_4 Spezies miteinander, haben C_4 Pflanzen unter geeigneten Bedingungen höhere Photosynthese-Kapazitäten bei geringerem Wasser- und Stickstoffverbrauch (Black, 1973; Ruan et al., 2012). Der Mechanismus der C_4 Photosynthese lässt sich als ein biochemischer CO_2 Konzentrationsmechanismus in RuBisCO Nähe beschreiben, wodurch die Effizienz der Kohlenstoffassimilation über den Calvin-Benson-Bassham-Zyklus stark gesteigert wird (Andrews und Lorimer, 1987; Furbank und Hatch, 1987). Obwohl weniger C_4 Spezies im Vergleich zu C_3 Spezies existieren, machen C_4 Spezies ein Viertel des gesamten primär fixierten Kohlenstoffs auf unserem Planeten aus (Still et al., 2003; Edwards et al., 2010; Sage und Zhu, 2011). Aus diesem Grund ist die Erforschung der C_4 Photosynthese und ihre Übertragung auf C_3 Nutzpflanzen von höchstem Interesse und Hoffnungsträger für eine zweite „Grüne Revolution" (von Caemmerer et al., 2012).

Zum Verständnis der C_4 Photosynthese ist das Wissen über den transkriptionellen und genetischen Bauplan derselben unerlässlich. Die aktuelle Entwicklung der „Next-Generation Sequencing"-Technologien eröffnet neue Möglichkeiten, im großen Maßstab Transkriptom-Studien durchzuführen.

Ziel der hier vorliegenden Doktorarbeit war es, den transkriptionellen Fingerabdruck der C_4 Photosynthese zu erforschen, um die Etablierung der C_4 Photosynthese und ihre notwendige Gen-Regulation während der Blattentwicklung zu verstehen. Hierfür wurde ein systembiologischer Ansatz gewählt, bei dem die Transkriptome von Vertretern der Cleomaceae und Cyperaceae verglichen wurden.

Die hierbei verwendeten Methoden sind in *Manuskript 1*, einem Methoden-Review über RNA-Sequenzierung vorgestellt und zusammengefasst (Külahoglu und Bräutigam, 2014).

Durch den Vergleich von zwei Transkriptom-Atlanten der nah verwandten *Gynandra gynandropsis* (C_4) und *Tarenaya hassleriana* (C_3) Spezies, konnten zwei Hauptkomponenten der C_4 spezifischen Kranz-Anatomie entschlüsselt werden (*Manuskript 2*; Külahoglu et al., 2014). Die vergrößerten Bündelscheidenzellen der C_4 Anatomie konnten mit höheren Ploidie-Graden durch verlängerte Endoreplikation in Verbindung gebracht werden. Die zweite Hauptkomponente der C_4 Anatomie, die höhere Venendichte, steht im Zusammenhang mit

einer Verzögerung der Mesophyllentwicklung durch die spätere Genexpression von Photosynthese, Chloroplasten-Entwicklung, generellen Blattentwicklungs- und Zellzyklus-relevanten Genen. Frühere Studien in *A. thaliana* (Scarpella et al., 2004), verknüpften die Verzögerung der Mesophyllentwicklung mit höherer Venendichte. Ähnliches konnten wir bei *G. gynandra* beobachten. Vergleiche zwischen den Genexpressionsmustern der C_4 Zyklus Gene in den beiden Transkriptom-Atlanten zeigten, dass die C_4 Zyklus Gene von verschiedenen ursprünglichen Expressionsdomänen in der C_3 Spezies für den C_4 Zyklus rekrutiert wurden. Dies macht die Anwesenheit eines generellen Hauptregulators der C_4 Photosynthese eher unwahrscheinlich.

Die C_4 Photosynthese als komplexes adaptives Merkmal erlaubt die Anpassung an warme trockene Klimazonen. Da C_4 Pflanzen evolutionär einen hohen Grad an Spezialisierung durchlaufen haben, wird angenommen, dass die phänotypische Plastizität aufgrund ihrer notwendigen komplexen Blattanatomie und Biochemie eingeschränkt ist (McKown und Sage, 2006).

Eine Ausnahme bildet die C_4 Photosynthese betreibende Feuchtgebiet-Spezies *Eleocharis retroflexa*, die eine außergewöhnliche phänotypische Plastizität bezüglich ihrer Anatomie und ihres Kohlenstoffwechsels besitzt (*Manuskript 3*). Vergleiche der Halm Transkriptome von *E. retroflexa*, die entweder unter Wasser oder an Land kultiviert wurden, ließen erkennen, dass diese Spezies überraschend flexibel in ihrer Nutzung des C_4 Zyklus auf transkriptioneller Ebene ist. Darüberhinaus reagierten *E. retroflexa* Halme auf Transkriptom-Ebene rasch auf kleinste Umweltreize wie zum Beispiel Wasserentzug und Trockenheit. Unter Wasser hingegen verliert *E. retroflexa* seine Kranzanatomie und verhält sich wie eine Wasserpflanze. Hierbei wird die Photorespiration hochreguliert, um den C_4 Zyklus wie bei C_3-C_4 Intermediaten zu unterstützen (Sage et al., 2012). Folglich ist *E. retroflexa* ein interessanter Vertreter einer hochplastischen anpassungsfähigen C_4 Spezies, was im Gegensatz zur allgemeinen Beobachtung steht, dass C_4 Spezies bezüglich Stoffwechsel und Anatomie weniger flexibel sind.

Die hier aufgeführten vergleichenden Transkriptom-Studien zeigen neue Verbindungen zwischen Anatomie und Transkriptom des hochkomplexen adaptiven Merkmals der C_4 Photosynthese.

Literatur

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.

Black, C.C. (1973). Photosynthetic carbon fixation in relation to net CO_2 uptake. Annual Review of Plant Physiology and Plant Molecular Biology **24,** 253-286.

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 **160,** 165-177.

Edwards, E.J., Osborne, C.P., Stroemberg, C.A.E., Smith, S.A., Bond, W.J., Christin, P.-A., Cousins, A.B., Duvall, M.R., Fox, D.L., Freckleton, R.P., Ghannoum, O., Hartwell, J., Huang, Y., Janis, C.M., Keeley, J.E., Kellogg, E.A., Knapp, A.K., Leakey, A.D.B., Nelson, D.M., Saarela, J.M., Sage, R.F., Sala, O.E., Salamin, N., Still, C.J., Tipple, B., and Consortium, C.G. (2010). The Origins of C_4 Grasslands: Integrating Evolutionary and Ecosystem Science. Science **328,** 587-591.

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 Physiology **85,** 958-964.

Külahoglu, C., and Bräutigam, A. (2014). Quantitative transcriptome analysis using RNA-seq. Methods Mol Biol **1158,** 71-91.

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. (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.

Ruan, C.-J., Shao, H.-B., and da Silva, J.A.T. (2012). A critical review on the improvement of photosynthetic carbon assimilation in C_3 plants using genetic engineering. Critical Reviews in Biotechnology **32,** 1-21.

Sage, R.F., and 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., and Zhu, X.-G. (2011). Exploiting the engine of C_4 photosynthesis. Journal of Experimental Botany **62,** 2989-3000.

Sage, R.F., Christin, P.A., and Edwards, E.J. (2011). The C_4 plant lineages of planet Earth. Journal of Experimental Botany **62,** 3155-3169.

Sage, R.F., Sage, T.L., and Kocacinar, F. (2012). Photorespiration and the Evolution of C_4 Photosynthesis. In Annual Review of Plant Biology, Vol 63, S.S. Merchant, ed, pp. 19-47.

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.

Still, C.J., Berry, J.A., Collatz, G.J., and DeFries, R.S. (2003). Global distribution of C_3 and C_4 vegetation: Carbon cycle implications. Global Biogeochemical Cycles **17**.

von Caemmerer, S., Quick, W.P., and Furbank, R.T. (2012). The Development of C_4 Rice: Current Progress and Future Challenges. Science **336,** 1671-1672.

III. INTRODUCTION

Next-generation sequencing – a powerful tool for transcriptome analysis of non-model species without a sequenced genome

Next-generation sequencing (NGS) technologies and co-development of computational power, software and algorithms enabled biologists to gather sequence data within hours or days, which in the past took several years. This opened opportunities for analyzing species with or without sequenced genomes on a large scale and in a time and cost effective manner (Bräutigam and Gowik, 2010; Strickler et al., 2012). These new technologies can be employed to answer biological questions in genomics (Cheng et al., 2013), transcriptomics (Schmid et al., 2005; Benedito et al., 2008; Li et al., 2010; Pick et al., 2011; Sekhon et al., 2011), interactomics (Jothi et al., 2008), and methylome sequencing (Cokus et al., 2008). Especially RNA-sequencing (RNA-seq) is extensively used to shed light on various areas in molecular plant research, such as gene function characterization (Novaes et al., 2008; Alagna et al., 2009; Barakat et al., 2009; Dassanayake et al., 2009; Wang et al., 2009; Bräutigam et al., 2010; Sekhon et al., 2011), novel transcript identification (Denoeud et al., 2008; Hansey et al., 2012), genome-wide single-nucleotide polymorphism (SNP) analyses (Hamilton et al., 2011; Childs et al., 2014) and alternative transcript splicing discovery (Hansey et al., 2012). The great depth of sequencing provided by Illumina or Solid platforms with more than 200 Million reads per run allow even the detection of rare transcripts. RNA-seq has been shown to be especially useful for generating a molecular basis for non-model species. There, only the expressed coding sequences of the genome are qualitatively and quantitatively sequenced and used for unigene database assembly and subsequent transcript quantification. This approach has been used in many recent studies of non-model plant systems, to name a few (Novaes et al., 2008; Barakat et al., 2009; Dassanayake et al., 2009; Wang et al., 2009; Bräutigam et al., 2010; Angeloni et al., 2011; Gowik et al., 2011; Gongora-Castillo et al., 2012).

In the studies presented in this PhD thesis, next-generation sequencing technologies were employed to analyze non-model plant species with regard to the adaptive complex trait termed C_4 photosynthesis.

Energy conversion in C_4 photosynthesis plants

Plants use CO_2 as substrate to generate organic compounds through biochemical pathways using sun light as primary energy source. The key photosynthetic enzyme for carbon fixation is the RuBisCO (ribulose-1,5-bisphosphate carboxylase-oxygenase), which catalyzes the productive carboxylation or futile oxygenation of ribulose-1,5-bisphosphate (RBP), using carbon dioxide (CO_2) or oxygen (O_2) as substrates, respectively (Bowes et al., 1971). The

oxygenation of RBP through the RuBisCO leads to the generation of toxic 2-phosphoglycolic acid (2-PGA; Anderson, 1971) that is removed by an energy-consuming process called photorespiration. The terms C_4 and C_3 photosynthesis derive from the number of carbon atoms of the product resulting from the first step of carbon fixation. In at least 65 lineages of 19 different angiosperm families, plants have evolved C_4 carbon concentrating mechanisms from C_3 ancestors spanning mono- and dicotyledonous plants (Sage et al., 2011). This mechanism allows C_4 plants to concentrate CO_2 at the site of RuBisCO (Hatch, 1987) and mainly circumvent the photorespiratory pathway. This complex trait enables C_4 plants to gain biomass more rapidly, live in dry, high light environments, and under low CO_2 conditions (Hatch, 1987; Osborne and Freckleton, 2009). Compared to C_3 species, C_4 plants have up to 80% reduced photorespiratory rates (Kanai and Edwards, 1999; Sage et al., 2012). By elevating the CO_2 concentration in bundle sheath (BS) cells, the RuBisCO has 2-to 5-fold higher catalytic activities (Seemann et al., 1984; Sage, 2002; Ghannoum and Way, 2011; Sage and Zhu, 2011). Additionally, the elevated photosynthetic capacity at warmer temperatures, allows C_4 plants to realize 1.3-4 times higher water use efficiencies and nitrogen usage efficiencies than C_3 plants (Long, 1999; Sage and Pearcy, 2000; Kocacinar et al., 2008; Ghannoum and Way, 2011). In part, this is due to C_4 plants requiring 50-80% less RuBisCO protein for a given photosynthetic rate (Sage et al., 1987; Sage and Pearcy, 2000). Comparing C_4 and C_3 species during daytime at temperatures between 25-30°C, C_4 plants have more than 50% higher radiation usage efficiencies (Long, 1999).

In terms of energy consumption, C_4 photosynthesis is most feasible under high light and warm temperature conditions. For each fixed CO_2 molecule, C_4 plants require up to 5 ATP and 2 NADPH, meanwhile C_3 photosynthesis requires 3 ATP and 2 NADPH for each fixed CO_2 molecule (Kanai and Edwards, 1999). In cooler climates, most C_4 plants are less productive compared to acclimatized C_3 species (Sage and Zhu, 2011).

C_4 photosynthesis – the basic biochemical pathway

The biochemical reactions of the C_4 cycle are a complex combination of both, biochemical and anatomical specialization; elevating CO_2 levels at the site of the RuBisCO. All C_4 lineages generate high CO_2 concentrations in the proximity of RuBisCO. Usually, the C_4 pathway is divided into two cell types on cellular level, the mesophyll (M) and the bundle sheath (BS) cells (Hatch, 1987; Sage, 2004). However, C_4 photosynthesis can also function within a single cell (Reiskind et al., 1989; Keeley, 1998; Voznesenskaya et al., 2003). In M cells, CO_2 isconverted to bicarbonate (HCO_3^-) via the carbonic anhydrase (CA) and initially fixed by the phospho*enol*pyruvate carboxylase (PEPC) using phospho*enol*pyruvate (PEP) as CO_2 acceptor. The first CO_2 fixation step leads to the generation of oxaloacetate (OAA), a four carbon containing organic acid. Depending on the C_4 photosynthesis subtype, OAA is

directly converted into a more stable C_4 acid, malate or aspartate. These C_4 acids then diffuse into the BS cells, where they are decarboxylated by one of three decarboxylases, *i.e.* the NAD-dependent malic enzyme (NAD-ME), the NADP-dependent malic enzyme (NADP-ME) or the PEP carboxykinase (PEPCK; Hatch, 1987). These decarboxylating enzymes are used to define the different subtypes of C_4 photosynthesis. However, a recent study suggests that, a sole PEPCK C_4 subtype is not beneficial energy-wise and only the NADP-ME and NAD-ME C_4 subtypes should be considered as distinct subtypes (Wang et al., 2013), because both run an inherent supplementary PEPCK cycle (Muhaidat et al., 2007; Furbank, 2011; Pick et al., 2011; Sommer et al., 2012; Muhaidat and McKown, 2013).

In the NAD-ME subtype, OAA will be transaminated to aspartate, which is transported to the mitochondria of BS cells and converted back to OAA via the aspartate aminotransferase (AspAT; Figure 1). There, NAD-dependent malate dehydrogenase (MDH) generates malate, which will be decarboxylated by the NAD-ME releasing CO_2 and pyruvate in the BS cells (Figure 1).

In the PEPCK based C_4 cycle most OAA is decarboxylated in the BS cytosol by PEPCK, yielding PEP and CO_2 in the BS cells (Figure 1). At the same time, NAD-ME activity in the mitochondria of BS cells can provide NADH for ATP production, fueling PEPCK activity (Kanai and Edwards, 1999) and helping to balance amino acid groups between M and BS cells by cycling alanine back to M cells (Furbank, 2011; Sommer et al., 2012). The decarboxylation reactions driven by either NAD-ME or PEPCK activity concentrate CO_2 in the BS cells (Figure 1).

In C_4 plants the RuBisCO operates exclusively in the BS cells, re-fixing the released CO_2. The remaining three-carbon compound –in case of the NAD-ME subtype, alanine; and in case of PEPCK, pyruvate– diffuses back to the mesophyll. There, alanine is converted back to pyruvate by the alanine aminotransferase and the primary CO_2 acceptor PEP is regenerated via pyruvate orthophosphate dikinase (PPDK) activity (Hatch and Slack, 1968). The enzymes taking part in the C_4 biochemical pathways are localized to the cytosol, chloroplasts and mitochondria. Thus, metabolite transport between the cell compartments and organelles plays an important role in the functionality of C_4 photosynthesis.

The main studies presented in this thesis will focus on the NAD-ME/PEPCK subtype of C_4 photosynthesis, featuring species from the Cleomaceae and Cyperaceae families (Marshall et al., 2007; Besnard et al., 2009).

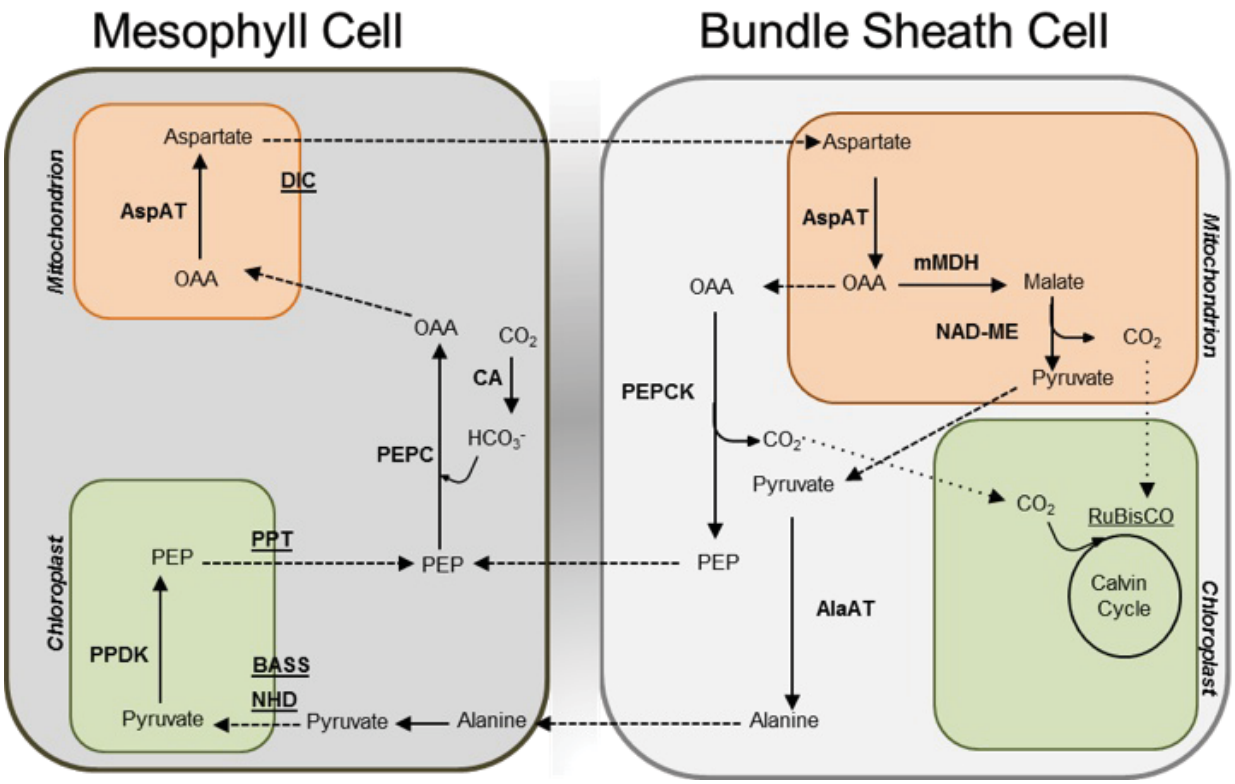


Figure 1. Schematic overview of the NAD-ME/PEPCK subtype C_4 cycle.
Normal printed terms represent metabolites; underlined and bold printed terms indicate relevant C_4 cycle transporters and bold printed terms indicate soluble C_4 cycle enzymes. Abbreviations of metabolites: **PEP**, phospho*enol*pyruvate; **OAA**, oxaloacetate. Abbreviations of enzymes: **RuBisCO**, ribulose 1,5bisphosphate carboxylase/oxygenase, **PEPC**, phospho*enol*pyruvate carboxylase; **CA**, carbonic anhydrase; **DIC**, dicarboxylate carrier; **AspAT**, aspartate aminotransferase; **mMDH**, mitochondrial malate dehydrogenase; **NAD-ME**, NAD-dependent malic enzyme; **AlaAT**, alanine aminotransferase; **PEPCK**, phospho*enol*pyruvate carboxykinase.
Abbreviations of transporters: **BASS**, bile acid:sodium symporter; **NHD**, sodium:hydrogen antiporter; **PPDK**, pyruvate orthophosphate dikinase; **PPT**, phosphate/phospho*enol*pyruvate translocator. Images were adapted from *Manuscript 3* and Sommer et al., 2012.

The C_4 leaf – Kranz anatomy and venation patterning

Close distances between M and BS cells are crucial for the functionality of the C_4 pathway. The vascular tissue is encompassed by the BS cells, which are in turn surrounded by M cells (Brown, 1975; Figure 2). The development of this so-called Kranz anatomy can be considered as one major step towards C_4 evolution. It guarantees the direct contact between M and BS cells, ensuring C_4 photosynthesis functionality. Therefore, leaves of C_4 plants develop differently from those of C_3 species.

C_4 related key changes in leaf structure comprise of proliferation of vascular tissue (narrow venation pattern), thinner leaves, abundant connections between the M and the BS cells via plasmodesmata to allow optimal exchange of the C_4 cycle metabolites, enlarged BS cells with thick cell walls and enhanced chloroplast accumulation (Brown et al., 2005; Figure 2). Putative signaling circuits that could be involved in establishment of Kranz anatomy have been identified recently (reviewed by Fouracre et al., 2014). The vascular pattern in both, C_3

and C_4 leaves seems to be initiated and established before differentiation and onset of photosynthesis (Dengler and Nelson, 1999; Sud and Dengler, 2000; Scarpella et al., 2006; McKown and Dengler, 2009). The initiation of procambial fate follows auxin flux, because exogenous application of auxin to leaf primordia has been reported to induce vein formation (Sachs, 1989; Scarpella et al., 2006). In *A. thaliana* procambial fate follows the expression domains of the auxin exporter PINOID-FORMED1 (PIN1) and the auxin response factor MONOPTEROS (Hardtke and Berleth, 1998; Scarpella et al., 2006; Sawchuk et al., 2007; Wenzel et al., 2007). After procambial fate determination, it is stabilized by the transcription factor ARABIDOPSIS THALIANA HOMEOBOX (ATHB8; Kang and Dengler, 2004; Scarpella et al., 2004), which is directly regulated by MONOPTEROS (Donner et al., 2009).

It is hypothesized that in C_4 plants, the vascular system itself provides the information, which initiates the fate of BS and M cells (Langdale and Nelson, 1991; Nelson and Langdale, 1992; Nelson and Dengler, 1997). The similarity of root endodermis and the BS cell layer in leaves has been thought to be equivalent (Esau, 1965; Nelson, 2011) and also controlled by the endodermis inducing regulatory *SHORTROOT/SCARECROW* (*SHR/SCR*) module of the root stele. Mutation of the *scr* gene in *Zea mays* perturbed Kranz anatomy and venation patterning (Slewinski et al., 2012). Moreover, *SCR* and *SHR* are significantly up-regulated during Kranz anatomy initiation in foliar leaves of *Z. mays* (Wang et al., 2013). Besides the *SCR/SHR* regulatory module, no other signal for BS/M cell fate induction has been found to date. Screening of *Z. mays* mutants revealed possible regulators for BS/M cell specification, *e.g.* bundle sheath defective2 (bsd2), golden2, golden2-like and high chlorophyll fluorescence136 (hcf136; Brutnell et al., 1999; Rossini et al., 2001; Covshoff et al., 2008; Fouracre et al., 2014). Two of the most promising BS/M cells specification regulators are golden2 and its paralog golden2-like 1, which have been implicated in controlling either BS chloroplast development or M chloroplast regulation, respectively (Hall et al., 1998; Rossini et al., 2001). These proteins are not compartmentalized to BS and M in C_3 species (Rossini et al., 2001; Fitter et al., 2002; Yasumura et al., 2005).

C_4 monocotyledons and eudicotyledons display higher vein numbers and orders (Ueno et al., 2006; McKown and Dengler, 2010). Both groups show higher number of procambial induction, followed by midvein and secondary order vein development, right after leaf emergence and unfolding. Comparative analysis of leaf development and vascularization between dicotyledonous C_4 and C_3 leaves in the genus *Flaveria* revealed, that differences in vein patterning originate from changes in the minor, but not in the major veins. In the C_4 *Flaveria* species, this happened through the formation of an extra minor vein order, *i.e.* seven vein orders formed in the *Flaveria bidentis* (C_4), whereas only six vein orders were initiated in the leaves of *Flaveria robusta* (C_3). Furthermore, the higher vein orders of minor veins occurred earlier in the C_4 leaves (McKown and Dengler, 2010).

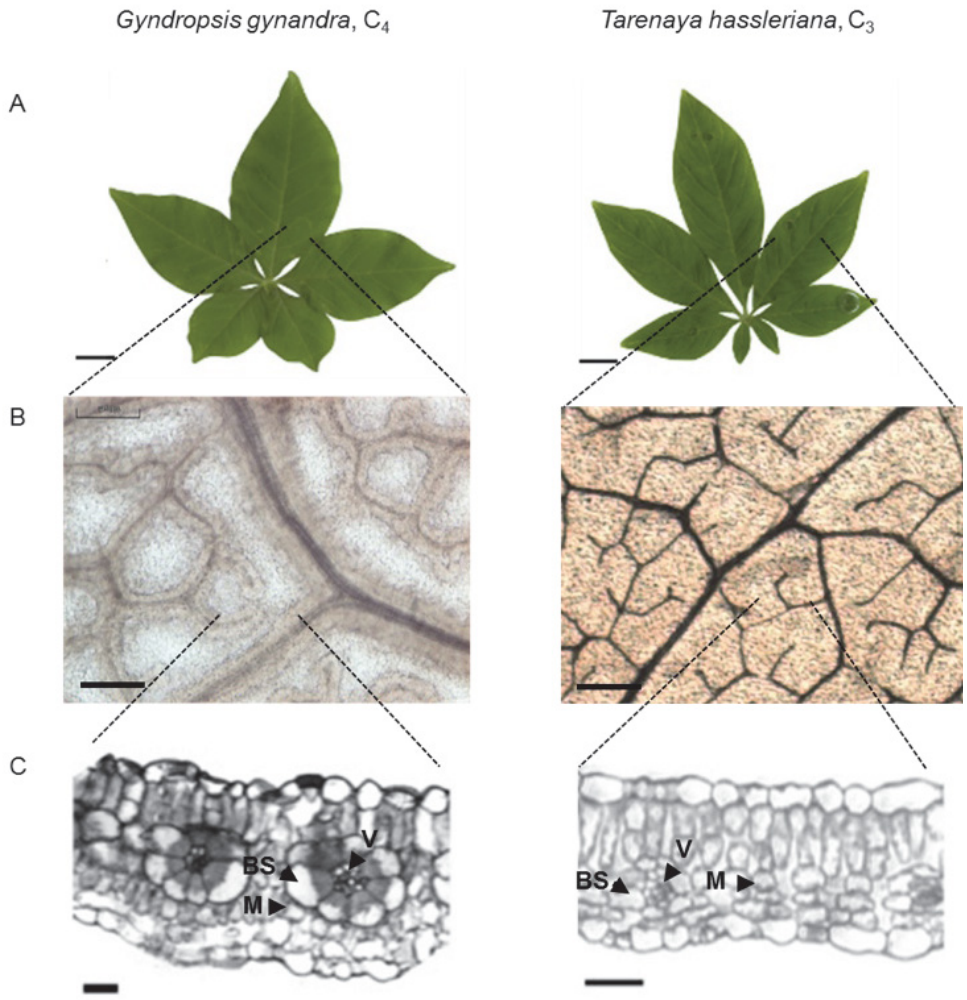


Figure 2. Overview of mature dicotyledonous Cleomaceae leaf architecture of *G. gynandra* (C_4 species; left) and *T. hassleriana* (C_3 species; right).
(A) Overview of mature leaves analyzed (stage 5); scale bar: 1 cm. **(B)** Vein patterning of the first third of leaves visualized with Safranine staining; scale bar 20µm. **(C)** Semi-thin cross-sections stained with Toluidine Blue. Cell types are indicated by closed arrows; **BS** bundle sheath; **M** mesophyll; **V** vein; scale bar 20 µm. Images were adapted from Külahoglu et al., 2014; *Manuscript 2*.

The evolution from C_3 to C_4 photosynthesis

Selective pressure resulting from a massive decline in atmospheric CO_2 levels 55 to 40 million years ago and accompanied by limited water availability, encouraged the evolution from C_3 C_4in some Angiosperms lineages (reviewed by Edwards et al., 2010). The evolution from C_3 to C_4 photosynthesis was a step-wise process with the acquisition of certain traits (Figure 3; Sage, 2004; McKown and Dengler, 2010; Sage et al., 2012). It is thought that higher vein density was one of the earliest acquired traits and a pre-condition for C_4 plant evolution (Sage, 2001; Sage, 2004). Recent analyses of anatomy and phylogeny within the C_4 and C_3 grasses led to new insights, regarding the factors that may have facilitated the repeated evolution of Kranz anatomy in the grass lineages (Christin et al., 2013; Griffiths et al., 2013; reviewed by Fouracre et al., 2014). This pre-conditioning towards Kranz anatomy

encompasses a shift in vein distance, BS cell size and organellar composition of BS and M cells (Sage, 2001; McKown and Dengler, 2007; Sage et al., 2012). Interestingly, close vein spacing has been found in C_3 plants, which are closely related to C_4 species, such as *Cleome* (Marshall et al., 2007; Voznesenskaya et al., 2007), *Flaveria* (McKown and Dengler, 2007) and *Heliotropium* (Muhaidat et al., 2011). The increased vein density is thought to be supporting leaf hydraulics in hot and arid environments (Sack and Scoffoni, 2013). In the grasses of the PACMAD clade, the frequent occurrence of C_4 photosynthesis has been linked to increased BS cell size (Christin et al., 2011).

Although the transition from C_3 to C_4 photosynthesis required several changes in leaf development, biochemistry and cellular organization (Figure 3), it is estimated that C_4 photosynthesis has evolved at least 36 times in eudicots and 26 times in monocots (Sage et al., 2011). These multiple independent and parallel origins of C_4 photosynthesis indicate that many, if not all, of the genes necessary for C_4 photosynthesis are already present in C_3 species. This also suggests that C_4 photosynthesis is a good example of the convergent evolution of a complex trait, making it an exceptional system for evolutionary studies (Sage et al., 2011). The evolution from a C_3 to a C_4 plant was suggested as “relatively easy” in terms of genetics (Westhoff and Gowik, 2010). The first step towards C_4 evolution is the presence of C_4 compatible anatomy, which allows implementation of the C_4 biochemical pathway in a C_3 background (Figure 3). Another early step in C_4 evolution and its prerequisite is the establishment of a photorespiratory CO_2 pump (C_2 cycle) by expressing the glycine decarboxylase complex (GDC) exclusively in BS cells (Sage et al., 2012). The C_2 cycle is based on glycine shuttling from M to BS cells (Edwards and Ku, 1987; Sage et al., 2012) being present in C_3-C_4 intermediate species (reviewed by Sage et al., 2012).

A recent study modeling the fitness landscape of C_4 evolution revealed that the evolution of the C_4 biochemistry has no local fitness maxima besides the C_4 endpoint. Each step poses an increase of the photosynthetic efficiency (Heckmann et al., 2013). Hence, C_3-C_4 intermediate species are not evolutionary dead-ends, but rather evolving towards full C_4 photosynthesis as selective pressure persists (Heckmann et al., 2013).

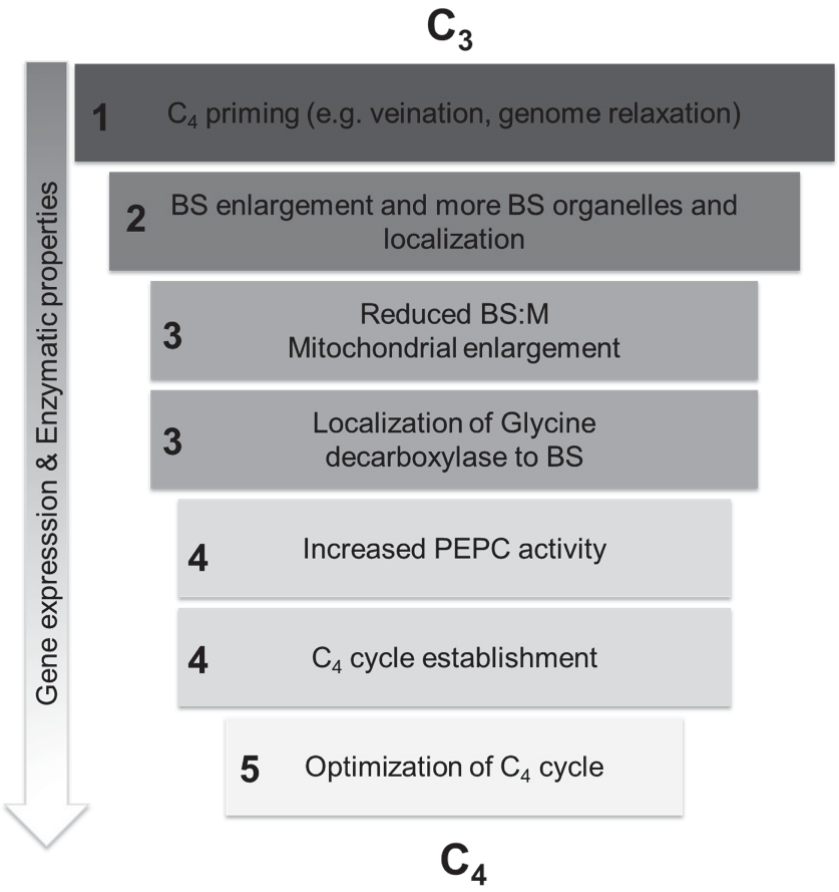


Figure 3. Conceptual model of C_4 photosynthesis evolution.
The model proposes five main steps of C_4 evolution indicated by numbers. **(1)** Preconditioning. **(2)** Proto-Kranz anatomy evolution. **(3)** Evolution of the photorespiratory CO_2 pump (C_2 photosynthesis). **(4)** Establishment of C_4 cycle and exclusive localization of the Calvin-Benson-Bassham cycle to the BS. **(5)** Optimization phase, where leaf anatomy and C_4 cycle are modified to maximize C_4 photosynthesis efficiency. **BS,** bundle sheath; **M,** mesophyll; **PEPC** phospho*enol*pyruvate carboxylase. This model was modified from (Sage et al., 2011; Westhoff and Gowik, 2010).

Transcriptional regulation of C_4 photosynthesis

System biological approaches and usage of “omics” technologies have been shown to be useful for shedding light on the transcriptional regulation of C_4 photosynthesis. These studies helped to elucidate the C_4 cycle integration in *Z. mays* (Li et al., 2010; Pick et al., 2011; Chang et al., 2012; Liu et al., 2013), *Panicum* (Bräutigam et al., 2014), *Flaveria* (Gowik et al., 2011; Mallmann et al., 2014) and *Cleome* (Bräutigam et al., 2010; Aubry et al., 2014), as well as BS and M cell specificity (Li et al., 2010; Pick et al., 2011; Chang et al., 2012; Aubry et al., 2014) and vein initiation (Wang et al., 2013).

Two recent publications analyzed C_4 photosynthesis at a systems level by comparing the transcriptome of closely related C_3 and C_4 species in the genus *Cleome* and *Flaveria* (Bräutigam et al., 2010; Gowik et al., 2011). Out of 13,662 transcripts approximately 603 were differentially expressed at a significant level between the mature leaves of C_3 and C_4 species in the Cleomaceae (Bräutigam et al., 2010). The comparative leaf transcriptome

analyses of the genus *Flaveria*, between a C_3 and C_4 species, has revealed 213 genes that exhibited significant differences (Gowik et al., 2011). The majority of the identified changes in transcript abundance could be related to the C_3-C_4 specific photosynthetic modes, indicating that these are not derived from the phylogenetic distances of the sequenced *Flaveria* species (Gowik et al., 2011). Excluding the C_4 cycle itself, the largest proportion of distinct regulated transcripts can be assigned to photosynthesis in both studies. Further functional groups of genes altered as a whole included the nitrogen amino acid metabolism and the translational machinery of chloroplast and the cytosol (Bräutigam et al., 2010; Gowik et al., 2011).

A recent study from Aubry and colleagues (2014) investigated the BS and M specific transcriptomes of the *G. gynandra* leaf. By comparing the expression profiles to *Z. mays*, they could isolate a shared set of homologous transcriptional regulators between *Z. mays* and *G. gynandra*, possibly inducing C_4 photosynthesis and maintaining cell-specific gene expression required for running the C_4 cycle. The complexity of Kranz development and the C_4 biochemistry suggest that several regulators must have been recruited for C_4 photosynthesis, which renders the existence of a master-switch unlikely (Westhoff and Gowik, 2010). However, recent results suggest that expression of the C_4 cycle enzymes are co-regulated together with photosynthesis (Pick et al., 2011; Aubry et al., 2014).

References

Alagna, F., D'Agostino, N., Torchia, L., Servili, M., Rao, R., Pietrella, M., Giuliano, G., Chiusano, M., Baldoni, L., and Perrotta, G. (2009). Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genomics **10,** 399.

Angeloni, F., Wagemaker, C.A.M., Jetten, M.S.M., den Camp, H.J.M.O., Janssen-Megens, E.M., Francoijs, K.J., Stunnenberg, H.G., and Ouborg, N.J. (2011). *De novo* transcriptome characterization and development of genomic tools for *Scabiosa columbaria L.* using next-generation sequencing techniques. Molecular Ecology Resources **11,** 662-674.

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

Barakat, A., DiLoreto, D.S., Zhang, Y., Smith, C., Baier, K., Powell, W.A., Wheeler, N., Sederoff, R., and Carlson, J.E. (2009). Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. BMC Plant Biology **9,** 51.

Benedito, V.A., Torres-Jerez, I., Murray, J.D., Andriankaja, A., Allen, S., Kakar, K., Wandrey, M., Verdier, J., Zuber, H., Ott, T., Moreau, S., Niebel, A., Frickey, T., Weiller, G., He, J., Dai, X., Zhao, P.X., Tang, Y., and Udvardi, M.K. (2008). A gene expression atlas of the model legume *Medicago truncatula*. Plant Journal **55,** 504-513.

Besnard, G., Muasya, A.M., Russier, F., Roalson, E.H., Salamin, N., and Christin, P.-A. (2009). Phylogenomics of C_4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. Molecular Biology and Evolution **26,** 1909-1919.

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

Bräutigam, A., and Gowik, U. (2010). What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biology **12,** 831-841.

Bräutigam, A., Schliesky, S., Külahoglu, C., Osborne, C.P., and Weber, A.P. (2014). Towards an integrative model of C_4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEPCK C_4 species. Journal of Experimental Botany **65,** 3579-3593.

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. (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.

Brown, N.J., Parsley, K., and Hibberd, J.M. (2005). The future of C_4 research – maize, *Flaveria* or *Cleome*? Trends in Plant Science **10,** 215-221.

Brown, W.V. (1975). Variations in anatomy, associations and origins of Kranz tissue. American Journal of Botany **62,** 395-402.

Brutnell, T.P., Sawers, R.J.H., Mant, A., and Langdale, J.A. (1999). BUNDLE SHEATH DEFECTIVE2, a novel protein required for post-translational regulation of the rbcL gene of maize. Plant Cell **11,** 849-864.

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 **160,** 165-177.

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., Bräutigam, 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 *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of Crucifers. Plant Cell **25,** 2813-2830.

Childs, K.L., Nandety, A., Hirsch, C.N., Gongora-Castillo, E., Schmutz, J., Kaeppler, S.M., Casler, M.D., and Buell, C.R. (2014). Generation of transcript assemblies and identification of Single Nucleotide Polymorphisms from seven lowland and upland cultivars of switchgrass. Plant Genome **7**. Doi: 10.3835/plantgenome2013.12.0041.

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 U S A **110,** 1381-1386.

Cokus S.J., Feng S.H., Zhang X.Y., Chen Z.G., Merriman B., Haudenschild C.D., Pradhan S., Nelson S.F., Pellegrini M., Jacobsen S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature **452**, 215–219.

Covshoff, S., Majeran, W., Liu, P., Kolkman, J.M., van Wijk, K.J., and Brutnell, T.P. (2008). Deregulation of maize C_4 photosynthetic development in a mesophyll cell-defective mutant. Plant Physiology **146,** 1469-1481.

Dassanayake, M., Haas, J.S., Bohnert, H.J., and Cheeseman, J.M. (2009). Shedding light on an extremophile lifestyle through transcriptomics. New Phytologist **183,** 764-775.

Dengler, N., and Nelson, T. (1999). Leaf structure and development in C_4 plants. C_4 plant biology, pp. 471-495 Academic Press: San Diego, CA.

Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., Jaillon, O., and Artiguenave, F. (2008). Annotating genomes with massive-scale RNA sequencing. Genome Biology **9**.

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., Osborne, C.P., Stroemberg, C.A.E., Smith, S.A., Bond, W.J., Christin, P.-A., Cousins, A.B., Duvall, M.R., Fox, D.L., Freckleton, R.P., Ghannoum, O., Hartwell, J., Huang, Y., Janis, C.M., Keeley, J.E., Kellogg, E.A., Knapp, A.K., Leakey, A.D.B., Nelson, D.M., Saarela, J.M., Sage, R.F., Sala, O.E., Salamin, N., Still, C.J., Tipple, B., and Consortium, C.G. (2010). The origins of C_4 grasslands: integrating evolutionary and ecosystem science. Science **328,** 587-591.

Edwards, G.E., and Ku, M.S. (1987). Biochemistry of C_3-C_4 intermediates. New York: Academic Press **In M.D. Hatch & N.K. Boardmann (Eds.),** pp.275-325.

Esau, K. (1965). Plant anatomy. Wiley & Sons, New York.

Fitter, D.W., Martin, D.J., Copley, M.J., Scotland, R.W., and Langdale, J.A. (2002). GLK gene pairs regulate chloroplast development in diverse plant species. Plant Journal **31,** 713-727.

Fouracre, J.P., Ando, S., and Langdale, J.A. (2014). Cracking the Kranz enigma with systems biology. Journal of Experimental Botany **65,** 3327-3339.

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,** 3103-3108.

Ghannoum, O., and Way, D.A. (2011). On the role of ecological adaptation and geographic distribution in the response of trees to climate change. Tree Physiology **31,** 1273-1276.

Gongora-Castillo, E., Fedewa, G., Yeo, Y., Chappell, J., DellaPenna, D., and Buell, C.R. (2012). Genomic approaches for interrogating the biochemistry of medicinal plant species. In Natural Product Biosynthesis by Microorganisms and Plants, Pt C, D.A. Hopwood, ed, pp. 139-159.

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

Griffiths, H., Weller, G., Toy, L.F., and Dennis, R.J. (2013). You're so vein: bundle sheath physiology, phylogeny and evolution in C_3 and C_4 plants. Plant, Cell & Environment **36,** 249-261.

Hall, L.N., Rossini, L., Cribb, L., and Langdale, J.A. (1998). GOLDEN 2: A novel transcriptional regulator of cellular differentiation in the maize leaf. Plant Cell **10,** 925-936.

Hamilton, J.P., Hansey, C.N., Whitty, B.R., Stoffel, K., Massa, A.N., Van Deynze, A., De Jong, W.S., Douches, D.S., and Buell, C.R. (2011). Single nucleotide polymorphism discovery in elite North American potato germplasm. BMC Genomics **12,** 302.

Hansey, C.N., Vaillancourt, B., Sekhon, R.S., de Leon, N., Kaeppler, S.M., and Buell, C.R. (2012). Maize (*Zea mays L.*) genome diversity as revealed by RNA-sequencing. PLoS ONE **7,** e33071.

Hardtke, C.S., and Berleth, T. (1998). The Arabidopsis gene MONOPTEROS encodes a transcription factor mediating embryo axis formation and vascular development. Embo Journal **17,** 1405-1411.

Hatch, M.D. (1987). C_4 photosynthesis – a unique bled of modified biochmistry, anatomy and ultrastructure. Biochimica Et Biophysica Acta **895,** 81-106.

Hatch, M.D., and Slack, C.R. (1968). A new enzyme for interconversion of pyruvate and phosphopyruvate and its role in C_4 dicarboxylic acid pathway of photosynthesis. Biochemical Journal **106,** 141-146.

Heckmann, D., Schulze, S., Denton, A., Gowik, U., Westhoff, P., Weber, A.P., and Lercher, M.J. (2013). Predicting C_4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. Cell **153,** 1579-1588.

Jothi R., Cuddapah S., Barski A., Cui K., Zhao K. (2008) Genome- wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. Nucleic Acids Research **36**, 5221–5231.

Kanai, R., and Edwards, G.E. (1999). The biochemistry of C_4 photosynthesis. C_4 plant biology**,** 49-87.

Kang, J., and Dengler, N. (2004). Vein pattern development in adult leaves of Arabidopsis thaliana. International Journal of Plant Sciences **165,** 231-242.

Keeley, J.E. (1998). C_4 photosynthetic modifications in the evolutionary transition from land to water in aquatic grasses. Oecologia **116,** 85-97.

Kocacinar, F., McKown, A.D., Sage, T.L., and Sage, R.F. (2008). Photosynthetic pathway influences xylem structure and function in *Flaveria* (Asteraceae). Plant Cell and Environment **31,** 1363-1376.

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

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,** 1060-1067.

Liu, W.Y., Chang, Y.M., Chen, S.C.C., Lu, C.H., Wu, Y.H., Lu, M.Y.J., Chen, D.R., Shih, A.C.C., Sheue, C.R., Huang, H.C., Yu, C.P., Lin, H.H., Shiu, S.H., Ku, M.S.B., and Li, W.H. (2013). Anatomical and transcriptional dynamics of maize embryonic leaves during seed germination. Proc Natl Acad Sci U S A **110,** 3979-3984.

Long, S.P. (1999). Environmental responses. C_4 plant biology**,** 215-249.

Mallmann, J., Heckmann, D., Bräutigam, A., Lercher, M.J., Weber, A.P., Westhoff, P., and Gowik, U. (2014). The role of photorespiration during the evolution of C_4 photosynthesis in the genus Flaveria. Elife**,** e02478.

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. The Plant Journal **51,** 886-896.

McKown, A.D., and Dengler, N.G. (2007). Key innovations in the evolution of Kranz anatomy and C_4 vein pattern in *Flaveria* (Asteraceae). American Journal of Botany **94,** 382-399.

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.

McKown, A.D., and Dengler, N.G. (2010). Vein patterning and evolution in C_4 plants. Botany-Botanique **88,** 775-786.

Muhaidat, R., and McKown, A.D. (2013). Significant involvement of PEPCK in carbon assimilation of C_4 eudicots. Annals of Botany **111,** 577-589.

Muhaidat, R., Sage, R.F., and Dengler, N.G. (2007). Diversity of Kranz anatomy and biochemistry in C_4 eudicots. American Journal of Botany **94,** 362-381.

Muhaidat, R., Sage, T.L., Frohlich, M., 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 and Environment **34,** 1723-1736.

Nelson, T. (2011). The grass leaf developmental gradient as a platform for a systems understanding of the anatomical specialization of C_4 leaves. Journal of Experimental Botany **62,** 3039-3048.

Nelson, T., and Langdale, J.A. (1992). Developmental genetics of C_4 photosynthesis. Annual Review of Plant Physiology and Plant Molecular Biology **43,** 25-47.

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

Novaes, E., Drost, D.R., Farmerie, W.G., Pappas, G.J., Grattapaglia, D., Sederoff, R.R., and Kirst, M. (2008). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics **9,** 312.

Osborne, C.P., and Freckleton, R.P. (2009). Ecological selection pressures for C_4 photosynthesis in the grasses. Proceedings of the Royal Society Biological Sciences **276,** 1753-1760.

Pick, T.R., Bräutigam, A., Schlueter, 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 C_4 model and provides candidates for regulation. Plant Cell **23,** 4208-4220.

Reiskind, J.B., Berg, R.H., Salvucci, M.E., and Bowes, G. (1989). Immunogold localization of primary carboxylases in leaves of aquatic and a C_3-C_4 intermediate species. Plant Science **61,** 43-52.

Rossini, L., Cribb, L., Martin, D.J., and Langdale, J.A. (2001). The maize Golden2 gene defines a novel class of transcriptional regulators in plants. Plant Cell **13,** 1231-1244.

Roth-Nebelsick, A., Uhl, D., Mosbrugger, V., and Kerp, H. (2001). Evolution and function of leaf venation architecture: A review. Annals of Botany **87,** 553-566.

Sachs, T. (1989). The development of vascular networks during leaf development. Curr. Top. Plant Biochem. Physiol.**8**, 168-183.

Sawchuk, M.G., Head, P., Donner, T.J., and Scarpella, E. (2007). Time-lapse imaging of Arabidopsis leaf development shows dynamic patterns of procambium formation. New Phytologist **176,** 560-571.

Sack, L., and Scoffoni, C. (2013). Leaf venation: structure, function, development, evolution, ecology and applications in the past, present and future. New Phytologist **198,** 983-1000.

Sage, R. (2001). Environmental and evolutionary preconditions for the origin and diversification of the C_4 photosynthetic syndrome. Plant Biology **3,** 202-213.

Sage, R.F. (2002). Variation in the k(cat) of RuBisCO in C_3 and C_4 plants and some implications for photosynthetic performance at high and low temperature. Journal of Experimental Botany **53,** 609-620.

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

Sage, R.F., and Pearcy, R.W. (2000). The physiological ecology of C_4 photosynthesis. In Photosynthesis (Springer), pp. 497-532.

Sage, R.F., and Zhu, X.-G. (2011). Exploiting the engine of C_4 photosynthesis. Journal of Experimental Botany **62,** 2989-3000.

Sage, R.F., Pearcy, R.W., and Seemann, J.R. (1987). The nitrogen use efficiency of C_3 and C_4 plants. 3 leaf nitrogen effects on the activity of carboxylating enzymes in

Chenopodium album (L.) and *Amaranthus retroflexus* (L). Plant Physiology **85,** 355-359.

Sage, R.F., Christin, P.A., and Edwards, E.J. (2011). The C_4 plant lineages of planet Earth. Journal of Experimental Botany **62,** 3155-3169.

Sage, R.F., Sage, T.L., and Kocacinar, F. (2012). Photorespiration and the evolution of C_4 photosynthesis. In Annual Review of Plant Biology **63**, S.S. Merchant, ed, pp. 19-47.

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 & Development **20,** 1015-1027.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of *Arabidopsis thaliana* development. Nature Genetics **37,** 501-506.

Seemann, J.R., Badger, M.R., and Berry, J.A. (1984). Variations in the specific activity of Ribulose-1,5-Bisphosphate Carboxylase between species utilizing differing photosynthetic pathways. Plant Physiology **74,** 791-794.

Sekhon, R.S., Lin, H., Childs, K.L., Hansey, C.N., Buell, C.R., de Leon, N., and Kaeppler, S.M. (2011). Genome-wide atlas of transcription during maize development. The Plant Journal **66,** 553-563.

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

Sommer, M., Bräutigam, A., and 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.

Strickler, S.R., Bombarely, A., and Mueller, L.A. (2012). Designing a transcriptome next-generation sequencing project for a nonmodel plant species. American Journal of Botany **99,** 257-266.

Sud, R.M., and Dengler, N.G. (2000). Cell lineage of vein formation in variegated leaves of the C_4 grass *Stenotaphrum secundatum*. Annals of Botany **86,** 99-112.

Ueno, O., Kawano, Y., Wakayama, M., and Takeda, T. (2006). Leaf vascular systems in C_3 and C_4 grasses: a two-dimensional analysis. Annals of Botany **97,** 611-621.

Voznesenskaya, E.V., Edwards, G.E., Kiirats, O., Artyusheva, E.G., and Franceschi, V.R. (2003). Development of biochemical specialization and organelle partitioning in the single-cell C_4 system in leaves of *Borszczowia aralocaspica* (Chenopodiaceae). American Journal of Botany **90,** 1669-1680.

Voznesenskaya, E.V., Koteyeva, N.K., Chuong, S.D.X., Ivanova, A.N., Barroca, J., Craven, L.A., and Edwards, G.E. (2007). Physiological, anatomical and biochemical characterisation of photosynthetic types in genus *Cleome* (Cleomaceae). Functional Plant Biology **34,** 247.

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 C_4 Kranz anatomy. Plant Journal **75,** 656-670.

Wang, W., Wang, Y., Zhang, Q., Qi, Y., and Guo, D. (2009). Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. BMC Genomics **10,** 465.

Wenzel, C.L., Schuetz, M., Yu, Q., and Mattsson, J. (2007). Dynamics of MONOPTEROS and PIN-FORMED1 expression during leaf vein pattern formation in Arabidopsis thaliana. Plant Journal **49,** 387-398.

Westhoff, P., and Gowik, U. (2010). Evolution of C_4 photosynthesis - Looking for the master switch. Plant Physiology **154,** 598-601.

Yasumura, Y., Moylan, E.C., and Langdale, J.A. (2005). A conserved transcription factor mediates nuclear control of organelle biogenesis in anciently diverged land plants. Plant Cell **17,** 1894-1907.

IV. AIM OF THIS PHD THESIS

Understanding the highly productive trait of C_4 photosynthesis and applying it to C_3 crop species is a major effort in plant breeding. The reproduction of the C_4 engine in C_3 crop plants is hoped to lead to a second "green revolution" (Sage and Zhu, 2011; von Caemmerer et al., 2012). Engineering of the C_4 trait is one of the most ambitious projects in plant sciences and successful manipulation would enormously enhance agricultural production and provide mankind with an increase in chemical energy and food (Sage and Zhu, 2011). However, this goal can only be achieved, if detailed understanding of the C_4 trait is gained. In order to rebuild the C_4 engine we need a comprehensive transcriptomic and genetic blueprint of:

(i) the genetic and transcriptional regulation leading to C_4 leaf anatomy
(ii) C_4 cycle metabolite transport and integration
(iii) bundle sheath and mesophyll cell specification.

The emergence of next-generation sequencing technologies facilitated genomic and transcriptional comparisons on a large scale in a time efficient manner. The molecular biological methods and bioinformatics tools used for generating the presented transcriptomes are introduced by a method review (*Manuscript 1;* **V.1 FIRST-AUTHORED MANUSCRIPTS**; Külahoglu and Bräutigam, 2014).

Recent transcriptome studies have shown that a limited subset of genes is differentially regulated between related dicotyledonous C_4 and C_3 species (Bräutigam et al., 2010; Gowik et al., 2011). Still, the regulation of Kranz anatomy and establishment of C_4 photosynthesis in the leaf remains largely unknown (reviewed by Fouracre et al., 2014). Cell specific and leaf developmental gradients between closely related C_3 and C_4 species are being employed to identify C_4 specific gene regulation and how "C_4-ness" is established in the leaf.

The aim of the presented studies in this PhD thesis is to understand the transcriptional differences, regulation and dynamics in phototrophic and heterotrophic tissues between C_4 and C_3 plant species. Special focus lays on developmental leaf gradients and culm transcriptomes for identifying C_4 specific gene regulative modules and understanding how "C_4-ness" is established in the leaf. The analyzed species were selected to facilitate transcriptome-level comparisons of C_4 with other photosynthetic traits.

In *Manuscript 2* (**V.2 FIRST-AUTHORED MANUSCRIPTS**), the selected C_3 and C_4 species in the Cleomaceae are closely related to each other and to the model plant *A. thaliana* (Marshall et al., 2007). Two comparative transcriptome atlases, including three developmental gradients (seed, seedling and leaf) were generated from *Gynandropsis gynandra* (C_4 species) and *Tarenaya hassleriana* (C_3 species; *Manuscript 2*; Külahoglu et al.,

2014). In the addendum (**VI. ADDENDUM**) to the **V. FIRST-AUTHORED MANUSCRIPTS** *Manuscript 3* the analyzed species of the Cyperaceae family is able to change its photosynthetic mode within one life cycle and switch from C_4 photosynthesis to a C_4-like state (Ueno and Wakayama, 2004). This sedge, *Eleocharis retroflexa* is able to adjust its carbon concentrating mechanism depending on the environmental conditions (Ueno and Wakayama, 2004). The culm transcriptomes of the amphibious sedge *E. retroflexa* grown in an aquatic and terrestrial environment were obtained to address, which physiological and transcriptional programs enable *E. retroflexa* to thrive during submergence and on soil, while being a C_4 species. We assessed how the change of environment affects the proposed photosynthetic modes in the genus *Eleocharis* (C_4-like and C_4 photosynthesis; *Manuscript 3*).

References

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. (2010). An mRNA Blueprint for C4 Photosynthesis Derived from Comparative Transcriptomics of Closely Related C3 and C4 Species. Plant Physiology **155,** 142-156.

Fouracre, J.P., Ando, S., and Langdale, J.A. (2014). Cracking the Kranz enigma with systems biology. Journal of Experimental Botany **65,** 3327-3339.

Gowik, U., Bräutigam, A., Weber, K.L., Weber, A.P., and Westhoff, P. (2011). Evolution of C4 photosynthesis in the genus Flaveria: how many and which genes does it take to make C4? The Plant Cell **23,** 2087-2105.

Külahoglu, C., and Bräutigam, A. (2014). Quantitative transcriptome analysis using RNA-seq. Methods Mol Biol **1158,** 71-91.

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. (2014). Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C3 and C4 Plant Species. The Plant Cell **26,** 3243-3260.

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 C3 to C4 photosynthesis. The Plant Journal **51,** 886-896.

Sage, R.F., and Zhu, X.-G. (2011). Exploiting the engine of C-4 photosynthesis. Journal of Experimental Botany **62,** 2989-3000.

Ueno, O., and Wakayama, M. (2004). Cellular expression of C3 and C4 photosynthetic enzymes in the amphibious sedge Eleocharis retroflexa ssp. chaetaria. J Plant Res **117,** 433-441.

von Caemmerer, S., Quick, W.P., and Furbank, R.T. (2012). The Development of C-4 Rice: Current Progress and Future Challenges. Science **336,** 1671-1672.

V.1 FIRST-AUTHORED MANUSCRIPTS

Manuscript 1

Quantitative Transcriptome Analysis using RNA-seq

Canan Külahoglu* *and Andrea Bräutigam*

Published in Methods in Molecular Biology (2014), Vol.1158, pp. 71-91
DOI 10.1007/978-1-4939-0700-7_5

Impact Factor: 1.29

**First Author*

Main findings:

This technical review presents the workflow for RNA-sequencing experiments in plants. It starts from experimental planning to best laboratory practices of established protocols to biological data extraction. It provides protocols for RNA extraction, DNAse treatment and evaluation of the Bioanalyser results. Furthermore, bioinformatical tools (*i.e.* fastQC, fastx, Bowtie, TopHat, Cufflinks) are explained with easy to use command line instructions and examples to help the researcher to interpret results. Different tools for subsequent biological information are presented helping the researcher to proceed with data interpretation.

Contributions:

- Establishment of laboratory protocols
- Bioinformatic data mining (except for **chapter 3.8**)
- Writing and editing of the manuscript

Quantitative Transcriptome Analysis Using RNA-seq

Canan Külahoglu and Andrea Bräutigam

Abstract

RNA-seq has emerged as the technology of choice to quantify gene expression. This technology is a convenient accurate tool to quantify diurnal changes in gene expression, gene discovery, differential use of promoters, and splice variants for all genes expressed in a single tissue. Thus, RNA-seq experiments provide sequence information and absolute expression values about transcripts in addition to relative quantification available with microarrays or qRT-PCR. The depth of information by sequencing requires careful assessment of RNA intactness and DNA contamination. Although the RNA-seq is comparatively recent, a standard analysis framework has emerged with the packages of Bowtie2, TopHat, and CuffLinks. With rising popularity of RNA-seq tools have become manageable for researchers without much bioinformatical knowledge or programming skills. Here, we present a workflow for a RNA-seq experiment from experimental planning to biological data extraction.

Key words: Next-generation sequencing, RNA-seq, Gene expression quantification, Circadian rhythm

1. Introduction

High-throughput RNA-sequencing (RNA-seq) enables the researcher to quantify gene expression, discover new splice variants and (polyadenylated) transcripts within a single assay. It has been successfully applied to a variety of plants **[1–3]**. Unlike an analysis with microarrays, RNA-seq detected differential splicing events along the developmental gradient in maize **[2]**. Knowledge about absolute expression levels provided by RNA-seq enables the identification of abundant genes in contrast to just identifying relative changes **[4, 5]**. RNA-seq data can be combined with microarrays to both identify abundant genes and to characterize their relative expression pattern **[6]**.

Compared to microarrays the RNA-seq experiment poses unique challenges. The RNA-seq experiment starts with planning a suitable strategy. After sampling total RNA is isolated from the samples in sequencing grade quality and contaminating DNA is removed. RNA is converted to cDNA libraries for sequencing.

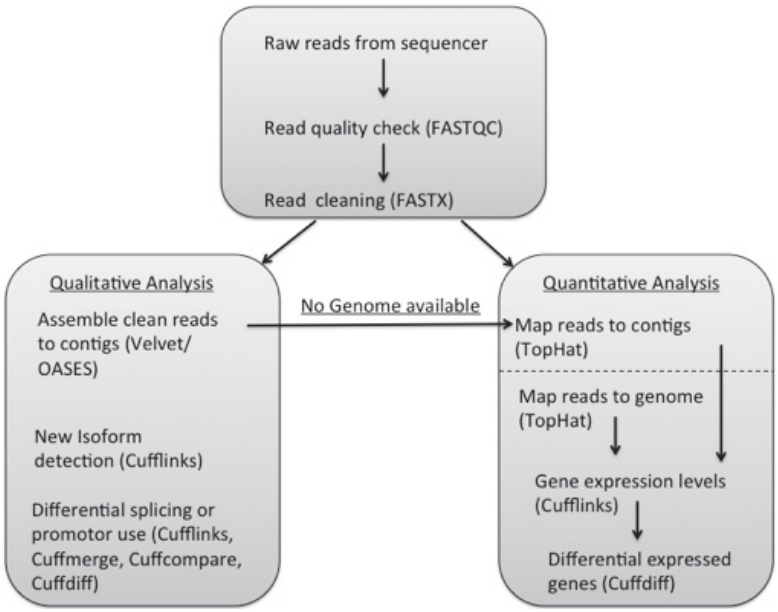


Figure 1. Overview of RNA-seq data analysis and biological information extraction. Software used in this review is written in *parentheses*

After the wetlab procedure of preparing RNA and the sequencing itself, analysis of the sequenced reads requires a bioinformatic pipeline of read cleaning, read mapping, statistics, and biological interpretation. In this protocol we introduce a pipeline, which requires minimal prior bioinformatic knowledge, since it relies on public domain software packages. A selection of possible methods to extract biological data from the experiment is also presented (summarized in **Figure 1**).

2. Material

2.1 Sampling of Plant Material

1. Liquid nitrogen.
2. Dewar.
3. Scissors or scalpel.
4. Aluminum foil bags or tubes with the lid punctured.

2.2 Grinding

1. Mortar.
2. Pestle.
3. Styrofoam grinding platform.
1. For easier grinding, the lid of a conventional Styrofoam shipping box can be hollowed out to accommodate the mortar, thus relieving the person of grinding by holding the mortar in place.
4. Liquid nitrogen transfer vessel.
2. To pre-cool the mortar efficiently and to transfer fresh liquid nitrogen during grinding, a transfer vessel can be fashioned out of a plastic 10 mL pipet and a 50 mL tube using tape.

2.3 RNA Extraction

In order to have optimal sequencing results it very important to avoid RNA degradation by RNAses (*see* **Note 1**). When preparing solutions for the described protocols, make sure to use RNAse-free water (*see* **Note 2**).

2.3.1 RNA Extraction Method I: Guanidinium Isothiocyanate Extraction Method

This standard method is modified from **[7, 8]**. It delivers around 200 ng/µl clean RNA per 100 mg fresh weight from tissues, which do not contain high phenolic, carbohydrate, or lipid contents. The main chaotropic reagent is guanidinium isothiocyanate, which effectively disrupts tissues and inactivates RNAses, while keeping the RNA intact.

1. *Water-saturated phenol*: dissolve 100 g phenol crystals in distilled water at 65 °C. Aspirate the upper water phase and store up to 1 month at 4 °C (*see* **Note 3**). Acidic phenol is also available commercially.
2. *1 N acetic acid*: dilute 2.8 mL of 17.4 M acetic acid with 47.2 mL DEPC-treated water.
3. *3 M sodium acetate*: dissolve 20.41 g sodium acetate trihydrate (Mr=136.08 g/mol) in 30 mL water, adjust the pH to 6.0 with 3 M acetic acid, fill up with water to the final volume of 50 mL, add 0.1 % DEPC, and sterilize by autoclaving.
4. *5 M lithium chloride*: dissolve 10.59 g lithium chloride (Mr=42.39 g/mol) in 50 mL water, add 0.1 % DEPC, and sterilize by autoclaving.
5. *RNase-ALL stock solution*: 4 M Guanidine isothiocyanate, 25 mM Na-acetate, 0.5 % (w/v) *N*-laurosylsarcosine, 0.7 % (v/v) mercaptoethanol, pH 7.0. Dissolve 47.2 g guanidinium isothiocyanate (Mr = 118.16g/mol) and 0.5g *N*-laurosylsarcosine sodium salt (Mr = 293.4 g/mol) in 30 mL water at 65 °C. Then add 830 µL 3 M Sodium acetate (pH 5.2). Adjust the pH to 7.0 with 2 N NaOH and fill up with water to the final volume to 100 mL.
6. *Chloroform–Isoamyl alcohol 24/1 (v/v) solution*: mix 24 mL of chloroform with 1 mL of isoamyl alcohol.
7. *DEPC-water*: mix H_2O with 0.1 % DEPC, stir and autoclave [alternative: purchase RNAse- and DNAse-free water]

8. *80 % p.a. EtOH*.

2.3.2 RNA Extraction Method II: Modified CTAB Protocol with Silica Columns

This protocol is recommended for recalcitrant plant tissues with high secondary metabolites, starch, and lipid composition. It is a cetyltrimethylammonium bromide (CTAB)-based method in combination with silica columns of RNAeasy Plant Mini kit (Qiagen, Germany) for effective RNA isolation **[9]**.

1. *CTAB-PVP buffer*: 2 % (w/v) CTAB, 2 % (w/v) Polyvinylpyrrolidone (PVP-40), 100 mM Tris–HCl (pH 8.0), 25 mM EDTA, 2 M NaCl. Add 2 % β-mercaptoethanol before usage.
2. *Chloroform–Isoamyl alcohol 24/1 (v/v) solution*: mix 24 mL of chloroform with 1 mL of isoamyl alcohol.
3. 96 % EtOH: use fresh p.a. EtOH.
4. Silica column-based commercial kit (e.g., Qiagen RNeasy Plant Kit, Qiagen, Germany).

2.4 RNA Quality Control

1. Photometer or NanoDrop (Thermo Scientific).
2. Access to Bioanalyzer (Agilent).
3. PCR reagents (PCR-machine, PCR-microtubes, dNTPs, TAQ-Polymerase, Primer, suitable reaction buffer).

2.5 Sequencing

Presumed to be outsourced to intramural or commercial supplier.

2.6 Read Analysis

1. Sequencing reads in .FASTQ file format.
2. Standard desktop computer >8 GB RAM, Linux environment preferred.
3. Programs of **Table 1** installed.

2.7 Read Mapping and Quantification

1. Cleaned reads from Subheading **3.7**.
2. Fasta file of the reference sequence.
3. `.gff/.gtf` file of annotation for the reference sequence.
4. Standard desktop computer with Linux-based operating system and >2 GB RAM.
5. Programs from **Table 2** installed.

Table 1 Overview of read analysis and cleaning software recommended in this review.

Program	Version	Environment	Source
FASTQC	v0.10.1	Install all dependencies	www.bioinformatics.babraham.ac.uk/projects
FASTX	v0.0.13	Install all dependencies	http://hannonlab.cshl.edu/fastx_toolkit/

Table 2 Overview of software needed for read mapping and expression statistics.

Program	Version	Environment	Source
BAMtools	v1.0.2	Install all dependencies	https://github.com/pezmaster31/bamtools
Bowtie	v0.12.8	Install all dependencies	http://bowtie-bio.sourceforge.net
Cufflinks	v2.0.2	Install all dependencies	http://cufflinks.cbcb.umd.edu/downloads/
TopHat	v1.4.1	Install all dependencies	http://tophat.cbcb.umd.edu/downloads/

2.8 Extracting Biological Information

1. Tab delimited text file (or Excel File) containing the gene identifier in the first column and the resulting read counts in subsequent columns with one header row.
2. Standard desktop computer system with >8 GB RAM (for hierarchical clustering analysis of genes >64 GB RAM).
3. Microsoft Office Excel.
4. MapMan Software.
5. MultiExperiment Viewer Software.
6. VirtualPlant Web access.

3. Methods

Designing a RNA-seq experiment is similar in principle to designing a microarray experiment. Each sampling point needs to be analyzed in biological replicates. The statistical methods require replicates to assess variation prior to determination of differentially expressed genes. While some tools estimate variation based on different samples (and use this assessment to guide their decisions), most tools require replication for proper assessment. The large number of reads is no substitute for replication. For some RNA-seq technologies such as Illumina HiSeq 2000 samples can be multiplexed up to 12-fold, that is, up to 12 samples can be barcoded and run on a single lane enabling replication at almost the same sequencing cost. Sampling time points for circadian experiments will depend on the analysis to be done post-sequencing and hence the individual experiment. If a model with predictive qualities or transcription factor network analysis is the goal, sampling time points should be defined with the help of the statistician or bioinformatician who will partake in the modeling effort. Finally, the strategy ought to be aligned with the available computer power. RNA-seq experiments create vast amounts of data to be stored and analyzed and hence require the infrastructure to support the experiment.

3.1 Sampling Plant Material for RNA-seq

1. Prepare the aluminum foil bags or puncture the lids of tubes.
2. Set up a dewar with liquid nitrogen.
3. Cut off the tissue, transfer it to the bag or tube and immediately shock-freeze. Any delay in freezing or thawing after initial freezing will lead to RNA degradation
4. Either proceed to grinding immediately or store at −80 °C.

3.2 Grinding the Samples

1. Set up the mortar and pestle for grinding.
2. Set up the dewar with clean liquid nitrogen.
3. Pre-cool the mortar and pestle by twice evaporating liquid nitrogen.
4. Grind the tissues under liquid nitrogen to a fine powder. Alternative grinding methods such as bead mills are suitable if the samples remain deep frozen at all times.
5. Store powder at −80 °C or proceed immediately to RNA extraction.

3.3 RNA Extraction

We will describe two alternative RNA extraction methods. The guanidinium isothiocyanate-based extraction method delivers sequencing quality RNA for most tissue types. For tissues with a high secondary-metabolite, carbohydrate, and fatty-acid content such as seed tissue, a CTAB-based extraction is recommended.

Conduct all procedures at room temperature unless specified otherwise. Keep RNA samples on ice during all processes. When working with toxic, volatile chemicals, such as chloroform and phenol, work under a fume hood.

3.3.1 Method I: Guanidinium Isothiocyanate Based Extraction Method

1. Pre-cool centrifuge to 10 °C.
2. Prepare RNase-ALL working solution by mixing acidic phenol with RNAase-ALL solution 1:1 (v:v). When using phenol, make sure the phenolic phase beneath the aqueous phase is used.
3. Homogenization: Transfer 100 mg of ground plant material to 2 mL microtube and add immediately 1 mL RNase-ALL work- ing solution (add 1 mL RNase-ALL working solution per 100 mg fresh tissue) and vortex tube 5 s. Incubate tubes on ice for at least 15 min, while vortexting sample every 5 min for 5 s.
4. Protein extraction: Add 300 μL of chloroform–isoamyl alcohol 24:1 (v/v) solution and vortex tube for 10 s. Incubate on ice for at least 5 min. Centrifuge at 10,000×*g* for 10 min at 10 °C. Transfer the upper aqueous phase (~700 μL to 1 mL) into a fresh 2 mL microtube. Add 700 μL water-saturated Phenol solution (acidic phenol). Invert several times and add subsequently 300 μL of chloroform–isoamyl alcohol 24:1 (v/v) solution to remove residual phenol. Vortex for 10 s and centrifuge at 10,000 × *g* for 10 min at 10 °C. Transfer the upper aqueous phase (~700 μL to 1 mL) into a fresh micro 2 mL tube. Add 300 μL of chloroform–isoamyl alcohol 24:1 (v/v) solution to remove residual phenol.
5. First RNA precipitation: Transfer the upper (aqueous) phase (~700 μL) containing the extracted RNA into a fresh 1.5 mL tube. Add 1/20 volume of 1 N acetic acid (~35 μL), invert several times, add 0.7–1.0 volume of 100 % ethanol (~490–700 μL) and vortex. Centrifuge at 10,000×*g* for 20 min at 10 °C and carefully remove the supernatant. Resuspend the RNA pellet with 1 mL of 3 M sodium acetate (pH 6.0). Centrifuge at 12,000 × *g* for 10 min at 10 °C and remove the supernatant. Add 1 mL of 80 % ethanol and resuspend the RNA pellet. Centrifuge at 12,000×*g* for 10 min at 10 °C and remove the supernatant. Add again 1 mL of 80 % ethanol and resuspend the RNA pellet. Centrifuge at 12,000×*g* for 10 min at 10 °C and remove the supernatant. Air-dry the pellet for 10–15 min at RT (set the tube upside down on a paper towel). Resuspend the pellet in 500 μL RNase-free water (in case of pooling: resuspend in 250 μL RNase-free water and combine after heat treatment).
6. Second precipitation: Incubate at 60 °C for 5 min to ensure complete solubilization. Centrifuge at 12,000×*g* for 1 min at 10 °C. Transfer supernatant into a fresh tube (combine your two samples). Add an equal volume of 5 M lithium chloride (~500 μL), mix well, and incubate overnight at 4 °C. Centrifuge at 12,000 × *g* for 10 min at 10 °C and remove the supernatant. Resuspend the pellet of RNA with 1 mL 80 % ethanol. Centrifuge at 12,000×*g* for 10 min at 10 °C and remove the supernatant. Resuspend the pellet again with 1 mL 80 % ethanol. Centrifuge at 12,000 rpm for 10 min at 10 °C and remove the supernatant. Air-dry the pellet for 10–15 min at RT (set the tube upside down on a paper towel). Add 50 μL DEPC-treated water. Incubate at 60 °C for 10–15 min to ensure complete solubilization. Centrifuge shortly.
7. Store RNA at −80 °C for long-term storage or on ice if continuing with Subheading **3.4**.

3.3.2 Method II: Modified CTAB Protocol with Silica Columns

If the material is recalcitrant during standard RNA isolation, a modified CTAB protocol combined with silica columns may yield intact, clean RNA:

1. Prepare CTAB buffer working solution: 50 % (v/v) CTAB 233 buffer stock solution, 2 % (v/v) BME, and 50 % (v/v) acidic phenol in a 50 mL polypropylene tube (Falcon). Pre-heat the CTAB working solution to 65 °C in water bath and make sure not to close the tube completely to prevent phenol spills.
2. *Homogenization*: Add 1 mL pre-heated CTAB buffer working solution per 100 mg frozen ground tissue in 2 mL microtube. Vortex tube for 5 s and be careful opening tube letting gases evaporate. Incubate samples in water bath at 65 °C for 15–30 min. Vortex samples every 5 min to help tissue disruption.
3. *Protein extraction*: Add an equal volume of chloroform–isoamyl alcohol (24:1) to each sample and vortex each sample for at least 10 s. Centrifuge samples at 10,000×*g* at 10 °C for 20 min. Transfer the aqueous (upper) supernatant (1 mL per tube) to a new 2 mL microtube

and add an equal volume of chloroform–isoamyl alcohol (24:1). Vortex samples for 10 s and centrifuge at 10,000 × *g* 4 °C for 10 min. Take care not to touch the interphase separating the aqueous (upper) and non- aqueous (lower) phase, when transferring the supernatant to a new 1.5 mL tube.

4. *RNA precipitation*: add to the supernatant 0.5 vol 96 % EtOH and invert tube immediately. Load supernatant–ethanol mixture quickly onto RNA binding silica columns (Qiagen RNAeasy plant kit or similar kit; column max volume 0.75 mL). Spin loaded column at 10,000×*g* for 30 s. Leftover supernatant–ethanol mixture are loaded onto the same column, processing the entire sample. Follow the kit protocol for the subsequent washing and desalting steps (*see* **Note 4**). Furthermore, we recommend performing the on-column DNAse digest available for the Qiagen RNA extraction kit.
5. *RNA elution*: add 50 µL of RNAse-free water on column and incubate sample for 1 min at room temperature. Put column in fresh RNAse-free tube and spin it for 1 min at 10,000 × *g*. Store RNA on ice or at −80 °C until continuing with Subheading ***3.4***.

3.4 Quality Assessment of RNA

1. Determine the RNA concentration, the protein contamination and the carbohydrate and ethanol contamination using a spectrophotometer. The 260/230 nm absorbance represents carbohydrate or ethanol contaminations and the 260/280 nm represents pro- tein contamination. Clean RNA ranges above 1.8–2.0 in both ratios. Values of the 260/230 nm and 260/280 nm ratios below 1.6 indicate a contamination of the extraction and it is not recommended to use this material for library preparation and subsequent sequencing.
2. Test the integrity of the RNA using a 2100 Bioanalyzer (Agilent) **[10]**. The electropherograms with intact RNA should show clear distinct peaks for the 28S and 18S ribosomal RNA and almost no peaks in the smaller RNA size range. Intact RNA is also indicated by the RNA Integrity Number (RIN), which describes numerically the overall RNA quality beyond the standard method of the ratio of the two ribosomal peaks **[11]**.

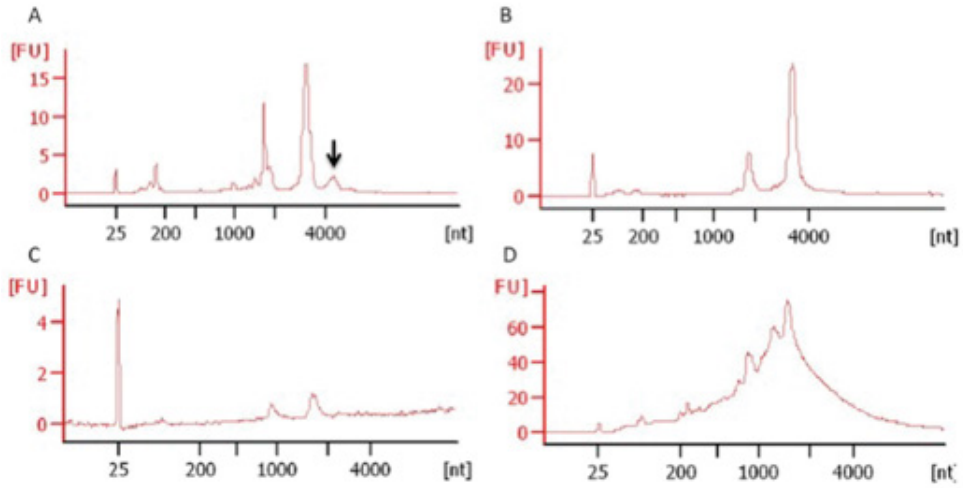


Figure 2. Bioanalyzer RNA 6000 Nanochip electropherograms of RNA samples with different qualities. *Upper panel* shows non-degraded high-quality RNA with characteristic 18 and 28s rRNA peaks before DNAseI treatment with gDNA peak indicated by *arrow* (**a**) and after the digest (**b**) with a RIN of 9. *Lower panel* shows two samples of degraded low concentrated RNA (**c**, **d**) with a RIN of below 3

For successful sequencing library preparation, it is recommended not to use RNA samples with a RIN below 8. On the RNA 6000 Nanochip 1–5 µL of RNA concentrations between 25 and 250 ng/µL can be loaded. A typical electropherogram is shown in **Figure 2**.

3. If DNAse treatment of the RNA samples is required dilute the RNA to 100 ng/µL in 20 µL volume and add 2 µL 10× reaction buffer and 1 µL of DNAse. Incubate reaction at 37 °C for 10 min. Inactivate DNAse by adding 2 µL RNAse-free (50 mM) EDTA and incubate tube at 75 °C for 5 min (*see* **Note 5**).

4. Test the DNA contamination by PCR on diluted RNA using RNA spiked with DNA as the control. DNA contamination is detected as an additional peak in the electropherogram (**Figure 2a**). A PCR using primers designed on genomic DNA on the RNA itself as the template will also detect DNA contamination when compared with a genomic DNA spiked control (*see* **Note 6**).
5. Retest the integrity of the RNA after DNAse digest using a 2100 Bioanalyzer chip.

3.5 Sequencing

Different sequencing technologies are currently available that differ in output, read length and price. The preparation protocols for libraries vary depending on the technology. The pipeline described below is customized for Illumina HiSeq 2000 sequencing data.

3.6 Read Analysis

The pipeline for read analysis is modular. Different modules can be replaced with alternative software if desired. The steps are: **(1)** read quality assessment, **(2)** read trimming and filtering to remove low quality reads, **(3)** read mapping, **(4)** parsing the read mapping to extract quantitative information, **(5)** statistical analysis and **(6)** extraction of biological information (**Figure 1**). If the experiment was conducted in a species without a sequenced genome, read quantification is prefaced with transcriptome assembly **[12, 13]**. For the pipeline described below six programs running on the command line are required (**Table 1**). The programs are well documented and described. Thus, only minimal bioinformatics expertise is required to successfully *run* the programs.

3.6.1 Read Quality Assessment

For read quality analysis run `FastQC` (http://www.bioinformatics. babraham.ac.uk/projects/fastqc/). The `FastQC` tool delivers a quality assessment of your data. Check read quality to estimate parameter settings for read cleaning:

1. Input: Read output for Illumina data; for each sample read files are divided in four Mio reads sub-files in the `FASTQ` format.
   This format also encodes the quality information for each read by the Q score, which is similar to the Phred score known from SANGER sequencing **[14, 15]**. The Phred quality scoring schemes estimates the error probability of the individual sequenced bases. So a Q score of 30 (Q30) assigned to a base is equivalent to one erroneous base call in 1,000 times.
2. Load data and run analysis via `FastQC` program with the following terminal command

```
fastqc -f format -o directory read_library. fastq/.fasta
```

   `FastQC` generates its output as HTML file containing all image files mentioned in the section below. For viewing the `FastQC` output open it in a Web browser.
3. Assess HTML file with regard to the listed attributes of the reads. **Figure 3** shows the different quality traits of sequencing data.

a) *Per sequence quality*: boxplot of base sequence quality generated by FastQC (**Figure 3a**). Usually for RNA-seq data the read quality drops toward the sequence end. Bases with a Q score below 20 should be trimmed by the `FASTX` toolkit.

b) *Quality score per sequence*: distribution of mean quality score of all bases per sequence, which should display one high peak at the *x*-axis end, meaning the average Q scores within the reads are equally distributed (**Figure 3b**). If you see multiple peaks, your sample has different pools of read quality. Instead of trimming, these reads can be subsequently filtered be the `FASTX` quality filter.

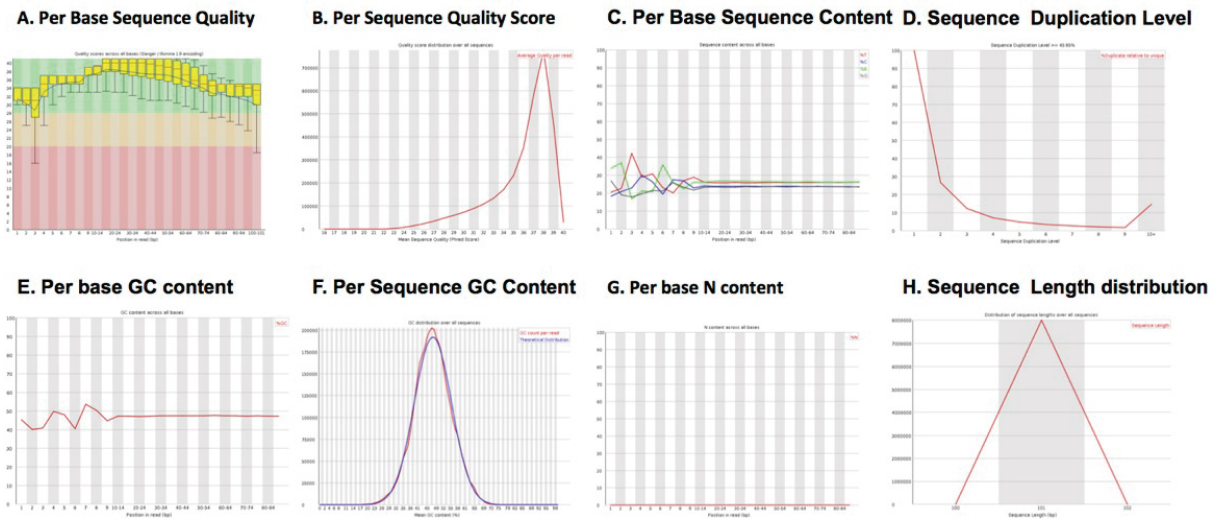


Figure 3. FastQC report of Illumina reads. Sequence characteristics used for read quality check.

c) *Per base sequence content*: plot of base sequence content. All bases should be equally present in a non-biased library. However, it is typical for RNA-seq data to have a base content bias in the first 1–9 bases due to the sequencing primers. A really biased library, which should be not used would show peaks for the individual bases that go up to 80 % (**Figure 3c**).

d) *Per base GC content*: similar to the base sequence content plot, though just plotting the GC content. This is biased as well at the sequencing start (1–10th bases), due to the primers used for sequencing (**Figure 3d**).

e) *Per sequence GC content*: distribution of GC content across all reads per sample. This graph should show a normal distribution. The blue curve is a theoretical normal distribution calculated from the mean and standard deviation of the loaded data, the red line plots the actual sample GC distribution. An indicator of sample contamination is a secondary peak in the sample GC distribution. These samples should be analyzed with caution and should be checked for contaminating sequenced DNA by a UNIREF blast (**Figure 3e**).

f) *Per base N content*: in this figure the uncalled bases per library are plotted. For high quality data the line should be flat. Peaks indicate *N* insertion in the sequence, which can be trimmed or filtered out by the artifact filter from the `FASTX` toolkit (**Figure 3f**).

g) *Length distribution*: plot of the library length distribution. This should be ranging depending on the number of sequencing cycles around 50 or 100 bp (**Figure 3g**).

h) *Sequence duplication level*: display of library uniqueness. `FastQC` usually warns the user of RNA-seq data, since sequences occur more than once in a non-normalized RNA-seq library. Over-sequencing of the highly expressed genes is needed to ensure capture of the lowly expressed ones (**Figure 3h**).

i) *Overrepresented sequences*: in this plot the library is analyzed for individual sequences that are overly represented, *e.g.*, adapter primer contamination. For an overrepresentation tag the sequences have to represent more than 0.1 % of the library (not shown).

j) *Kmer Plot/content*: this analysis helps the user to spot unusual enrichments of sequences, which are not aligned with the reads, *e.g.*, adapter sequences that start at variable points within the reads (not shown).

3.6.2 Read Cleaning

The FASTX-Toolkit contains a set of command line tools for read file FASTA/FASTQ preprocessing **[16]**. Satisfactory Illumina sequence runs will have less than 20 % of reads removed by this pipeline. To access all functionalities of each program type in the terminal *e.g.* (*see* **Note 7**):

```
fastx_clipper –h
```

This will print the help screen on the terminal, where all functions and flags are briefly explained.

1. Input: read files from Sequencing Center in .FASTQ format

2. OPTIONAL: If adaptor sequences are present, run the fastx_clipper to remove them from the reads.

```
fastx_clipper -v –Q33 -a ADAPTER –i <read_library.fastq> -o
<clipped_output_file>
```

The `–a` flag indicates the adapter used for this sample, `–i` stands for input (`.fastq` file) and `–o` for output. Use the `–v` (verbose) for more explanatory program output. If you have Illumina generated reads use `–Q33`.

Output: The Fastx clipper will deliver a `.fastq` read file as named by the user.

3. Run the `fastx_trimmer` to trim bases from the reads, which have an overall low PHRED score and thereby high error rate (**Figure 3a**).

```
fastx_trimmer -v -Q33 -f N -l N –i read_library.fastq –o
output_directory/trimmed_ read_library.fastq
```

The flags `–f` indicate the first base `–l` the last base to keep from your reads, the `–Q33` indicates that you have Illumina NextGen reads.

Output: `.fastq` file containing trimmed reads. Use this file for the next step.

4. Proceed with the `fastx_artifacts_filter` to filter out reads with more than three uncalled bases in a stretch (**Figure 3g**).

```
fastx_artifacts_filter -v -Q33
-i output_directory/trimmed_read_library.fastq
-o output_directory/filtered_read_library.fastq
```

The flags `–i` stands for input file and `–o` for output file. With Illumina data the `–Q33` needs to be added to the command.

Output: filtered read file in `.fastq` format, named as indicated by user after `–o` flag.

5. Run `fastq_quality_trimmer` to cut bases from the end of the reads depending on the quality threshold set by the user.

```
fastq_quality_trimmer -v –Q33 -t N -l N
-i output_directory/filtered_read_library.fastq
-o output_directory/filtered2_read_library.fastq
```

It removes also reads, which are shorter than a certain length or which have a lower overall base PHRED score than at 50 % of the bases. Specify with `–t` the desired PHRED quality threshold of your reads and with `–l` the minimum read length tolerated by the program. Reads that are shorter than this length will be discarded after trimming.

Output: quality trimmed read file in `.fastq` format, named as indicated by user after `–o` flag.

6. Concatenate cleaned output fastq files from **step 5** to one sample file for each library (*see* **Note 8**).

```
cat subreadfiles>output_sample_file
```

3.7 Read Mapping and Quantification

3.7.1 Read Mapping

Read mapping using TopHat requires the reference genome/transcriptome, an annotation file in `.gff/.gtf` format and the read files. TopHat was especially customized for the needs of short RNA-seq reads longer than 75 bp from the Illumina Genome Analyzer. It aligns the reads in two consecutive steps, using first the Bowtie "engine" to map reads that match continuously to the reference sequence followed by a second step which identifies gapped alignment **[17, 18]**. You can use

either paired-end or single reads for a TopHat run, mixing both read types is not supported yet.

`TopHat` delivers several output files. The mapping output file `accepted_hits.bam` is in the bam format. The splicing junctions reported by `TopHat` are stored in the `junctions.bed` file, which is a UCSC BED track (http://genome.ucsc.edu/FAQ/FAQformat.html) of the junctions with each of it consisting of two connected BED blocks. Each BED block is as long as the maximal overhang of reads spanning the junction. The number of alignments spanning the junctions is described by the score. TopHat also reports the insertions and deletion sites of the alignment. These are saved in the `insertions.bed` and `deletions.bed` UCSC BED tracks. The unmapped reads are stored in the `unmapped_reads.fa.gz` file.

1. Create Bowtie index from assembly or reference genome. This is usually a `.FASTA` file.

```
bowtie-build –f <fasta_file.fa> <output_file>
```

   The reference index comprises of several number files with the .ebwt file ending. Bowtie employs a technique derived from data-compression called Burrows-Wheeler transformation.

2. Run TopHat with genome `.gff/-gtf`, index and reads (default settings).

```
tophat -p N -G *.GTF/GFF
-o output_directory bowtie_index read_file.FASTA/FASTQ
```

   Per default TopHat allows up to two base mismatches and three indels per read alignment. In case you are working with a polymorphic or heterozygous population you can increase the allowed mismatched bases up to five per read by using the `–N` flag. The number of CPUs used by TopHat can be set by the `–p` flag. The annotation file provided for the genome is loaded by using the `–G` flag in your code. If the first run using a prebuilt index fails restart for another time before compiling the index yourself.

3. Use the terminal and bam_tools to read out the mapping statistics **[20]**

bamtools count -in accepted_hits.bam

   This command counts all mapped reads

```
bamtools filter -in accepted_hits.bam -tag "NH:1"|bamtools count
```

   This command counts all uniquely mapped reads

```
gunzip unmapped_left.*
grep -c '^@HWI' unmapped_left.*
```

   This command counts all unmapped reads

3.7.2 Read Quantification

The `Cufflinks` package contains four tools for RNA-seq analysis: **(1)** `Cufflinks` assembles and quantifies transcripts, **(2)** `Cuffcompare` compares transcript assemblies to annotation, **(3)** `Cuffmerge` merges two or more transcript assemblies into one, and **(4)** `Cuffdiff` detects differentially expressed genes and transcripts, including different isoforms and promoter usage **[21, 22]**.In order to calculate the expression level of each transcript, `Cufflinks` counts the reads mapping to each transcript and normalizes these to the length of the transcript. To make expression counts comparable from one sequencing run to the other the expression is normalized to millions of mapped reads. Both normalization steps are incorporated in `Cufflinks`. To quantify paired end reads, `Cufflinks` uses a probabilistic model resulting in FPKM output (expressed fragments per kilobase per million); single end reads are quantified as RPKM (reads per kilobase per million). `Cufflinks` assembles the reads for isoform detection, with the output stored in the following files `isoforms.fpkm_tracking skipped.gtf transcripts.gtf`. `Cuffdiff` is part of the `Cufflinks` package and it first calculates the expression of two or more samples and then tests

statistically whether gene expression changes across the samples are significant. Additionally, it is able to identify differentially spliced isoforms.

1. Quantify reads by:

```
cufflinks -p 2 –G *.GTF/GFF -o output_directory accepted_hits.bam
```

This step requires the `TopHat` output `accepted_hits.bam` files of the libraries and annotation file in the `.gtf/gff` format. The `–p` flag indicates the number of CPUs Cufflinks is allowed to use.

Output: The software output delivers for expression quantification fragments per kilobase of transcript per million mapped reads (FPKM or RPKM for single end reads) in the `genes.fpkm_tracking` file. Export this file tab delimited file to Excel. This is the output needed for Subheading ***3.9***.

2. Estimate significantly differentially expressed reads by

```
cuffdiff -o output_directory *.GTF/GFF
sample1_replicate1.bam[,...,sample1_replicateM.bam]
sample2_replicate1.bam[,...,sample2_replicateM.bam]
```

`Cuffdiff` requires the `.gtf/.gff` reference annotation file and `accepted_hits.bam` files from the `TopHat` mapping step. Biological replicates of each condition are separated by comma, and each different condition is separated by a white space. The samples can be loaded as `.bam` or `.sam` files.

3. Export output files to Excel or any preferred tabular calculation software, *e.g.* output file for differential expression analysis `gene_exp.diff`.

3.8 Extracting Biological Information

Only a brief introduction into a limited selection of possible analyses can be given due to space constraints. All software developed to analyze microarray-generated data can be adapted to use with RNA-seq experiments. The programs were chosen since they operate in graphical user interfaces familiar to biologists.

3.8.1 Excel

Even though programming languages are a more useful way to operate with large datasets, `Excel` is a versatile program, which is familiar to most biologists.

Converting File Formats

Excel is capable of opening files in which values are separated by a common delimiter; in most cases this will be a tabulator. Conversely, by choosing “Save as” `Excel` can save files using different delimiters as needed for downstream programs.

Transferring Information from Different Sources

Given two sources of information, for example **(1)** a file with expression values each headed by a gene identifier and **(2)** a list of genes known to be involved in circadian processes, these information can be combined in one file for easy access and interpretation. The combination of large files takes significant computing time (*see also* **Note 9**).

1. Open both files in `Excel`, combine into one file on different tabs.
2. Ensure that the gene identifiers have the same format in each table (e.g., capital and small letters, gene identifier given as the locus or the gene isoform); given a difference, alter all identifiers at once using functions such as “replace,” or “LEFT” or “RIGHT.”
3. Use the VLOOKUP function to match the information. This function requires **(1)** the common identifier between both tables, **(2)** the table from which to gather the new information with the common identifier in the first column, **(3)** the number of the column from which the

information is to be returned, and **(4)** FALSE to disable searching for similar but not identical matches.

4. After the function was applied to match the new information to each identifier, copy the columns and reinsert as values only to save computing power.

Using this simple function, the expression information can be augmented with descriptions from TAIR (or similar), with localization information, with expression information from previous experiments, essentially with any information that is given as a table with the same gene identifier format. Information from different sources, such as RPKM files and significance tests can be combined.

Testing Enrichment

For many experiments, it is desirable to assess whether a list of genes (*i.e.*, those that are differentially regulated in the experiment, or those that are members of one expression pattern cluster, see below;

now called *group A*) is enriched in genes from a second list (*i.e.*, a previous experiment, or a list of genes known to be involved in a trait from a publication, now called *group B*). Excel can be used to make this assessment although this analysis can be carried out faster using programming languages such as R **[23]**.

1. Input: a table that indicates membership in *group A* and *group B* for each gene.
2. Using the COUNTIFS function, count the number of genes that belong to **(1)** both *A* and *B*, **(2)** to *A*, and **(3)** to *B*.
3. Calculate a contingency table with the four fields in *A* and *B*, not in *A*/in *B*, in *A*/not in *B*, and not *in A*/not in *B*.
4. Calculate whether there is significant enrichment using a Fishers Exact Test or a chi square test (if the numbers are large) **[24]**.
5. If multiple tests are carried out at the same time, the alpha level of significance requires adjustment to multiple hypothesis testing.

The spectrum of possible corrections ranges from Bonferroni (very conservative) **[25]** to Benjamini and Hochberg False Discovery Rate (FDR, rather permissive) **[26]**.

3.8.2 Mapman

Mapman is a tool to visualize data on previously determined pathways **[27, 28]**.

1. Download the software from http://mapman.gabipd.org/web/guest/mapman; download the Mapman mapping for your species of interest from the MapMan store. The Mapman program requires three input files to run: **(1)** the mapping in which each gene is associated with a function, provided by the software or the Mapman store, **(2)** the pathway file in which the pathway image is associated with the genes in the pathway, provided by the software, and **(3)** the experiment file provided by the user.

2. Produce your input file by opening the file generated by the Cufflinks-pipeline. Make sure it has row identifiers identical to those in the Mapman mapping file, which can be inspected in the text version of the mapping file. Remove all information except the expression information. Consider whether you want to transform the expression information into ratios or normalize it for display.

3. Display a pathway of interest by clicking at the mapping file, the pathway file and the experiment file.

Mapman provides numerous pathways build into the software, which can all be inspected for each dataset. It is mostly a visualization tool. If given a subset of significantly changed genes, it can calculate enrichment.

3.8.3 MultiExperiment Viewer

MultiExperiment Viewer is a tool to analyze large datasets in both supervised and unsupervised methods **[29]**. The software is Java-based and freely available.

1. Download the software from http://www.tm4.org/mev/
2. Open the batch file and adjust the RAM available for the application to 2/3 of the RAM available on your computer. For example, if you run on 8 GB of RAM alter the parameter –Xmx to –Xmx6000 in the last line of the script. Save the alteration.
3. Run MeV.
4. Load the data by clicking the appropriate button.
5. Run a principle component analysis on the samples. The data reduction tool PCA will capture only the variation between samples and thus replicates will appear closer in a PCA compared to different samples. Based on the closeness of points in the PCA, similarity and difference between samples can be visualized.
6. Run a hierarchical clustering on the samples. Hierarchical clustering will sort similar samples close together on a tree based on similarity matrices. Euclidean and Pearson distance are most frequently used and reflect absolute differences (Euclidean) and trends (Pearson). Biological replicates are expected to cluster closer together than different samples.
7. Run a k-means clustering on the genes. For this clustering, a predetermined number of expected clusters have to be entered by the user. The program will fit the genes into clusters to achieve the most homogenous cluster appearance possible given the number of possible clusters available to the program.

MeV is a powerful software suite that can be used to analyze data beyond simple analyses of whether biological replicates are indeed similar or visualize the general trends in the data.

3.8.4 VirtualPlant

The VirtualPlant Platform http://virtualplant.bio.nyu.edu/cgi-bin/vpweb/ was designed to extract information from large datasets **[30]**.

1. Upload gene list(s) to VirtualPlant.
2. Click analyze, the menu will open and you can choose your type of analysis.
3. Use the functions “union,” “intersect,” and “symmetric difference” to identify membership and membership overlap.
4. Use the function Biomap to determine whether functions and/or categories are overrepresented in one of the lists.
5. VirtualPlant is connected to Cytoscape to visualize the data on known networks **[31]**.

To analyze data that includes more conditions that a comparison of two states clustering algorithms are required to identify the genes that have similar expression patterns **[32]**. Hierarchical clustering and k-means clustering are two of the early tools for this purpose. Recently, a range of additional algorithms has been developed **[33, 34]**. It is advisable to determine the tools needed for analysis of the data while planning the experiment to ensure that the requirements of the algorithms to be used are met.

4. Notes

1. Use RNAse-free plasticware. Wear gloves at all times when handling with RNA and all solutions and equipment for RNA work. Treat all working surfaces with RNAse inhibiting reagents such as RNAseExitus (AppliChem). Bake glassware and spatulas for RNA work at 180 °C for >2 h prior to use or use RNAseExitus solution (AppliChem).
2. Prepare all RNA extraction solutions and buffer in RNA-free water treated previously with 0.1 % (v/v) Diethylpyrocarbonate (DEPC). DEPC inactivates RNases. When adding DEPC to the solution, make sure to use a syringe, which is submerged in the solution. Before autoclaving, it is important to stir the solution on a magnetic stirrer for at least 2 h, without closing the container tightly. After autoclaving, DEPC is broken down to CO_2 and H_2O.
3. The acidic pH is critical factor to ensure the separation of RNA from DNA and proteins. Never use buffered phenol instead of water-saturated phenol.
4. After adding Wash Buffer 1 invert closed microtube with column several times to ensure that all remnants from prior steps are washed away.
5. During DNAse treatment do not exceed the incubation time of 30 min at 37 °C. Before

DNAse inactivation at elevated temperature make sure to add 5 mM EDTA (end concentration).

6. In case of persistent gDNA contamination of your RNA sample, do multiple DNAse treatments.
7. View basic help page by typing -h.
8. Sub-read files, which had a bad `FastQC` report are re-checked by `FastQC` again after the `FASTX` cleaning. If the sub read files now pass the `FastQC` pipeline, include them when concatenating the high quality cleaned sub read files to one sample.
9. `Excel` is a suboptimal tool to analyze large datasets. Scripts will do the same job much faster. However, given the fact that most biologists are not skilled programmers, `Excel` delivers the same results but will take up to minutes to compute. Once computation is finished it is advisable to copy the resulting columns and reinsert the information as values to prevent `Excel` from crashing.

5. References

1. Jiao YL, Tausta SL, Gandotra N et al (2009) A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. Nat Genet **41**:258–263

2. Li PH, Ponnala L, Gandotra N et al (2011) The developmental dynamics of the maize leaf transcriptome. Nat Genet **42**:1060–1067

3. Davidson RM, Gowda M, Moghe G et al (2012) Comparative transcriptomics of three poaceae species reveals patterns of gene expression evolution. Plant J 7**1**:492–502

4. Bräutigam A, Kajala K, Wullenweber J et al (2011) An mRNA blueprint for C_4 photosynthesis derived from comparative transcriptomics of closely related C_3 and C_4 species. Plant Physiol **155**:142–156

5. Gowik U, Bräutigam A, Weber KL et al (2011) Evolution of C_4 photosynthesis in the genus Flaveria: How many and which genes does it take to make C_4? Plant Cell **23**:2087–2105

6. Pick TR, Bräutigam A, Schlüter U et al (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

7. Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate phenol chloroform extraction. Anal Biochem **162**:156–159

8. Chomczynski P, Sacchi N (2006) The single- step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. Nat Protoc **1**: 581–585

9. Sangha JS, Gu K, Kaur J et al (2010) An improved method for RNA isolation and cDNA library construction from immature seeds of *Jatropha Curcas l.* BMC Res Notes **3**:126

10. Müller O, Hahnenberger K, Dittmann M et al (2000) A microfluidic system for high-speed reproducible DNA sizing and quantitation. Electrophoresis **21**:128–134

11. Schröder A, Mueller O, Stocker S et al (2006) The RIN: an RNA Integrity Number for assigning integrity values to RNA measurements. BMC Mol Biol **7**:3

12. Bräutigam A, Gowik U (2010) What can Next-Generation Sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol **12**:831–841

13. Schliesky S, Gowik U, Weber APM et al (2012) RNA-seq assembly - Are we there yet? Front Plant Sci **3**:220

14. Ewing B, Green P (1998) Base-calling of automated sequencer traces using PHRED. II.

Error probabilities. Genome Res **8**:186–194

15. Ewing B, Hillier L, Wendl MC et al (1998) Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. Genome Res **8**:175–185

16. Blankenberg D, Gordon A, Von Kuster G et al (2010) Manipulation of FASTQ data with galaxy. Bioinformatics **26**:1783–1785

17. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10**:R25

18. Trapnell C, Pachter L, Salzberg SL (2009) Tophat: discovering splice junctions with RNA-seq. Bioinformatics **25**:1105–1111

19. Ferragina P, Manzini G (2001) An experimental study of a compressed index. Inf Sci **135**: 13–28

20. Li H, Handsaker B, Wysoker A et al (2009) The sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078–2079

21. Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **28**:511–515

22. Trapnell C, Roberts A, Goff L et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with Tophat and Cufflinks. Nat Protoc **7**:562–578

23. Gentleman RC, Carey VJ, Bates DM et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5**:R80

24. Fisher RA (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. J R Stat Soc **85**:87–94

25. Dunn OJ (1961) Multiple comparisons among means. J Am Stat Assoc **56**:52–64

26. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate - A practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol **57**:289–300

27. Thimm O, Blasing O, Gibon Y et al (2004) Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J **37**:914–939

28. Usadel B, Nagel A, Thimm O et al (2005) Extension of the visualization tool Mapman to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. Plant Physiol **138**: 1195–1204

29. Howe E, Holton K, Nair S et al (2010) Mev: Multiexperiment viewer

30. Katari MS, Nowicki SD, Aceituno FF et al (2010) Virtualplant: a Software Platform to support Systems Biology Research. Plant Physiol **152**:500–515

31. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res **13**:2498–2504

33. Langfelder P, Horvath S (2012) Fast R functions for robust correlations and hierarchical clustering. J Stat Softw **46**:1–17

33. Allen JD, Xie Y, Chen M et al (2012) Comparing statistical methods for constructing large scale gene networks. PLoS One **7**:e29348

34. Jay JJ, Eblen JD, Zhang Y et al (2012) A systematic comparison of genome-scale clustering algorithms. BMC Bioinformatics **13**:7

V.2 FIRST-AUTHORED MANUSCRIPTS

Manuscript 2

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*

Published in Plant Cell (2014), Vol. 26, No.8, pp. 3243-60, doi: 10.1105/tpc.114.123752.

Impact Factor: 10.65

**First Author*

Main findings:

Comparative transcriptome atlases between two closely related Cleomaceae species, *G. gynandra* and *T. hassleriana,* featuring 18 tissues (with 3 developmental gradients) shed light on the global transcriptional architecture of C_4 plants and provides an exhaustive resource for research on C_4 photosynthesis in dicotyledons. The transcriptomes of the different tissues between the C_4 and C_3 species display a similar tissue specific signature. Analysis of the C_4 cycle gene expression patterns revealed, that the C_4 cycle genes were recruited to C_4 photosynthesis from different expression domains in the C_3 species, ranging from general house-keeping patterns to specific heterotrophic tissues. A structure related expression module recruited from the C_3 root to the C_4 leaf was isolated and comparison of gene expression dynamics between leaf development of C_3 and C_4 leaf provided details of the genetic features of Kranz anatomy. Mesophyll differentiation is delayed in the C_4 leaf facilitating extended vein formation. The enlarged bundle sheath cells typical for Kranz anatomy were linked to developmental factors, endoreduplication and larger bundle sheath nuclei.

Contributions:

- Defining the leaf gradient stages and preliminary experiments
- Plant Cultivation and sampling of all plant material
- RNA extraction and library preparation
- Expression quantification
- Bioinformatic data analysis of transcriptome
- Data analysis
- Microscopic analyses
- Semi-thin leaf cross-sections
- Enzyme assays
- Flow cytometry of leaf gradient nuclei
- Writing and editing of Manuscript

This article is a *Plant Cell* Advance Online Publication. The date of its first appearance online is the official date of publication. The article has been edited and the authors have corrected proofs, but minor changes could be made before the final version is published. Posting this version online reduces the time to publication by several weeks.

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 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

[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

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

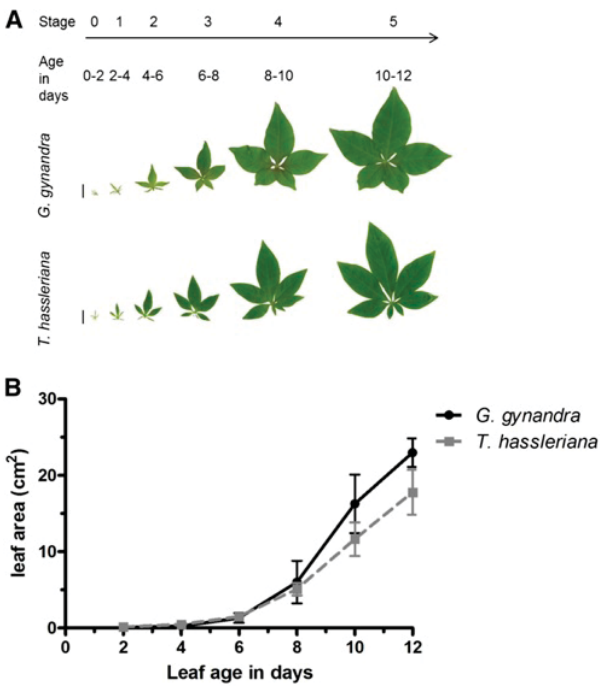


Figure 1. Overview of Leaf Shape and Expansion Rate in *G. gynandra* and *T. hassleriana*.

(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.

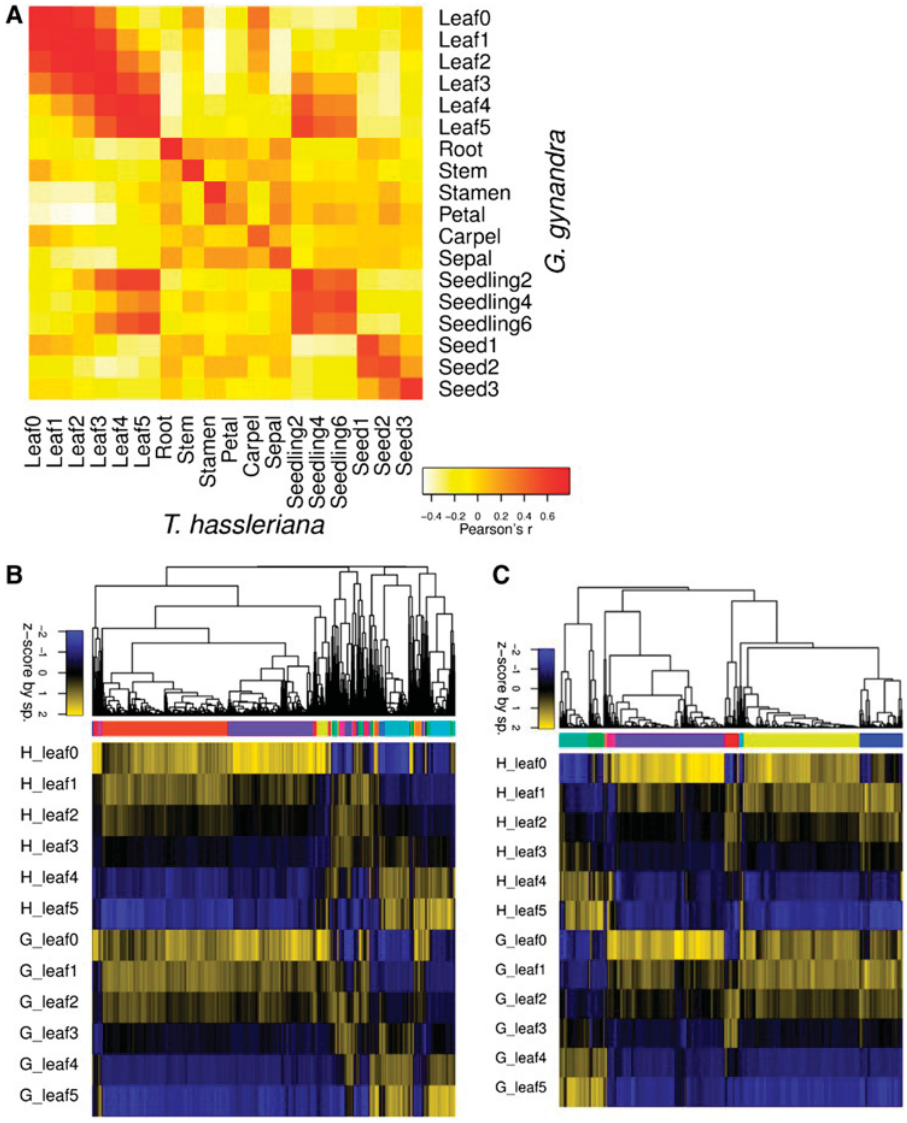


Figure 2. Comparative Tissue Dynamics and Gene Expression Pattern between *G. gynandra* and *T. hassleriana*.

(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

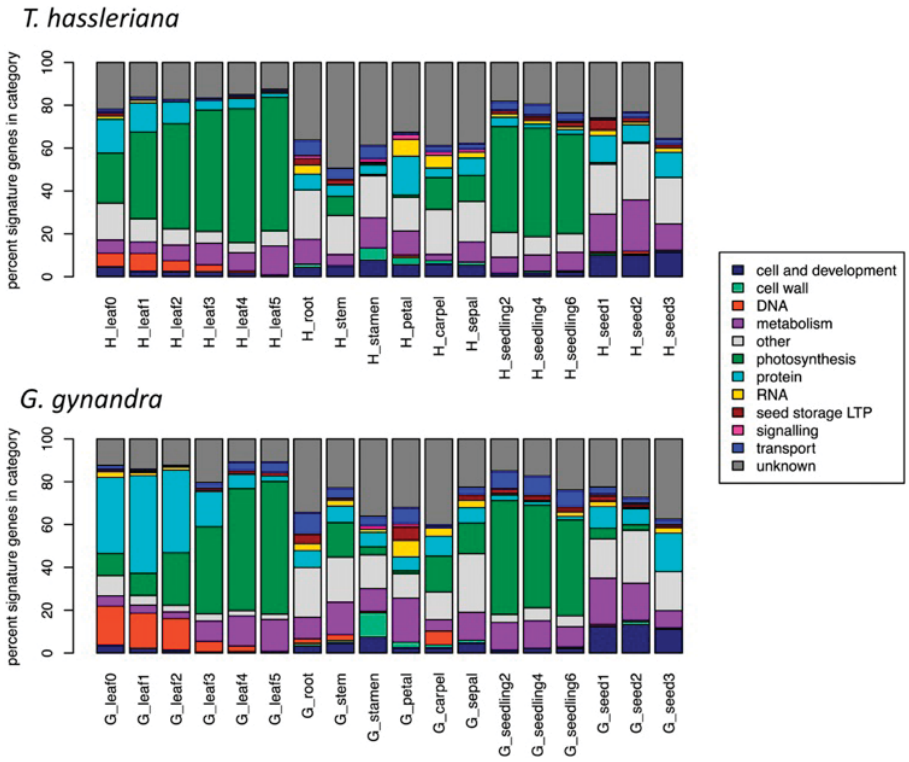


Figure 3. Distribution of Signature Genes in Each Tissue in *G. gynandra* and *T. hassleriana*.

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).

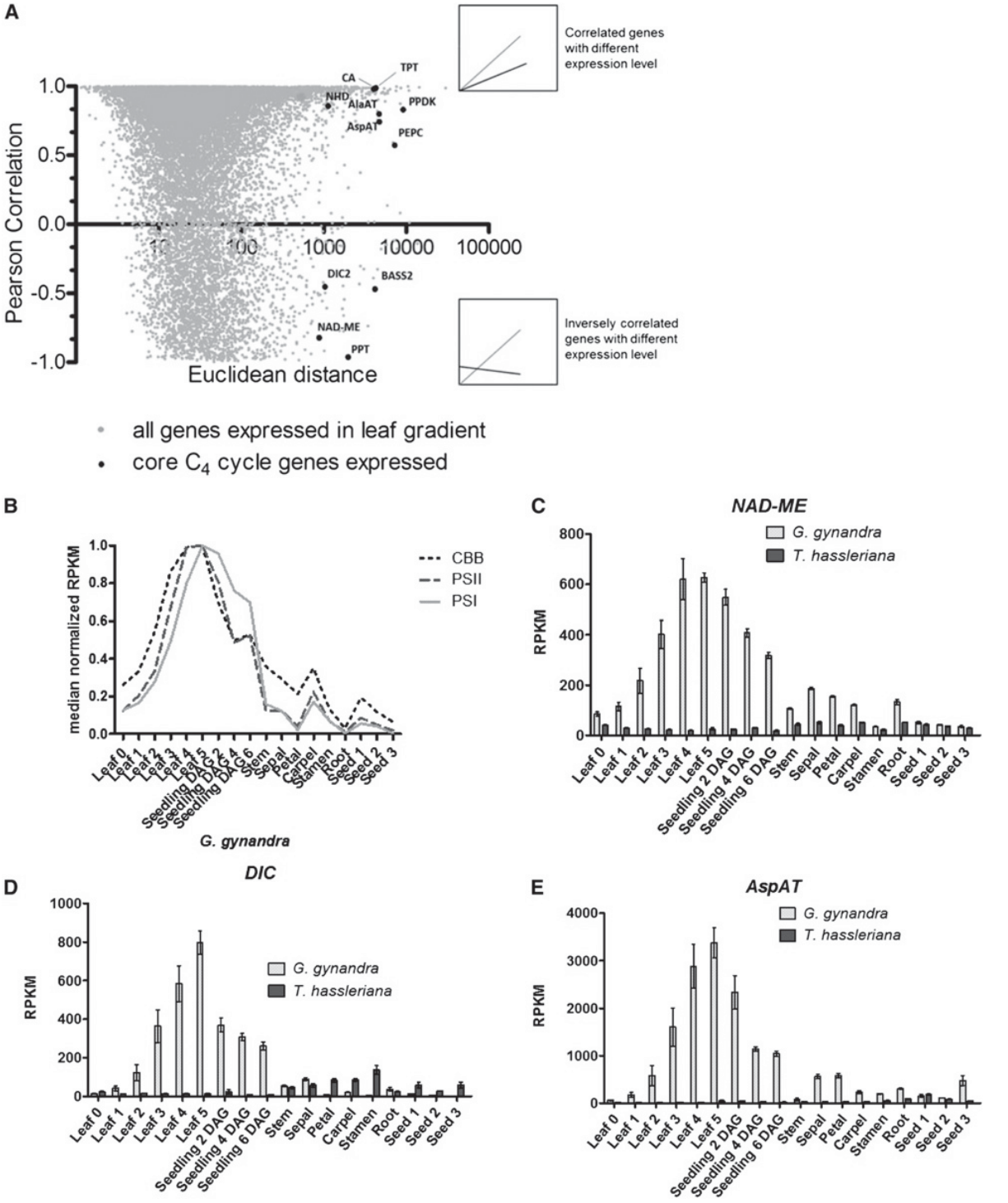


Figure 4. Comparison of Gene Expression Dynamics within the Leaf Gradient of Both Species.

(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

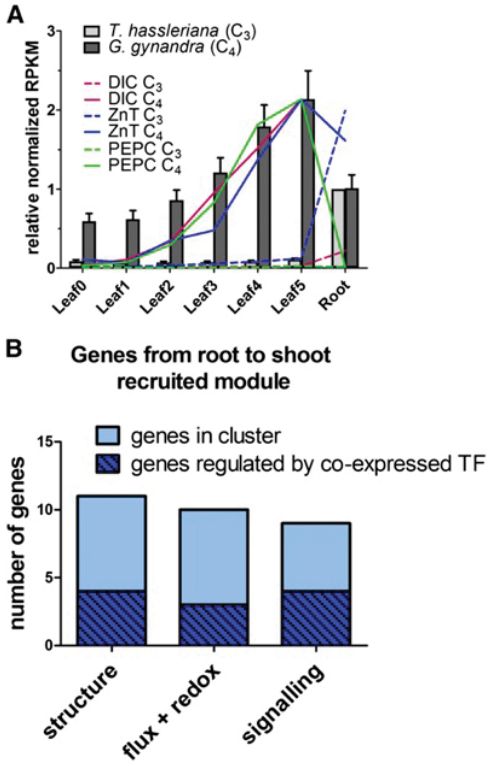


Figure 5. Recruitment of Genes from the Root to Leaf Expression Domain in the C_4 Plant *G. gynandra*.

(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, $n = 3$.

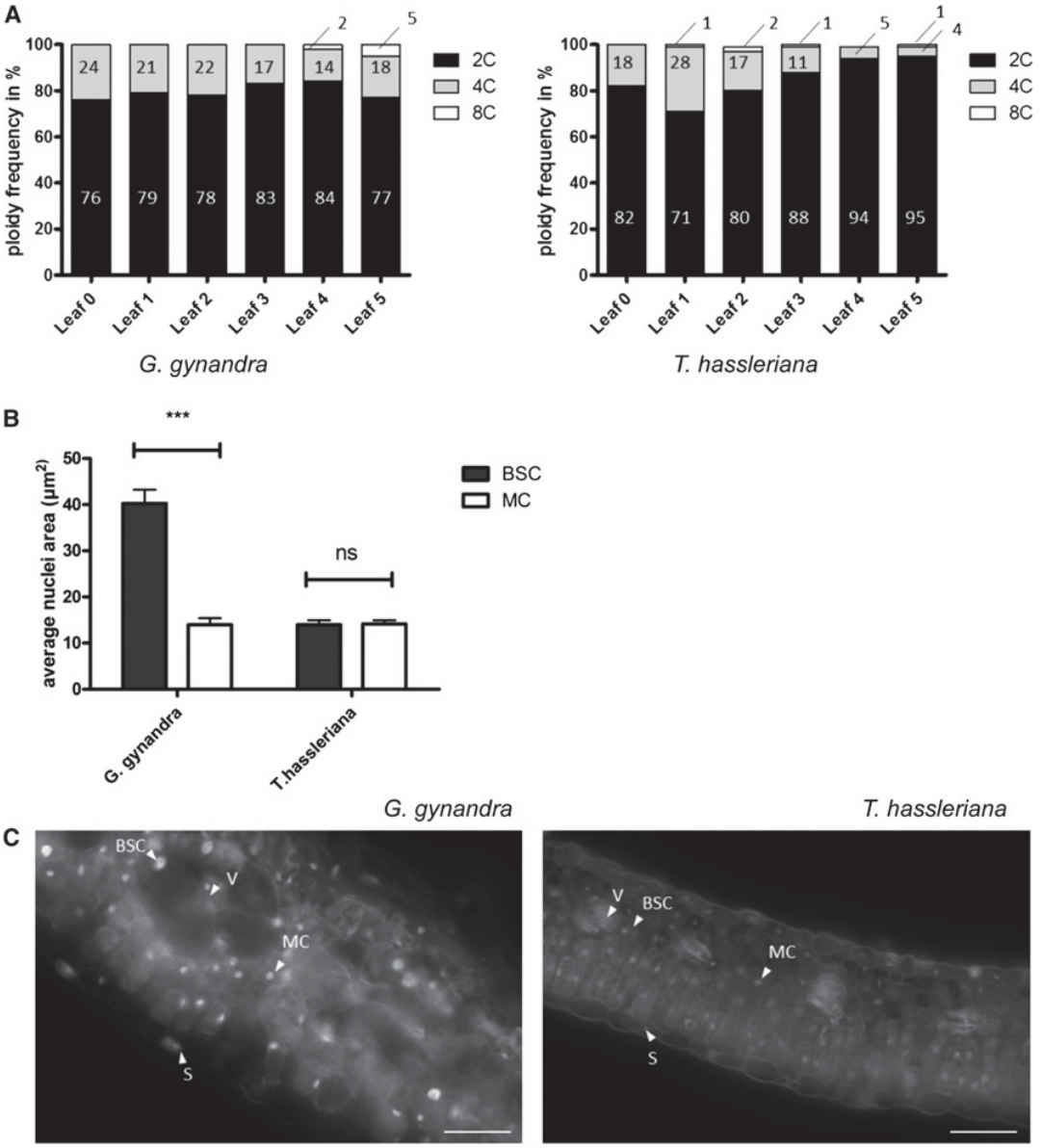


Figure 6. Distribution of Ploidy Levels during Leaf Development and Nuclei Area of BSC and MC between *G. gynandra* and *T. hassleriana*.

(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

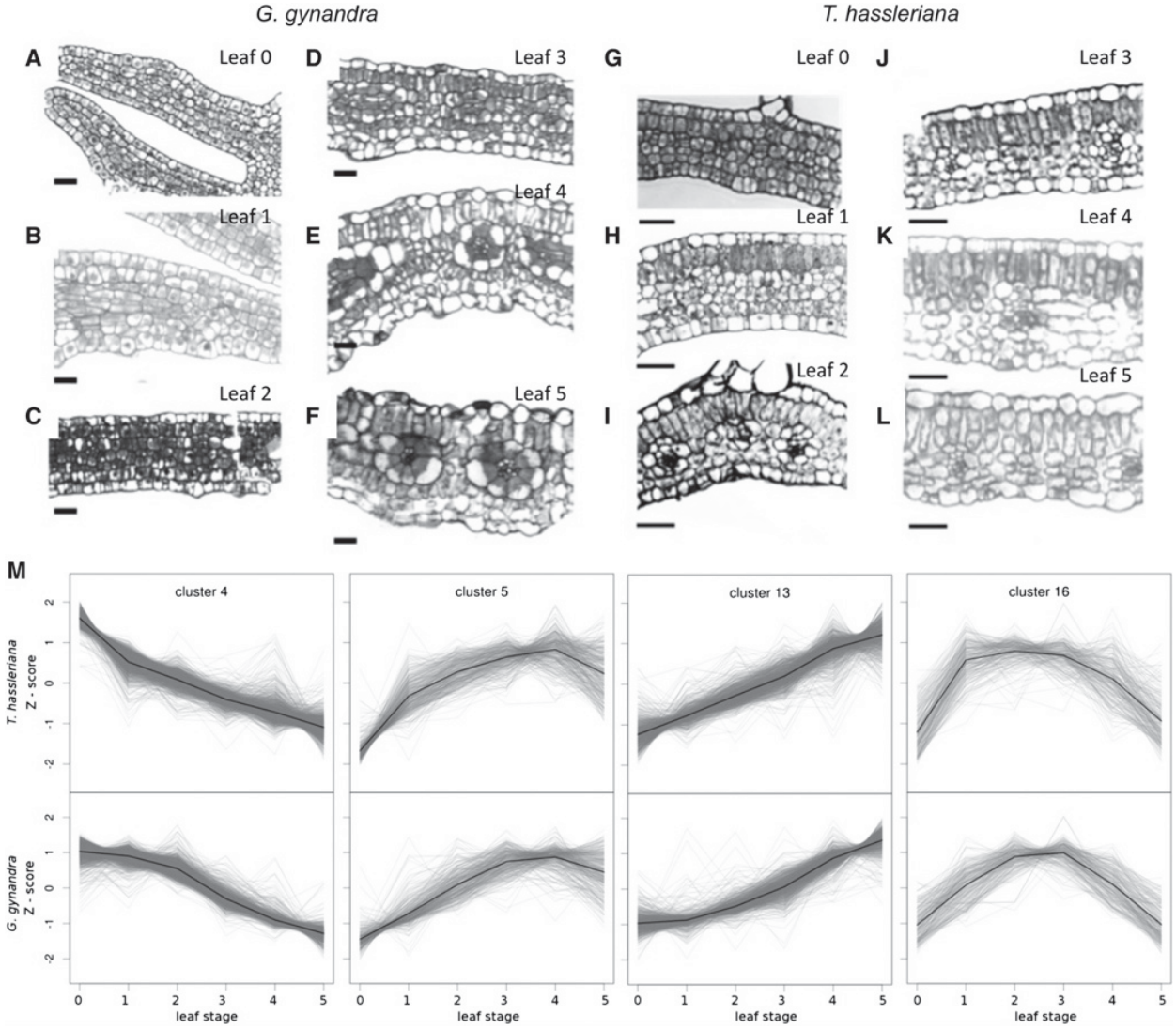


Figure 7. Analysis of Shifted Gene Expression Pattern and Leaf Anatomy during Leaf Ontogeny.

(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:** 140–50.

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.

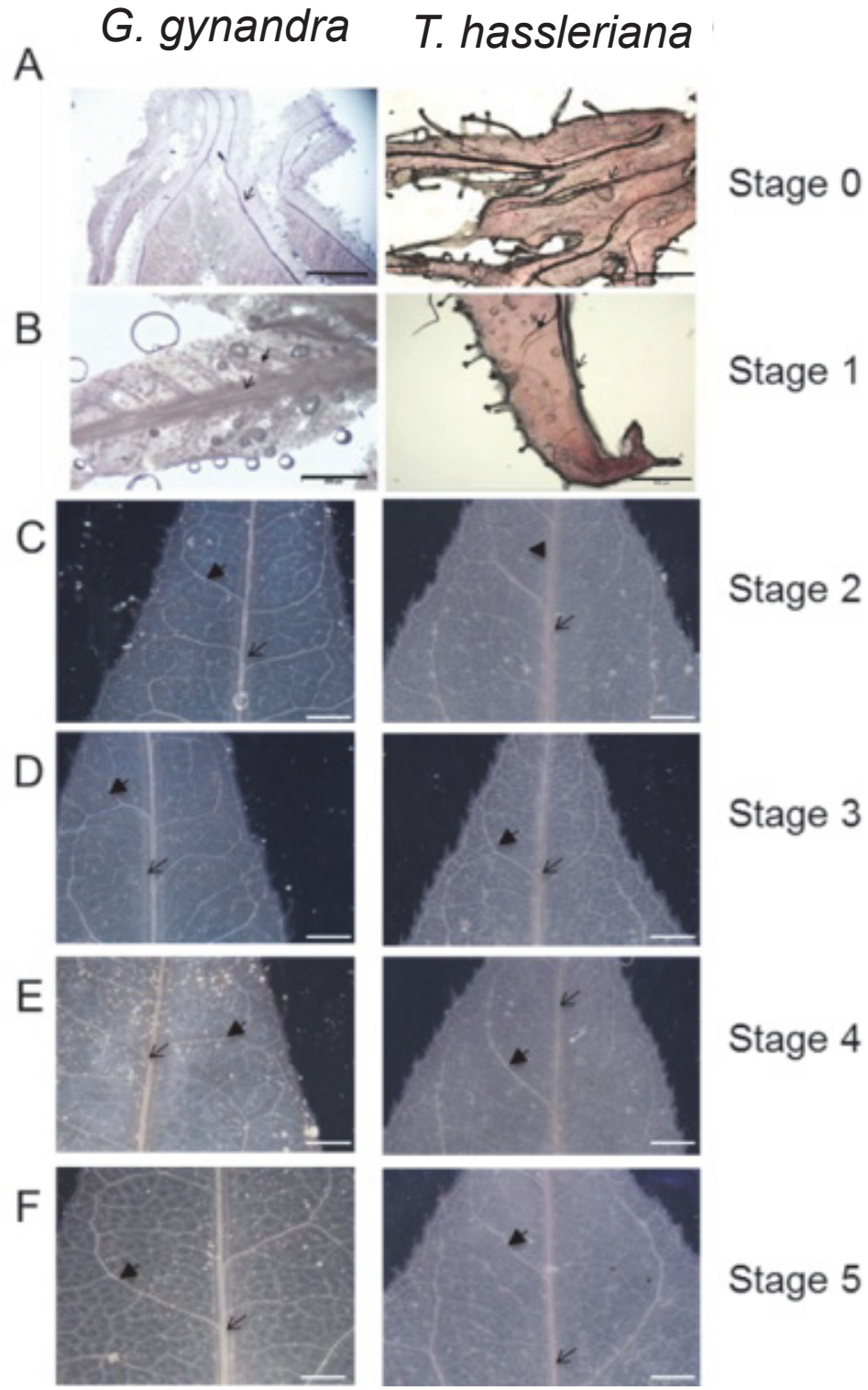
Xu, W., Purugganan, M.M., Polisensky, D.H., Antosiewicz, D.M., Fry, S.C., and Braam, J. (1995). Arabidopsis TCH4, regulated by hormones and the environment, encodes a xyloglucan endotransglycosylase. Plant Cell **7:** 1555–1567.

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.

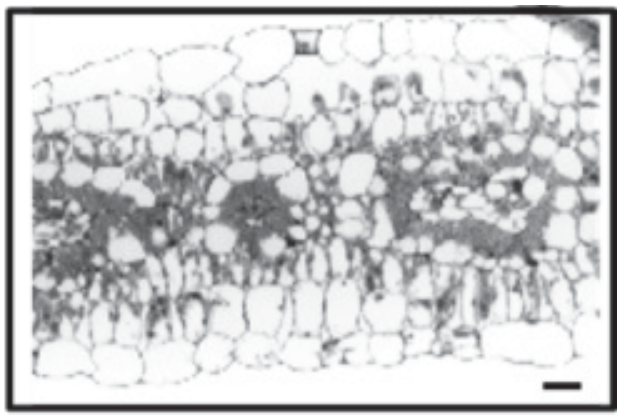
Zhiponova, M.K., et al. (2013). Brassinosteroid production and signaling differentially control cell division and expansion in the leaf. New Phytol. **197:** 490–502.

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.

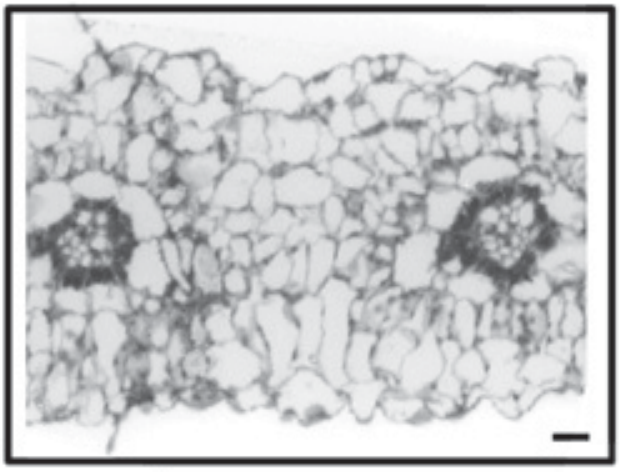


Supplemental Figure 1. Venation patterning during leaf development of *G. gynandra* and *T. hassleriana*.
(A-B) Cleared safranine stained leaves of stage 0 and 1 (n=3; scale bar 0.5 mm)
(C-F) Cleared leaves of stage 2, 3, 4 and 5 respectively (n=3; scale bar 1 mm)
Open arrows indicate the midvein (1°)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

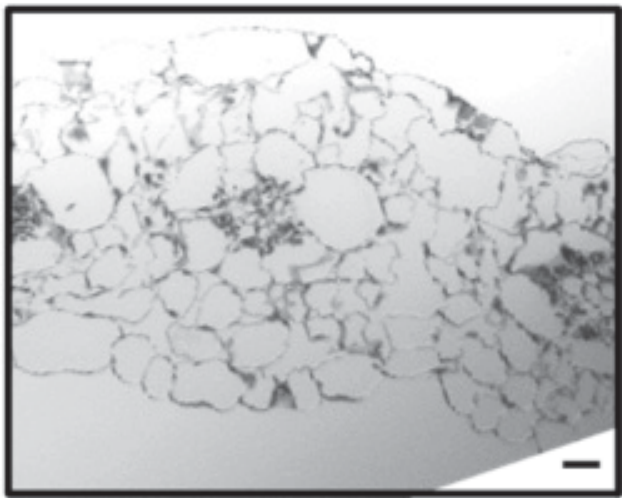
A



B



C



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)

A

G. gynandra *T. hassleriana*

B

G. gynandra *T. hassleriana*

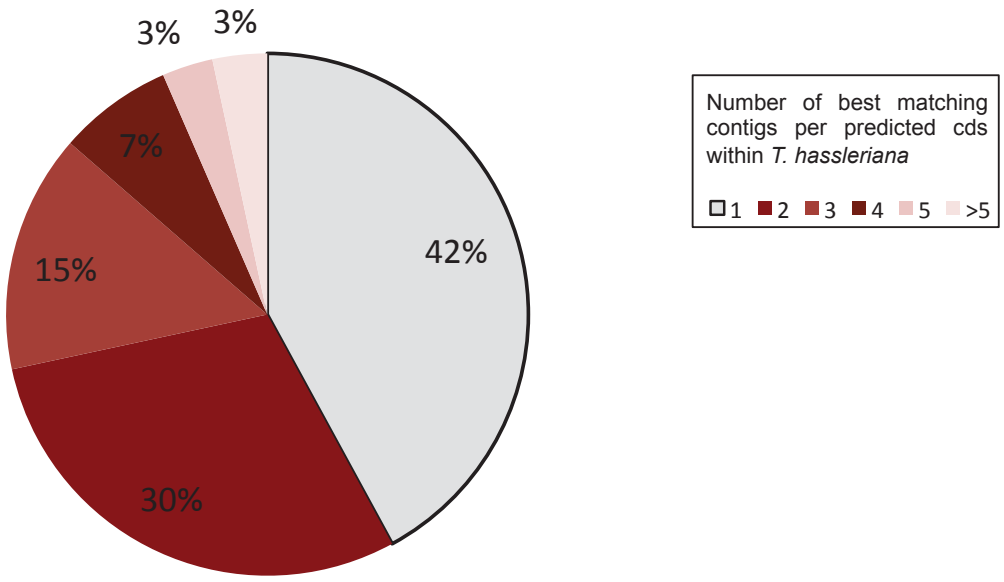
1

2

3

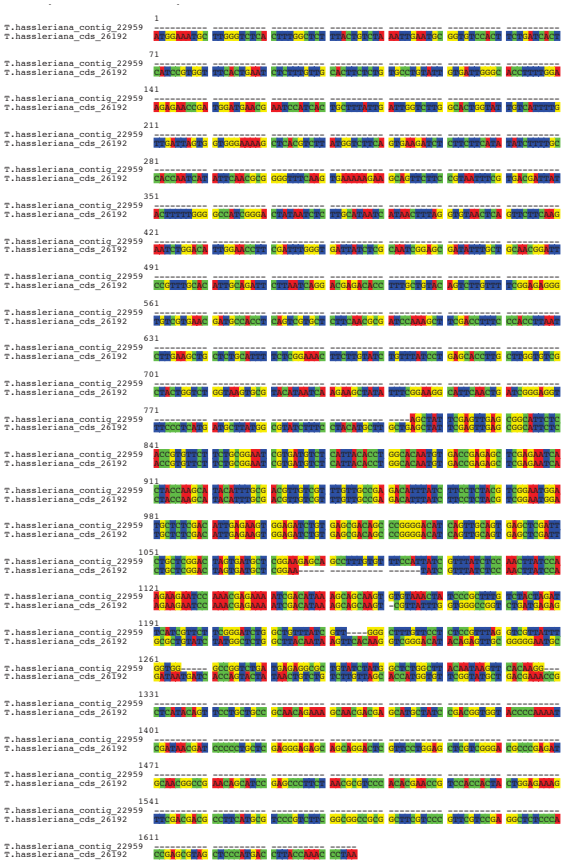
Supplemental Figure 3. Images of tissues harvested for RNA-seq in *G. gynandra* and *T. hassleriana*. (A) Photographic image of *G. gynandra* and *T. hassleriana* 8-week old plants, from which leaf gradient, stem and root system were harvested **(B) Seed coat development from harvested developmental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

A



B

Fragmented contig aligned to *T. hassleriana* cds

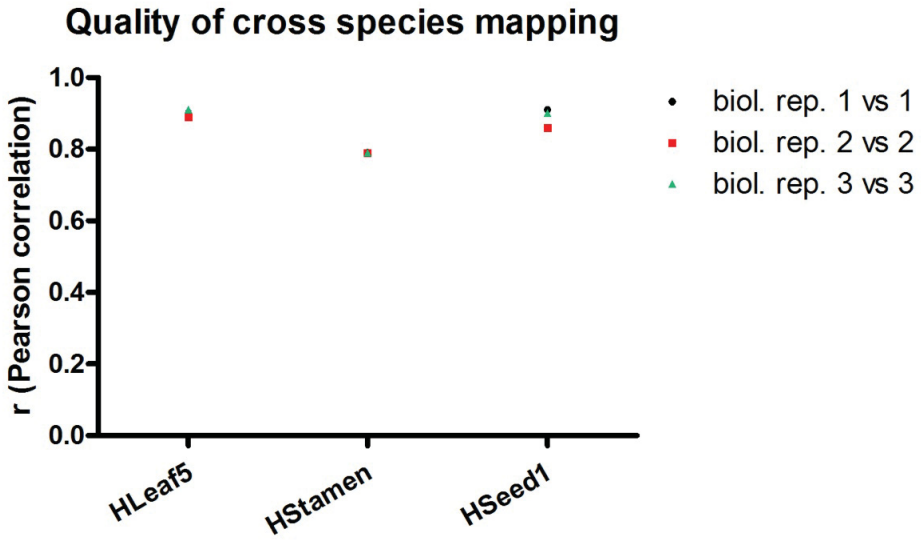


C

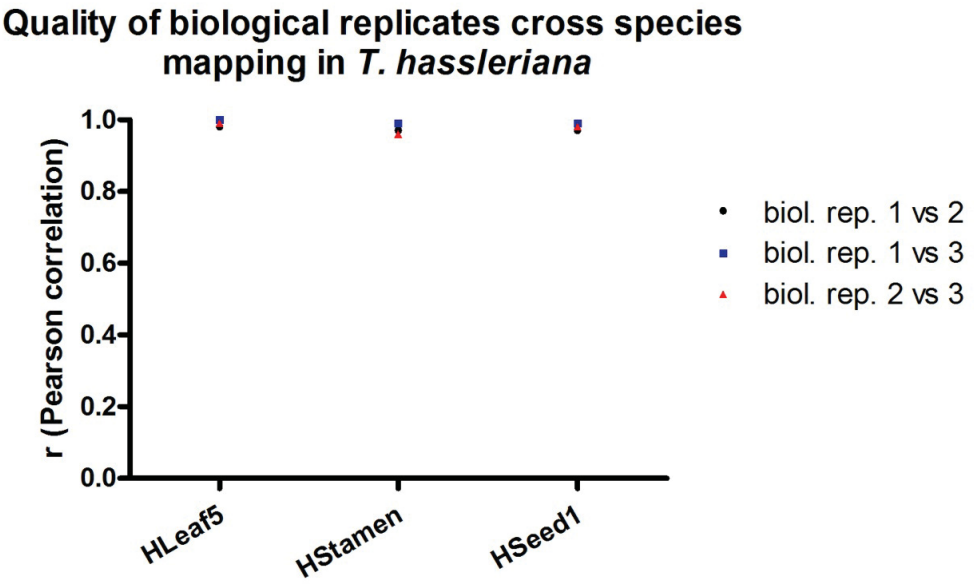
Fused/Hybrid contig aligned to *T. hassleriana* cds

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).

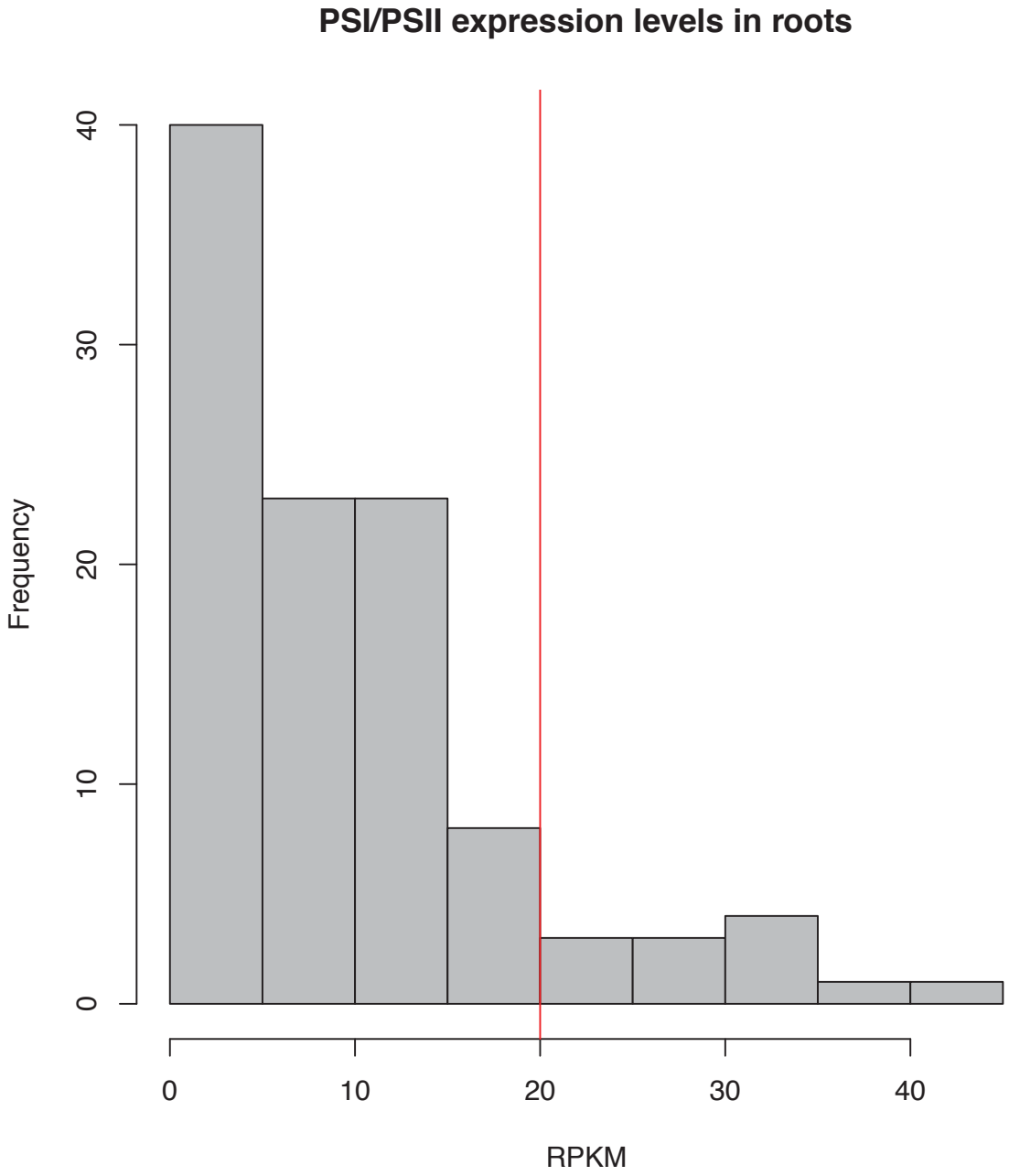
A



B



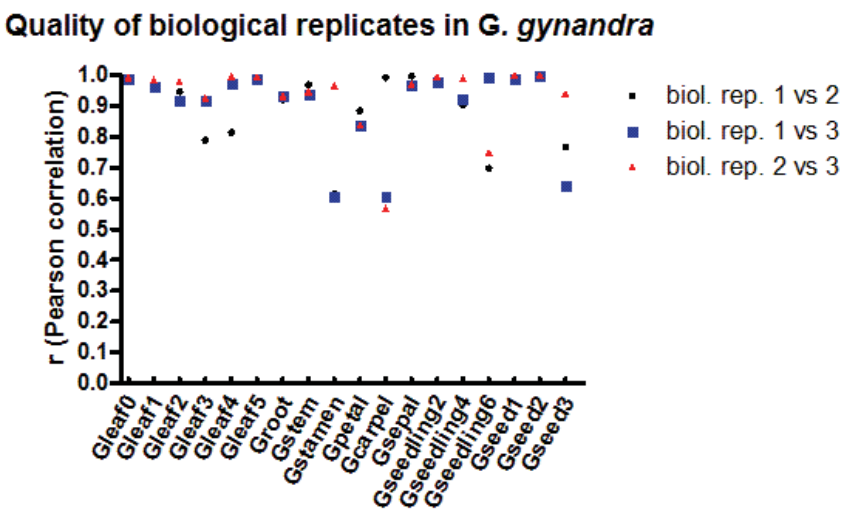
Supplemental Figure 5. Quality assessment of the biological replicates of *T. hassleriana* libraries mapped to *A. thaliana* and mapping similarity of *T. hassleriana* libraries mapped to *A. thaliana* and to its own cds.
(A) Pair-wise Pearson's correlation (*r*) was calculated for all three pairs of 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*.



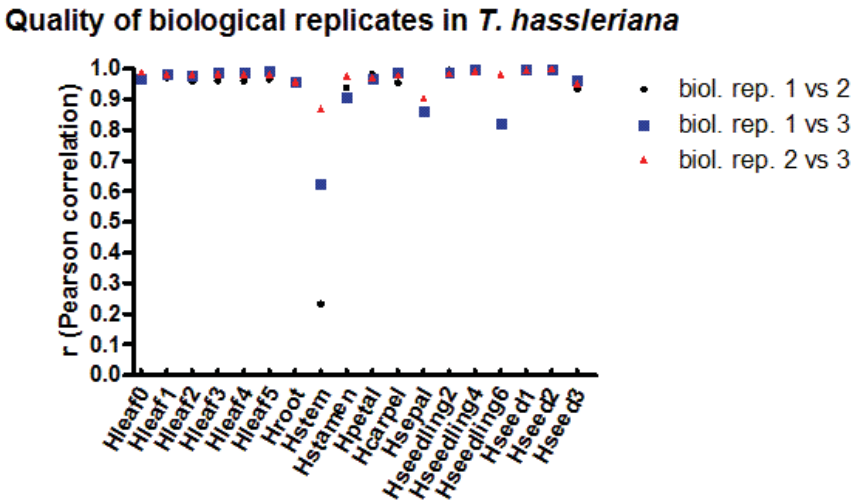
Supplemental Figure 6. Determination of base line gene expression via a histogram of photosystem (PS) I and II transcript abundances reads per mappable million (RPKM) in the *G. gynandra* root.

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.

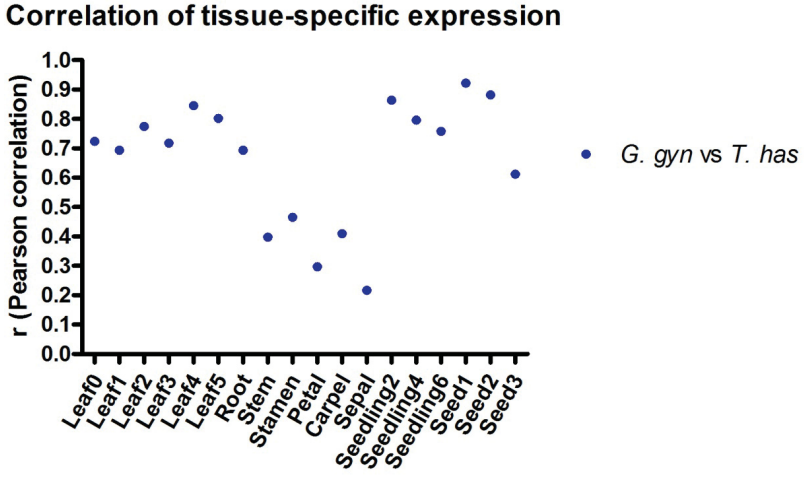
A



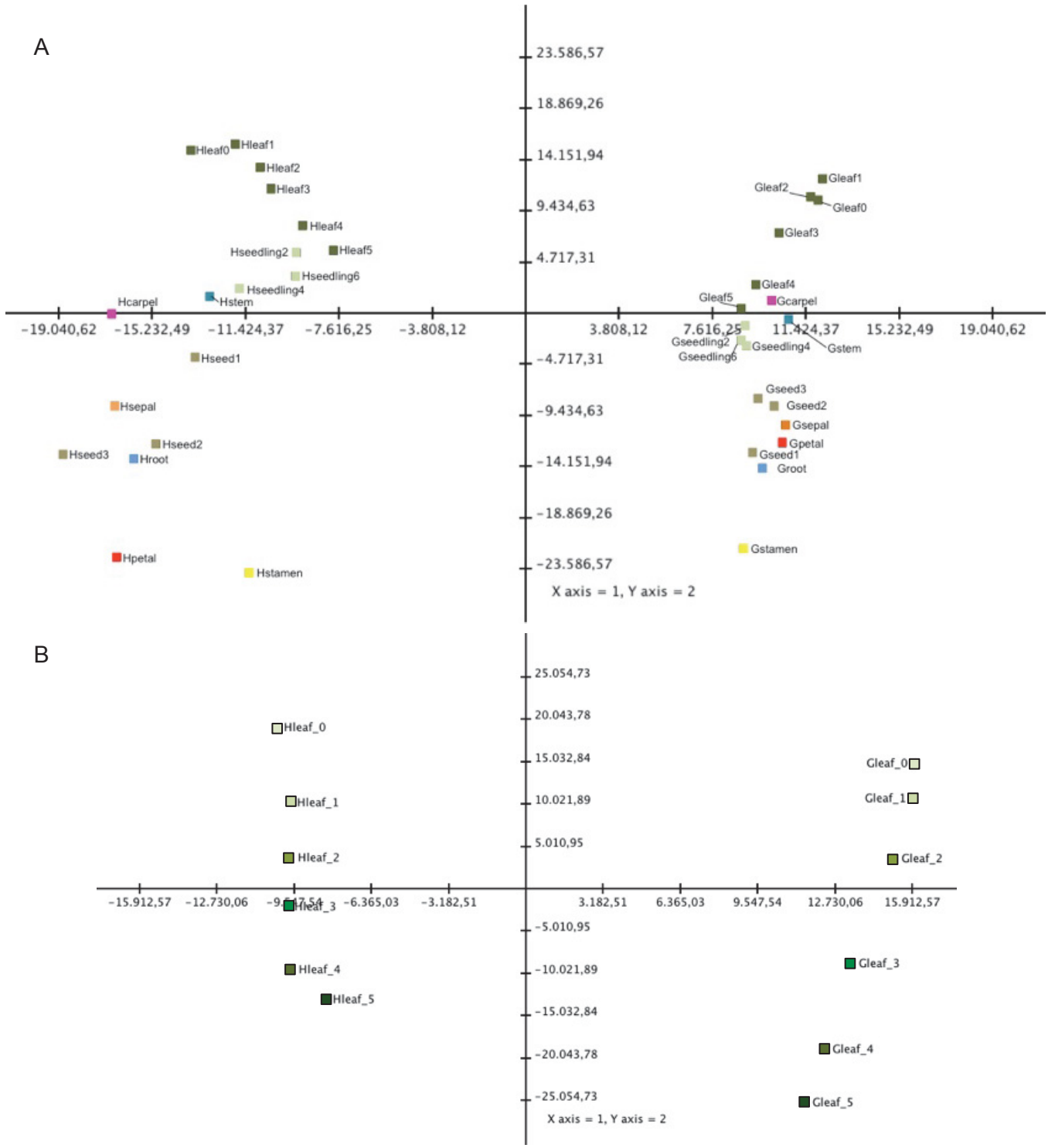
B



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*.

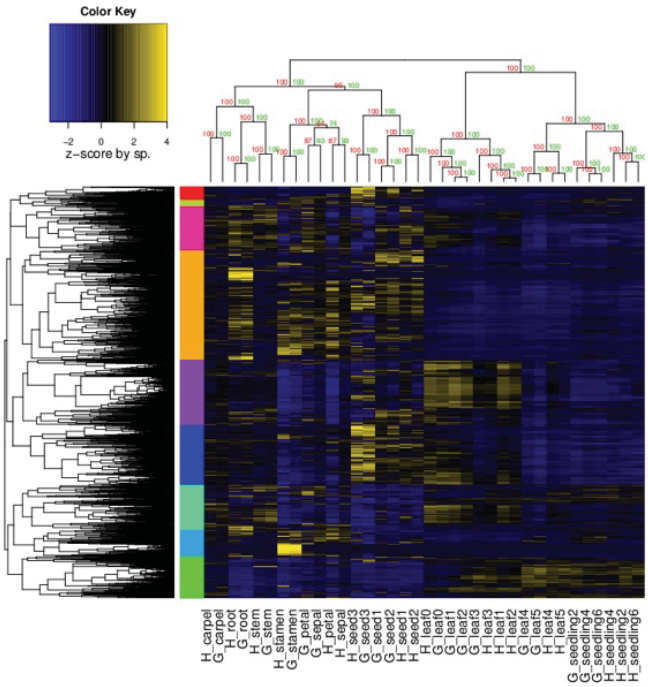


Supplemental Figure 8. Principle component analysis between *G. gynandra* and *T. hassleriana.*

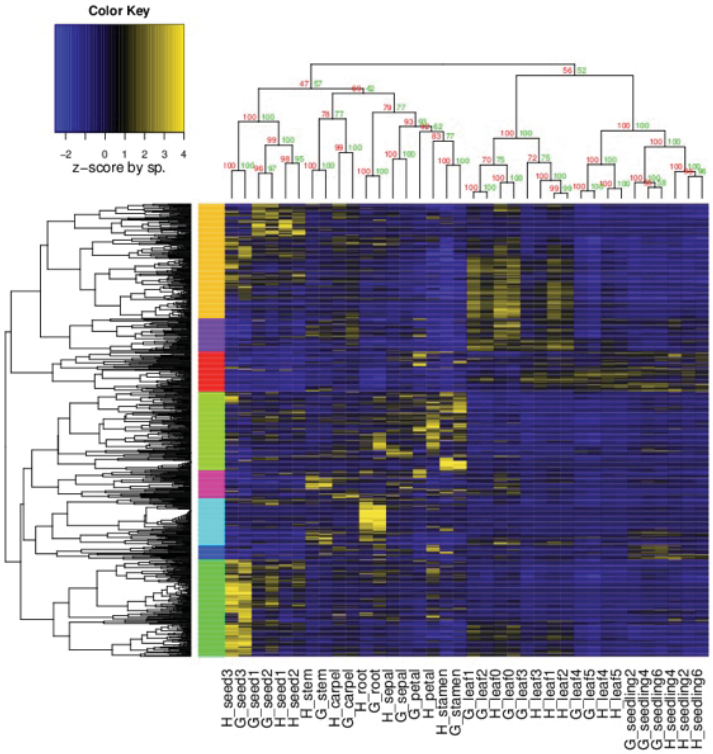
(A) Plot shows all averaged tissues from *G. gynandra* (G) and *T. hassleriana* (H) sequenced (n=3). The first component describes 15% of all data variablility seperating both species. The second component (14%) separates samples by tissue identity within each species. Tissues are indicated by color 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.

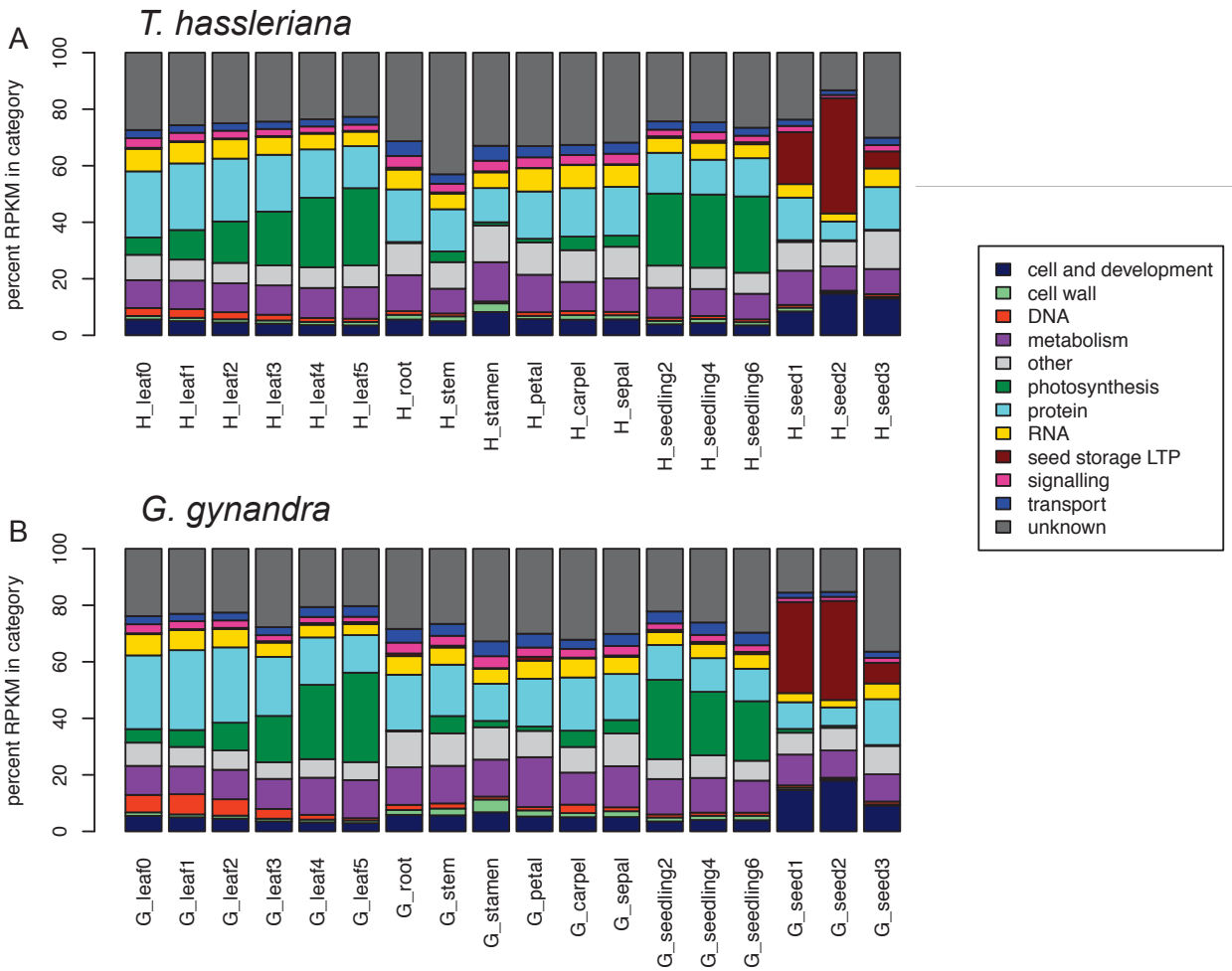
A



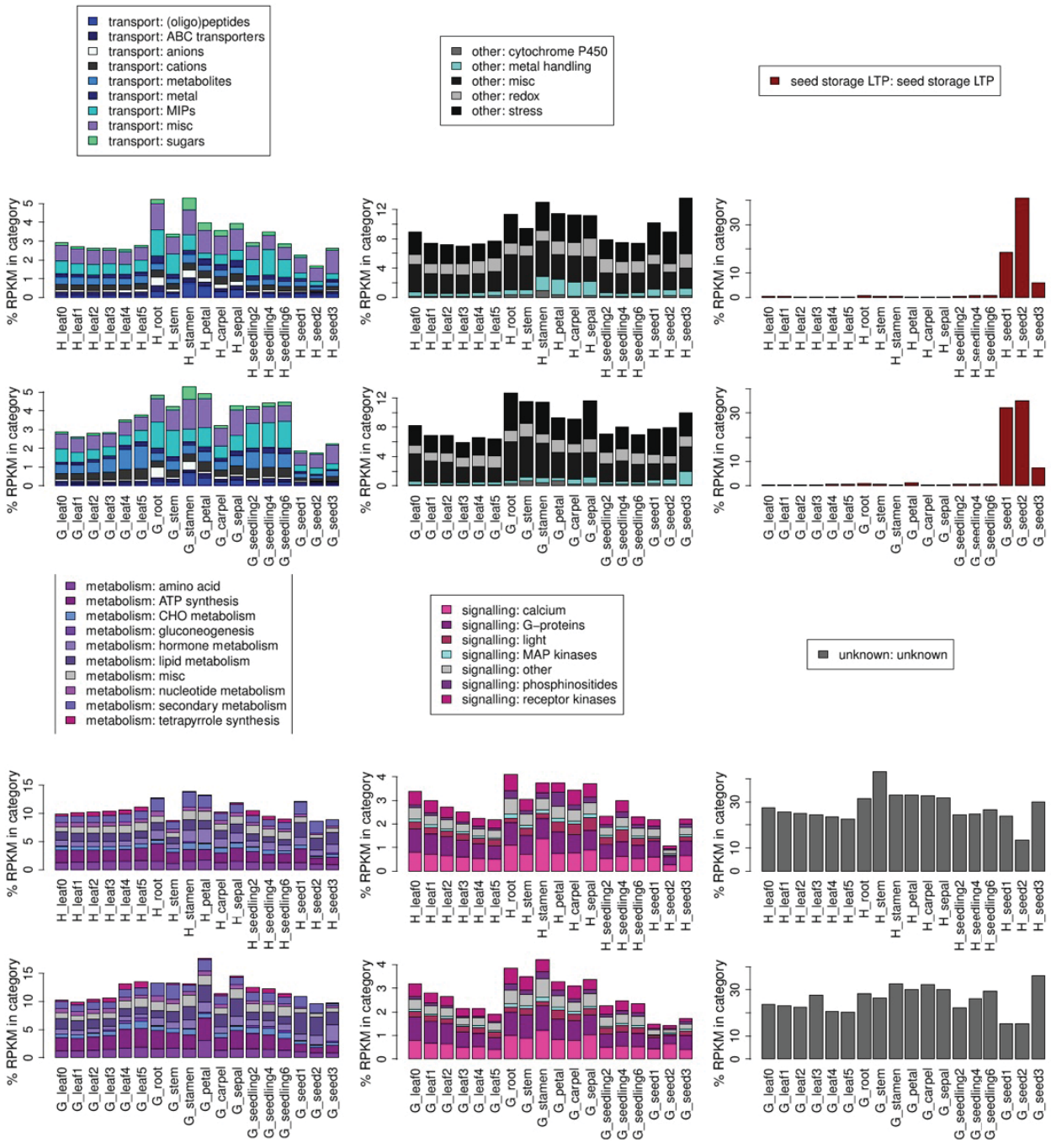
B



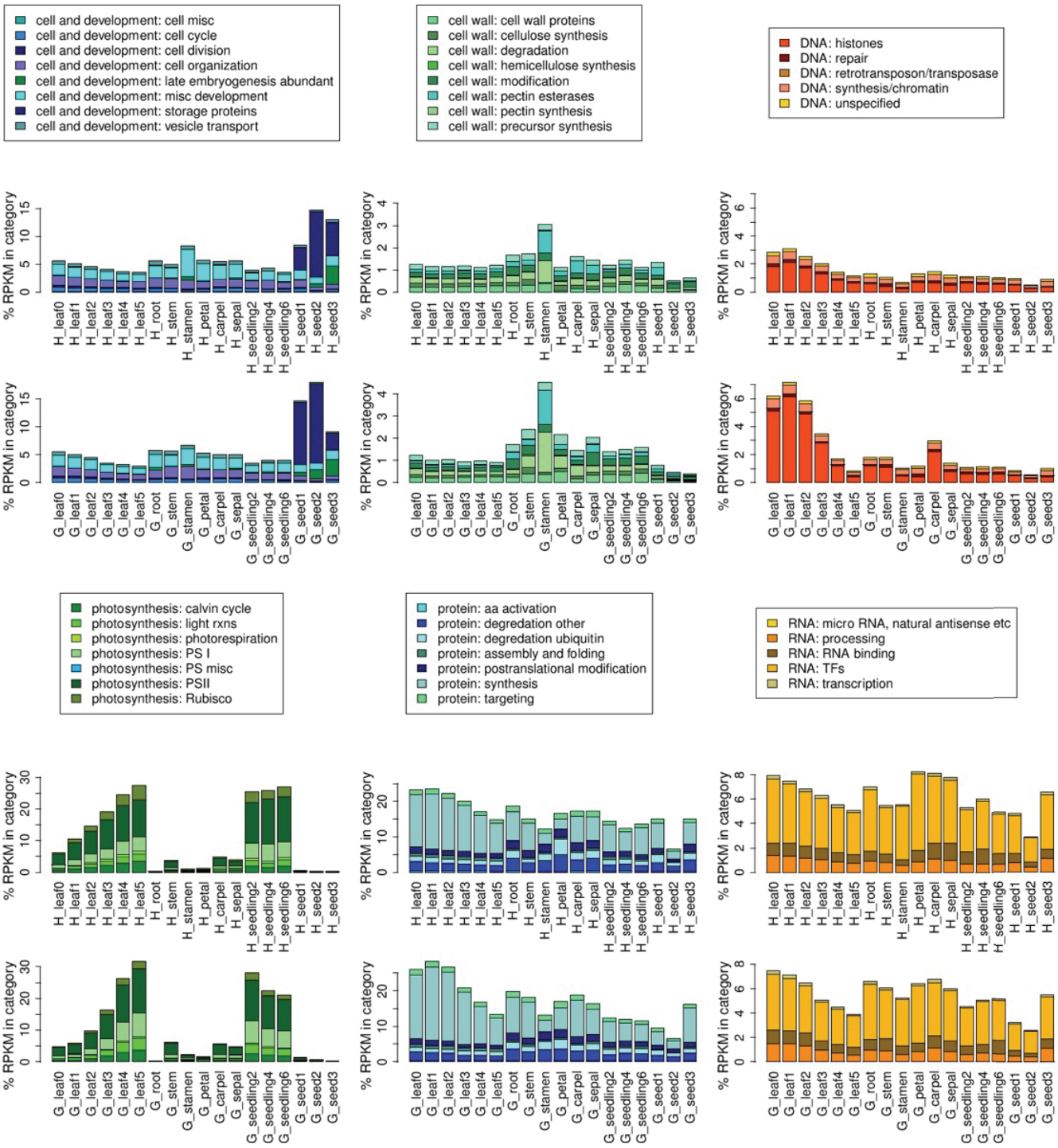
Supplemental Figure 9. Hierarchical cluster analysis with bootstrapped samples of *G. gynandra* and *T. hassleriana.* Numbers above the nodes show the approximately unbiased p-value (red) and the bootstrap probability (green). Blue is lowest expression and yellow highest expression. Left-hand vertical 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.



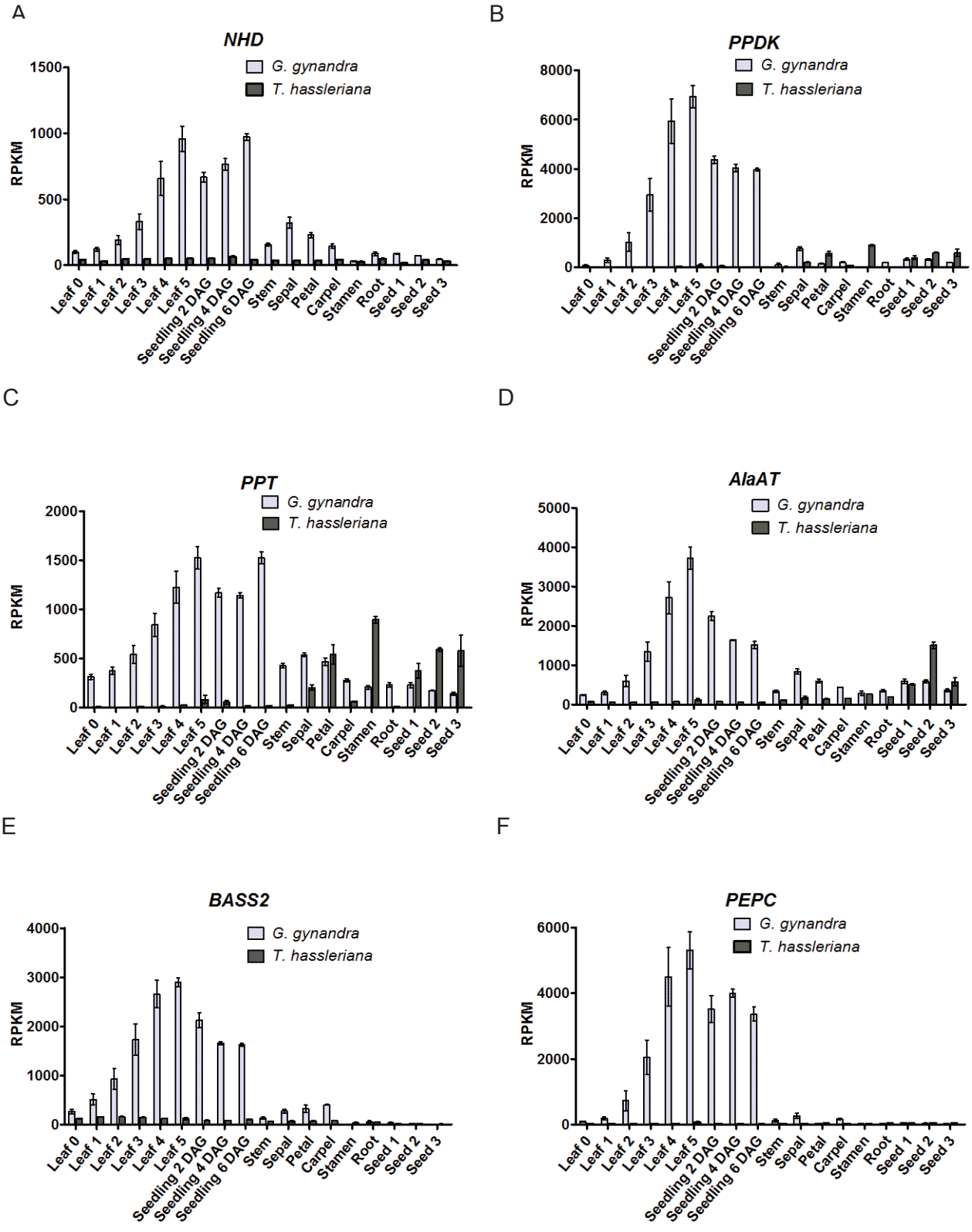
Supplemental Figure 10. Transcriptional investment of each tissue compared in both species. Cumulative average RPKMs in percent for basal Mapman categories for each tissue in *G. gynandra* and *T. hassleriana*.



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.

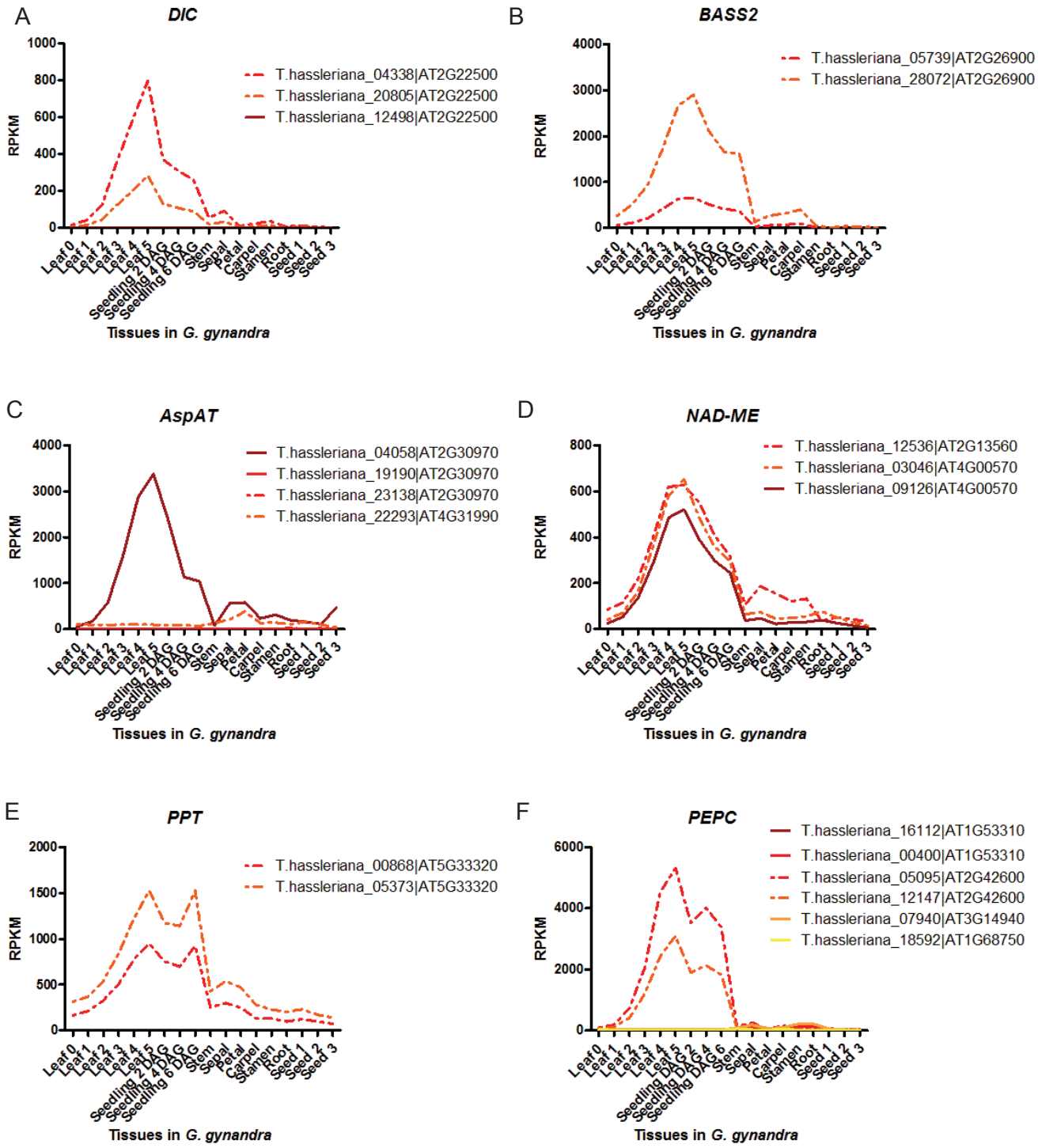


Supplemental Figure 11.2. Transcriptional investment at secondary Mapman category of each tissue compared in both species (Part 2). Distribution of the Mapman categories in each tissue in *G. gynandra* and *T. hassleriana*. Plot shows percent of average RPKMs of the 12 customized secondary Mapman bins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

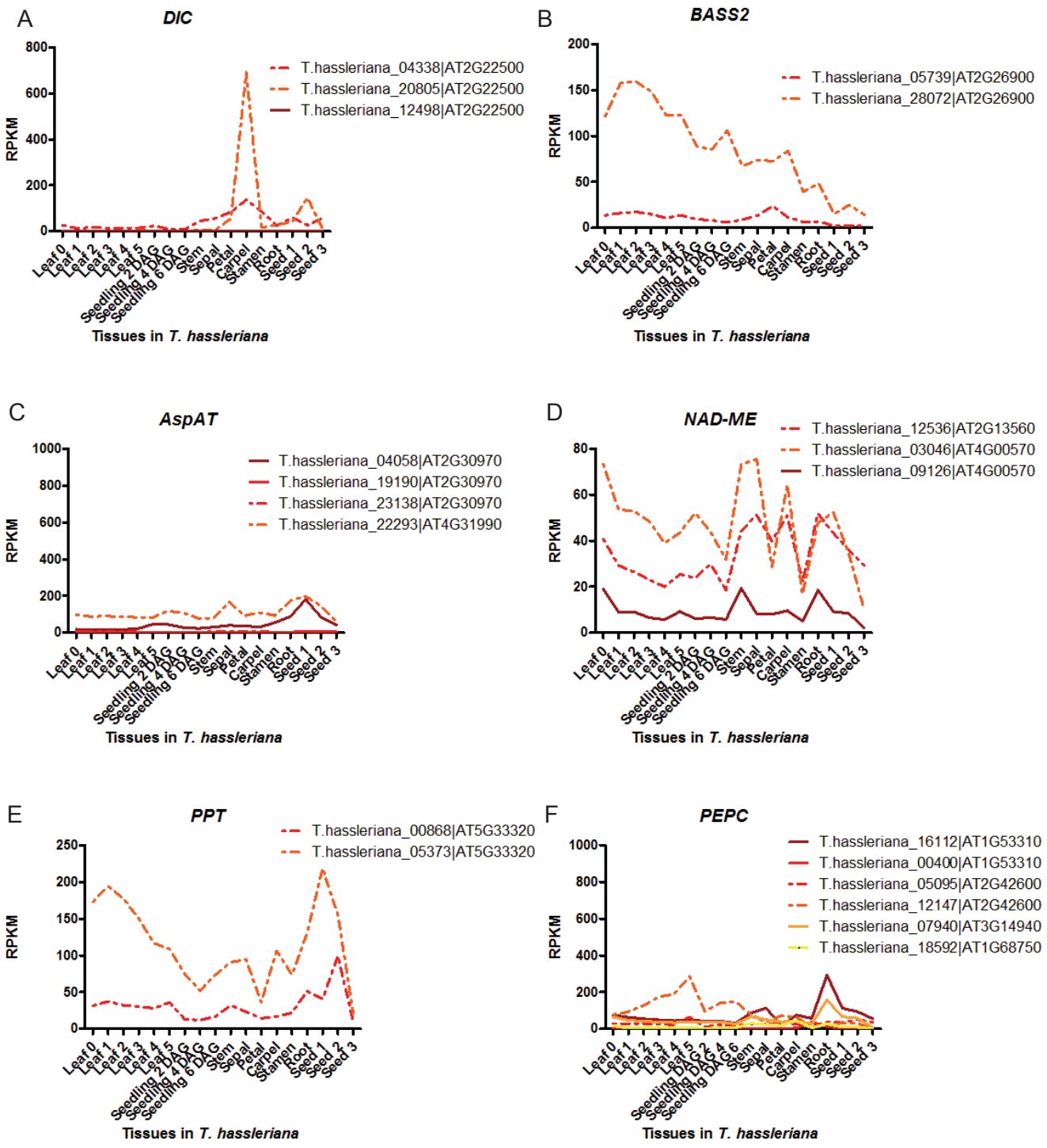


Supplemental Figure 12. Comparison of gene expression dynamics within the leaf gradient of both species.

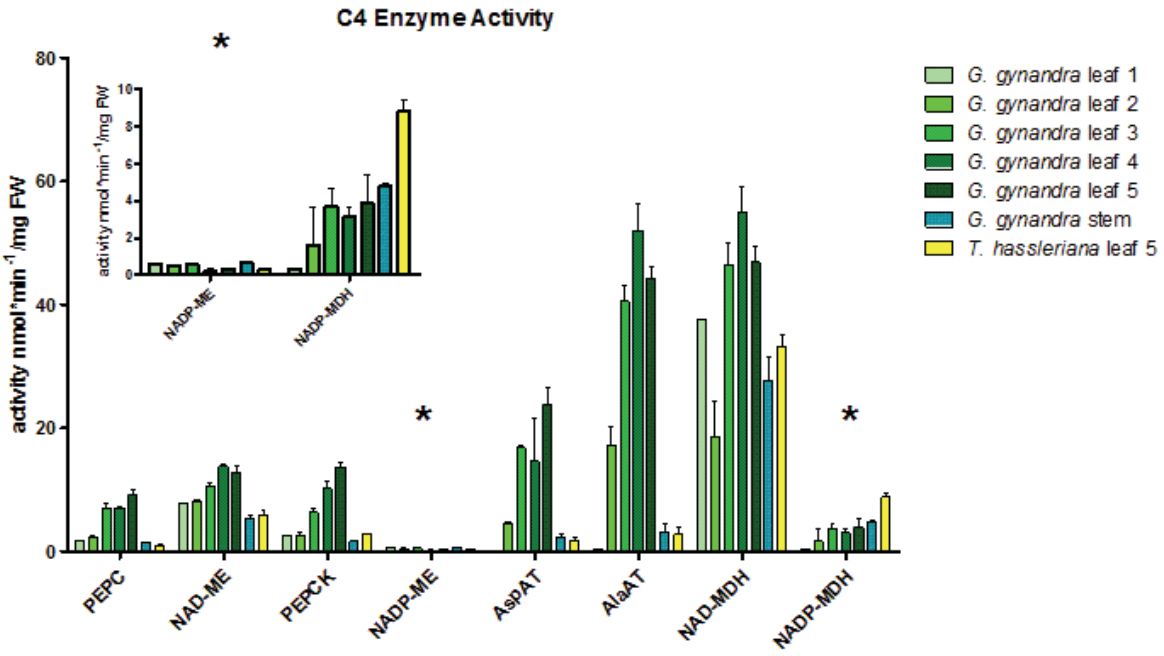
(A-F) Average expression pattern of highest abundant putative ortholog of C_4 cycle genes (*NHD, PPDK, PPT, AlaAT, BASS2, PEPC*) in photo- and heterotrophic tissues in *G. gynandra* (light grey) and *T. hassleriana* (dark grey); (n=3 ± SE, standard error)



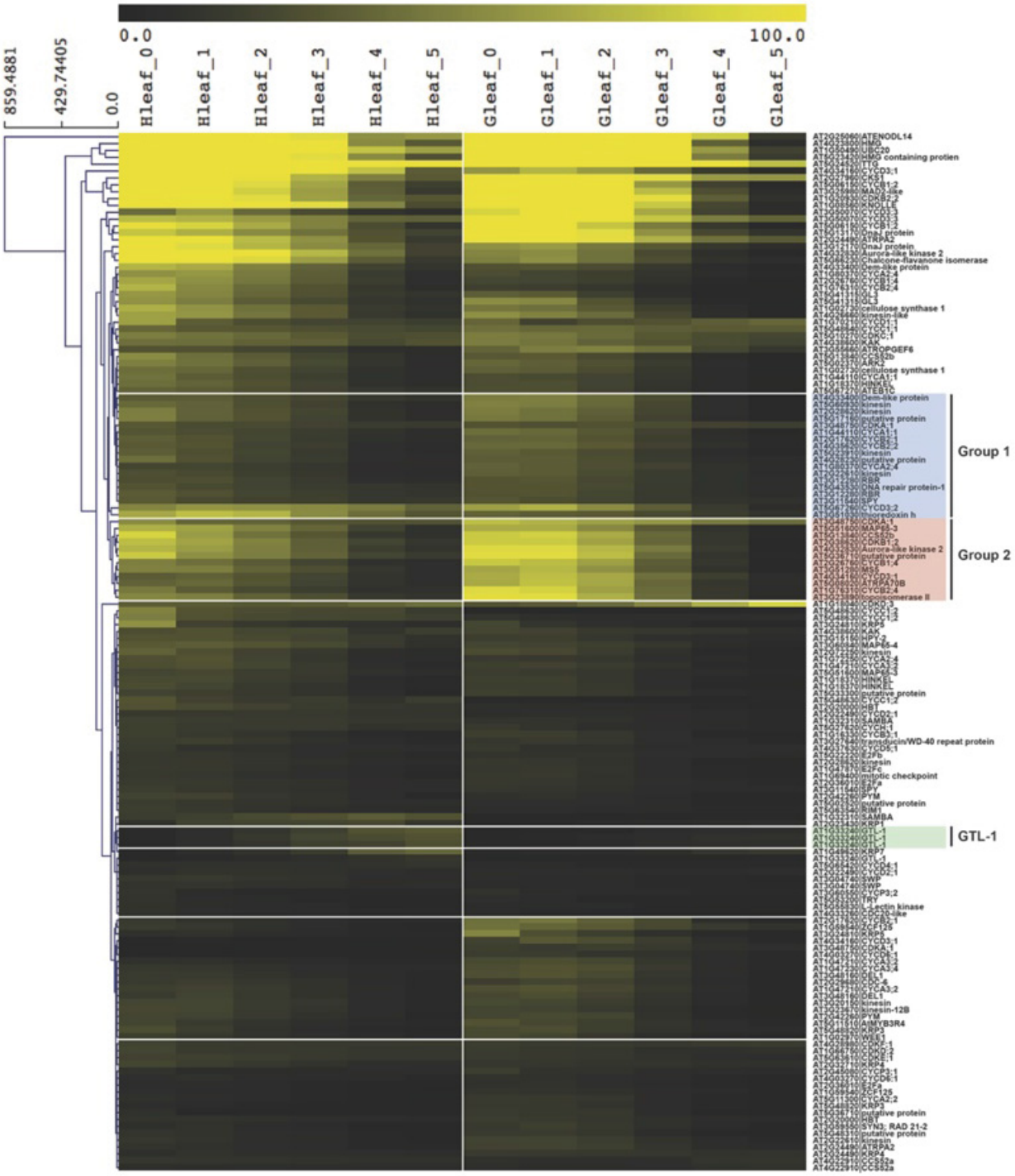
Supplemental Figure online 13. Plot of all C_4 gene putative orthologs expression pattern (RPKM) in *G. gynandra,* that were annotated as C_4 genes with AGI identifier and respective *T. hassleriana* ID. **(A-F)** Average expression pattern of putative ortholog of C_4 cycle genes (*DIC, BASS2, AspAT,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



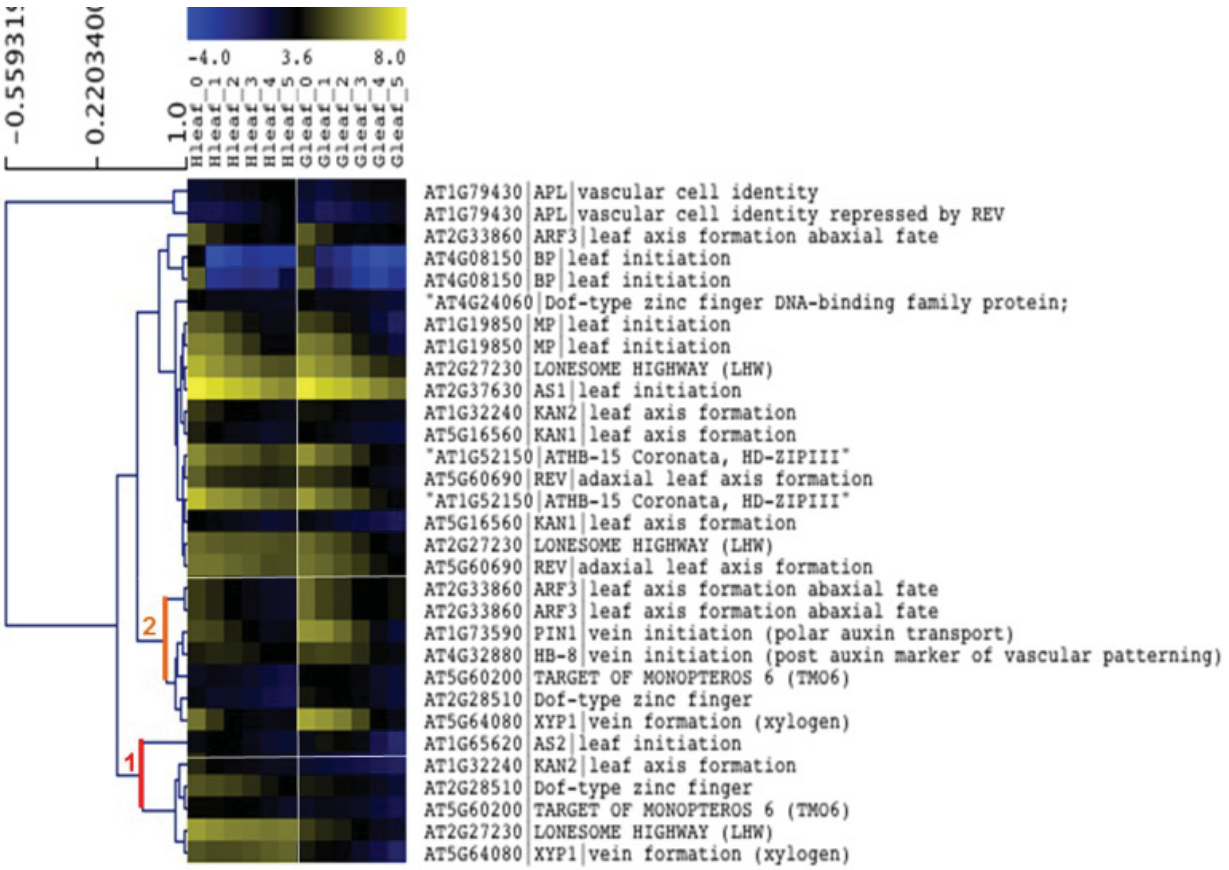
Supplemental Figure online 14. Plot of all C_4 gene putative orthologs expression pattern (RPKM) in *T. hassleriana,* that were annotated as C_4 genes with AGI identifier and respective *T. hassleriana* ID. **(A-F)** Average expression pattern of putative ortholog of C_4 cycle genes (*DIC, BASS2, AspAT,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

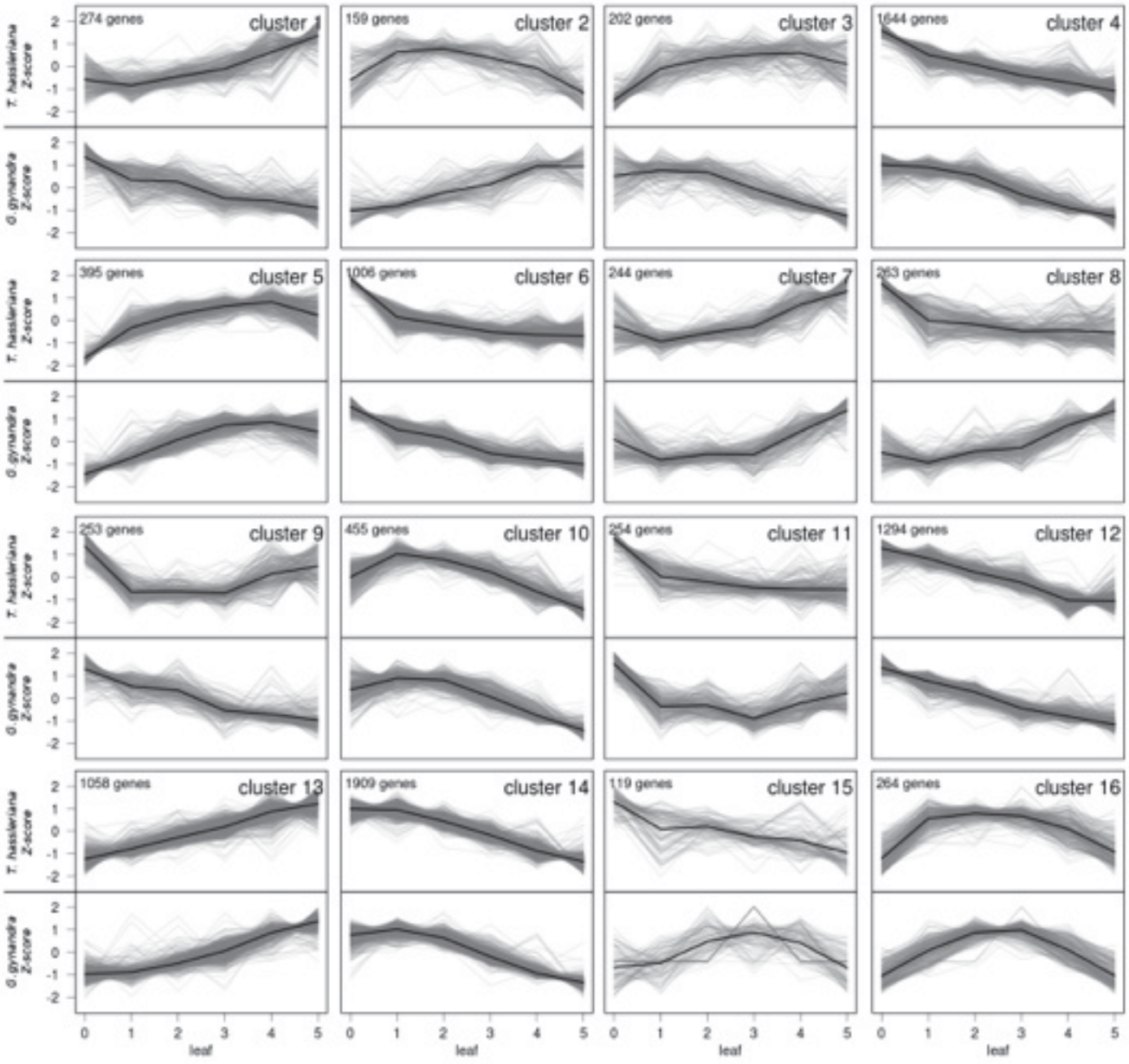


Supplemental Figure 16. Hierarchical clustering of average RPKM with Euclidean distance of core cell cycle genes in *T. hassleriana* and *G. gynandra*. Core cell cycle genes were extracted from (Vandepoele et al., 2002; Beemster et al., 2005). Deregulated cluster of interest are marked with blue and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

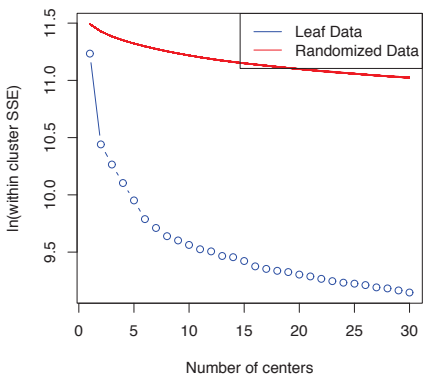


Supplemental Figure 17. Hierarchical clustering with Pearson's correlation of leaf developmental factors. Averaged transcript abundances (RPKM) of leaf gradient sample of transcriptional regulators involved in axial and vasculature fate determination were clustered. Group 1 (orange) and group 2 (red)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

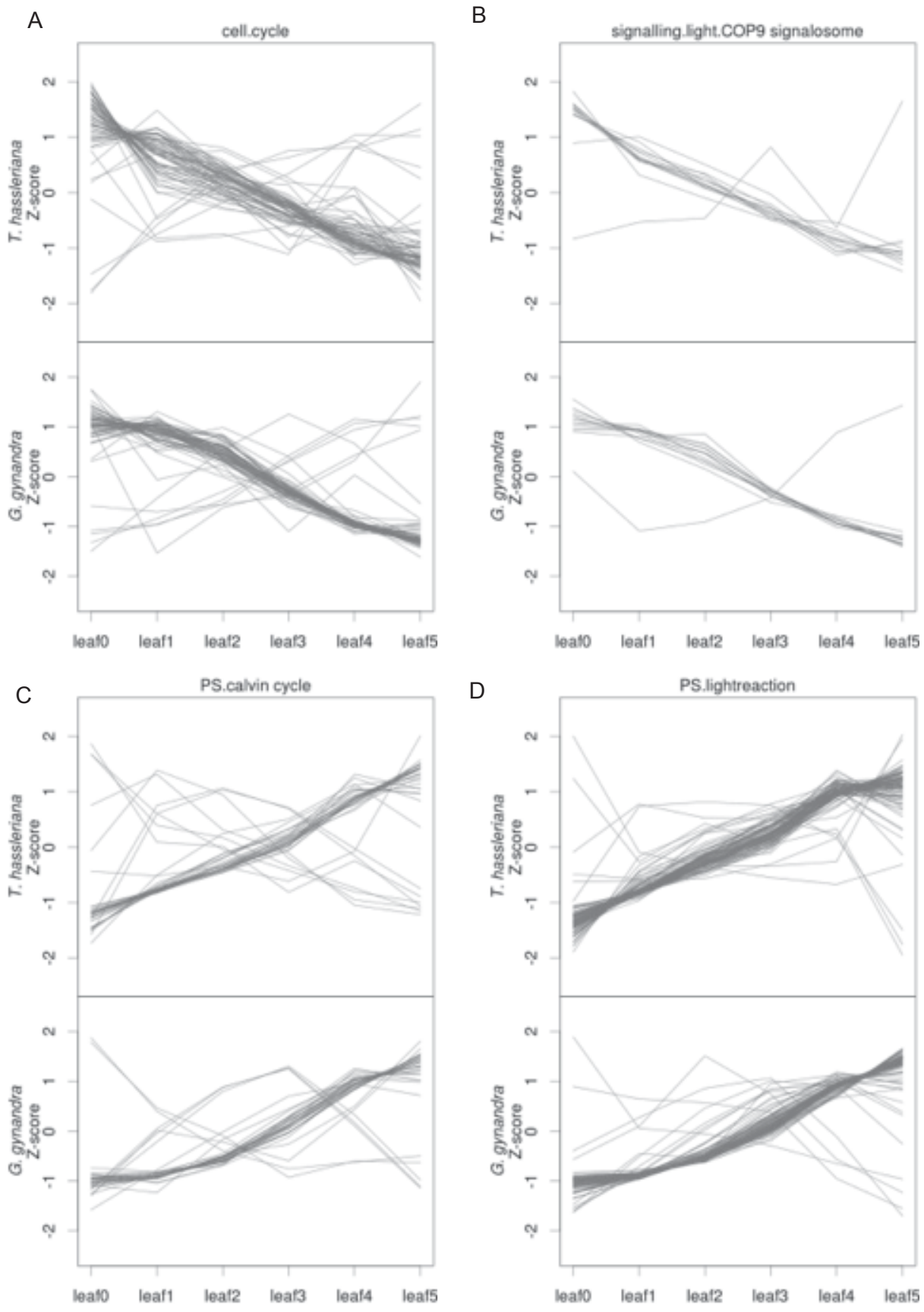
A



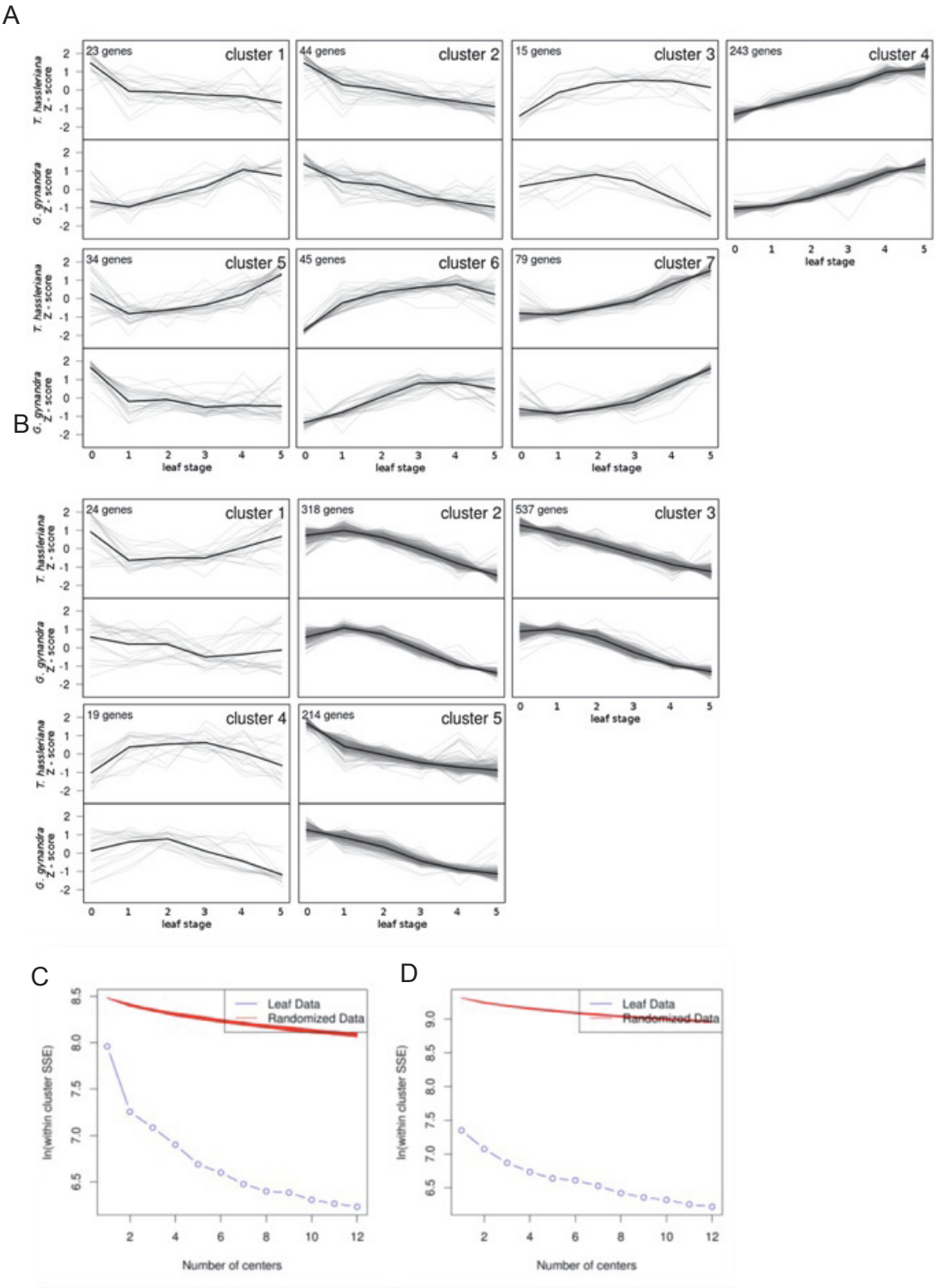
B



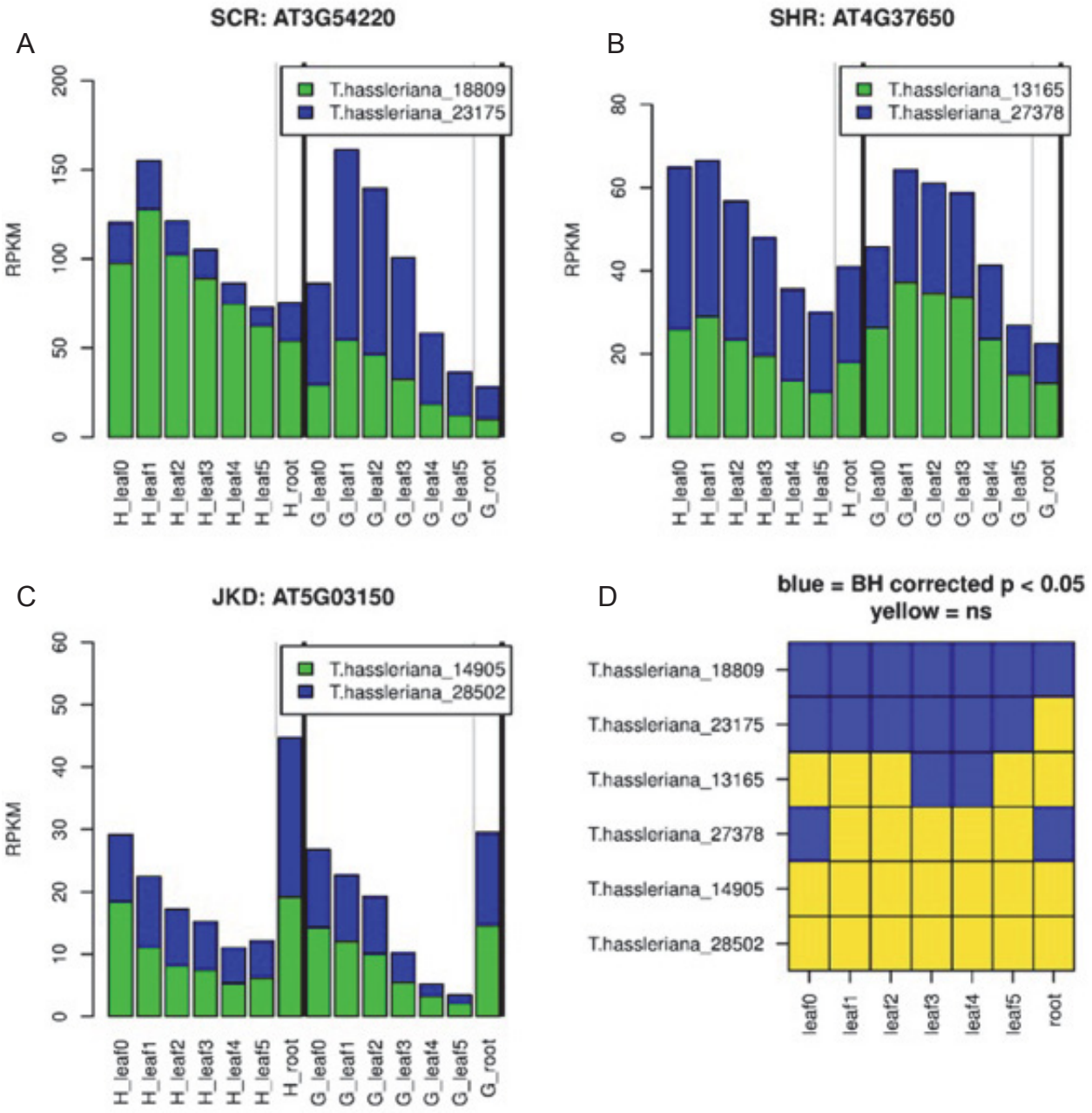
Supplemental Figure 18. *K*-means clustering of leaf gradient expression data and quality assessment. **(A)** *K*-means clustering of transcript abundances (RPKM) of leaf stage averages (*n*=3) between *T. hassleriana* and *G. gynandra* shown as species-scaled *Z*-scores. Size of each cluster is indicated 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 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.

VI. ADDENDUM PREPARED FOR PUBLICATION

Manuscript 3

Plasticity of C_4 Photosynthesis in the amphibious sedge *Eleocharis retroflexa*

Canan Külahoglu*, *Simon Schliesky, Manuel Sommer, Alisandra K. Denton, Andreas Hussner, Robin C. Buell, Andrea Bräutigam and Andreas P.M. Weber*

Submission ready version November 2014

Impact Factor: N.A.

**First author*

Main findings:

The transcriptome contig resource for the amphibious sedge *Eleocharis retroflexa* is presented in this study. The dataset includes the transcriptomes of aquatic and terrestrial *E. retroflexa* culms complemented with physiological data, anatomy and enzyme activity measurements. *E. retroflexa* plants undergo structural and metabolic changes during submergence in water by higher investment in light capture harvesting components and photosynthetic apparatus and less expression of cell wall related genes. Also the photosynthetic mode and carbon assimilation is adjusted by stronger C_4 cycle signature in the terrestrial culms, while the aquatic culms display enhanced Calvin-Benson-Bassham cycle and photorespiration similar to C_3-C_4 intermediate species. In terrestrial culms the C_4 cycle has great plasticity depending on water availability and drought stress on transcriptional level.

Contributions:

- Plant Cultivation and sampling of all plant material
- Establishing RNA extraction protocol, RNA extraction and Illumina library preparation
- Expression quantification
- Bioinformatic data analysis (except contig assembly, differential expression analysis (Edge R) and cumulative relative expression plots)
- Data analysis
- Microscopic analyses
- Enzyme assays and physiological data
- Quantitative real-time PCR
- Writing of Manuscript

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.

Abstract

C_4 photosynthesis is a complex adaptive trait, facilitating the adaption to hot and arid environments. It has been hypothesized that evolution of the C_4 trait 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 a remarkable phenotypic plasticity in their mode of photosynthetic carbon assimilation. In particular *Eleocharis retroflexa*, which is classified as the most C_4-like species amongst the known *Eleocharis* C_4 performing sedges, is able to thrive under submerged conditions by reconfiguration of its culm anatomy and potentially its photosynthetic mode. The underlying molecular mechanisms permitting adaptation to environmental change through metabolic plasticity are however unknown to date. To begin to unravel the physiological and transcriptional programs that enable *E. retroflexa* to thrive during submergence and on soil, we employed deep RNA-sequencing of aquatically and terrestrially grown culms and contextualized these molecular data with physiological parameters and enzyme activity measurements. We found that *E. retroflexa* undergoes structural and metabolic rewiring during submergence by adapting its culms fully to the new habitat and adjusting its carbon metabolism. While the aquatic *E. retroflexa* culm transcriptome reflects characteristics described for flooding tolerant plants, the carbon metabolism displays a typical C_3-C_4 intermediate signature, featuring high abundance of photorespiratory transcripts. At the same time the C_4 cycle is maintained. Owing to its metabolic plasticity, *E. retroflexa* represents an interesting model to unravel the molecular mechanisms of adaptation to changing environments by phenotypic plasticity.

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 and McKown 2006).

The CO_2 concentrating mechanism of C_4 photosynthesis is considered as 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 photosynthesis performing plants is thought to reduce their potential for phenotypic plasticity and photosynthetic acclimation to variable environments, as compared to C_3 plants (reviewed by Sage and McKown 2006).

C_4 photosynthesis requires a distinct anatomical and biochemical infrastructure for optimal functionality (Hatch 1987). Generally spoken, 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 next to the ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO; Bowes et al. 1971; Furbank and 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 phospho*enol*pyruvate 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 and 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 phospho*enol*pyruvate carboxykinase (PEPCK), followed by assimilation into carbohydrates by the Calvin-Benson-Bassham cycle (CBBC; Hatch 1987; Hatch and 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 and 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 those 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 aquatic and terrestrial habitat (Ueno 2004; Ueno and 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*, *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 and Wakayama 2004). However, the 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 submerged terrestrial wetland species are employing to overcome the new challenges of the aquatic environment. The challenges of the aquatic environment are, for example, physical restrictions on light availability, gas exchange, and nutrient availability (Krause-Jensen and Sand-Jensen 1998; Pedersen et al. 2013). In particular, the gas diffusion rates in water are 10^4-fold slower than in air. The limitation of CO_2 and O_2 gas flow severely 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 and Voesenek 2009). As an adaptation to flooding, some species grow out of water by shoot elongation to reestablish aerial photosynthesis (Setter and Laureles 1996), whereas others develop specialized “aquatic leaves” specifically accustomed to the wet habitat (Bailey-Serres and 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. 2003; Mommer et al. 2006; Mommer et al. 2005b; Sand-Jensen and 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 and 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).

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; group2: every two days 250ml water; group3: every four days 150ml water; group4: 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 and algal growth was suppressed by co-cultivation of shrimp. Temperature in aquariums was constant at approx. 25°C 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.

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 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 and 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 database (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. Unigene database was filtered for contigs that either match an *S. italica* identifier or a plant identifier in Uniref100. This resulted in a unigene database of 27,021 contigs and reduced possible contamination through non-plant contigs.

The final unigene database 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.

Gene expression profiling

Expression abundances were determined by mapping the single-end read libraries (each replicate for each condition) independently against *S. italica* primary transcript coding sequences V2.1 (Bennetzen et al. 2012) using BLAT V35 (Kent 2002) in dnax mode (nucleotide sequence of query and reference are translated in six frames to protein) and counting the best mapping hit based on e-value for each read uniquely. Default BLAT parameters were used for mapping. 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. Differential 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 Bonferroni-Holms correction (Holm 1979).

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 own genome mapping, we mapped (i) *T. hassleriana* mature leaf reads (Külahoglu 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 (MEV4; http://www.tm4.org/mev/; Saeed et al. 2006; Saeed et al. 2003) 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 and 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).

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).

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). Mean normalized expression (MNE) was calculated via the ΔΔCT method after 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) in three biological replicates with three technical replicates per sample. Chlorophyll measurements were performed according to 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 δ^{13C} measurements followed the two-point method described by Coplen et al. (2006).

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 and Madan 1999), producing a contig database of 43,817 unigenes with an N50 of 984 bases (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 porphyrin and 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 database (Table 1), however, annotation of the unigene database with *S. italica* revealed known issues of *de novo* assemblies (Schliesky et al. 2012). Around 58% of the annotated *E. retroflexa* unigenes were matched to a *S. italica* identifier, which was assigned as 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. C_4 cycle genes, such as the alanine aminotransferase (AlaAT) and the triose-phosphate transporter (TPT) were absent in the unigene database. For comparison *E. retroflexa* reads were aligned to the *S. italica* gene models using the Blat algorithm. In total, 21,679 *S. italica* identifiers were matched by 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 database, 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. These contigs were annotated as plant specific 16S, 18S and 26S rRNA subunits. On the basis of these results and previous experience (Bräutigam 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 minor 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). In contrast Pearson's correlation between the biological replicates (average mean r = 0.87, Supplemental Table 3) was similar to the Pearson's correlation between averages of the transcriptomes of the two conditions (average mean r=0.86; Supplemental Table 3). These results indicated that one third of the gene expression was changed by the culm's habitat. Despite this 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 all 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 the genes called as differentially expressed and the remaining gene set. In total, 8% of the whole transcriptome (1,356 genes) was differentially regulated between the 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). If the variation among the samples was random, it would be randomly distributed between the significantly changed transcripts and the remaining transcripts. We compared the variability among replicates of all 708 significantly changed transcripts. There was a significant enrichment between genes with two fold variation between replicates and genes that were differentially expressed genes (Fisher's exact test P-

value<0.001; Figure 2C). Thus, the genes that were related to habitat acclimatization were the ones showing greater variation between replicates than other genes.

Functional changes in the 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

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 habitus of aquatic *E. vivipara* plants (Supplemental Figure 4; Ueno 2001). Thus, to trace down the differences between aquatic and terrestrial culm structure the respective transcriptomes were analyzed for pathway enrichment of cell wall related categories (Figure 3A). All mentioned changes were statistically tested for significance. The 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 ENDOTRANSGLYCOSE 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 monitor 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 the cuticle waxes. In the aquatic culms no stomata could be detected. Thus, the anatomical and structural changes within culm anatomy were traceable 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 hence assessed the transcriptomes for consequences of the submerged lifestyle on photosynthesis and light capture. Most of the Mapman annotated transcripts for light reactions including photosystem 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 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 and Wakayama 2004). The transcriptomes of the terrestrial and aquatic culms now enabled a detailed analysis 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 the terrestrial and between 234-9,972 RPM in the aquatic culms (Supplemental Table 4). In the 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:HYDR OGEN ANTIPORTER/PHOSPHO*ENOL*PYRUVATE TRAN-SLOCATOR* (*BASS2/NHD/PPT*), importing substrates (phosphate and pyruvate) for pyruvate, phosphate dikinase (PPDK) activity and exporting phospho*enol*pyruvate (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 (δ^{13C}) with -15.76 ‰ was comparable to that of other C_4 plants (Figure 6B; Cernusak et al. 2013).

The aquatically grown culms had decreased C_4 cycle enzyme expression, though the C_4 photosynthesis signature was still higher as 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; Supplemental 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 HYDROXYMETHYLTRANSFERASE* (*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 ‰ (δ^{13C} value) showed stronger discrimination against ^{13}C, which is closer to the range of C_3 plants (Figure 6B; Cernusak et al. 2013).

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

During its lifecycle, *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 in 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; Group2: every two days 250ml water; Group3: every four days 150ml water; Group4: 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). Moreover, 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 group1 to group2 had no significant effect on gene expression (Figure 8A-C). However, limiting the water availability to watering every fourth day (group3) 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). The 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 group1 and group3 (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).

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 and Wakayama 2004; Uchino et al., 1995; Ueno 2004). In general, C_4 plants have been proposed to display less plasticity in the range of growth habitats and phenotypic plasticity (Sage and McKown 2006). To determine, which transcriptional programs enable *E. retroflexa* to acclimatize to terrestrial and aquatic lifestyle so rapidly and successfully, meanwhile performing C_4/C_4-like photosynthesis, we generated two 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 database 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). With 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* (Bräutigam 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 and Wang 2011; Schliesky et al. 2012).

Cross-species mapping introduces less bias in expression level dynamics than mapping to contigs

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 successfully used 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).

To evaluate whether cross-species mapping itself leads to any systematic bias on estimating expression levels, we visualized the fraction of reads mapping to higher and lower expressed sequences in the recently sequenced species *Tarenaya hassleriana* to its own genome (Cheng et al. 2013) versus the genome of the model species *A. thaliana* (Lamesch et al. 2012; Supplemental Figure 2B). This comparison shows that, in contrast to contig

mapping, mapping to a cross-species reference does not cause visual alterations in the mapping dynamics.

Transcriptional changes between the terrestrial and the aquatic culms are closely related to the habitat switch and photosynthetic mode

To put the amount of changes seen between the aquatic and terrestrial culms (mean *r*=0.86; Supplemental Table 3) in context, the correlation between the two transcriptomes is 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, comparing 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 transcriptomes 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 being detected 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 transcript 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 transcriptomes of *E. retroflexa* show high plasticity and variability in transcripts linked to the habitat acclimatization

Biological variation of transcript abundances between each of 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 microenvironmental 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 gibberilic acid to terrestrial culms can trigger submerged (C_3) culm anatomy featuring small BSCs and the absence of stomata (Ueno 2001). However, no enhanced gibberilic acid signaling could be detected in the aquatic transcriptomes (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 and Voesenek 2008; Jackson, 2008). In the 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 their aquatic habitat at the time of their harvest.

Instead we find significant changes related to histone modification and DNA methyltransferases (DNMT; Supplemental Figure 7; Supplemental Dataset 7). Five *ALIFIN-LIKE* transcriptional regulators are 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 the 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 and 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 habitus (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 and 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). Less investment 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 and 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 and 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 and 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 and Poulson 2003; Frost-Christensen and 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 and 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 et al., 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* (Bräutigam 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 and Bernacchi 2003). Aquatic acclimated culms are thinner and have reduced cuticles to decrease the internal diffusion path for CO_2 to the chloroplasts (Maberly and Madsen 2002; Madsen and Sandjensen 1991). Aquatic *E. retroflexa* culms visually appear to be thinner, have less dry matter (Figure 3B) and their transcriptional investment in cuticular 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. 2003).

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 and 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 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 higher solubility catalyzed by carbonic anhydrases (CA; reviewed by Pedersen et al. 2013). In *E. retroflexa* one *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 C_3-C_4 intermediate, but lack complete BSC compartmentation of RuBisCO (Edwards and 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 and 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 and Wakayama 2004). However, the aquatic culms show an untypically 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 and Spence 1989). Similar conditions can occur for aquatic adapted plants, when they grow as dense a canopy, which leads CO_2 depletion and O_2 supersaturation of the water during 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 and 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 associates 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 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).

So far no examples of classic two-cell BSC/MC C_4 photosynthesis have been discovered for aquatic plant species. With the variety of mechanisms evolved to 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 transcriptomes provides a unigene database for further molecular studies. The transcriptomes of the terrestrial and aquatic culms now 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.

Tables

Table 1. Sequencing and Mapping statistics and transcriptome dynamics of *E. retroflexa* reads aligned to *S. italica* and *E. retroflexa* reference.

Eleocharis samples	Aquatic 1	Aquatic 2	Terrestrial 1	Terrestrial 2	Terrestrial 3
raw reads	30469686	34239945	22011612	30043954	25010427
cleaned reads	28,203,253	33,838,855	21,729,249	29,391,792	24,629,693
mapped reads to *S. italica*	7,768,905	12,337,799	7,648,795	8,182,178	7,585,432
mapped reads to unigene database (>200 bases)	7,593,124	23,674,037	18,779,815	24,787,798	17,606,148
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	4,882	5,263	5,826	5,807	5,495
Number of *S. italica* IDs > 1,000 RPM	132	143	135	136	136
Number of *S. italica* ID matching	19,298	20,248	19,893	20,041	19,814
Number of unigenes matching	34,971	38,548	38,324	38,352	38,489
***S.italica* IDs covered by reads (%)**	54.4	57.1	56.1	56.5	55.9
***E. retroflexa* unigenes covered by reads (%)**	79.8	88.0	87.5	87.5	87.8
Transcript number >0 RPM aligned to *S. italica*	19,298	20,248	19,893	20,041	19,814
Transcript number >1 RPM aligned to *S. italica*	14,875	15,378	16,001	15,910	15,654
Transcript number >20 RPM aligned to *S. italica*	4,882	5,263	5,826	5,807	5,495
Transcript number >1,000 RPM aligned to *S. italica*	132	143	135	136	136

Figures

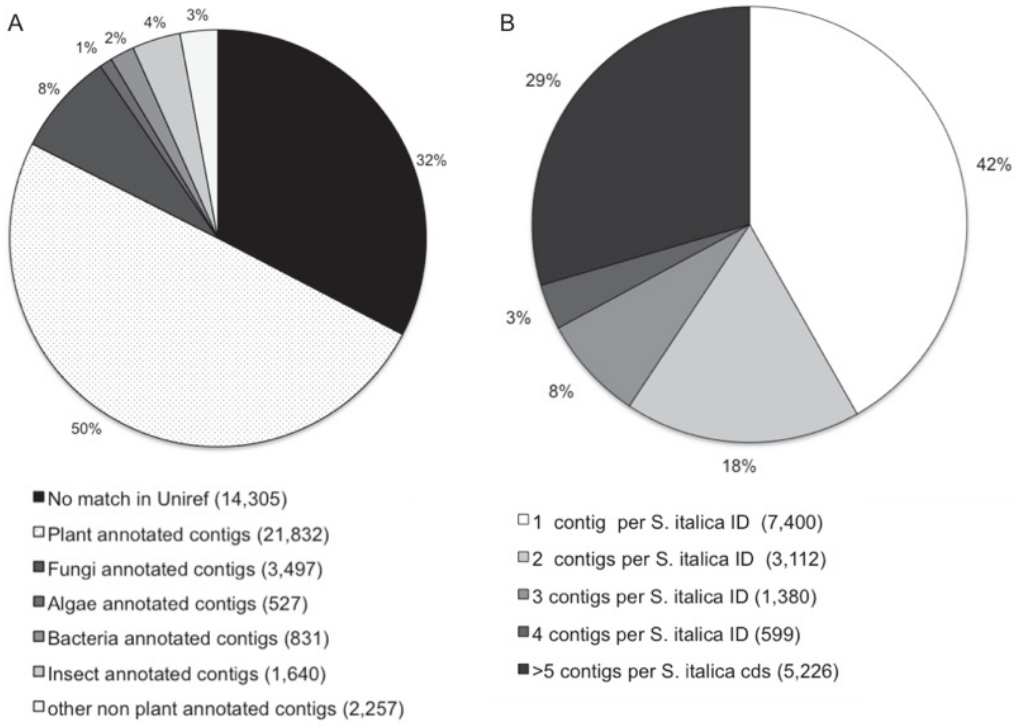


Figure 1. Annotation of *E. retroflexa* contig database.
(A) Annotation of contigs against Uniref 100. Distribution of contigs annotated by Uniref100 falling into major species categories of plant, bacteria, algae, fungi and other non-plant annotated contigs. Total contig number is indicated in parentheses. **(B)** Annotation of contigs against *S. italica* identifier. Percentage of contig number per predicted identifier shows redundancy of assembled contigs. Number of best matching contigs per predicted *S. italica* identifier is indicated in parentheses.

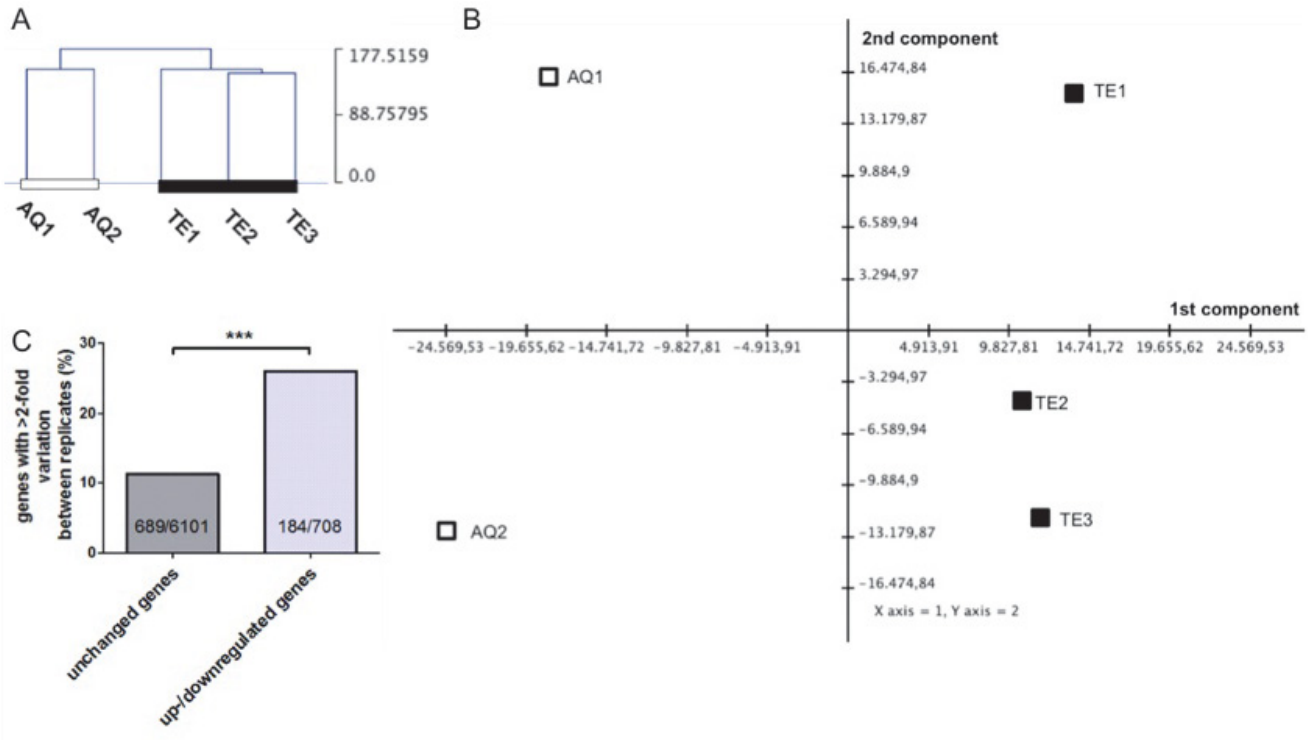


Figure 2. Transcriptome dynamics and variability between samples and habitat.
(A) Hierarchical clustering of all sequenced *E. retroflexa* samples. *E. retroflexa* transcriptomes (>1 RPM filtered) were clustered, after normalizing to z-scores per row, with Euclidean distance and average linkage. **AQ**, aquatic samples; **TE**, terrestrial samples. **(B)** Principle component analysis (PCA) between aquatic and terrestrial *E. retroflexa* transcriptomes. Plot shows all sequenced samples from aquatic (white boxes) and terrestrial (black boxes) *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 replicates and enrichment analysis of significant changed transcripts. Fisher's exact test ***P-value<0.001.

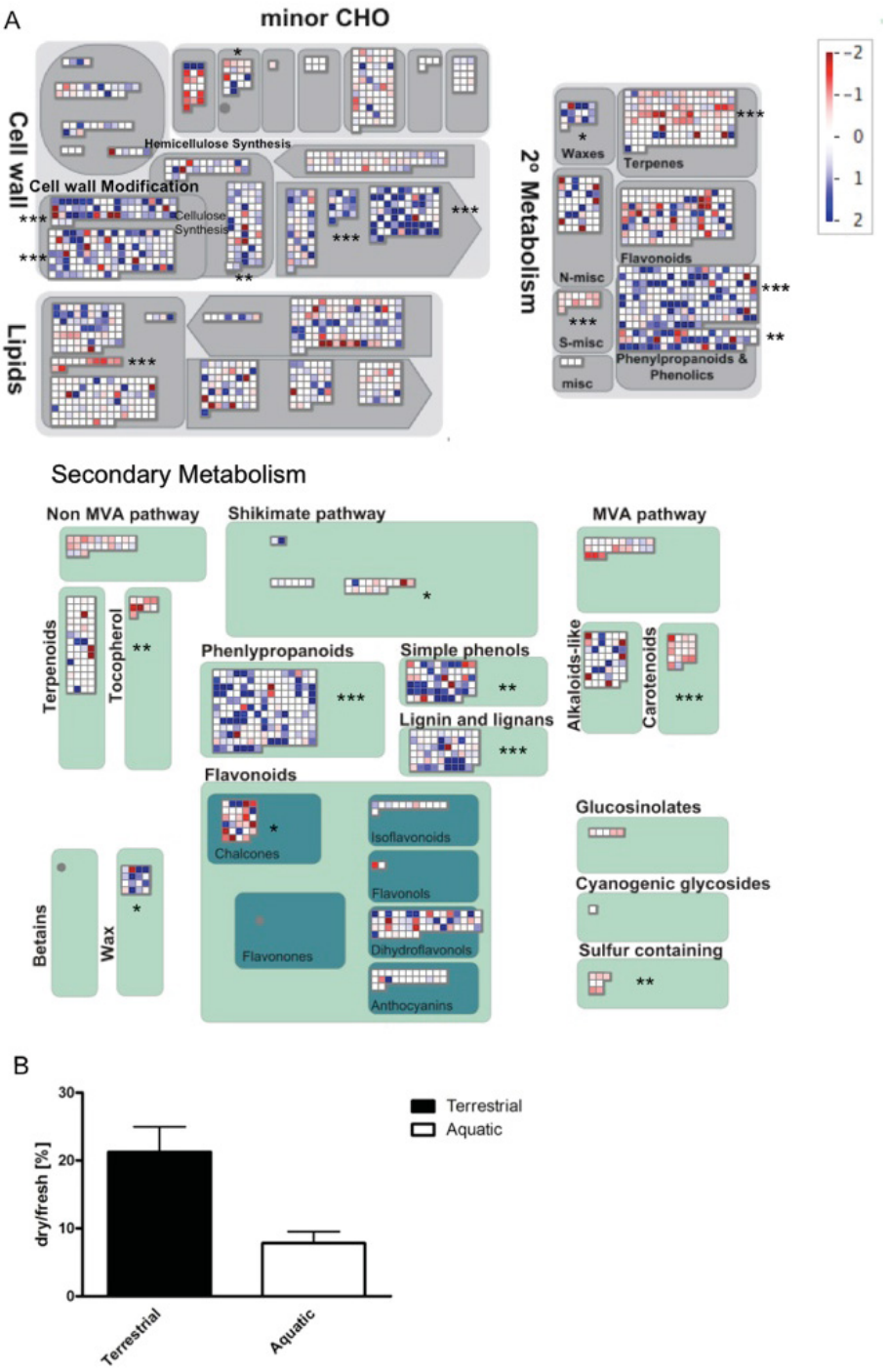


Figure 3. Structural differences between aquatic and terrestrial *E. retroflexa* culms.
(A) Overview of secondary and carbon metabolism transcriptome levels (Mapman, TAIR10) in *E. retroflexa* culms. Heatmaps depict log2 fold-changes of aquatic versus terrestrial transcript levels in RPM. Red (ratio<0) represents an increase of transcript fold change in aquatic culms. Blue indicates (ratio>0) an increase of transcript fold changes 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 2006). **(B)** Comparison of water content and biomass between terrestrial (black) and aquatic culms as ratio dry weight (DW) against fresh weight (FW) in percent. n= 3 biological replicates; error bars ± SE, standard error.

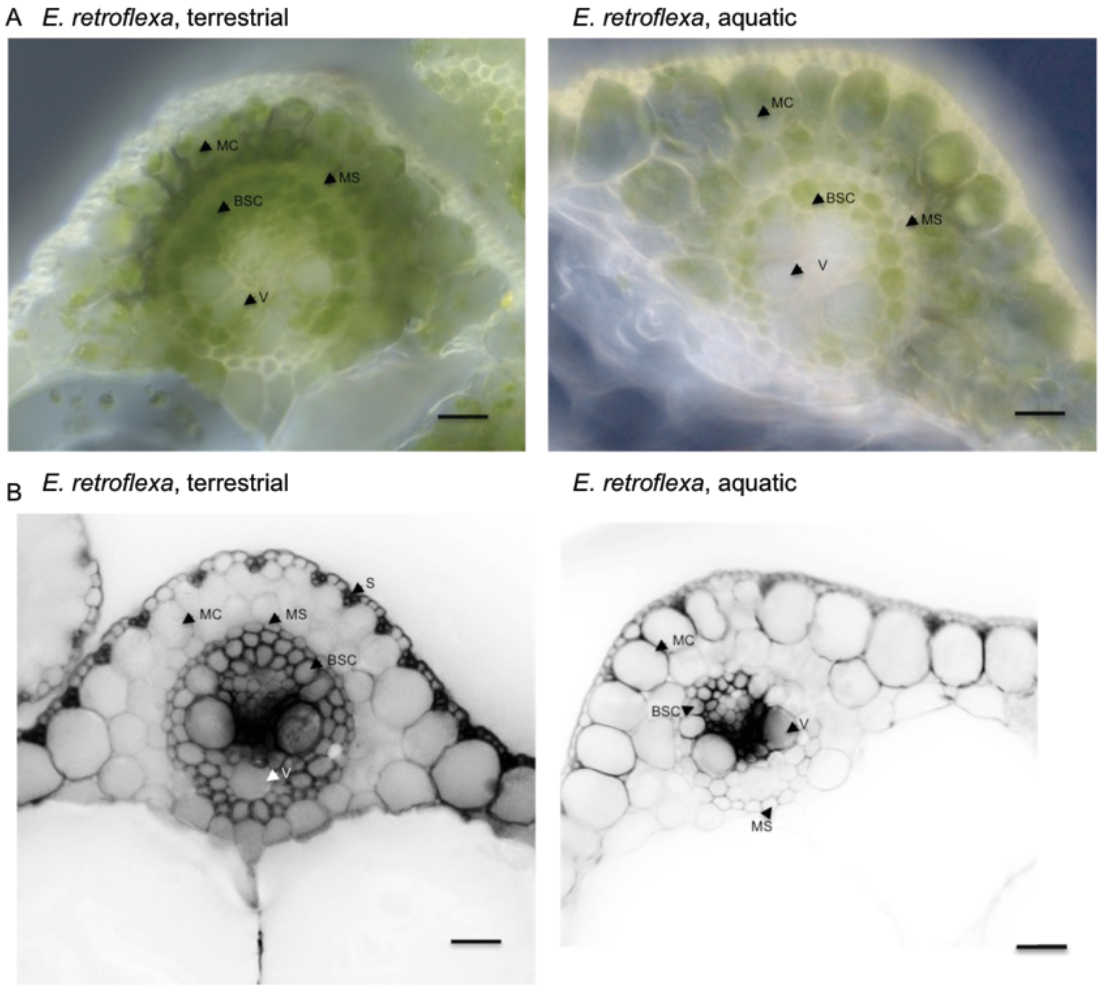


Figure 4. Culm anatomy of mature *E. retroflexa* plants grown under terrestrially (left) and aquatically (right) conditions.
(A) Microscopic images of *E. retroflexa* cross-sectioned culms grown on soil and under water. **(B)** Autofluorescence microscopic images of aquatic and terrestrial *E. retroflexa* cross-sectioned culms. Scale bar: 20µm. Cell types are indicated by closed arrows. **BSC**: bundle sheath cell; **MC**: mesophyll cell; **MS**: mestome sheath; **V**: vein.

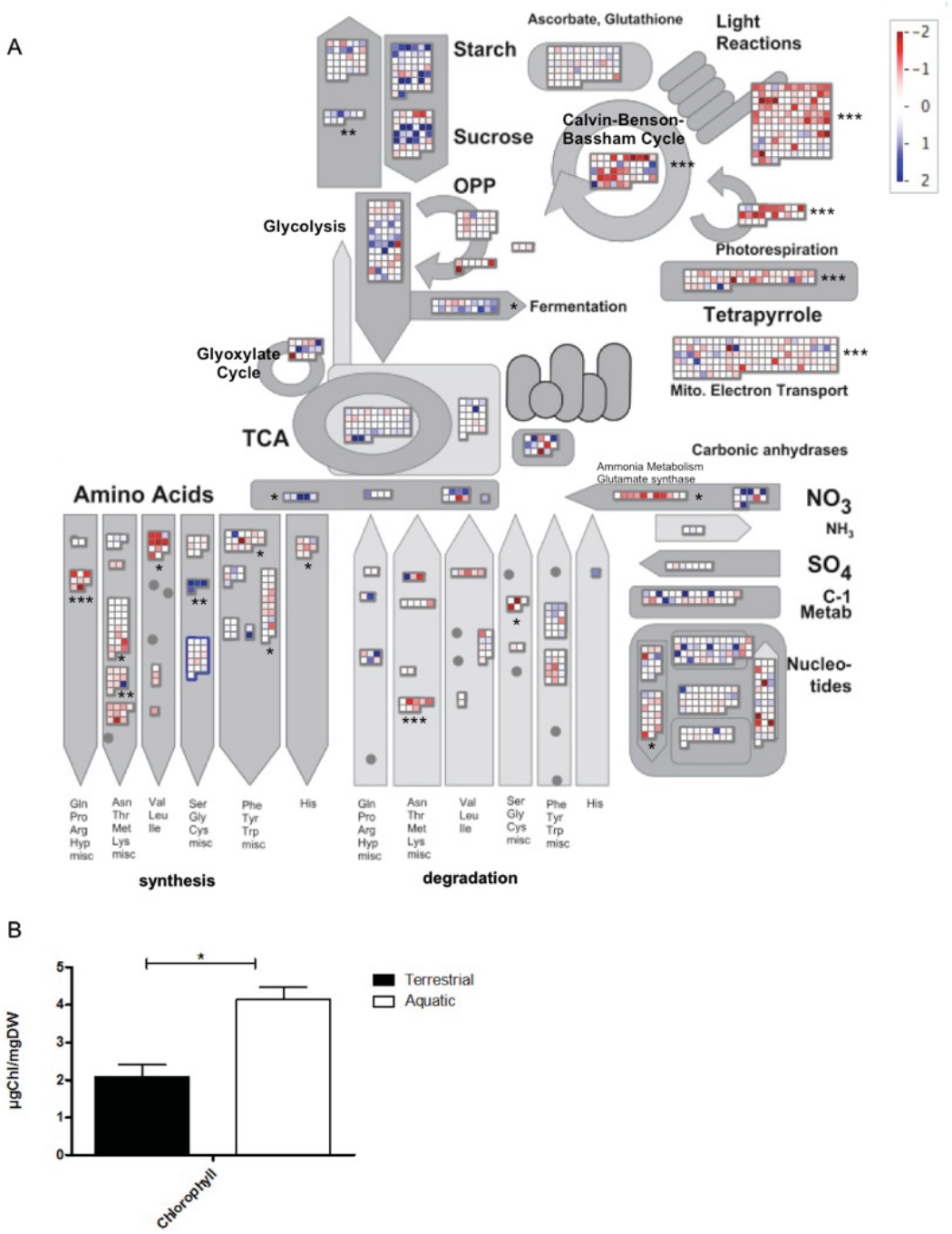


Figure 5. Transcriptional and physiological differences between aquatic and terrestrial *E. retroflexa* culms.

(A) Overview of central metabolism transcript levels (Mapman, TAIR10) in *E. retroflexa* culms. Heatmaps show log2 fold-changes of aquatic versus terrestrial transcript levels in RPM. Red (ratio<0) represents an increase of transcript levels in aquatic and blue (ratio>0) an increase of transcript levels in terrestrial culms. Asterisks indicate significant fold-changes. Wilcoxon Rank Test calculated significant fold-changes. *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 2006). **(B)** Total chlorophyll content in terrestrial (black) and aquatic 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 aquatic samples (**P*-value<0.05).

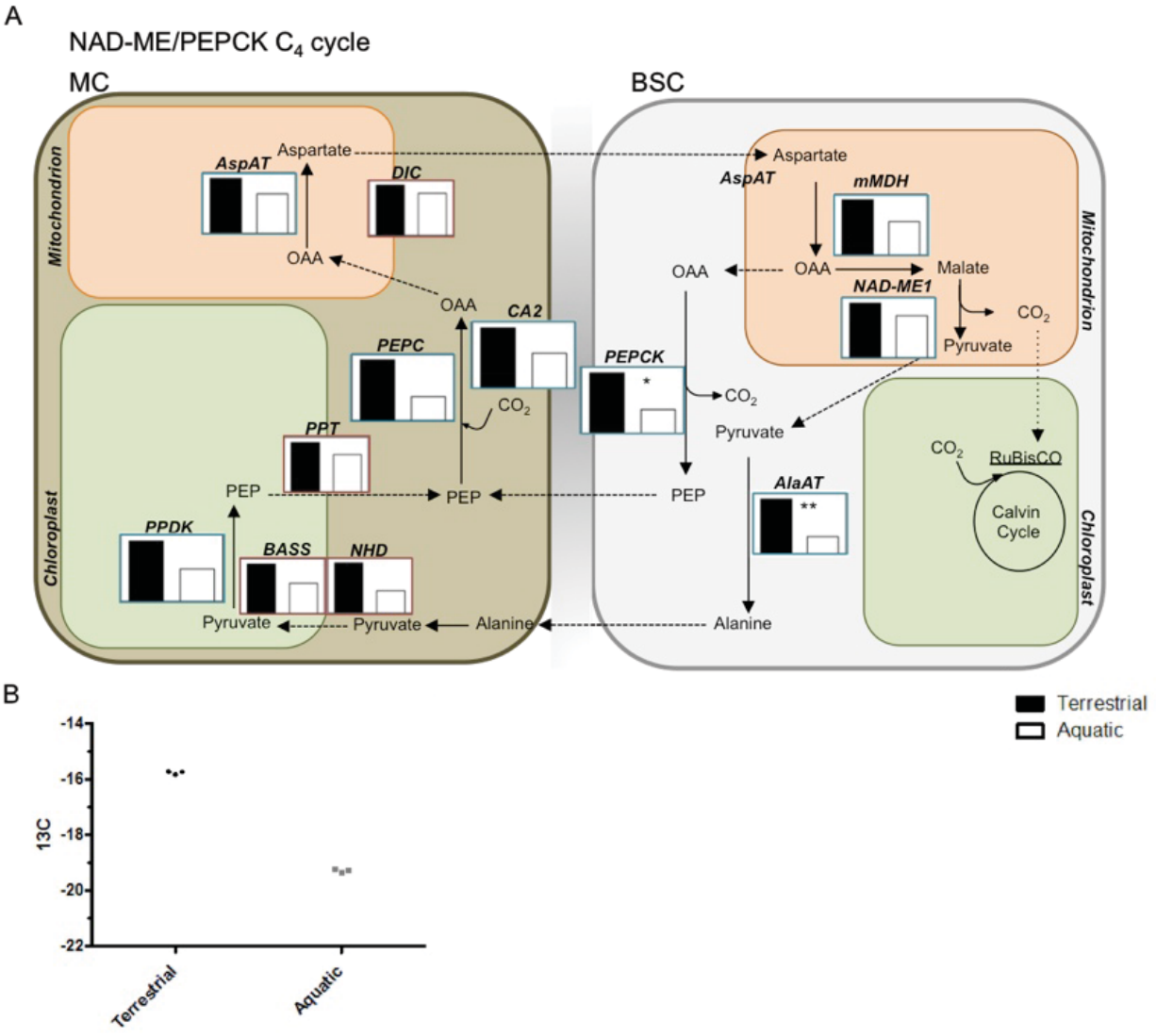


Figure 6. C_4 is altered between terrestrial and aquatic *E. retroflexa* culms.
(A) Schematic and simplified overview of the C_4 cycle known for NAD-ME/PEPCK subtype plants (adapted from Sommer et al. 2012). Relative transcript abundances between terrestrial (black) and aquatic (white) transcriptomes 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 BH 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***: *PHOSPHO*ENOL*PYRUVATE CARBOXYLASE*; ***CA2***: *CARBONIC ANHYDRASE2*; ***DIC***: *DICARBOXY -LATE CARRIER*; ***AspAT***: *ASPARTATE AMINOTRANSFERASE*; ***mMDH***: *mitochondrial MALATE DEHYDROGENASE*; ***NAD-ME1***: *NAD-dependent MALIC ENZYME1*; ***AlaAT***: *ALANINE AMINOTRANSFERASE*; ***PEPCK***: *PHOSPHO*ENOL*PYRUVATE CARBOXYKINASE*; ***BASS***: *BILE ACID:SODIUM SYMPORTER*; ***NHD***: *SODIUM:HYDROGEN ANTIPORTER*; ***PPDK***: *PYRUVATE ORTHOPHOSPHATE DIKINASE*; ***PPT***: *PHOSPHATE/PHOSPHO*ENOL*PYRUVATE TRANSLOCATO R* **(B)** $^{13}C/^{12}C$ isotope ratio of terrestrial (black) and aquatic (grey) culms; n=3.

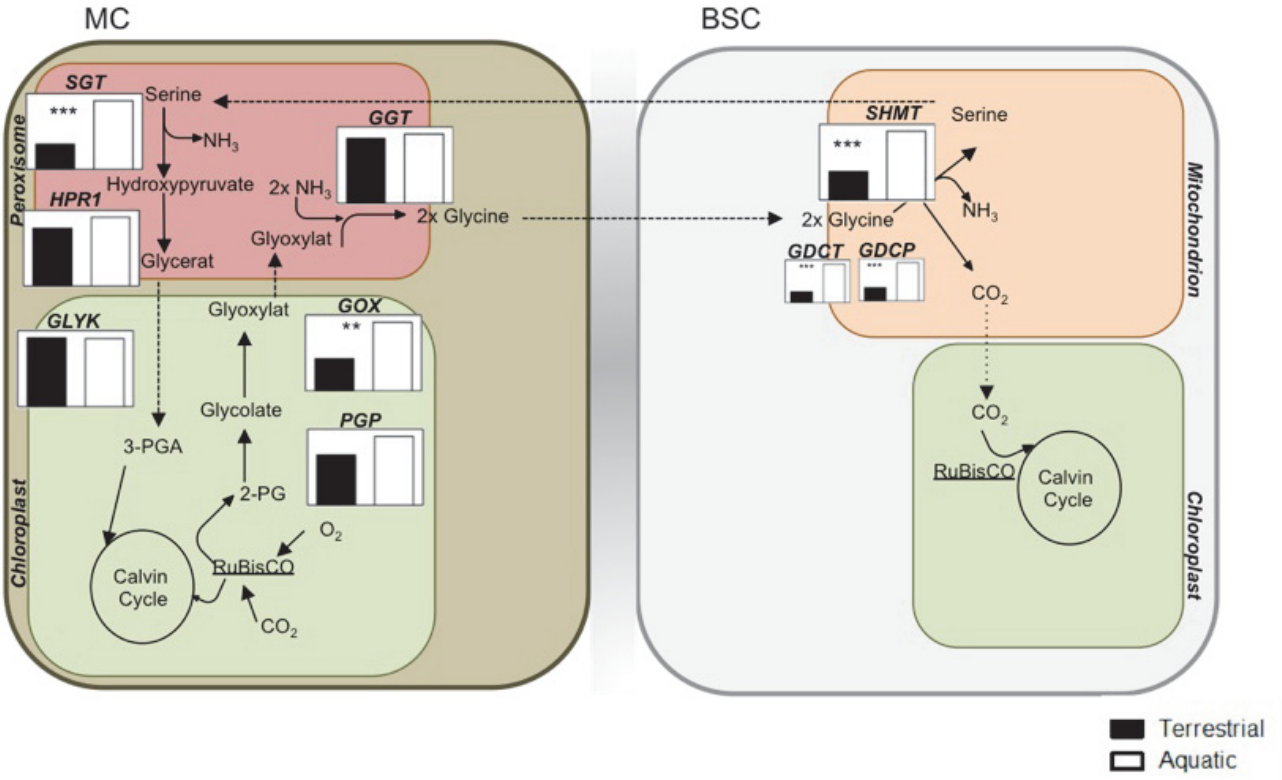


Figure 7. Photorespiration is enhanced in the aquatic culms.
Schematic overview of the photorespiratory pathway known for C_3-C_4 intermediate plants (adapted from Gowik et al. 2012). Relative transcript abundances between terrestrial (black) and aquatic (white) transcriptomes are shown in small insets 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 BH 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 HYDROXYMETHYLTRANSFERASE*; ***SGT*****:** *SERINE:GLYOXYLATE AMINOTRANSFERASE*; ***HPR1*****:** *NADH-dependent HYDROXYPYRUVATE REDUCTASE*; ***GLYK*****:** *GLYCERATE KINASE*

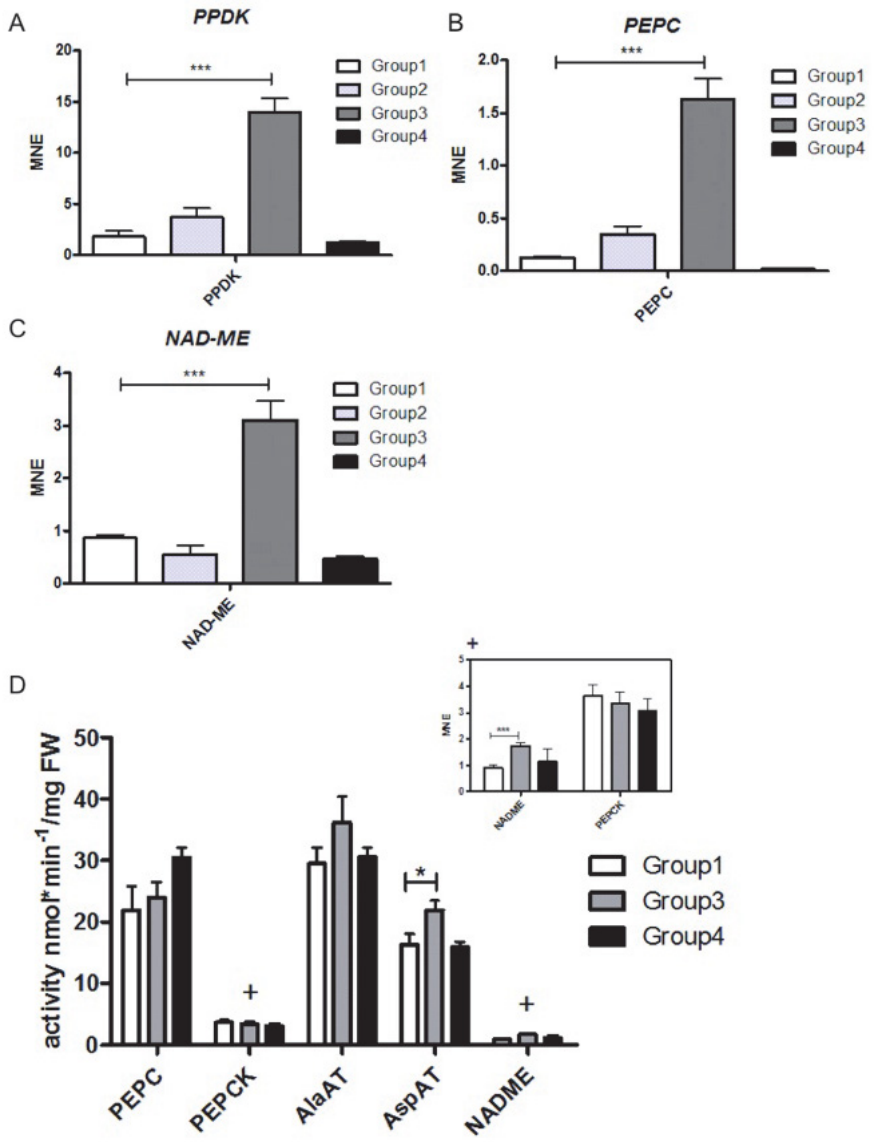


Figure 8. Analysis of metabolic plasticity of *E. retroflexa* terrestrial culms under increasing water deprivation.

(A-C) Transcriptional pattern of selected C_4 and photorespiratory genes in water deprived terrestrial *E. retroflexa*. Quantitative real-time PCR was performed with 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); Group2: every two days 250ml water (light grey); Group3: every four days 150ml water (dark grey); Group4: 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 terrestrial *E. retroflexa*. 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); Group3 every four days 150 ml water (dark grey); Group4: 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 insets.

Datasets

The following datasets are stored on the external medium enclosed to this thesis:

Dataset 1. Annotated transcriptome expression data (RPM) of *E. retroflexa* aquatic and terrestrial culms including EdgeR analysis of differentially expressed genes and annotation.

Dataset 2. CAP3 assembled filtered *E. retroflexa* unigene database.

Supplemental Material

Supplemental Datasets

The following supplemental datasets are stored on the external medium enclosed to this thesis:

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.

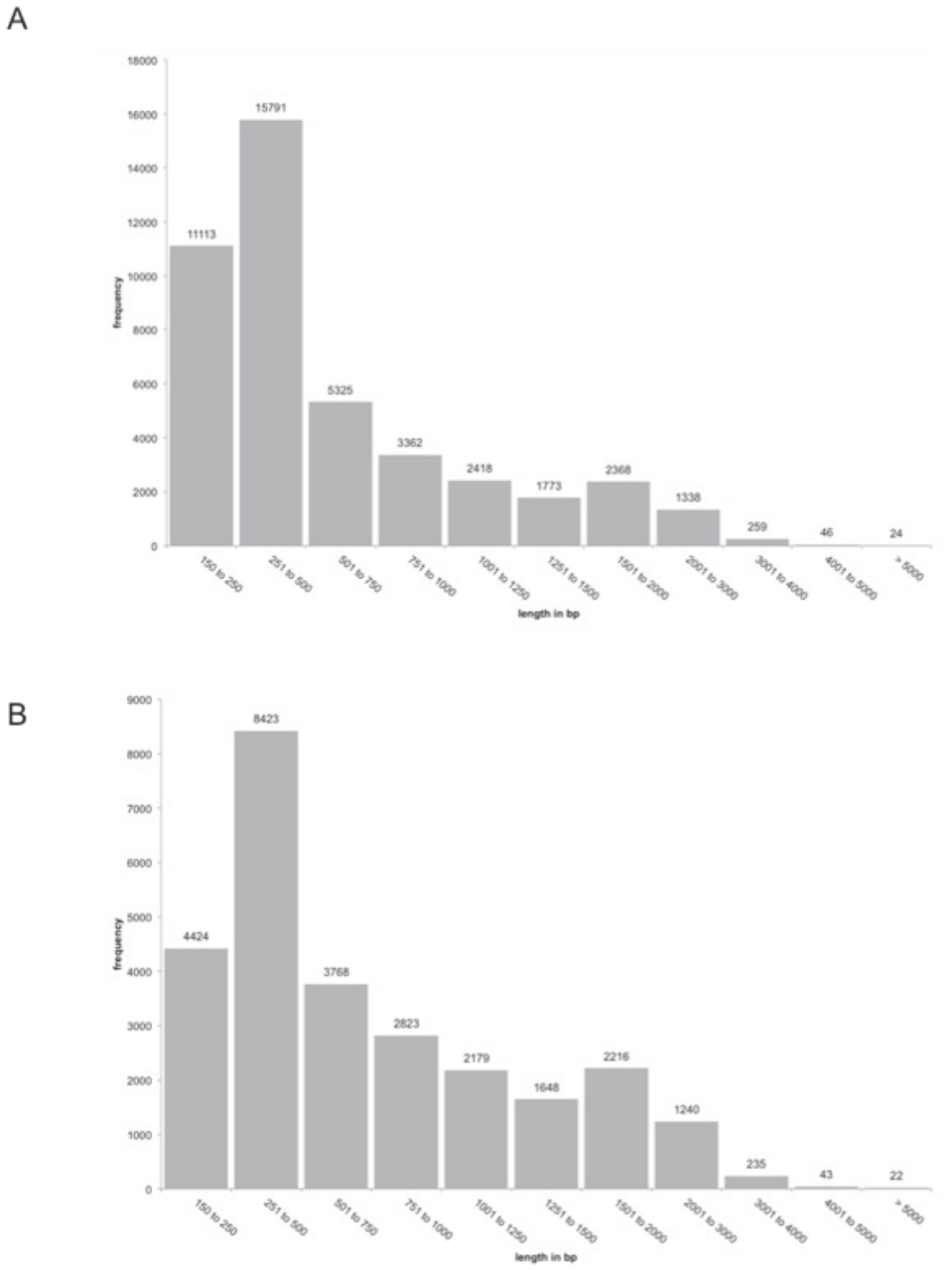
Supplemental Dataset 4. *E retroflexa* contigs annotated with Uniref100 via tblastx based on highest bitscore.

Supplemental Dataset 5. *E. retroflexa* contigs annotated against *S. italica* V2.1 primary transcripts.

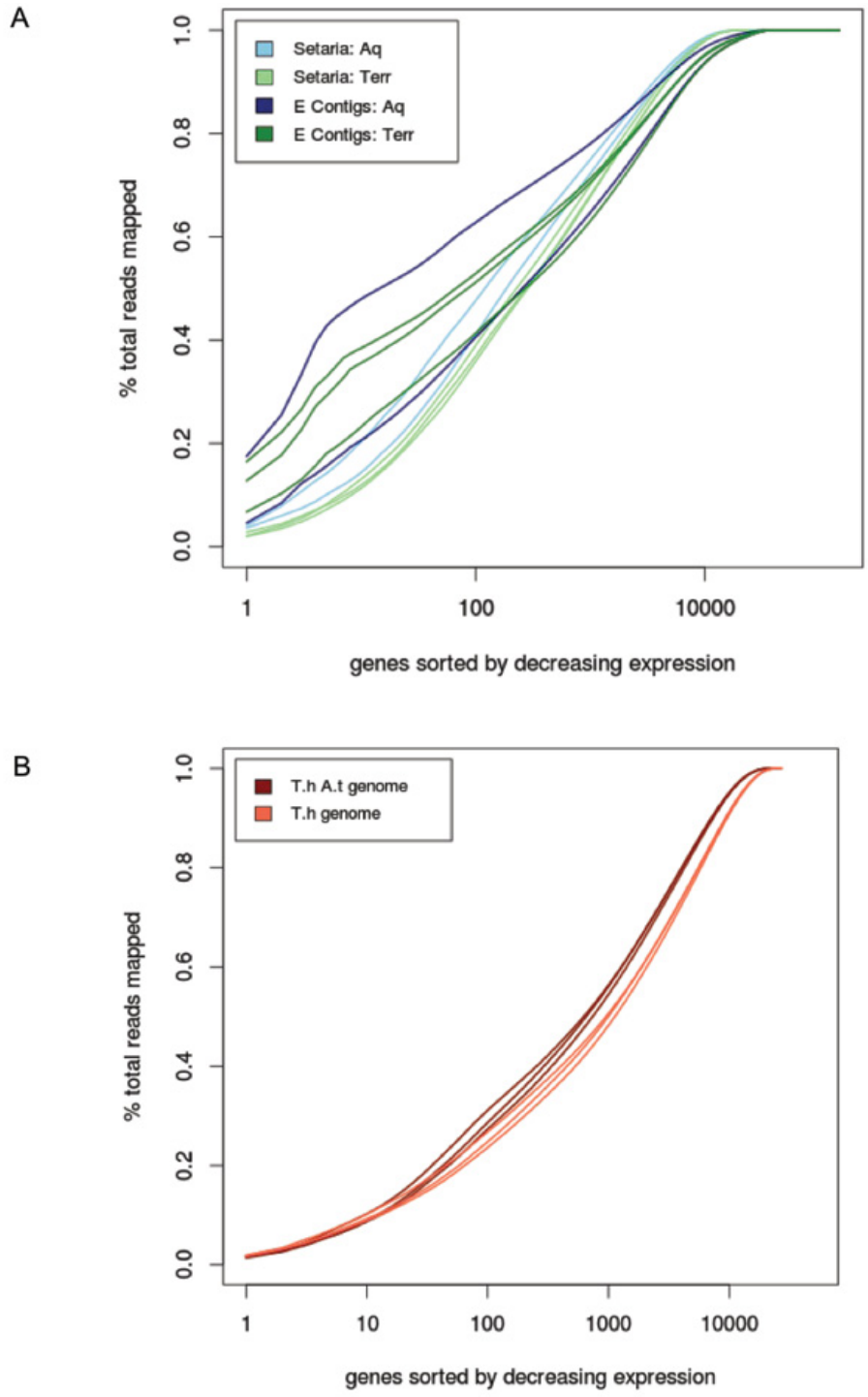
Supplemental Dataset 6. Mapman category enrichment analysis with Fisher's Exact test.

Supplemental Dataset 7. Wilcoxon rank sum test of Mapman categories.

Supplemental Figures

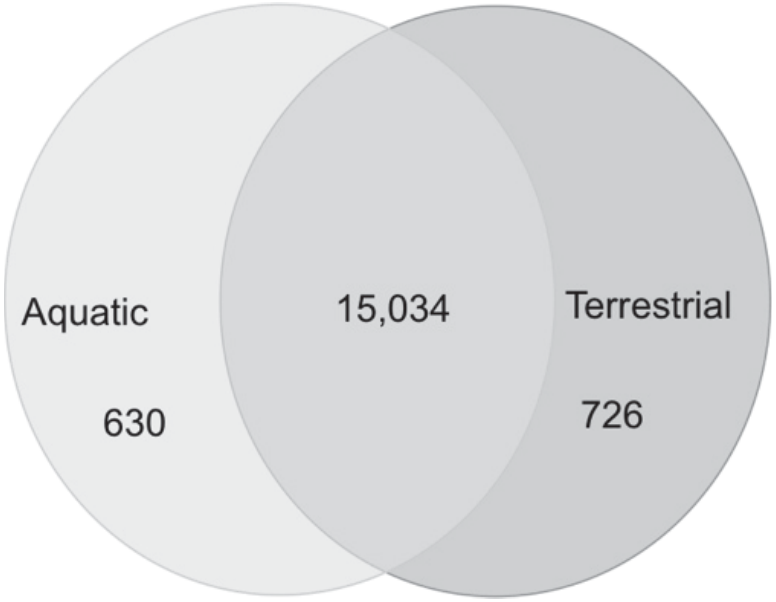


Supplemental Figure 1. Histograms of *E. retroflexa* unigene database.
(A) Frequency distribution of *E. retroflexa* contigs (43,817) with length over 200 bp assembled with CAP3. **(B)** Frequency distribution of filtered *E. retroflexa* contig length of contigs (27,021), that were annotated by either *S. italica* or UniRef100 as *Viridiplantae*.

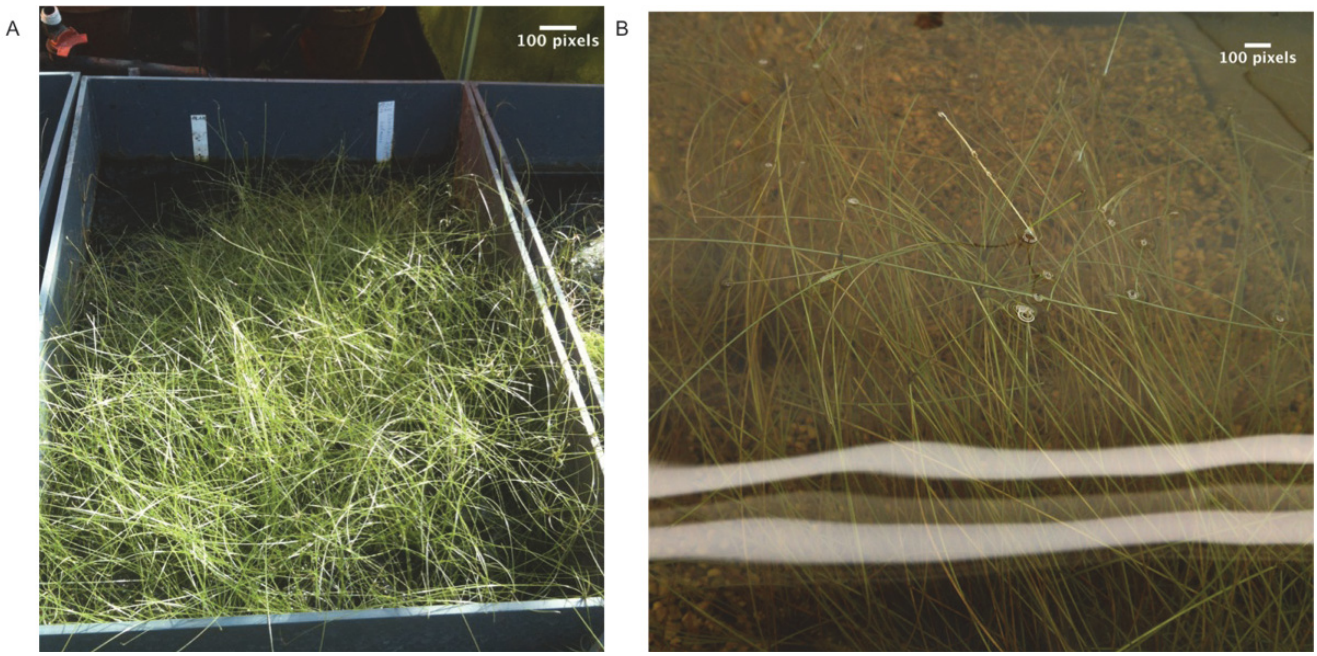


Supplemental Figure 2. Cumulative relative expression plots of reads mapped against various references by BLAT.

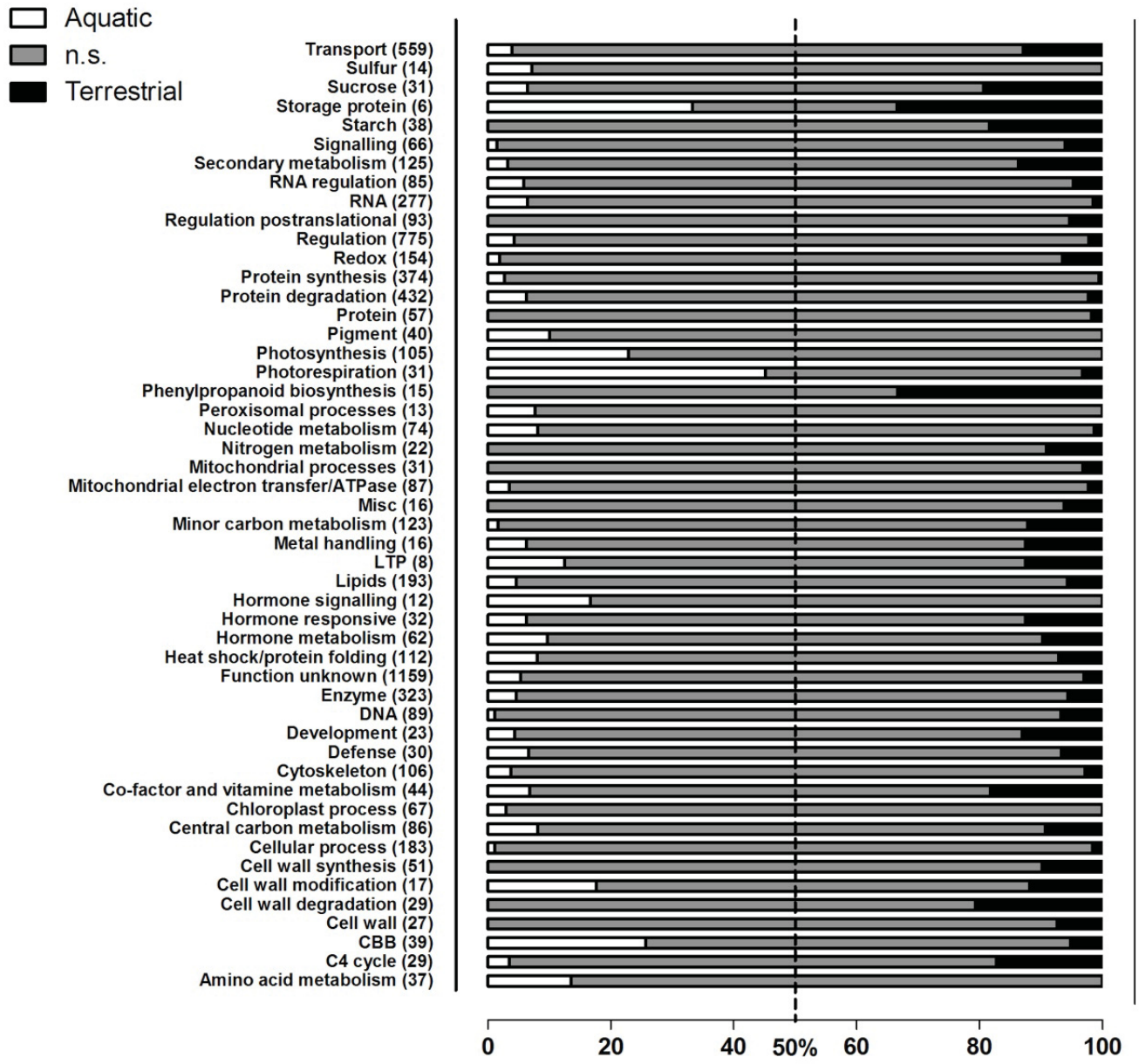
Raw count files were sorted descending by matched reads per contig/cds and the amount of reads relative to total mapped reads per sample was summed up. **(A)** aquatic (blue; Aq) and terrestrial (green; Terr) *E. retroflexa* reads mapped to *E. retroflexa* contigs (darker color; E contigs) or *S. italica (Setaria)* primary transcript (lighter color). **(B)** *Tarenaya hassleriana (T. h)* mature leaf reads mapped to the *T. hassleriana* primary transcripts (lighter color) and the *Arabidopsis thaliana (A. t)* reference (dark color).



Supplemental Figure 3. Comparison aquatic and terrestrial *E. retroflexa* transcriptomes.
Numbers indicate transcripts shared between averaged *E. retroflexa aquatic and terrestrial* transcriptomes (n=3, terrestrial, green; n=2, aquatic, blue) or significantly higher abundant. *P*-value<0.05; *P*-values were Benjamini-Hochberg FDR corrected.

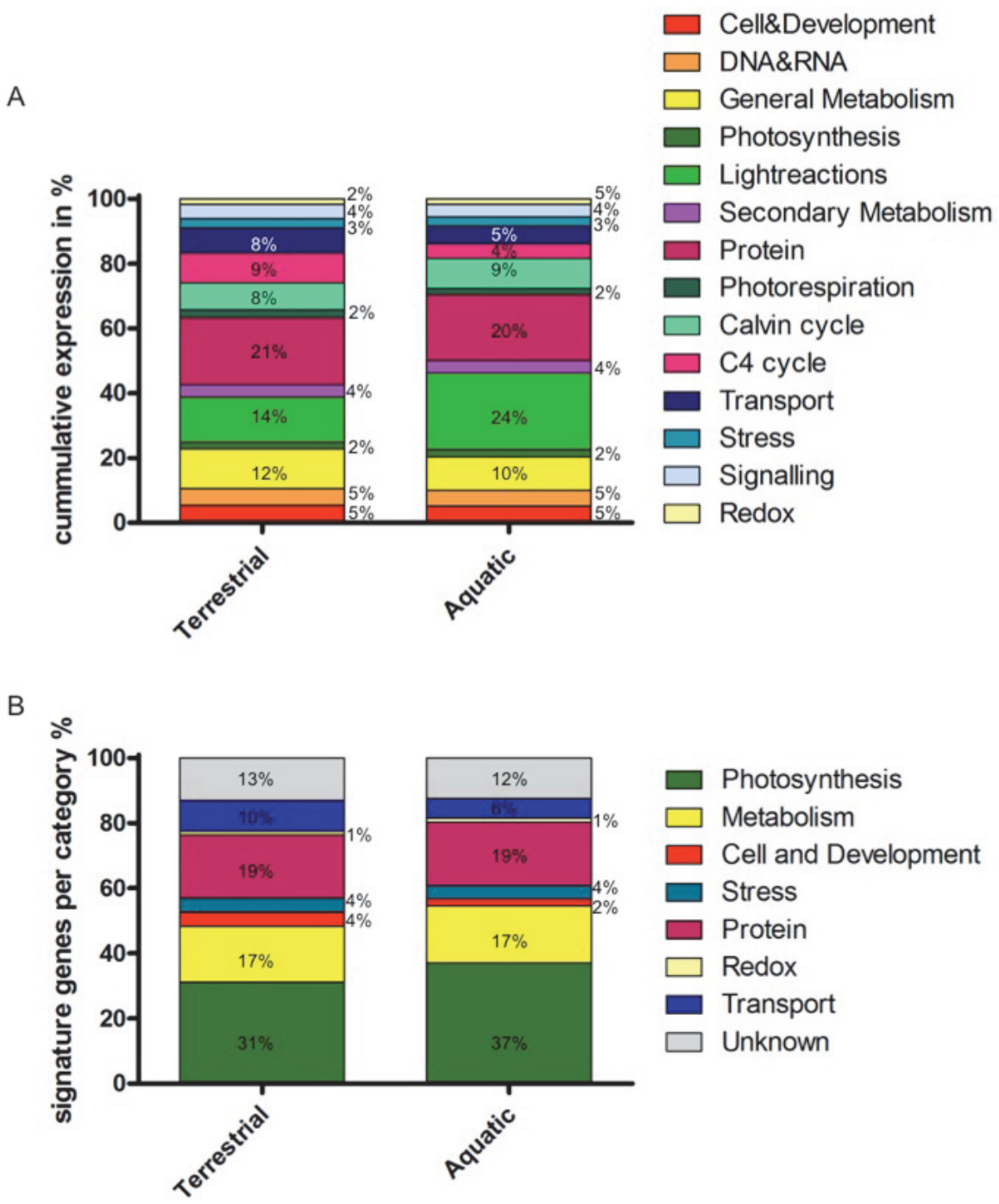


Supplemental Figure 4. Photographic images of *E. retroflexa* during cultivation.
(A) Phenotype of 6-week-old terrestrial *E. retroflexa* grown in swamp boxes. **(B)** Phenotype of submerged E. *retroflexa* plants grown in aquaria after 4 weeks of cultivation. Scale bar represents 100 pixels.



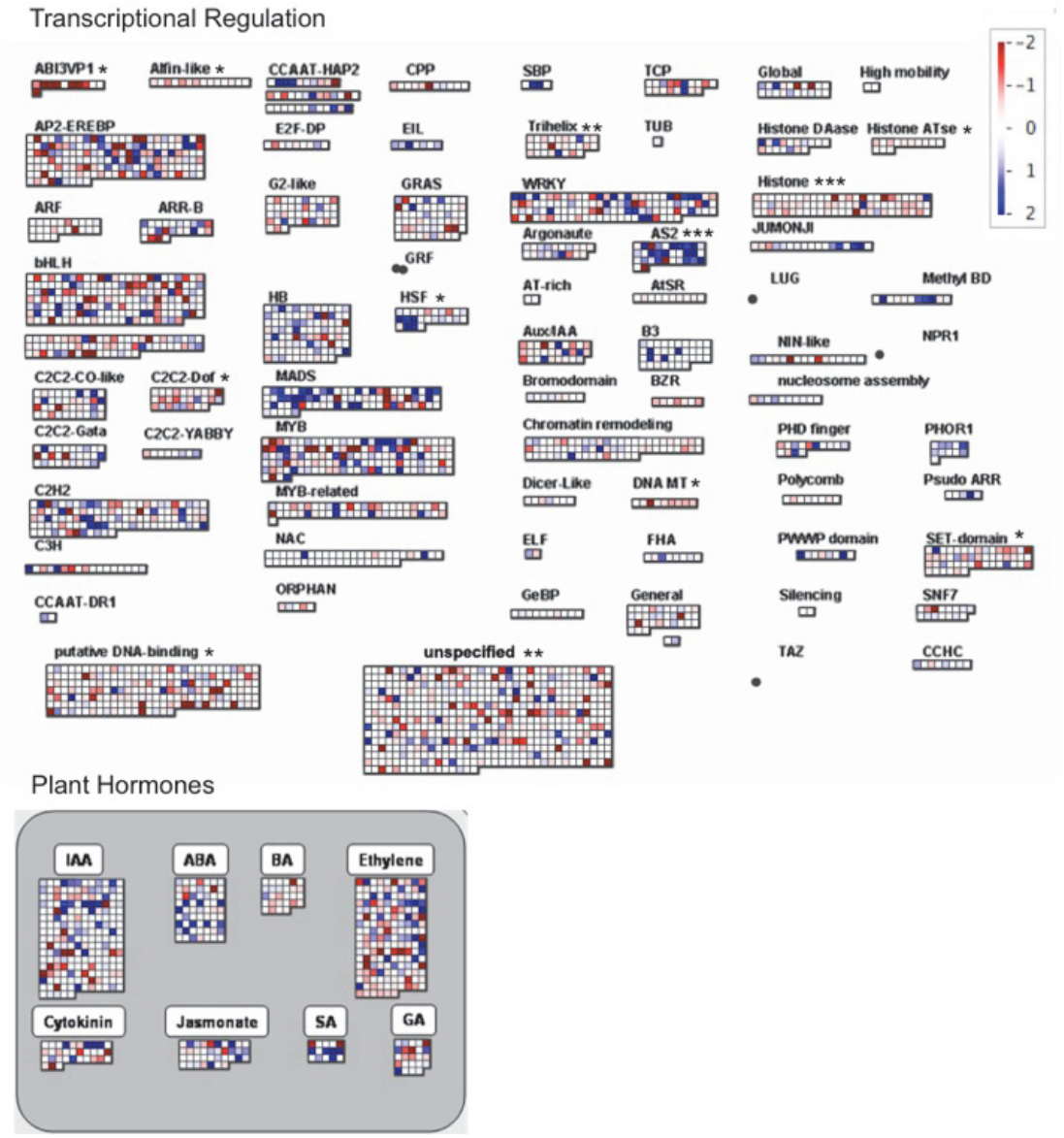
Supplemental Figure 5. Quantitative transcript abundance patterns between aquatic and terrestrial *E. retroflexa* culms.

Relative number of transcripts (in percent) that are significantly up-regulated (BH-corrected *P*-value<0.05) in either aquatic (white) or terrestrial (black) transcriptome per custom Mapman-derived category. Numbers of all changed genes per category are indicated in parentheses.



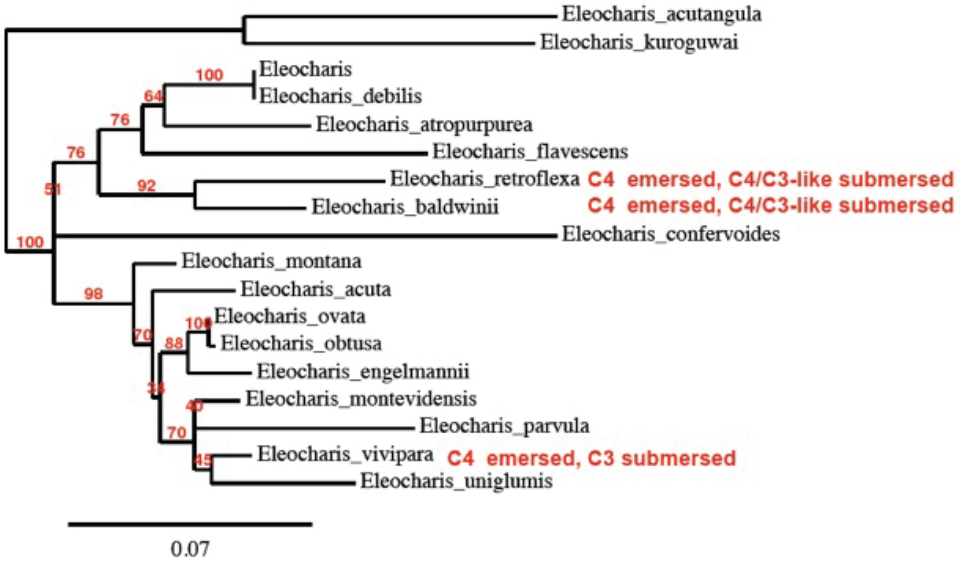
Supplemental Figure 6. Transcriptional investment of aquatic and terrestrial *E. retroflexa* culms.

(A) Cumulative average RPMs in percent of custom basal Mapman categories for each tissue in *E. retroflexa.* **(B)** Distribution signature genes in *E. retroflexa* terrestrial and aquatic culms. Percentage of signature genes expressed over 1,000 RPM falling in each basal Mapman category for every averaged tissue.



Supplemental Figure 7. Transcriptional regulation and plant hormone expression patterns in *E. retroflexa* culms grown in water and on soil.

Overview of gene expression dynamics of the Mapman categories plant hormones and transcriptional regulation in *E. retroflexa* culms. Heatmaps depict transcriptional log2 fold-changes aquatic versus terrestrial RPM. Red (ratio<0) represents an increase of gene expression in aquatic and blue (ratio>0) an increase of transcript accumulation in terrestrial culms. Significant fold-changes were calculated by Wilcoxon Rank Test and are indicated by asterisk. *P*-values (**P*-value<0.05; ***P*-value < 0.01; ****P*-value <0.001) were Benjamini-Yekutieli FDR corrected.



Supplemental Figure 8. Phylogeny of *E. retroflexa* based on internal transcribed spacer (ITS) sequences.
Phylogenetic tree is based on ITS sequence similarity. Alignments were performed with Clustal W (Larkin et al. 2007) and tree was calculated with PhyML. Red numbers indicate branch support by 100 times bootstrapping.

Supplemental Tables

Supplemental Table 1. Overview of *E. retroflexa* qRT-PCR primers. Primers were designed on basis of unigene database and optimized for 60°C annealing temperature. All sequences are displayed in 5'-3' orientation.

Target	contig ID	Forward Primer	Reverse Primer	length (bp)
PEPC1	contig_19711	CTCTCTCTTGTGCGCCTT	GCCATTCCTGACGCTTCT	127
NAD-ME2	contig_23316	CCTCTTTCCTTCCATATCTAGCATT	TCACATCACCATGCCCCTC	107
UBQ10	contig_15500	GAGGTTGATCTTTGCGGGT	GGGTTGACTCCTTCTGGATG	77
PPDK	contig_34949	ATTGGAGGGAAGGGGAGAT	GCAAAACACGACACAAAACAG	123

Supplemental Table 2. Overview of *E. retroflexa* CAP3 assembly statistics and annotation of contigs.

	Number of contigs
Total number of contigs from CAP3 assembly	149,303
Number of contigs > 200bp	43,817
Number of contigs with matching plant ID	**27,021**
N50 all contigs	432
N50 of contigs > 200bp	984
N50 of contigs with matching plant ID	**1,199**
Number of contigs aligning to *S. italica*	25,386
Number of *S.italica* genes matching at least 1 contig	13,204
Number of uniquely matched *S. italica* IDs by contigs	7,400
Number of contigs matching Uniref100 IDs	29,512
Number of contigs matching Viridiplantae IDs	21,832

Supplemental Table 3. Pearson's correlation (*r*) of transcriptome data. Aquatic (AQ) and terrestrial (TE) *E. retroflexa* culms mapped to *S. italica* of each individual replicate per condition and between averaged samples of *Tarenaya hassleriana* (T. has) and *Gynandropsis gynandra* (G. gyn) transcriptomes (Külahoglu et al. 2014).

Samples	Pearson's correlation *(r)*
AQ 1 vs 2	0.88
TE 1 vs 2	0.84
TE 1 vs 3	0.86
TE 2 vs 3	0.95
mean AQ vs TE	0.88
T. has leaf vs root	0.09
T. has leaf vs petal	0.11
T. has leaf vs G. gyn leaf	0.80

Supplemental Table 4. Averaged C_4 cycle transcripts (RPMs) of *E. retroflexa* aquatic (AQ) and terrestrial (TE) transcriptomes.

Name	mean AQ	mean TE
NADME	1898	1755
PEPC	9972	25622
PEPCK	235	593
PPDK	8593	15916
mMDH	625	1043
AlaAT	1407	4510
AspAT	1744	3835
CA2	2232	3847
DIC	842	1010
NHD	621	1424
BASS	2030	3329
PPT	644	853
TPT	768	1121

Acknowledgements

Work in the authors' laboratory was supported by grants of the Deutsche Forschungsgemeinschaft (EXC 1028, IRTG 1525, and WE 2231/9-1 to APMW). We are grateful to the HHU Biomedical Research Center (BMFZ) for support with RNA-Seq analysis and to the MSU High Performance Computing Cluster (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, performed CAP3 assembly and bioinformatics data analysis; 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.

References

Agarie, S., Kai, M., Takatsuji, H. and 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. and Hatch, M. (1990) Enzymes of C_4 photosynthesis. In 'Methods in Plant Biochemistry, Volume 3,.(Ed. PJ Lea) pp. 39-72. Academic Press: San Diego, CA.

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

Bailey, S., Horton, P. and 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., et al. (2005) The universal protein resource (UniProt). *Nucleic Acids Research* 33: D154-D159.

Bennetzen, J.L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A.C., et al. (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. and 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., et al. (2010) Manipulation of FASTQ data with Galaxy. *Bioinformatics* 26: 1783-1785.

Boeger, M.R.T. and 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. and 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. and 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., et al. (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. and 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. and 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. and 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. and Farquhar, G.D. (2013) Environmental and physiological determinants of carbon isotope discrimination in terrestrial plants. *New Phytol* 200: 950-965.

Chen, T., Zhu, X.-G. and 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.

Chen, T.Y., Ye, R.J., Fan, X.L., Li, X.H. and Lin, Y.J. (2012) Identification of C_4 photosynthesis metabolism and regulatory-associated genes in *Eleocharis vivipara* by SSH. *Photosynthesis Research* 112: 215-215.

Cheng, S., van den Bergh, E., Zeng, P., Zhong, X., Xu, J., Liu, X., et al. (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. and 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. and 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., et al. (2006) New guidelines for delta C-13 measurements. *Anal. Chem.* 78: 2439-2441.

De Vos, M., Van Oosten, V.R., Van Poecke, R.M., Van Pelt, J.A., Pozo, M.J., Mueller, M.J., et al. (2005) Signal signature and transcriptome changes of *Arabidopsis* during pathogen and insect attack. *Mol Plant Microbe Interact* 18: 923-937.
Edwards, G.E. and Ku, M.S. (1987) Biochemistry of C_3-C_4 intermediates. *New York: Academic Press* In M.D. Hatch & N.K. Boardmann (Eds.): pp.275-325.
Frost-Christensen, H., JØRgensen, L.B. and 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. and Sand-Jensen, K. (1995) Comparative kinetics of photosynthesis in floating and submerged *Potamogeton* leaves. *Aquatic Botany* 51: 121-134.
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 Physiology* 85: 958-964.
Ge, X., Chen, H., Wang, H., Shi, A. and Liu, K. (2014) De novo assembly and annotation of *Salvia splendens* transcriptome using the Illumina platform. *PLoS ONE* 9: e87693.
Gowik, U., Bräutigam, A., Weber, K.L., Weber, A.P.M. and Westhoff, P. (2011) Evolution of C_4 Photosynthesis in the Genus Flaveria: How Many and Which Genes Does It Take to Make C_4? *Plant Cell* 23: 2087-2105.
Haberlandt, G. (1904) Physiologische pflanzenanatomie. W. Engelmann.
Hatch, M.D. (1987) C_4 photosynthesis – a uniqe blend of modified biochemistry, anatomy and ultrastructure. *Biochimica Et Biophysica Acta* 895: 81-106.
Hatch, M.D. and Slack, C.R. (1970) Photosynthetic Co2-Fixation Pathways. *Ann Rev Plant Physio* 21: 141-&.
Heckmann, D., Schulze, S., Denton, A., Gowik, U., Westhoff, P., Weber, A.P., et al. (2013) Predicting C_4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell* 153: 1579-1588.
Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6: 65-70.
Holmes, M. and Klein, W. (1987) The light and temperature environments. *SPEC. PUBL. BR. ECOL. SOC. 1987.*
Huang, X.Q. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Research* 9: 868-877.
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.
Jackson, M.B. (2008) Ethylene-promoted elongation: an adaptation to submergence stress. *Ann. Bot.* 101:229–48
Jahnke, L.S., Eighmy, T.T. and Fagerberg, W.R. (1991) Studies of Elodea nuttalii grown under photorespirator conditions. 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. and 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. and 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., et al. (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. and 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., et al. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* 40: D1202-1210.
Langdale, J.A. and 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., et al. (2007) Clustal W and clustal X version 2.0. *Bioinformatics* 23: 2947-2948.

Lee, S.C., Mustroph, A., Sasidharan, R., Vashisht, D., Pedersen, O., Oosumi, T., et al. (2011) Molecular characterization of the submergence response of the *Arabidopsis thaliana* ecotype Columbia. *New Phytol* 190: 457-471.
Lee, W.Y., Lee, D., Chung, W.-I. and 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. and 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. and Madsen, T.V. (2002) Freshwater angiosperm carbon concentrating mechanisms: processes and patterns. *Functional Plant Biology* 29: 393-405.
Maberly, S.C. and Spence, D.H.N. (1989) Photosynthesis and photorespiration in fresh-water organisms – amphibious plants. *Aquatic Botany* 34: 267-286.
Madsen, T.V. and Sandjensen, 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., et al. (2014) The role of photorespiration during the evolution of C_4 photosynthesis in the genus Flaveria. *Elife*: e02478.
Martin, J.A. and Wang, Z. (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics* 12: 671-682.
McElwain, E.F., Bohnert, H.J. and 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. and 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. and 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. and 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. and 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. and 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., et al. (2009) Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis. *Proc Natl Acad Sci U S A* 106: 18843-18848.
Papanicolaou, A., Stierli, R., Ffrench-Constant, R.H. and Heckel, D.G. (2009) Next generation transcriptomes for next generation genomes using est2assembly. *BMC Bioinformatics* 10: 447.
Pedersen, O., Colmer, T.D. and Sand-Jensen, K. (2013) Underwater photosynthesis of submerged plants - recent advances and methods. *Front Plant Sci* 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. and Kriedemann, P.E. (1989) Determination of accurate extinction coefficients and simultaneous-equations for asssaysing 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. and 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. and 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. and 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., et al. (2006) TM4 microarray software suite. In *DNA Microarrays, Part B: Databases and Statistics*. Edited by Kimmel, A. and Oluver, B. pp. 134-93.
Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., et al. (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. and McKown, A.D. (2006) Is C_4 photosynthesis less phenotypically plastic than C_3 photosynthesis? *Journal of Experimental Botany* 57: 303-317.
Sand-Jensen, K. (1989) Environmental variables and their effect on photosynthesis of aquatic plant communities. *Aquatic Botany* 34: 5-25.
Sand-Jensen, K. and 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. and Brautigam, A. (2012) RNA-Seq Assembly - Are We There Yet? *Front Plant Sci* 3: 220.
Schmittgen, T.D. and Livak, K.J. (2008) Analyzing real-time PCR data by the comparative C-T method. *Nat Protoc* 3: 1101-1108.
Schulze, S., Mallmann, J., Burscheidt, J., Koczor, M., Streubel, M., Bauwe, H., et al. (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. and 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. and 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., et al. (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. and 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. and 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. and Miyachi, S. (1988) Photosynthetic characteristics of an amphibious plant, *Eleocharis vivipara*-Expression of C_4 and C_3 modes in contrasting environments. *Proc Natl Acad Sci U S A* 85: 6733-6737.
Ueno, O. and Wakayama, M. (2004) Cellular expression of C_3 and C_4 photosynthetic enzymes in the amphibious sedge Eleocharis retroflexa ssp. chaetaria. *J Plant Res* 117: 433-441.
Usadel, B., Nagel, A., Thimm, O., Redestig, H., Blaesing, O.E., Palacios-Rojas, N., et al. (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.W.P.M. and 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, 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 C_4 Kranz anatomy. *Plant Journal* 75: 656-670.
Waters, M.T., Wang, P., Korkaric, M., Capper, R.G., Saunders, N.J. and Langdale, J.A. (2009) GLK transcription factors coordinate expression of the photosynthetic apparatus in Arabidopsis. *The Plant Cell* 21: 1109-1128.
Yekutieli, D. and 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.

VII. CONCLUSION AND OUTLOOK

Photosynthesis enables plants to fix inorganic carbon namely CO_2 in the form of sugars by using energy from sunlight, increasing carbon sequestration and storage of ecosystems. One especially efficient form of photosynthetic energy conversion is C_4 photosynthesis, which in comparison to C_3 photosynthesis facilitates plants to achieve higher photosynthetic capacity, higher water usage and nitrogen use efficiencies in high light, arid and warm environments (Black, 1973; Ruan et al., 2012). There are relatively few C_4 plant species (7,500) compared to C_3 plants (250,000); however, C_4 performing plants account for around a quarter of primary carbon fixation on our planet (Still et al., 2003; Edwards et al., 2010; Sage and Zhu, 2011). Due to the high photosynthetic efficiency of C_4 plants, there has been a major effort to reproduce the trait of C_4 photosynthesis in C_3 crop plants (von Caemmerer et al., 2012). For re-engineering the C_4 trait in C_3 plants, mechanistic understanding of the parts that make up the system is required (Denton et al., 2013). The first step towards rebuilding a trait is to fully understand all components required, just like a genomic and transcriptomic blueprint. With the emergence of next-generation sequencing technologies, genomic and transcriptional comparisons are possible on a large scale. Thus, cell specific and leaf developmental gradients between closely related C_3 and C_4 species are being employed to identify C_4 specific gene regulation and how "C_4-ness" is established in the leaf. We developed a workflow for RNA-seq laboratory practices and data analysis (**V.1 FIRST-AUTHORED MANUSCRIPTS:** *Manuscript 1*; Külahoglu and Bräutigam, 2014), which was applied for the data sets presented.

Within the scope of this PhD thesis, we took a systems biological approach for detailed transcriptome comparisons within the Cleomaceae and Cyperaceae families. The species were chosen to facilitate transcriptome-level comparisons of the C_4 trait with other photosynthetic traits. The C_3 and C_4 species in the Cleomaceae are closely related to each other and to the model plant *A. thaliana* (Marshall et al., 2007); while the analyzed Cyperaceae species is able to change its photosynthetic mode within one life cycle and switch from C_4 to C_4-like photosynthesis (Ueno and Wakayama, 2004). *Manuscript 2* (**V.2 FIRST-AUTHORED MANUSCRIPTS**) incorporates a detailed comparison of developmental transcriptome atlases in *G. gynandra* (C_4) and *T. hassleriana* (C_3), with physiological data (Külahoglu et al., 2014). The analysis of the core C_4 enzyme expression patterns across heterotrophic and phototrophic tissues between C_4 and C_3 species supports the hypothesis that the C_4 relevant genes were recruited to the C_4 cycle on a small scale or in a modular fashion from a variety of expression domains in the C_3 plant. These results render the existence of a C_4 master regulator highly unlikely. However, the C_4 cycle gene expression seems to be co-regulated together with photosynthesis in *G.gynandra*, as it has been observed in *Z. mays*. (Pick et al., 2011). The bundle sheath of the leaf and the root

endodermis, the mesophyll and the root cortex, have been proposed as analogous cell types (reviewed by Slewinski, 2013). A set of genes, which are exclusively expressed in the C_3 root of *T. hassleriana*, have been recruited in the C_4 species from the root to the mature leaf of *G. gynandra*. This finding supports the hypothesis that the scarecrow (SCR) and shortroot (SHR) regulatory network has been coopted into developing C_4 leaves from the root expression domain (Slewinski et al., 2012; Wang et al., 2013). In *Manuscript 2*, two key features of Kranz anatomy could be connected by integration of expression and anatomical data. We could associate the larger bundle sheath cell size in C_4 species with higher ploidy levels in these cells, possibly related to extended endocycling. Further, leaf differentiation is generally delayed in the C_4 species *G. gynandra*, which was connected to a delay in photosynthetic gene, chloroplast biogenesis, developmental factors and distinct cell cycle gene expression. Since mesophyll differentiation has been shown to end vascular tissue initiation in *A. thaliana* (Scarpella et al., 2004), this delay is contributing to the higher vein density through longer initiation of vein orders in *G. gynandra*. Similarly, in the C_4 monocotyledon *Setaria viridis* the vein order formation takes longer compared to the C_3 species rice (Fouracre et al., 2014), leading to denser venation in *S. viridis*. We hypothesize that the observed delay in mesophyll and chloroplast differentiation could be a consequence of extended repression of photomorphogenesis by the COP9 signalosome in the developing C_4 leaf.

Manuscript 3 (**VI. ADDENDUM**) gives a detailed analysis for understanding the transcriptional adaptability and components of the C_4 trait by comparing *Eleocharis retroflexa* transcriptomes grown under water and on soil. These plants are able to adjust their photosynthesis type depending on environmental conditions (Ueno and Wakayama, 2004). Using the *Eleocharis* system we could analyze the C_4 trait and C_3-C_4 intermediate traits within one species, focusing only on the gene expression profiles changed through the mode of photosynthesis and environment, without the evolutionary “noise” that can hamper interspecies comparisons. *E. retroflexa* culms are surprisingly flexible in using the C_4 cycle and are able to rapidly react to micro-environmental changes, *e.g.* water deprivation and drought. As it has been shown before, abscisic acid signaling plays a major role in terrestrial C_4 establishment in the genus *Eleocharis* (Ueno, 1997, 1998). While submerged *E. retroflexa* culm transcriptomes reflect characteristics known for flooding tolerant plants, the carbon metabolism displays a typical C_3-C_4 intermediate signature, with up-regulation of photorespiratory transcripts and maintenance of the C_4 cycle (Sage et al., 2012). Thus, *E. retroflexa* represents an interesting model for studying the molecular mechanisms of adaption in highly dynamic environments. This species contrasts the broad observation that C_4 species display less plasticity in phenotype and metabolism, which is due to their high level of anatomical and metabolic specialization (Sage and McKown, 2006).

Prior studies in the C_4 monocotyledon species *Z. mays* have analyzed anatomy and transcriptional dynamics during C_4 leaf development and trait establishment (Li et al., 2010; Pick et al., 2011; Chang et al., 2012; Wang et al., 2013). These earlier studies helped to understand the C_4 cycle integration and also lead to the isolation of new C_4 cycle relevant transporter and potential regulators. A recent study featured leaf maturation between two independent origins of C_4 photosynthesis – *G. gynandra* and *Z. mays*. This different approach revealed that the C_4 trait encompasses – besides parallel evolution of *cis*-elements and gene evolution – also homologous transcription factors following analogous temporal and spatial patterns in independent C_4 lineages (Aubry et al., 2014). Hence, comparative transcriptomics linked with physiological and anatomical data are providing new insights into C_4 and C_3 specific leaf ontology and give first hints on how Kranz anatomy might be regulated.

Comparative transcriptome studies within one species under varying conditions (*Manuscript 3*), between closely (*Manuscript 2*) or even more distantly related species can shed light on the molecular dynamics of the highly complex adaptive trait of C_4 photosynthesis. A subsequent step to these systemic analyses is the molecular and biochemical analysis of potential C_4 leaf *trans*-acting regulatory candidates for their function in C_4 photosynthesis establishment and maintenance. Further, future comparative analyses of closely related C_3 and C_4 species' genomes could elucidate, which photosynthetic *cis*-regulatory elements and genetic signatures are needed for C_4 evolution, and understand how these elements orchestrate C_4 specific expression modules leading to a functional C_4 plant.

References

Aubry, S., Kelly, S., Kumpers, B.M., Smith-Unna, R.D., and Hibberd, J.M. (2014). Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis. PLoS Genet **10,** e1004365.

Black, C.C. (1973). Photosynthetic carbon fixation in relation to net CO_2 uptake. Annual Review of Plant Physiology and Plant Molecular Biology **24,** 253-286.

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 **160,** 165-177.

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

Edwards, E.J., Osborne, C.P., Stroemberg, C.A.E., Smith, S.A., Bond, W.J., Christin, P.-A., Cousins, A.B., Duvall, M.R., Fox, D.L., Freckleton, R.P., Ghannoum, O., Hartwell, J., Huang, Y., Janis, C.M., Keeley, J.E., Kellogg, E.A., Knapp, A.K., Leakey, A.D.B., Nelson, D.M., Saarela, J.M., Sage, R.F., Sala, O.E., Salamin, N., Still, C.J., Tipple, B., and Consortium, C.G. (2010). The Origins of C_4 Grasslands: Integrating Evolutionary and Ecosystem Science. Science **328,** 587-591.

Fouracre, J.P., Ando, S., and Langdale, J.A. (2014). Cracking the Kranz enigma with systems biology. Journal of Experimental Botany **65,** 3327-3339.

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. (2014). Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C3 and C4 Plant Species. The Plant Cell **26,** 3243-3260.

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,** 1060-1067.

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. The Plant Journal **51,** 886-896.

Pick, T.R., Braeutigam, A., Schlueter, 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 C-4 Model and Provides Candidates for Regulation. Plant Cell **23,** 4208-4220.

Ruan, C.-J., Shao, H.-B., and da Silva, J.A.T. (2012). A critical review on the improvement of photosynthetic carbon assimilation in C_3 plants using genetic engineering. Critical Reviews in Biotechnology **32,** 1-21.

Sage, R.F., and 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., and Zhu, X.-G. (2011). Exploiting the engine of C_4 photosynthesis. Journal of Experimental Botany **62,** 2989-3000.

Sage, R.F., Sage, T.L., and Kocacinar, F. (2012). Photorespiration and the Evolution of C_4 Photosynthesis. In Annual Review of Plant Biology, Vol 63, S.S. Merchant, ed, pp. 19-47.

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.

Still, C.J., Berry, J.A., Collatz, G.J., and DeFries, R.S. (2003). Global distribution of C-3 and C-4 vegetation: Carbon cycle implications. Global Biogeochemical Cycles **17**.

Ueno, O. (1997). Induction of Kranz anatomy and C_4-like biochemical traits in the submerged form of the amphibious sedge, *Eleocharis vivipara* by abscisic acid. Plant Physiology **114,** 1066-1066.

Ueno, O. (1998). Induction of Kranz anatomy and C_4-like biochemical characteristics in a submerged amphibious plant by abscisic acid. Plant Cell **10,** 571-583.

Ueno, O., and Wakayama, M. (2004). Cellular expression of C_3 and C_4 photosynthetic enzymes in the amphibious sedge *Eleocharis retroflexa* ssp. chaetaria. J Plant Res **117,** 433-441.

von Caemmerer, S., Quick, W.P., and Furbank, R.T. (2012). The Development of C_4 Rice: Current Progress and Future Challenges. Science **336,** 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 C_4 Kranz anatomy. Plant Journal **75,** 656-670.

VIII.1 CO-AUTHORED MANUSCRIPTS

Manuscript 4

The *Tarenaya hassleriana* Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers

*Shifeng Cheng, Erik van den Bergh, Peng Zeng, Xiao Zhong, Jiajia Xu, Xin Liu, Johannes Hofberger, Suzanne de Bruijn, Amey S. Bhide, **Canan Külahoglu**, Chao Bian, Jing Chen, Guangyi Fan, Kerstin Kaufmann, Jocelyn C. Hall, Annette Becker, Andrea Bräutigam, Andreas P.M. Weber, Chengcheng Shi, Zhijun Zheng, Wujiao Li, Mingju Lv, Yimin Tao, Junyi Wang, Hongfeng Zou, Zhiwu Quan, Julian M. Hibberd, Gengyun Zhang, Xin-Guang Zhu, Xun Xu, and M. Eric Schranz*

Published in Plant Cell (2013) 25: 8, pp. 2813–2830
doi: 10.1105/tpc.113.113480

Impact Factor: 10.65

Co-author

Main findings:

This study provides the genome of the *Cleomaceae* species *Tarenaya hassleriana.* Comparative analysis of sister lineage genomes reveals, that genome evolution by polyploidization and gene duplication has an effect on reproductive traits. *Tarenaya* has undergone an ancient genome triplication, which is independent of the Brassicaceae-specific genome duplication in *Arabidopsis thaliana* and nested triplication in *Brassica*. The Brassicaceae lineage retained twice as many floral relevant genes as *Tarenaya*, which likely lead to the morphological diversity in *Brassica*. Furthermore, the *T. hassleriana* genome provides a resource for future research of the Cleomaceae as well as the evolution of the Brassicaceae genomes.

Contributions:

- Providing transcriptome data of eight different *T. hassleriana* tissues and mapping it to genome
- Principle component analysis of transcriptomes
- Proof-reading of manuscript

The Plant Cell, Vol. 25: 2813–2830, August 2013, www.plantcell.org

LARGE-SCALE BIOLOGY ARTICLE

The *Tarenaya hassleriana* Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers[W][OPEN]

Shifeng Cheng,[a,1] Erik van den Bergh,[b,1] Peng Zeng,[a] Xiao Zhong,[a] Jiajia Xu,[c] Xin Liu,[a] Johannes Hofberger,[b] Suzanne de Bruijn,[d,e] Amey S. Bhide,[f] Canan Kuelahoglu,[g] Chao Bian,[a] Jing Chen,[a] Guangyi Fan,[a] Kerstin Kaufmann,[e] Jocelyn C. Hall,[h] Annette Becker,[f] Andrea Bräutigam,[g] Andreas P.M. Weber,[g] Chengcheng Shi,[a] Zhijun Zheng,[a] Wujiao Li,[a] Mingju Lv,[c] Yimin Tao,[c] Junyi Wang,[a] Hongfeng Zou,[a,i,j] Zhiwu Quan,[a,i,j] Julian M. Hibberd,[k] Gengyun Zhang,[a,i,j] Xin-Guang Zhu,[c] Xun Xu,[a] and M. Eric Schranz[b,2]

[a] Beijing Genomics Institute, 518083 Shenzhen, China
[b] Biosystematics Group, Wageningen University, 6708 PB Wageningen, The Netherlands
[c] Plant Systems Biology Group, Partner Institute of Computational Biology, Chinese Academy of Sciences/Max Planck Society, Shanghai 200031, China
[d] Molecular Biology Group, Wageningen University, 6708 PB Wageningen, The Netherlands
[e] Institute for Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany
[f] Plant Developmental Biology Group, Institute of Botany, Justus-Liebig-University, 35392 Giessen, Germany
[g] Institute of Plant Biochemistry, Center of Excellence on Plant Sciences, Heinrich-Heine-University, D-40225 Duesseldorf, Germany
[h] Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2E9
[i] State Key Laboratory of Agricultural Genomics, Beijing Genomics Institute, 518083 Shenzhen, China
[j] Key Laboratory of Genomics, Ministry of Agriculture, Beijing Genomics Institute, 518083 Shenzhen, China
[k] Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom

The Brassicaceae, including *Arabidopsis thaliana* and *Brassica* crops, is unmatched among plants in its wealth of genomic and functional molecular data and has long served as a model for understanding gene, genome, and trait evolution. However, genome information from a phylogenetic outgroup that is essential for inferring directionality of evolutionary change has been lacking. We therefore sequenced the genome of the spider flower (*Tarenaya hassleriana*) from the Brassicaceae sister family, the Cleomaceae. By comparative analysis of the two lineages, we show that genome evolution following ancient polyploidy and gene duplication events affect reproductively important traits. We found an ancient genome triplication in *Tarenaya* (Th-α) that is independent of the Brassicaceae-specific duplication (At-α) and nested *Brassica* (Br-α) triplication. To showcase the potential of sister lineage genome analysis, we investigated the state of floral developmental genes and show *Brassica* retains twice as many floral MADS (for MINICHROMOSOME MAINTENANCE1, AGAMOUS, DEFICIENS and SERUM RESPONSE FACTOR) genes as *Tarenaya* that likely contribute to morphological diversity in *Brassica*. We also performed synteny analysis of gene families that confer self-incompatibility in Brassicaceae and found that the critical *SERINE RECEPTOR KINASE* receptor gene is derived from a lineage-specific tandem duplication. The *T. hassleriana* genome will facilitate future research toward elucidating the evolutionary history of Brassicaceae genomes.

INTRODUCTION

Studies of the model plant *Arabidopsis thaliana* and its close relatives in the Brassicaceae family have provided fundamental insight into the processes and patterns of plant evolution and function (Koornneef and Meinke, 2010; Hu et al., 2011; Wang et al., 2011). Comparative analyses between Brassicaceae and crop species have had profound influences on plant improvement and production. For example, knowledge about the control and evolution of plant reproductive traits, such as floral and fruit development and self-incompatibility (SI) systems, can be directly related to plant fitness and yield (Tanksley, 2004; Shen et al., 2005). The Brassicaceae have also been a model for understanding the dynamics and impacts of ancient polyploidy (genome doubling), considering that the entire family has undergone a whole-genome duplication (named At-α) and the *Brassica* crops have had an additional genome triplication (Br-α) (Blanc et al., 2003; Thomas et al., 2006; Wang et al., 2011). Genes retained in multiple copies due to these ancient polyploidy events, in addition to more recent tandem duplications, have played important roles in the evolution and regulation of

[1] These authors contributed equally to this work.
[2] Address correspondence to eric.schranz@wur.nl.
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: M. Eric Schranz (eric.schranz@wur.nl).
[W] Online version contains Web-only data.
[OPEN] Articles can be viewed online without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.113.113480

key traits (Edger and Pires, 2009; Flagel and Wendel, 2009). However, the polyploid history of the Brassicaceae also complicates synteny and evolutionary inferences to distantly related crop species.

To fully exploit the fundamental trait and genome insights garnered from Brassicaceae systems and improve synteny analyses to more distant crops, we report the genome sequencing and analysis of *Tarenaya hassleriana* from the Brassicaceae sister family Cleomaceae. Currently, papaya (*Carica papaya*), a member of the order Brassicales, is the closest relative with a complete genome sequence; however, these two lineages diverged more than 70 to 110 million years ago (Ming et al., 2008; Beilstein et al., 2010). The Cleomaceae is the phylogenetic sister family to the Brassicaceae, with the two lineages having diverged only ~38 million years ago (Schranz and Mitchell-Olds, 2006; Couvreur et al., 2010). Brassicaceae and Cleomaceae share many traits (Hall et al., 2002; Iltis et al., 2011), such as a preponderance of herbaceous species, the same general floral ground plan (four sepals, four petals, six stamens, and two fused carpels), and a replum in the mostly dehiscent fruits, referred to as capsules. There are also a number of key differences. Most of the 300 Cleomaceae species are restricted to the semitropics and arid desert regions and lack a genetic pollen-pistil SI system, whereas most of the 3700 Brassicaceae species largely radiated into cold temperate regions and possess a genetically regulated SI system (Guo et al., 2011). Another striking distinction is in floral symmetry: Cleomaceae have mostly monosymmetric flowers and Brassicaceae have mostly disymmetric flowers (Endress, 1999; Patchell et al., 2011). Cleomaceae also exhibit greater variation in the basic floral plan with increases in stamen number, petal dimorphisms, and stalks to the ovary, whereas Brassicaceae exhibit greater diversity in fruit morphology and dehiscence capabilities (Franzke et al., 2011). Comparative analyses can be used to elucidate the genomic basis of these differences. The Cleomaceae species we sequenced is *T. hassleriana*, often referred to as the spider flower, which is widely grown as an ornamental species and used as an educational model (Marquard and Steinback, 2009). This species was formerly named *Cleome hassleriana* (often erroneously labeled as *Cleome spinosa*), but the genus *Cleome* has undergone recent taxonomic revisions (Iltis and Cochrane, 2007).

Brassicaceae and Cleomaceae have undergone independent ancient polyploidy events. At least five ancient polyploidy events have occurred in the evolutionary history of *Arabidopsis* (Bowers et al., 2003; Van de Peer et al., 2009), four of which are shared with Cleomaceae: ζ near the origin of seed plants (Jiao et al., 2011), ε near the origin of angiosperms (Jiao et al., 2011), the ancient hexaploidy At-γ shared by nearly all eudicots (Jaillon et al., 2007; Vekemans et al., 2012), and At-β restricted to part of the order Brassicales as it is lacking from the papaya genome (Ming et al., 2008). The most extensively studied ancient polyploidy event is the more recent At-α genome duplication (Arabidopsis Genome Initiative, 2000; Bowers et al., 2003; Schnable et al., 2012) and is shared by all Brassicaceae species (Schranz et al., 2012). The crop genus *Brassica* has all the ancient polyploidy events in common with *Arabidopsis* but also has undergone an additional and more recent whole-genome triplication (hexaploidy) event (Br-α) after its split with *Arabidopsis* around ~17 million years ago (Wang et al., 2011). Limited BAC and transcriptome sequencing revealed that *Tarenaya* lacked the At-α event and that it underwent an independent ancient genome triplication (Th-α) (Schranz and Mitchell-Olds, 2006; Barker et al., 2009). Thus, *Tareneya* provides a unique opportunity to contrast genome evolution from a common ancestor comparing three genomic equivalents in *Tarenaya*, two in *Arabidopsis*, and six in *Brassica*, and furthermore to contrast two independent ancient genome triplications (Th-α versus Br-α). We not only compare these polyploidy events and more recent tandem duplication events, but also show how they contributed to the genes regulating key reproductive traits (Van de Peer et al., 2009).

RESULTS

Genome Sequencing and Integration with Physical Map

The *T. hassleriana* genome is relatively small (~290 Mb: 2n = 20) and within the range of sequenced Brassicaceae species: *Schrenkiella parvula* (formerly *Thellungiella parvula*), 140 Mb (Dassanayake et al., 2011); *A. thaliana*, 157 Mb (Bennett et al., 2003); *Arabidopsis lyrata*, 207 Mb (Hu et al., 2011); *Capsella rubella*, 210 Mb (Slotte et al., 2013); *Aethionema arabicum*, 240 Mb (Haudry et al., 2013); *Sisimbrium irio*, 262 Mb (Haudry et al., 2013); *Eutrema salsugineum* (formerly *Thelungiella salsuginea*), 314 Mb (Wu et al., 2012; Yang et al., 2013); *Leavenworthia alabamica*, 316 Mb (Haudry et al., 2013); and *Brassica rapa* 485 Mb (Wang et al., 2011). To generate a high-quality draft genome assembly, we used both sequenced paired-end libraries and constructed a Bacterial Artificial Chromosome (BAC) based whole-genome profiling (WGP) physical map. We used the Illumina next-generation sequencing platform to generate ~70.2 Gb (245X genome-depth) raw data of paired-end reads ranging from 90 to 100 bp (see Supplemental Table 1 online) from seven libraries with various insert sizes (350 to 20 kb). Sequence data were filtered, yielding ~40 Gb of high-quality sequence (~139X coverage) (see Supplemental Table 2 online). Sequences were assembled using SOAPdenovo (Li et al., 2010) (version 2.21) (see Supplemental Table 3 online). We also sequenced and mapped over 4 Gb of transcriptome data to the assembly showing that >94% of the genic regions were covered (see Supplemental Table 4 online).

The physical map was made using the Keygene WGP fingerprinting technique (van Oeveren et al., 2011). We generated 192,000 BAC-based Illumina sequence tags from 19,200 BACs from two libraries (*Eco*RI and *Mse*I) with an average insert size of ~125 kb (giving a total of 32X genome equivalents) (see Supplemental Table 5 online). We identified 87,617 high-quality and unique WGP tag sequences that allowed us to uniquely identify 15,567 BACs (with an average of 40.3 tags per BAC). These were used to build a high-stringency map assembly using modified Finger Printed Contigs (FPC) software (Engler et al., 2003) to generate 786 contigs using data from 9396 BACs (see Supplemental Table 6 online). We integrated the WGP physical map scaffolds with the Short Oligonucleotide Analysis Package (SOAP) sequence scaffolds to produce 77 superscaffolds from the integration of 349 sequence scaffolds (integrating between two and 21 scaffolds per superscaffold with an average of 4.5

Table 1. Summary of the genome sequencing, assembly, and annotation.

Assembly			
	N50 (size/number)	N90 (size/number)	Total sizes
Contigs	21.58 kb/2761	2.7 kb/13591	222 Mb
Scaffolds	551.9 kb/98	64.8 kb/622	256.5 Mb
Superscaffolds	1.26Mb/40	7.4kb/1014	273Mb
Annotation			
	Glean	RNA-Seq supported	Homologous with *Arabidopsis*
# Genes	28917	20337	24245
	LTR	DNA transposons	Total size
TE sizes, Mb (%)	97.28 (38.19)	11.8 (4.62)	110 (43.3)

scaffolds per superscaffold) through Basic Local Alignment Search Tool (BLAST) mapping of the WGP anchors (tags). We evaluated the quality of the integration between physical map and superscaffolds by manually checking the ordering and orientation of connected scaffolds by analyzing collinearity relative to *A. lyrata* (example shown in Supplemental Figure 1 online), confirming that all *Tarenaya* superscaffolds have extensive and extended synteny. We further validated our assembly by comparing it with four previously published Sanger-sequenced BACs (Schranz and Mitchell-Olds, 2006; Navarro-Quezada, 2007), revealing nearly identical assemblies (see Supplemental Figure 2 online). The final assembly statistics of the integrated dataset are summarized (Table 1; see Supplemental Table 7 online). With this integration, the N50 was increased by more than 2.6-fold (N50 = 1.26 Mb) due to the merger of most of the de novo assembled scaffolds into superscaffolds.

Gene Annotation

Gene annotation was conducted using a pipeline that integrates de novo gene prediction, homology-based alignment, and RNA-seq data. In total, we conducted >4 Gb of RNA-seq, of which 77.4% of reads could be confidently mapped onto the genome (see Supplemental Table 8 online). To analyze the overall gene expression patterns and to provide basic gene expression information, transcriptomes were generated for several *T. hassleriana* tissues (see Supplemental Table 9 online). A principal component analysis showed that stamen, root, and seed profiles separate most, a result similar to the pattern detected in *A. thaliana* where these three tissue types separate most over the first two components (see Supplemental Figure 3 online) (Schmid et al., 2005).

The prediction of gene models by various techniques was summarized (see Supplemental Table 10 online), with a large range in the number of predicted genes. To be conservative, we used the models predicted by GLEAN to have a high posterior probability out of the various gene prediction techniques, which resulted in the identification of 28,917 highly supported gene models with an average transcript length of 2216 bp, coding sequence size of 1169 bp, and 5.27 exons per gene, both similar to that observed in *A. thaliana* and *B. rapa* (see Supplemental Figure 4 online). A total of 92.9% of gene models have a homolog match or conserved motif in at least one of the public protein databases, including Swissprot (McMillan and Martin, 2008), 71.1%; the Translated European Molecular Biology Laboratory database (Boeckmann et al., 2003), 92.5%; InterPro (Zdobnov and Apweiler, 2001), 74.9%; the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000), 55.2%; and Gene Ontology (GO) (Ashburner et al., 2000), 55.7% (see Supplemental Table 11 online), and 97.1% are represented among the public Expressed Sequence Tag (EST) collections or de novo Illumina mRNA-Seq data. In addition to protein-coding genes, we also identified 220 microRNA, 862 tRNA, and 685 small nuclear RNA genes in the *T. hassleriana* genome (see Supplemental Table 12 online). Orthologous clustering of proteomes predicted for *T. hassleriana* and three Brassicaceae species (*A. thaliana*, *A. lyrata*, and *B. rapa*) revealed 15,112 genes in 12,689 families in common (Figure 1). We found that

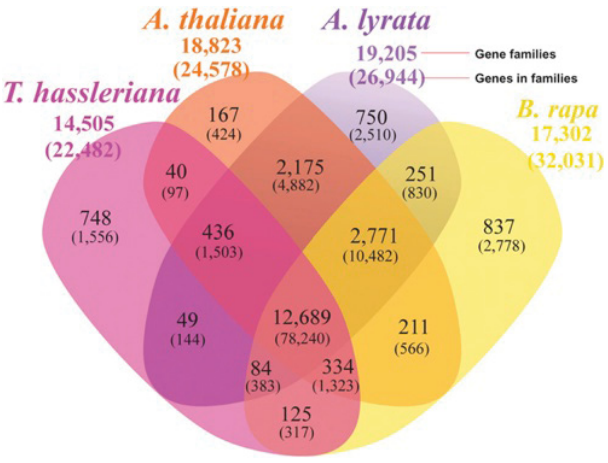


Figure 1. Venn Diagram Illustrating the Shared and Unique Gene Families from *T. hassleriana* (Cleomaceae), *A. thaliana*, *A. lyrata*, and *B. rapa* (Brassicaceae).

In total, we predicted 28,917 well-supported gene models for *T. hassleriana*, of which 22,482 could be placed into one of the 14,505 gene families. A total of 87.5% of gene families found in *Tarenaya* were present in all three Brassicaceae species and 7.4% in one or two Brassicaceae species, and only 5% of gene families were unique to *Tarenaya* with many of these associated genome-specific retrotransposons. Thus, comparative functional and evolutionary analysis of well-characterized *Arabidopsis* and Brassicaceae genes is feasible using *Tarenaya* as an outgroup.

20,926 *T. hassleriana* genes clustered with at least one of the three genomes. Furthermore, 1556 *Tarenaya*-specific genes in 748 families were identified, most of which were enriched for genes of unknown function and for which 34% have EST supported annotation.

To evaluate the status of our genomic assembly, we compared it to the nine published crucifer genomes (Haudry et al., 2013) (see Supplemental Table 13 online). Our *T. hassleriana* assembly has the highest assembled sequence versus expected genome size with an estimated 93% genome size completeness. For comparison, the *B. rapa* assembly only has 51.6% genome coverage. Our assembly also has the largest scaffold N50 (1.26 Mb) of the Illumina-only sequenced genomes (*T. hassleriana*, *Ae. arabicum*, *S. irio*, *Leavenworthia alabamica*, and *B. rapa*). The number of predicted genes is on par with that in the other crucifers and the number of *A. thaliana* orthologs is slightly lower than average, as expected from an organism outside the Brassicaceae. Eighty-eight percent of ultraconserved core eukaryotic genes (Parra et al., 2009) were found, suggesting a slightly lower covered gene space. This might be due to the fact that the percentage of transposable elements (TEs) is relatively high, which can cause difficulties assembling regions where TEs and genes are intermixed.

Because of our high genome sequencing coverage, we were able to identify more than 43% of the 293-Mb *T. hassleriana* genome as being composed of transposons. For comparison, only 31% of the 529 Mb *B. rapa* genome has been identified as transposons (see Supplemental Table 13 online). The discrepancy is because of the great difference in the coverage of the assemblies (93% versus 56%). Thus, the percentage of transposons in *Brassica* is largely underestimated. To examine the contiguity of the genome assembly (causes for gaps between contigs in the scaffolds), we analyzed the distribution of contigs versus transposon long-terminal repeats across the largest 491 scaffolds (see Supplemental Figure 5 online). By doing so, we can demonstrate that regions with a high density of long terminal repeats (LTRs) correspond to smaller contigs in scaffolds and thus generate regions of lower contiguity. Both de novo repeat identification and homology-based methods were applied to predict transposable elements (TEs) (see Supplemental Table 14 online). The majority of repetitive sequences were Class I long-terminal repeat retrotransposons, constituting 36.6% of the genome compared with 27.1% in *B. rapa*. The overall lower percentages of annotated transposons in *Brassica* are likely due to its lower genome sequence coverage (see Supplemental Table 13 online) because the *B. rapa* genome is nearly 200 Mb larger than *Tarenaya*. Most of the repeats were located in the intergenic regions.

Comparative Analysis of Ancient Polyploidy Events

The Brassicaceae-Cleomaceae system allowed us to compare genome evolution after several rounds of independent ancient polyploidy (Figure 2). We confirmed that the Cleomaceae polyploidy event (Th-α) occurred independently of, and more recently than, the Brassicaceae-specific duplication event detected in *Arabidopsis* and *Brassica* (At-α). We also detected the nested *B. rapa* ancient hexaploidy (triplication of the genome) event (Br-α) and show that it is of approximately the same age as Th-α. For all three taxa, we detected the diffuse signal of the older and shared events (At-β, At-γ, ε, and ζ) (Figure 2). The Th-α and Br-α events are of approximately the same age and, as discussed below, represent independent ancient hexaploidy events. Using whole genome intragenomic dot plots of *Tarenaya*, we showed many triplicated blocks (see Supplemental Figure 6 online), and analysis of syntenic depth with QuotaAlign (Tang et al., 2011) shows that 49.4% of genes are found at 3× coverage (see Supplemental Table 15 online). To illustrate the homologous relationships and the evolutionary history of triplicated/duplicated segments in Cleomaceae and Brassicaceae, we integrated intra- and intergenomic analyses (Figure 3). We analyzed synteny relationships

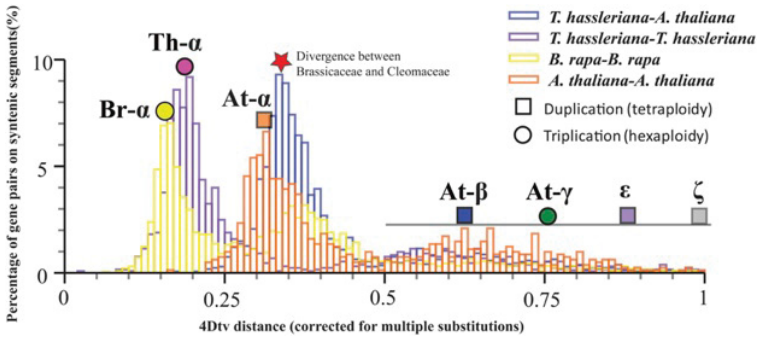


Figure 2. Relative Timing of the Polyploidy Events and Lineage Splitting Based on Divergence of Fourfold Degenerate Sites (4DTv) for Duplicated Genes within *A. thaliana*, *B. rapa*, and *T. hassleriana* and Orthologous Genes between *A. thaliana* and *T. hassleriana*.

All plots detect broad overlapping peaks between 0.5 and 1.0, representing shared older polyploidy events (At-β, At-γ, ε, and ζ). The divergence of the Brassicaceae-Cleomaceae lineages is seen by the differentiation of *Arabidopsis* and *Tarenaya* homologs at the peak centered at ~0.35 (highlighted by red star). The divergence between paralogs from the At-α duplication event occurred slightly after the lineage splitting and is detected by the peaks centered at ~0.3 for both *Arabidopsis* and *Brassica*. The At-α peak is lacking from *Tarenaya*, proving At-α is Brassicaceae specific. Nearly overlapping distributions between 0.15 and 0.25 were detected for *Brassica* and *Tarenaya*, representing the independent Br-α and Th-α ancient hexaploidy events, respectively.

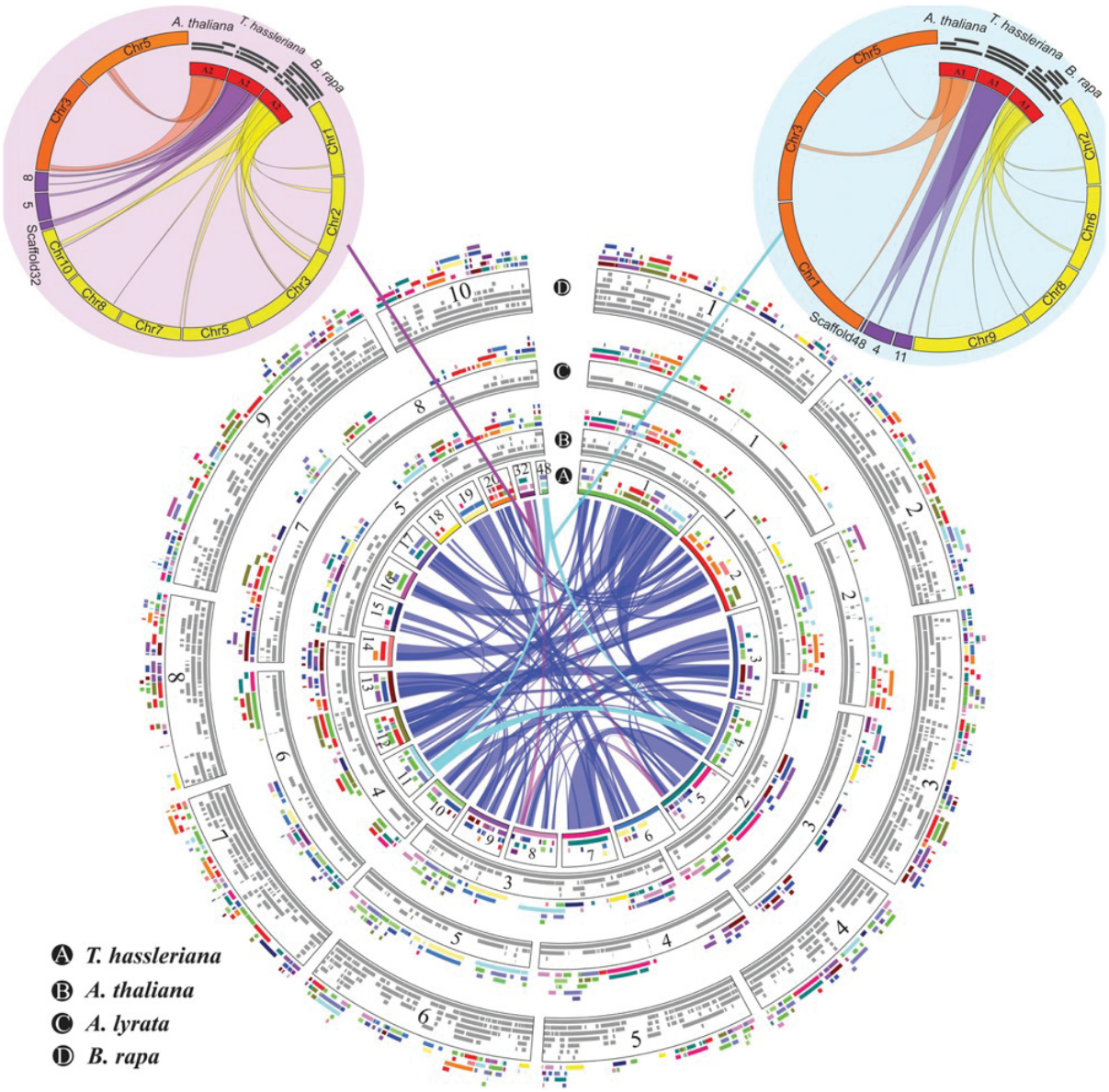


Figure 3. Homologous Genome Blocks within and between Genomes for Cleomaceae and Brassicaceae.

The largest 20 (plus additional two smaller scaffolds) color-coded superscaffolds of *T. hassleriana* are taken as the reference, such that any region homologous to the 22 scaffolds is colored accordingly. A, Self-alignment of *Tarenaya* superscaffolds, with the inner circle showing links of syntenic blocks. Over 47% of the genome is found in three copies, supporting the conclusion that it experienced an ancient hexaploidy event (Th-α = triplication). Rings within a genome (inner gray bars) and blocks homologous to *Tarenaya* (outer color-coded bars) for completed Brassicaceae genomes: B, *A. thaliana*; C, *A. lyrata*; and D, *B. rapa*. The inner gray bars show a clear pattern related to the ancient polyploidy events of the Brassicaceae (At-α = duplication) and nested Brassica-specific lineage (Br-α = triplication). The color-coded outer rings of homology relative to *Tarenaya* show a complex pattern due to the independent polyploidy events between families. The two small insets illustrate examples of the three *Tarenaya* to two *Arabidopsis* to six *Brassica* genome equivalents due to ancient polyploidy events.

both within and between genomes (see Supplemental Figures 6 to 9 online). For *T. hassleriana*, 86, 83, and 85% of the protein-coding genes were homologous to the genes in the *A. thaliana*, *A. lyrata*, and *B. rapa* genomes, respectively (see Supplemental Table 16 online). By making these comparisons of *Tarenaya* versus *A. thaliana*, *A. lyrata*, and *B. rapa* genomes (Figure 3), we found significant, 3:2, 3:2, 3:4, 3:5, and 3:6 homologous patterns, respectively, which is consistent with the polyploid history of the species. To illustrate this pattern, we highlighted two ancestral blocks (A1 and A2) (Figure 3, two small insets). Note that the three

Tarenaya blocks show almost perfect collinearity, whereas one of two *Arabidopsis* regions is broken across two chromosomes, suggesting a Brassicaceae-specific rearrangement(s) after At-α. Since synteny analysis has been extensively performed within Brassicaceae, we also show our results with the collinear blocks color-coded according to the current Brassicaceae conventions (Schranz et al., 2006) (see Supplemental Figure 10 online).

We inferred the putative "A ancestor" (pre-At-α) shared by *A. thaliana* and *A. lyrata*, the "B ancestor" of *B. rapa* (pre-Br-α but post-At-α ancestral genome state), and the "T ancestor" of *T. hassleriana* (pre-Th-α ancestral genome state) (see Methods). We compared our identified homologous replicated blocks within and between genomes in Brassicaceae species and *T. hassleriana* with the results of an earlier analysis of conserved At-α blocks (Blanc et al., 2003; Thomas et al., 2006). First, we reconstructed our version of the pre-At-α ancestor (A ancestor) of *A. thaliana* (version TAIR9), which resulted in 64 ancestral regions involving 19,976 protein-coding genes (see Supplemental Table 15 online). The corresponding relationships to the blocks identified by Wolfe and colleagues (Blanc et al., 2003) (that we refer to as "Wolfe blocks") are illustrated in Supplemental Figure 11 online. We used the same method on the minimized genomes of *A. lyrata*, *B. rapa*, and *T. hassleriana* and generated 61, 71, and 87 conserved ancestral blocks covering 24,373; 25,646; and 20,680 protein-coding genes, respectively, to represent the postulated A, B, and T ancestors (see Supplemental Figures 12 to 14 online). A table listing all genes included in each block for each genome is provided in the supplemental materials (see Supplemental Data Set 1 online). A comparison of the reconstructed T and A ancestor genomes revealed a 1:1 relationship, supporting our conclusion that the ancestral genome of the Brassicaceae and Cleomaceae was conserved before the independent duplication (At-α) and triplication (Th- α) events, respectively. Since *B. rapa* underwent a nested and specific triplication following the At-α, we see a 1:2 pattern when we compare the inferred T to B ancestors. A comparison of conserved ancestral genomic blocks across species is shown in Supplemental Figure 15 online. We then traced the extent of gene retention and fractionation in homologous blocks after polyploidy events. We partitioned the two subgenomes of *Arabidopsis*, three subgenomes of Brassica, and three subgenomes of *T. hassleriana*, respectively, by comparing them with the reference A ancestor of *A. lyrata* (see Supplemental Figures 16 to 19 online). These improvements to understanding genome evolution after independent ancient polyploidy events of Brassicaceae species will facilitate synteny analyses in distant crop species.

Comparative Analysis of Type II MADS Box Genes

The development of the four floral organ types and later the fruits is regulated by Type II MINICHROMOSOME MAINTENANCE1, AGAMOUS, DEFICIENS and SERUM RESPONSE FACTOR (MADS) domain proteins (Smaczniak et al., 2012) as described by the ABCDE model (Theissen, 2001). The types of MADS box genes that regulate development are remarkably well conserved across eudicots, with the molecular mechanisms of their action extensively studied in *Arabidopsis*. We found that the *Tarenaya* genome contains representatives of all the major Type II MADS box genes described in *Arabidopsis* and *Brassica* (see Supplemental Figure 20 online). We concentrated on the retention of the MADS box genes derived from At-α, Br-α, and Th-α and compared this with the polyploid origins of additional duplicates (At-β, At-γ, ε, T [tomato] [Tomato Genome Consortium, 2012], and Pt-α [poplar] [Tuskan et al., 2006]) (Figure 4). Theoretically, the At-α, Br-α, and Th-α events should have given rise to two *Arabidopsis*, six *Brassica*, and three *Tarenaya* gene copies (syntelogs) from a single ancestral gene. Of the 11 MADS box gene clades involved in floral, fruit, and inflorescence development that were likely present in the most recent common ancestor of Brassicaceae and Cleomaceae shown in Figure 4, we found that only three duplicate pairs are in fact maintained in *Arabidopsis* due to At-α: *APETALA1* (*AP1*)/*CAULIFLOWER* (*CAL*) (A-function), *SHATTERPROOF1* (*SHP1*)/*SHP2* (D-function), and *SEPALLATA1* (*SEP1*)/*SEP2* (E-function). Thus, *Arabidopsis* has only three of 11 possible replicates (27.3% syntelog retention). This implies that during early Brassicaceae evolution (before the split of *Arabidopsis-Brassica*), there were 14 gene lineages. From these 14 lineages then there would be 42 possible syntelogs in *Brassica* due to Br-α triplication, of which 28 copies are considered as additional syntelogs. We find a remarkable 19 of these additional gene copies (67.8% syntelog retention). This includes all three possible copies maintained for the following seven Brassica gene families: *SEP4*, *SEP3*, *AGAMOUS-LIKE79* (*AGL79*), *FRUITFULL*, *AP1*, *PISTILLATA* (*PI*), and *SHP1*. The only two genes to return to single copy in *Brassica* after Br-α are *CAL* and *SHP2*. When we also consider the At-α gene retention, then a maximum of four of six gene copies are found in the *SEP1/2* clade, *AP1*/*CAL* clade, and the *SHP1/2* clade. From the 11 ancestral loci, 33 syntelogs would be expected in *Tarenaya* due to the Th-α triplication, with 22 possible additional syntelogs. However, we only recovered six (27.3% syntelog retention) with no cases where all three possible copies are maintained (we do not count the additional tandem duplicate of Th-*AP3* here, which is discussed below). Thus, we find more than double the syntelog retention in *Brassica* than in *Tarenaya*, despite the fact that both are ancient hexaploids of approximately the same age. The greatest differential in gene copy retention between *Brassica* and *Tarenaya* is for the *AGL79* clade (3 versus 1) and the *SHP1/2* clade (4 versus 1). The *SHATTERPROOF* genes in Brassicaceae regulate various traits during carpel and fruit formation (Colombo et al., 2010). The single-copy nature of the *SHP* homolog in *Tarenaya* is thus notable, since this is the only gene that is duplicated in *Arabidopsis* due to At-α but has returned to single copy in *Tarenaya* (see Supplemental Figure 21 online). In Cleomaceae, fruit morphology is less diverse than in Brassicaceae, and we hypothesize that the retention of *SHP* genes plays an important role in the morphological variability of Brassicaceae.

Comparative Collinearity Analysis of Floral Developmental Regulators

To assess the contributions of ancient polyploidy and tandem gene duplications to floral regulatory gene diversification, we conducted a more detailed analysis of gene synteny and expression patterns. Almost all floral MADS box genes show conserved synteny between Brassicaceae and Cleomaceae. For example, the A-class genes show stable duplicate retention; the loci containing

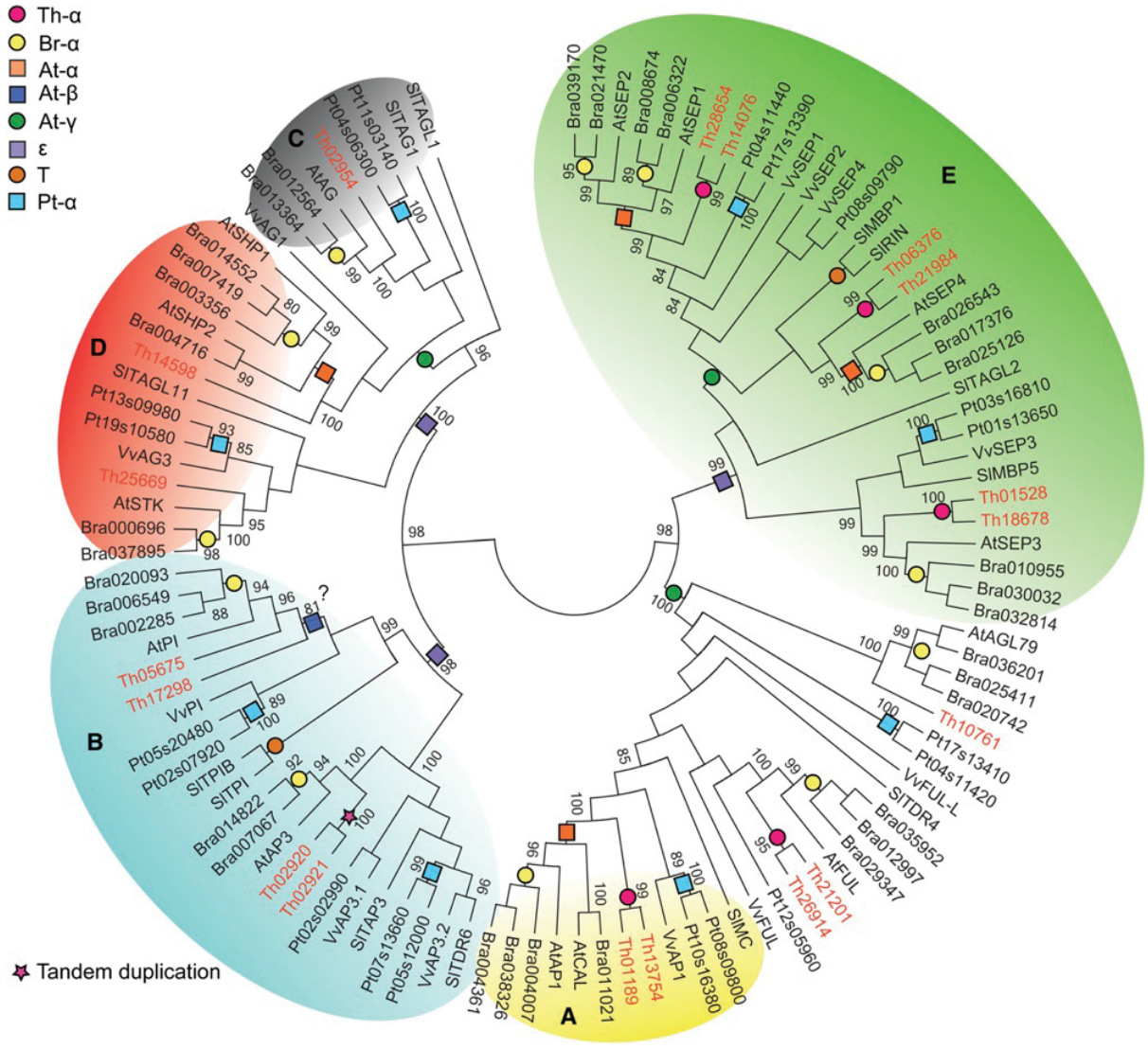


Figure 4. Phylogenetic Tree of Type-II MADS Box Transcription Factor Genes Involved in Floral Organ Specification.

The alignment used to create this tree is available in the online materials (see Supplemental Data Set 2 online). The major floral MADS box genes cluster into five groups corresponding to the five main functional types (AP1-like genes, shown in yellow; AP3/PI-like genes, B-type shown in blue; AG-like genes shown in gray; STK-like genes shown in red; and SEP-like genes in green) according to the ABC(DE) model of floral development. Species included are *A. thaliana* (At), grape (*Vitis vinifera*, Vv), tomato (*Solanum lycopersicum*, Sl), poplar (*Populus trichocarpa*, Pt), *B. rapa* (Br), and *T. hassleriana* (Th). *Tarenaya* genes are indicated in red. The colored squares (duplication events) and circles (triplication events) placed on nodes represent gene lineage expansion(s) that can be associated with particular ancient polyploid events: Th-α, At-α, Br-α, At-β, At-γ, ε, T (identified by tomato genome sequencing), and Pt-α (identified by poplar genome sequencing). Type-II MADS box genes are often retained after ancient polyploidy events. From the tree above, we calculated a 27.3% syntelog (homolog generated by a polyploidy event) retention after At-α, 27.3% syntelog retention after Th-α, and a much higher (67.8%) syntelog retention after Br-α, despite the fact that Th-α and Br-α triplications are of approximately the same age. The *Tarenaya* B-class genes show unusual patterns in that the *AP3* homologs (Ch02920 and Th02921) represent a recent tandem duplication, which is rare for floral MADS box genes, and there are two copies of *PI* homologs that are likely due to At-β with one lineage being lost in Brassicaceae. Shown is a maximum likelihood tree with 1000 replicate bootstrap values, of which the branches with a bootstrap value of >80 are presented, visualized topology only.

the *AP1* and *CAL* homologs in *Tarenaya*, *Arabidopsis*, and *Brassica* are syntenic to one another with little evidence of local rearrangements (see Supplemental Figure 22 online).

Compared with other MADS box genes, the B-class (*PI* and *AP3*) genomic regions show a more dynamic pattern (Figure 5). The split between *PI* and *AP3* is an old duplication due to the angiosperm ε polyploidy event (Figure 4), with almost no detectable collinearity between these regions (Figure 5). Comparison of B-class *Tarenaya* genomic regions with Brassicaceae allowed us to detect two B-class duplication events: A recent *Tarenaya* tandem duplication of *AP3* (Th02920 and Th02921) and an older *PI* duplication, likely due to At-β, which has been lost from Brassicaceae but is still retained in *Tarenaya* (Th17298) (Figure 5; see Supplemental Figures 23 and 24 online).

Strikingly, we also detected two gene transposition events: a Brassicaceae-specific AP3 transposition and a shared transposition event of one *PI* gene before the split of Brassicaceae and Cleomaceae (Figure 5). The Brassicaceae-specific AP3 transposition event also involved the flanking EMBRYO DEFECTIVE1967 (EMB1967) gene containing two conserved domains: the N-terminal region of microspherule protein (MCRS_N) and Forkhead-associated (FHA). Tarenaya AP3 is similarly flanked by a Forkhead-associated protein (Th02919) that has its highest match to EMB1967; however, the orientation of AP3 and the Forkhead genes is inverted between species (see Supplemental Figure 25 online) and is also detected in a distantly related Cleome species. In general, B-class genes are functionally highly conserved across angiosperms, whereas A-class gene function appears to be less conserved (Litt and Kramer, 2010). However, we have shown that it is in fact the B-class genes that have undergone transposition events.

Members of the TEOSINTE BRANCHED1, CYCLOIDEA, and PCF (TCP) gene family play an important role in the transition from polysymmetric to monosymmetric flowers (reviewed in Busch and Zachgo, 2009; Jabbour et al., 2009; Rosin and Kramer, 2009), including monosymmetric Brassicaceae (Busch and Zachgo, 2007; Busch et al., 2012). Cleomaceae floral morphology, especially in petal and stamen position, numbers, and asymmetry, is quite variable. However, the role of Tarenaya TCP homologs in monosymmetry has not yet been fully characterized. We find a pattern of conservation of genomic collinearity around the TCP1 locus between species (see Supplemental Figure 26 online). *Arabidopsis* contains only a single TCP1 locus, as does *A. lyrata*. Due to Br-α, Brassica has three syntenic copies of TCP1. We also can detect the syntenic regions in Brassicaceae species due to At-α (see Supplemental Figure 26 online) but find no At-α derived homologs of TCP genes, suggesting the loss of a TCP1 syntelog occurred early in Brassicaceae evolution. In Tarenaya, we find three genomic regions derived from Th-α, with two copies of TCP1 intact (Th21666 and Th24587) (see Supplemental Figure 26 online). The correlation between multiple copies of TCP members and monosymmetry has been noted across angiosperms (Rosin and Kramer, 2009).

Expression of A- and B-Class Homolog Genes

The expression of *T. hassleriana* homologs of *A. thaliana* major floral regulators was also analyzed with quantitative RT-PCR (qRT-PCR). The two putative *T. hassleriana* homologs of *CAL/AP1* (Th-*CAL/AP1-1* and Th-*CAL/AP1-2*) show similar expression in all bud stages, but Th-*CAL/AP1-1* with only half of the transcript abundance of Th-*CAL/AP1-2*. Th-*CAL/AP1-1* shows highest expression in sepals and ~10 times lower expression in petals. Expression of neither homolog was detectable in stamens, petals, gynoecia, capsules, roots, or leaves. Th-*PI-1*, the homolog of *PI* in *A. thaliana*, is collinear with the Brassicaceae gene order and is expressed mainly in petals and stamens, with less expression in younger and higher expression in older stages (see Supplemental Figure 27 online). The second *PI* homolog of *T. hassleriana*, Th-*PI-2*, is expressed at a much lower rate than Th-*PI-1*, ranging from around 50% of the Th-*PI-1* expression in stamens to only 10% of the Th-*PI-1* expression in late bud states.

The second two putative floral homeotic B-function genes, Th-*AP3-1* and Th-*AP3-2*, are highly similar in coding, 3′, and 5′ untranslated region sequence, and both are homologous to the *AP3* gene in *A. thaliana*. Th-*AP3-1* is expressed throughout the observed stages of bud development. In petals and stamens at anthesis, it is the most highly expressed gene of the MADS box genes analyzed (see Supplemental Figure 27 online). In petals, it is expressed at~200% and in stamens around 400% higher level than Th-*CAL*, Th-*PI-1*, and Th-*AP3-2*. Expression of Th-*AP3-2* in buds is around one-third lower than that of Th-*AP3-1*, and differential expression between both genes is detected in petals and stamens. While Th-*AP3-1* shows higher expression in stamens than in petals, Th-*AP3-2* has stronger expression in petals than in stamens (see Supplemental Figure 27 online). The divergence in expression of the B-class genes, along with the aforementioned gene transpositions, is indicative of the likely role in B-class gene functional differentiation and the regulation of the different floral morphologies between families. We acknowledge that these comparative analyses of *Tarenaya* to Brassicaceae are based on the analysis of the draft genome sequence of a single domesticated accession of one species; thus, future comparative analyses to other Cleomaceae species is needed for validation.

Evolution of the Brassicaceae SI Locus

Many Brassicaceae species possess a pollen-pistil recognition system that confers SI through the rejection of self-pollen (Boyes et al., 1997). This system is based on the interaction of the stigmatically expressed S-receptor kinase (encoded by *SRK*) with a small polymorphic peptide that is coded by the S-locus Cys-rich protein (*SCR*) gene, both located on the so-called S-locus of the genome. Many *SCR* alleles are likely derived from (partial) duplication and/or gene conversion from *SRK* alleles (Koornneef and Meinke, 2010; Guo et al., 2011). The cysteine-rich protein kinase (*CRK*) genes and *ARK* genes that belong to the S-locus are part of a larger family of receptor-like protein-kinases (*RLK*) genes, which have been shown to be mostly involved in oxidative stress and pathogen response (Chen et al., 2004; Wrzaczek et al., 2010). It should be noted that most ecotypes of *A. thaliana* are self-compatible due to pseudogenization of the *SRK* and/or *SCR* genes, a pattern seen in other self-compatible crucifers, but still an exception among Brassicaceae (Nasrallah et al., 2004). *T. hassleriana* does not possess a SI system, but it still contains an S-like locus that contains

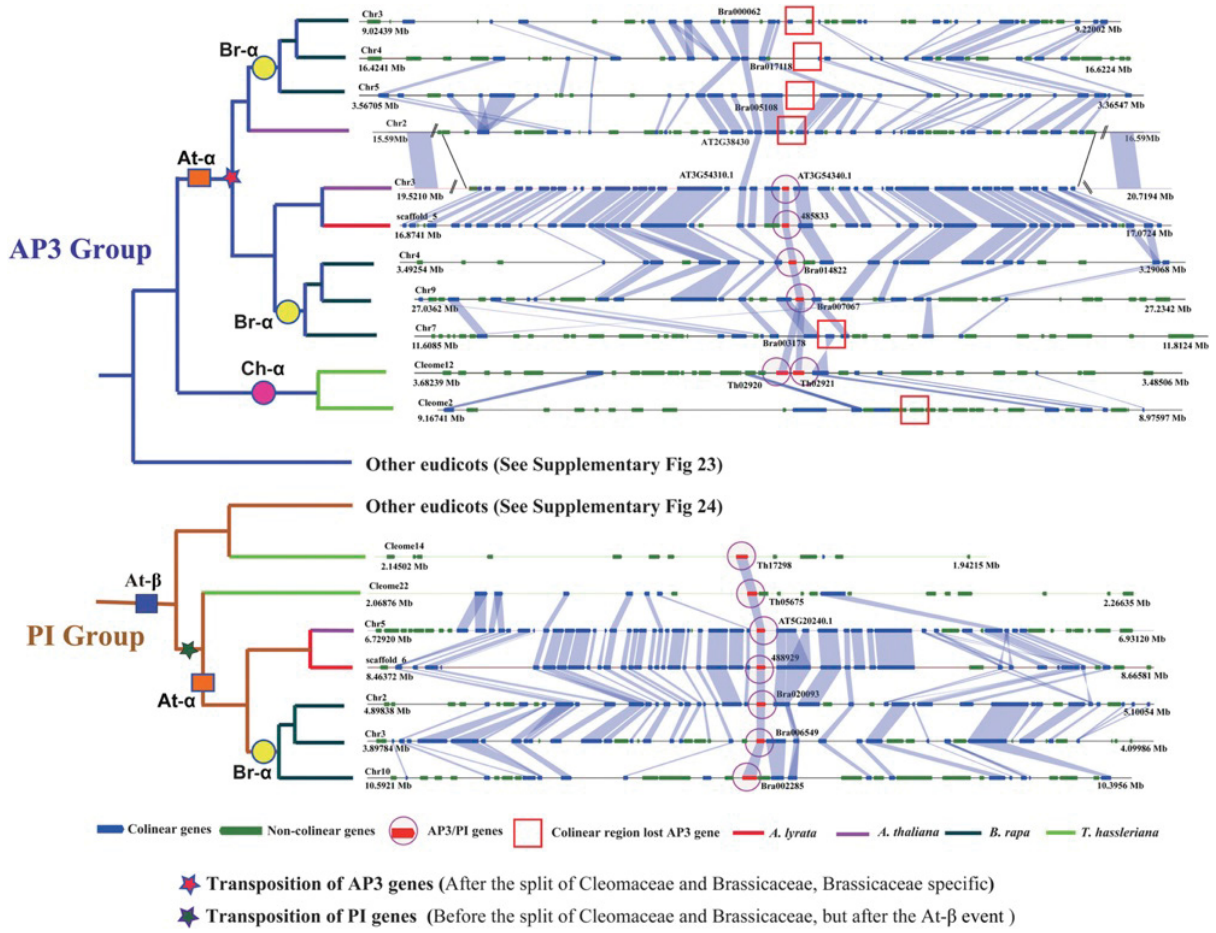


Figure 5. Collinearity Analysis of B-Class Type II MADS Box Gene (*AP3* and *PI*) Homologs Reveals Unusual Patterns of Gene Loss, Lineage-Specific Transpositions, and Local Tandem Duplications.

The placement of ancient polyploid events giving rise to gene duplicates is shown on appropriate nodes (At-β, At-α, Br-α, and Th-α). For the *AP3* group genes in Brassicaceae (shown by red bars), there is only a single locus retained in *A. thaliana* and *A. lyrata* and two retained Brassica syntelogs derived from Br-α. Collinear homoeologous regions derived from At-α are detectable in Brassicaceae genomes; however, the *AP3* syntelogs were lost (regions highlighted in red boxes). *T. hassleriana* has an unusual tandem duplication of *AP3* genes in one of two homoeologous regions derived from Th-α. The *AP3* genes and the neighboring Forkhead gene (*EMB1967*) are the only genes syntenic to the AP3 Brassicaceae region (see Supplemental Figure 25 online). The Cleomaceae AP3 region is syntenic with AP3 regions of all other eudicot genomes analyzed (see Supplemental Figure 23 online). Thus, we conclude that there was a lineage-specific transposition of AP3 and the neighboring Forkhead locus in the Brassicaceae. There is only a single copy of PI genes (red bars) in *A. thaliana* and *A. lyrata* and all three Br-α derived syntelogs in Brassicaceae. There is no detectable homoeologous region in Brassicaceae derived from At-α. In *Tarenaya*, we detected one syntenic *PI* gene and region to the Brassicaceae, but also a second region that is syntenic to other eudicots (see Supplemental Figure 24 online). We conclude that these two *Tarenaya PI* genes were generated due to the At-β ancient duplication event with the subsequent transposition of one locus into the region collinear between Brassicaceae and Cleomaceae and loss of the nontransposed locus from only the Brassicaceae lineage. The differences in genomic context and gene expression (see Supplemental Figure 27 online) may contribute to shifts in floral morphology and symmetry between families.

functional genes, and it is likely that this part of the genome is close to the ancestral state of this locus for Brassicaceae.

SRK and *ARK* on the S-locus are characterized by specific variations on the following protein domain compositions: B_lectin (B), S_locus_glyco (S), PAN-2 (Pa), Pkinase_Tyr (Pk), and Duf3403 (D1) and/or DUF3660 (D2) (Zhang et al., 2011). Using the Pfam database (Punta et al., 2012), we found that the S locus region in *C. rubella*, *B. rapa*, *A. lyrata*, and *A. thaliana* as well as the homologous region in *T. hassleriana* mostly contains *SRK* genes with a B-S-Pa-D1-Pk-D2 protein domain structure, followed by the B-S-Pa-Pk-D1/2 protein domain structure, which is more common across all SRK families (Figure 6; see Supplemental Figure 28 online). Three syntenic regions containing most of the genes of the S-locus were found in *Tarenaya*. One of these

contained a gene with the exact B-S-Pa-Pk-D1/2 protein domain structure: Th11131. Our analysis of domain structure further found that another gene, Th22785, had a B-S-Pa-Pk protein domain structure, an architecture shared with many angiosperm genes but not with SI-specific SRK alleles. We have also found two homologs of the SI-modifier gene, *Pub8*, in two of the three *Tarenaya* syntenic regions. One of the homologs, Th22784, is adjacent to the Th22785 locus. The other homolog, Th25331, is contained in the syntenic region that completely lacks any S_locus_glyco Pfam-containing proteins. Interestingly, we do not find a *Pub8* homolog in close proximity to Th11131. However, based on the alignment of all three regions, we can assume that the single-copy ancestral region contained homologs to *Pub8*, *ARK3*, and *B120* (Figure 6). To ensure that syntenic genes that were not found were not due to gaps in our assembly coverage, we manually confirmed that all relevant regions had no large gaps. In Supplemental Table 17 online, a summary of the coverage in these regions can be found. No large gaps were present except for one 2.5-kb gap in the region around Th22784. Based on the syntenic order of S-locus genes in other species, it is extremely unlikely that a gene is present in this gap, let alone any important candidate gene for this region.

We conclude that the syntenic *Tarenaya* gene Th11131, which is most closely related to *ARK3* with which it shares protein architecture, is similar to the ancestral version of SI locus genes in Brassicaceae. We hypothesize that the origin of the S-locus is due to a local rearrangement/duplication of an *ARK3*-like gene and subsequent expansion and diversification of the S-locus. Since *ARK3* in *Arabidopsis* is not expressed in the stigma, the regulatory domains needed for tissue-specific expression may potentially be derived from the concurrent duplication of elements from neighboring genes.

DISCUSSION

The completed genome of *A. thaliana* was a major milestone in plant biology and has provided a key tool for understanding plant gene function and genome organization (Arabidopsis Genome Initiative, 2000). Subsequently, there has been a great effort and interest to sequence other crucifer species to leverage the knowledge gained from *Arabidopsis* to other species in a comparative context. To date, there are more completed crucifer genomes (including the crop *B. rapa*; Wang et al., 2011) than any

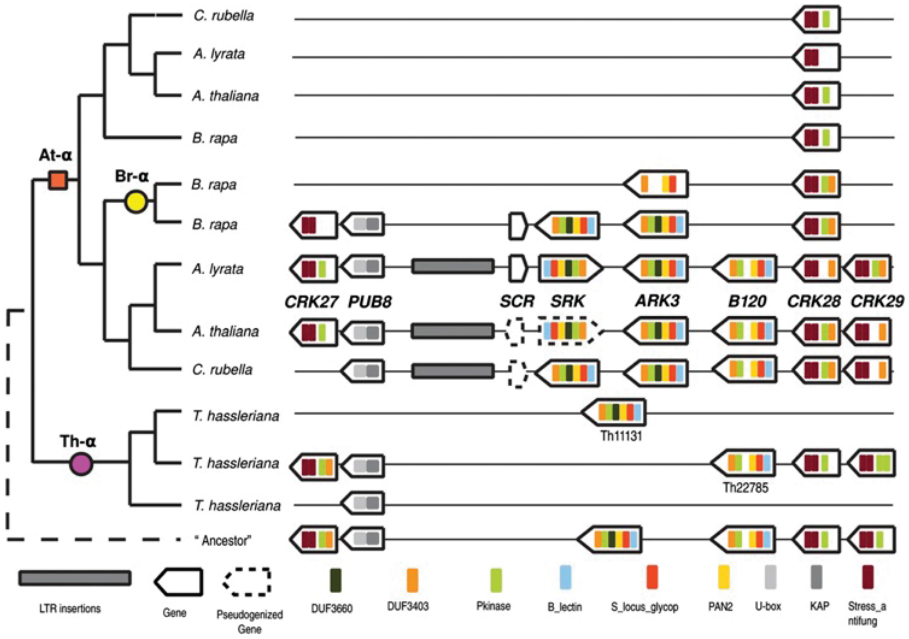


Figure 6. Synteny and Protein Domain Analysis of the Brassicaceae SI-Like Regions with Cleomaceae and Inference of the Ancestral Genomic Region.

All regions presented here are drawn in more detail (including genetic coordinates) in Supplemental Figure 28 online. Genes are marked by block arrows and color coded according to their protein domain composition as listed in the legend. In the case of pseudogenes, the block arrows have a dashed border. To find protein domain composition in pseudogenes, the longest open reading frame was translated in silico (see Methods). Gene orientation is shown by block arrows pointing left for gene orientation toward the 5′ end and pointing right for gene orientation toward the 3′end. The At-α duplication and the Br-α and Th-α triplications have been marked on the tree with an orange box (At-α) and yellow and purple circles (Br-α and Th-α, respectively). Each branch corresponds to a subgenome resulting from such a polyploid event. Theoretically, *B. rapa* should have six subgenomes, but only the regions showing synteny are listed here for clarity. The bottom branch represents a hypothetical layout of this genome region in the common ancestor of these species before the At-α, Br-α, and Th-α polyploid events. From our results, we conclude that an *ARK3*-like gene underwent a Brassicaceae-specific tandem gene duplication generating the key SI receptor *SRK*.

other eudicot plant family, and ambitious plans exist to sequence many more (such as the Brassica Map Alignment Project; Pires et al., 2013). We sequenced *T. hassleriana*, a phylogenetic outgroup of the Brassicaceae sister family, the Cleomaceae. We have shown that the vast majority of genes in *Tarenaya* have clear homologs within Brassicaceae. We provide several examples of how this sister-group genome can be used to elucidate patterns of gene, genome, and trait evolution within the Brassicaceae. Specifically, we focused on independent ancient polyploidy events, floral MADS box, and SI gene evolution. The genome of *T. hassleriana* will facilitate future research into the evolutionary and functional history of *Arabidopsis* genes and pathways.

While it has long been known that there are numerous recent polyploid plants (Jiao et al., 2011), with the arrival of the genomics era, it has become clear that there also is extensive evidence for ancient polyploidy across the tree of life (Kasahara, 2007; Jiao et al., 2011; Murat et al., 2012). Most ancient plant polyploid events that have been identified are ancient tetraploidy events (duplications such as At-α), but there are at least four published genome analyses of ancient plant hexaploidy (triplication events): at the base of the eudicots (At-γ) (Jaillon et al., 2007), in tomato (T) (2012), in Brassica (Br-α) (Wang et al., 2011), and this report of the ancient genome triplication in the *Tarenaya* (Tr-α). The Br-α and Tr-α events are of approximately the same age, allowing us to contrast independent ancient hexaploidy events from closely related lineages. The analysis of the retention of replicated genes (syntelogs) of the Type II MADS box genes provides a compelling example of what can be deduced by the comparison of these independent ancient triplications. We found that *Brassica* retains nearly twice as many Type-II MADS box genes as does *Tarenaya*. Genes retained after polyploidy often are dosage-sensitive gene complexes, whereby interacting partners must be maintained in the proper ratios (Edger and Pires, 2009). Considering the wealth of phenotypic diversity seen in the *Brassica* crops, we hypothesize that this great enrichment of morphological regulators in Brassica derived from Br-α may play a significant role. Also, the recently released genome of *L. alabamica* of the Brassicaceae also revealed an ancient hexaploidy (La-α) (Haudry et al. 2013). Future comparisons of the closely related La-α, Br-α, and Th-α events should be fruitful.

Comparative analyses can also be used to identify important gene transposition and deletion events. Type-II MADS box genes are remarkably resistant to gene transpositions and thus their collinearity is highly conserved across all angiosperms (Type-I MADS box genes are highly prone to transposition, but their functions are less known). When comparing *Tarenaya* to Brassicaceae, we found almost all Type-II MADS box genes are collinear, except for the B-class homologs of *AP3* and *PI*. Specifically, we established that there was a Brassicaceae-specific transposition of *AP3* and a rare tandem duplication of a floral MADS box gene of *AP3* homologs in *Tarenaya*. The transposition of *AP3* also involved a neighboring Forkhead gene, which may be coregulated and important for *AP3* function. We further demonstrate that *Tarenaya* has two homologs of *PI*: one that is syntenic with other eudicots and for which the locus is lost from Brassicaceae and one that is syntenic with Brassicaceae *PI* homologs. Both *PI* and *AP3* in *Tarenaya* have diverged in expression levels. MADS box B-class gene homologs in the *AP3* lineage as well as *TCP* members have been implicated in the establishment of monosymmetric flowers in monocots, including Orchidaceae species (Tsai et al., 2004, 2008; Mondragón-Palomino and Theissen, 2009; Bartlett and Specht, 2010; Preston and Hileman, 2012), *PI* genes have contributed to floral diversification in Asterids (Viaene et al., 2009), and B-class genes have contributed to *Aquilegia* floral diversification (Kramer et al., 2007). Thus, it is possible that both B-class and TCP genes have an impact on floral monosymmetry in Cleomaceae. Furthermore, B-class gene diversification has also been implicated in regulating floral gender shifts (Ackerman et al., 2008) and could similarly have diversified in Cleomaceae.

By comparing Brassicaceae genomes with the *Tarenaya* genome, we established that *SRK* in the S-locus occurred via a local rearrangement/duplication of an *ARK3*-like gene and subsequent expansion and diversification in Brassicaceae. In *Tarenaya*, we can clearly identify syntenic regions that contain homologous functional genes, including an *ARK3* and *CRK* homologs. The exact function of *ARK3* is not known, but based on gene expression analysis, this gene is thought to function during development of the sporophyte, perhaps in processes related to organ maturation, the establishment of growth pattern transitions (Dwyer et al., 1994), and/or involvement in pathogen responses (Pastuglia et al., 2002). However, it is not expressed in the stigma. Further research on Brassicaceae and Cleomaceae *ARK3* homologs is needed to establish the function of these genes. Interestingly, while Cleomaceae species do not have a SI system, many species, including *T. hassleriana*, are polygamous (trimonoecious) and can have flowers on the same inflorescence with different genders: male sterile, female sterile, or complete (Stout, 1923; Cruden and Lloyd, 1995; Machado et al., 2006), providing an alternative mechanism to reduce inbreeding.

Our results demonstrate the utility of the *Tarenaya* genome in complementing Brassicaceae genetic research in efforts to understand the function and evolution of genes and traits. The *Tarenaya* genome sequence will also pave the way for further studies of Cleomaceae traits not found in Brassicaceae, such as the evolution of C_4 photosynthesis (Brown et al., 2005; Marshall et al., 2007).

METHODS

Sample Preparation, Library Construction, Genome Sequencing, and Assembly

The purple-flowered *Tarenaya hassleriana* (Purple Queen) line selected for sequencing (ES1100) was first inbred by hand-pollination and floral bagging for four generations. Earlier generations of this line have previously been used for both BAC library construction and limited BAC sequencing (Schranz and Mitchell-Olds, 2006) and transcriptome sequencing and analysis (Barker et al., 2009); however, the material was referred to as being from *Cleome spinosa*. The two species are morphologically very similar, with only slight differences in stem-spine morphology, pubescence of sepals, ovary and capsules, and flower color; *C. spinosa* has only white flowers and the sepal and ovary are glandular-pubescent, whereas *T. hassleriana* can be white, pink, or purple and the sepals and ovary are glabrous (have no pubescence). Thus, many

commercial seed providers erroneously label their *T. hassleriana* material as *C. spinosa*.

We extracted DNA from leaves of *T. hassleriana* and constructed seven pair-end libraries with insert sizes of 350 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb, and 20 kb. Illumina Hisequation 2000 was then applied to sequence those DNA libraries, and 70.22-Gb raw data were generated. Low-quality reads, reads with adaptor sequences, and duplicated reads were filtered, and the remaining high quality data were used in the assembly. SOAPdenovo2.21 was applied to assemble the genome in the procedure of contig construction, scaffold construction, and gap closure. After gap closure, the assembly was broken down into contigs again according to the position of Ns in the assembly.

WGP and Physical Map Construction

The BAC library was prepared using leaf material of *T. hassleriana*. Two BAC libraries were subsequently generated: the first using *Hin*dIII (CLEH library) and the second using *Eco*RI (CLEE library). Average insert sizes were 145 kb for the CLEH library and 130 kb for the CLEE library. The vector used for library construction was pCC1BAC (Epicentre). For each library, 9600 clones were picked and arrayed into 384 well plates. Together, the two libraries equal ~8.8 genome equivalents (4.6 GE CLEH library and 4.2 GE CLEE library) at an estimated haploid genome size of 300 Mb. WGP was performed according to the methods detailed by van Oeveren et al. (2011). The resulting FPC map was used in further analysis.

FPC Map Assembly and Integration with de Novo Assembled Scaffolds

Sequence-based physical BAC maps were assembled using an improved version of FPC software (Keygene), capable of processing sequence-based BAC fingerprint (WGP) data instead of fragment mobility information as used in the original FPC software (Soderlund et al., 1997). The scaffolds from the SOAPdenovo assembly were then mapped to the physical contigs using nucleotide BLAST (Altschul et al., 1997). Hits were used only when they had a 100% identity match to the anchors. Subsequent filtering was performed to eliminate anchors with multiple hits and to establish superscaffold strand direction. The scaffolds were then ordered according to the mapped anchors and reassembled into superscaffolds.

RNA-Seq

A mixed sample from five tissues (buds, leaves, petioles, stems, and flowers) from *Tarenaya* flowering plants was used to isolate RNA. Total RNA was extracted using Trizol (Invitrogen). The isolated RNA was then treated with RNase-Free DNase and then subsequently with an Illumina mRNA-Seq Prep Kit, following the manufacturer's instructions. The insert size of the RNA libraries was ~200 bp, and the sequencing was done using Illumina GA II. Raw reads were filtered if there were adaptor contaminations and low quality (>10% bases with unknown quality). After filtering, all RNA reads were mapped back to the reference genome using Tophat Version 1.3.3 (Trapnell et al., 2009), implemented with bowtie66 version 0.12.7 (Langmead, 2010), and the transcripts were assembled according to the genome using Cufflinks (version 1.1.0) (Trapnell et al., 2012). Single libraries from eight different tissues were isolated with the Qiagen RNeasy plant mini kit and treated with RNase-free DNase. Illumina TruSeq Libraries were constructed according to the manufacturer's suggestions and sequenced. Raw reads were filtered to remove adaptors and low-quality bases and mapped to the predicted coding sequences using Cufflinks.

Plant Genome Sources

In all analyses where plant genomes are used, source database and version information can be found in Supplemental Table 18 online.

Gene Annotation

Gene models were predicted following several steps: (1) De novo gene prediction. De novo predictions were performed on repeat masked genome assembly. AUGUSTUS (version 2.03) (Stanke and Morgenstern, 2005), GlimmerHMM (version 3.02) (Majoros et al., 2004), and SNAP (version 2.0) were used to perform the de novo annotation. (2) Homology gene prediction. The protein sequences from *Arabidopsis thaliana*, *Brassica rapa*, *Carica papaya*, *Glycine max*, Theobroma cacao, and *Vitis vinifera* were mapped to the *Tarenaya* genome using tBLASTn, by an E-value cutoff of 10^{-5}, and then Genewise (version 2.2.0) (Birney et al., 2004) was used for gene annotation. (3) RNA aided annotation. All the RNA reads were mapped back to the reference genome by Tophat (version 1.0.14) (Trapnell et al., 2009), implemented with bowtie version 0.12.5, and the transcripts were assembled according to the genome using Cufflinks (version 0.8.2) (Trapnell et al., 2012). All the predictions were combined using GLEAN to produce the consensus gene sets.

The tRNA genes were identified by tRNAscan-SE (Lowe and Eddy, 1997). For rRNA identification, the *Arabidopsis* rRNA sequences were first downloaded from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/nuccore). Then, rRNAs in the database were aligned against the *Tarenaya* genome using BLASTn to identify possible rRNAs. Other noncoding RNAs, including microRNA and small nuclear RNA, were identified using INFERNAL (Nawrocki et al., 2009) by searching against the Rfam database.

Gene Family Clustering

OrthoMCL (version 1.4) (Li et al., 2003) was used with default parameters followed by an all-versus-all BLASTP (E-value $\leq$ 1e-5) process using the protein sequence data sets from six plant species. Splice variants were removed from the data set (the longest protein sequence prediction was usually maintained), and the internal stop codons and incompatible reading frames were filtered. All gene families were classified according to the presence or absence of genes for specific species, and it was determined which gene families were species specific or genus specific.

A total of 184204 sequences from *A. thaliana*, *Arabidopsis lyrata*, *B. rapa*, *C. papaya*, *Vitis vinifera*, and *T. hassleriana* were clustered into 24,591 gene families. Of these, 9395 contained sequences from all six genomes, 2492 from Brassicaceae (*A. thaliana*, *A. lyrata*, and *B. rapa*), 1176 from plants as outgroups only bearing the At-γ event (*C. papaya* and *V. vinifera*), and 748 clusters were specific to *Tarenaya*. Of the 28,917 protein-coding genes predicted for *Tarenaya*, 22,482 were clustered in a total of 14,505 groups. The 748 *Tarenaya*-specific clusters contained 1556 genes, of which 529 have at least one INTERPRO domain. Singletons make up a total of 6435 genes, of which 2926 have at least one INTERPRO domain. Interestingly, many gene families that have decreased in number of genes in *Tarenaya* show bigger genes than others, containing more exons and more TE insertions. However, more Gene Ontology annotations are enriched for these contraction gene families. These results indicated that these contraction gene families may be functionally constrained.

Repeat Annotation

Repeats of the *Tarenaya* genome were identified by a combination of homology-based and de novo approaches. In the homology-based method, databases of known repetitive sequences were used to search against the genome assembly. In this way, RepeatMasker-3.2.9 and RepeatProteinMask software (Chen, 2004) were used to build the homology database and search the repeat sequence.

Furthermore, in the de novo approach, three de novo software packages (Piler-DF-1.0 [Edgar and Myers, 2005], RepeatScout-1.0.5, and LTR-FINDER-1.0.5 [Xu and Wang, 2007]) were used to build a de novo repeat database of the *Tarenaya* genome. RepeatMasker was then used

to identify repeats using both the repeat database that we built in-house and Repbase. Finally, the de novo prediction and the homolog prediction of TEs according to the position in the genome were combined.

Phylogenetic Analysis and Species Divergence Time Estimation

The maximum likelihood phylogenetic tree of *T. hassleriana* and other plant genomes was constructed using whole-genome fourfold degenerate sites among species. *Oryza sativa* and *Sorghum bicolor* were taken as the monocot outgroups. The following steps were taken: First, all of the single-copy gene families were extracted from the OrthoMCL clustering results. Second, multiple sequence alignments were run for each single copy gene family using the protein-coding sequences. Third, for each aligned gene family, the CDS back-translation of the protein multiple alignments was performed from the original DNA sequences using in-house Perl scripts, and then the fourfold degenerate sites of orthologous genes in all single-copy gene families (concatenated into one supergene for each species) were extracted. The branch length represents the neutral divergence rate. The substitution model (GTR+gamma+I) and Mrbayes (Ronquist et al., 2012) were used to reconstruct the phylogenetic tree.

Synteny and Collinearity Analysis

First, a homology search within and between species was performed using BLASTP (E-value threshold 1e-7, top 20 hits). Tandem gene families and weak matches were removed using in-house Perl scripts for further analysis. Tandem gene families were defined as clusters of genes within 10 intervening genes from one another, and the longest model was maintained to represent each family. For the weak matches, only the top BLAST hits were retained by applying a C-score threshold of 0.8 [C-score(A, B) = score (A, B)/max(best score of A, best score of B)] (Putnam et al., 2007).

Then, based on these filtered BLAST results, the whole genome-wide sequence alignments within and between genomes using genes as anchors, which was to search syntenic blocks, were conducted by an in-house pipeline implementing Dynamic Programming (parameters: score_of_match, 50; penalty of mismatch, −5; penalty of indel, −5; penalty_of_extension_indel, −2; block_size, ≥5 gene pairs; gap_between_neighbor_blocks, 30 genes). The running time for each whole-genome sequence alignment using genes as anchors by Dynamic programming was ~5 to ~6 h. At the same time, i-Adhore 3.0 (Proost et al., 2012) (http://bioinformatics.psb.ugent.be/software) was used to identify syntenic and collinearity blocks (gap_size = 30, cluster_gap = 35, q_value = 0.75, prob_cutoff = 0.01, anchor_points = 5, alignment_method = gg4, level_2_only = false, table_type = family) within and between genomes, and all the syntenic blocks identified by i-adhore were contained in our results identified by the Dynamic programming method; however, the latter method is more sensitive and accurate. All the dot plot figures were plotted using a Support Vector Graphics package implemented perl scripts (see Supplemental Figures 1, 2, 6 to 9, and 11 to 14 online).

Ancestral Genome Reconstruction

Based on the paralogous duplicates within each genome, four minimized genomes were independently created for *A. thaliana*, *A. lyrata*, *B. rapa*, and *T. hassleriana*, by condensing local duplications to one gene, removing transposons, and including only genes within blocks defined by retained pairs. Each of the minimized genomes represents the ancestral state and predate the recent polyploidy event. At the same time, these four ancestral state genomes were compared with the Ken Wolfe's 45 ancestral blocks of *A. thaliana* (see Supplemental Figures 11 to 14 online).

Partition of the *T. hassleriana* Genome into Three Subgenomes Following the Recent Polyploidy Event

To avoid the potentially confounding results with the independent ancient polyploidy events, the ancestor genome of *A. lyrata* (A ancestor) was used as the reference, and collinear blocks were identified using i-adhore 3.0 by aligning the four proteomes (*A. thaliana*, *A. lyrata*, *B. rapa*, and *T. hassleriana*) against the A ancestor genome. From the last common ancestor (the A ancestor represents this genome state) of these four species, both *A. thaliana* and *A. lyrata* experienced a whole-genome duplication event (At-α), *B. rapa* experienced one whole-genome duplication event (At-α) and an additional whole-genome triplication event (Br-α), and *T. hassleriana* experienced an independently whole-genome triplication event (Th-α). Therefore, quote ratios of 2:1, 2:1, 3/4/5/6:1, and 3:1 were observed for *A. thaliana*, *A. lyrata*, *B. rapa*, and *T. hassleriana* genomes. For *T. hassleriana*, the triplicated blocks identified were chained into three subgenomes compared with the A ancestor using dynamic programming. The main criteria are that the chained triplicated blocks are (1) nonoverlapping in the *T. hassleriana* genome; (2) have no more than 10% overlap between their orthologous A ancestor regions (results were similar using 0% overlap in A ancestor); and (3) maximize coverage of the *T. hassleriana* genome (annotated gene space). A similar strategy was taken for other genomes based on their polyploidy level from the last recent common ancestor, as shown in Supplemental Figures 16 to 19 online. For *T. hassleriana*, a total of 16,770 (63.2%) *Tarenaya* genes are in the triplicated blocks compared with the A ancestor, 688 (2.6%) genes are in the duplicated blocks, which indicate that another copy was possibly lost after the triplication event, and only 135 (0.5%) have one syntenic ortholog, which means that a very small subset of triplicated genes lost two copies. However, 8913 (33.6%) *Tarenaya* genes are not in any replicated blocks, which indicates species-specific deletion in *A. lyrata* or species-specific gains in *Tarenaya* after their divergence along evolutionary time.

Interproscan and Gene Functional Annotation

Interproscan version 4.5 was used to scan protein sequences against the protein signatures from InterPro (Hunter et al., 2009) (version 22.0) to infer functions for the protein-coding genes. This was done for the entire target proteomes involved in our main text analysis, including *A. thaliana*, *A. lyrata*, *T. hassleriana*, *B. rapa*, *Solanum lycopersicum*, *V. vinifera*, *Prunus persica*, *Populus trichocarpa*, and *C. papaya*. InterPro integrates protein families, domains, and functional sites from different databases: Pfam, PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, and PANTHER. Interproscan integrates the searching algorithms of all these databases. In total, Interproscan identified 93,038 protein domains of 4733 distinct domain types. Seventy-five percent of the genes (21,829 out of 28,917 genes in total) have been assigned with at least one domain.

Genomic Analysis for Reproductive Traits

For the syntenic and protein domain analysis of SI genes, the genes annotated in *A. thaliana*, *C. rubella*, and *A. lyrata* were used as published in an extensive study into S-locus variation in *Arabidopsis* species (Guo et al., 2011). Homologs were then sought using top BLAST hits of these genes against *T. hassleriana* and *B. rapa*. All homology candidates were confirmed by manual inspection of the alignment in dot plots generated in MAFFT (Katoh et al., 2002) to confirm synteny of candidate regions. After compiling a definitive list of syntenic regions, the genes from these regions were analyzed through the PFAM online protein domain analysis program (Punta et al., 2012). Figure 6 was manually compiled from the results of this program.

qRT-PCR

For the qRT-PCR, total RNA was isolated from roots, leaves, three bud stages (1 to 5 mm, 5 to 10 mm, and 10 to 25 mm length), sepals, petals, stamens, carpels at anthesis, and three stages of siliques (10 mm, 10 to 30 mm, and 30 to 50 mm length) with the GeneJET plant RNA purification

mini kit (Fermentas). First-strand cDNA was synthesized using 500 ng total RNA with the RevertAid H Minus First Strand cDNA synthesis kit (Fermentas) using random hexamer primers.

The qRT-PCR experiments were performed according to the MIQE guidelines (Bustin et al., 2009). Exon spanning primers were generated using PerlPrimer 1.1.21. (Marshall, 2004). A primer efficiency test was performed and the primers were tested with genomic DNA to ensure cDNA specificity. Standard dose response curves were constructed for all genes using serial dilutions (1:50 to 1:50,000) of 10- to 25-mm-long bud cDNA template to calculate amplification efficiency. The qRT-PCR assay was performed with the LightCycler 480 II (Roche) and the data analyzed with the LCS480 1.5.0.39 software. Each reaction was composed of 10 μL of 2× DyNAmo Flash SYBR Green Mastermix (Biozym Scientific), 2 μL each of 10 μM forward and reverse primers, 1 μL water, and 5 μL of 1:100 diluted template cDNA. Each reaction was performed in biological duplicate and technical triplicate along with water and RNA controls for each primer pair. The *GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C SUBUNIT* and *ELONGATION FACTOR 1-ALPHA* genes served as internal controls. The following PCR program was used: 7 min at 95°C; 45 cycles of 10 s at 95°C, 15 s at 60°C, and 15 s at 72°C, followed by a melting curve of 5 s at 95°C, 1 min at 65°C, and 30 s at 97°C. The quantification cycles (C_q) were calculated according to the second derivative maximum algorithm. The raw C_q data were analyzed according to the Comparitive C_q method ($\Delta\Delta C_q$) (Schmittgen and Livak, 2008). Gene expression was first normalized relative to the expression of the two reference genes in the respective tissues. The expression was further normalized with the expression of the reference genes in 10- to 25-mm long buds, which acted as an interassay calibrator. The relative expression was then calculated with reference to the expression of *T. hassleriana CAL* in stage 3 buds.

Accession Numbers

Sequence data from this article can be found in the Arabidopsis Genome Initiative or GenBank/EMBL databases under the following accession numbers: Th-*CAL/AP1-1*, Th01189; Th-*CAL/AP1-2*, Th13754; Th-*PI-1*, Th05675; Th-*PI-2*, Th17298; Th-*AP3-1*, Th02920; Th-*AP3-2*, Th02921; *EMB196*, AT3G54350; and *TCP1*, AT1G67260. The genomic reads of *T. hassleriana*, as well as RNA sequencing data, have been deposited into the NCBI Short Read Archive under accession numbers SRA058749 and GSM1008474. The information for the raw reads data can be found in Supplemental Table 1 online. The genome sequence and annotation data set have been deposited into NCBI (project ID PRJNA175230 [superscaffolds]; the accession number is AOUI00000000).

Supplemental Data

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

Supplemental Figure 1. Evaluation of Superscaffold13 Integrated by the Combination of SOAPdenovo Assembled Scaffolds and the Physical Map as an Example of Our Verification Method.

Supplemental Figure 2. Comparison of the New *T. hassleriana* Assembly with Previously Sequenced BACs.

Supplemental Figure 3. Principal Component Analysis of Gene Expression.

Supplemental Figure 4. Comparison of the Distribution of Gene Features (Including mRNA Length, CDS Length, Exon Length, Intron Length) among Several Selected Angiosperms.

Supplemental Figure 5. Graph Representing Contig Length versus LTR (from Retrotransposons) Density in the *T. hassleriana* Assembly.

Supplemental Figure 6. Dot Plot Figure to Show Syntenic Duplicates within the *T. hassleriana* Genome.

Supplemental Figure 7. Dot Plot Figure to Show the Collinear Blocks between *A. thaliana* and *T. hassleriana* Genomes.

Supplemental Figure 8. Dot Plot Figure to Show the Collinear Blocks between *A. lyrata* and *T. hassleriana* Genomes.

Supplemental Figure 9. Dot Plot Figure to Show the Collinear Blocks between *B. rapa* and *T. hassleriana* Genomes.

Supplemental Figure 10. Homologous Blocks within and between Genomes Based on the Chromosome Layout of *A. lyrata*.

Supplemental Figure 11. Dot Plot Figure to Show the Corresponding Homologous Relationship between "Wolfe" Blocks and the 64 Ancestor Genome Blocks of *A. thaliana* Identified in This Article and Named "A Ancestor."

Supplemental Figure 12. Dot Plot Figure to Show the Corresponding Homologous Relationship between the "Wolfe" Blocks and the 61 Ancestor Genome Blocks of *A. lyrata* Identified in This Article, Forming the Alternative "A" Ancestor Used in Subgenome Analyses.

Supplemental Figure 13. Dot Plot Figure to Show the Corresponding Homologous Relationship between the "Wolfe" Blocks and the 71 Ancestor Genome Blocks of *B. rapa* Identified in This Article Named "B Ancestor."

Supplemental Figure 14. Dot Plot Figure to Show the Corresponding Homologous Relationship between "Wolfe" Blocks and the 87 Ancestor Genome Blocks of *Tarenaya hassleriana* Identified in This Article Named "C Ancestor."

Supplemental Figure 15. Multiple Homologous Relationships between the Ancestor Genome Blocks ("Wolfe" Ancestral Blocks, A Ancestral Blocks of *A. thaliana*, A Ancestral Blocks *of A. lyrata*, B Ancestral Blocks *of B. rapa*, C Ancestral Blocks of *T. hassleriana*) and the Reference Genome *A. thaliana*.

Supplemental Figure 16. The Genome of *A. lyrata* Is Partitioned into Two Subgenomes (Based on the At-α Event) by Comparing against the Reference Ancestor Genome of *A. lyrata*.

Supplemental Figure 17. The Genome of *A. thaliana* Is Partitioned into Two Subgenomes (Based on the At-α Event) by Comparing against the Reference Ancestor Genome of *A. lyrata*.

Supplemental Figure 18. The Genome of *B. rapa* Is Partitioned into Three Subgenomes (Based on the Br-α Event) by Comparing against the Reference Ancestor Genome of *A. lyrata*.

Supplemental Figure 19. The Genome of *T. hassleriana* Is Partitioned into Three Subgenomes (Based on the Ch-α Event) by Comparing against the Reference Ancestor Genome of *A. lyrata*.

Supplemental Figure 20. Complete ML Tree of MADS Type II Homologs in All Sequenced Angiosperms.

Supplemental Figure 21. *SHP1*, Two Gene Copies, and Synteny.

Supplemental Figure 22. *AP1/CAL* Gene Copies and Synteny.

Supplemental Figure 23. Synteny Plot of the *Tarenaya* Region Containing Th02920/Th02921 (*AP3* Genes).

Supplemental Figure 24. Synteny Plot of the *Tarenaya* Region Containing Th17298 (*PI* Gene).

Supplemental Figure 25. Simplified Synteny Plots of *AP3* and *PI* Genes in *Tarenaya*.

Supplemental Figure 26. *TCP1* Gene Copies and Synteny.

Supplemental Figure 27. Histogram of Gene Expression in Various *Tarenaya* Tissues.

Supplemental Figure 28. SI Locus Domain–Related Genes in Syntenic Block Context in Brassiaceae and *Tarenaya*.

Supplemental Table 1. Library Construction and Sequencing (Raw data).

Supplemental Table 2. Statistics of the Generation of the Clean Data.

Supplemental Table 3. Summary Statistics of the Preliminary Genome Assembly.

Supplemental Table 4. Evaluation of Gene Region Coverage by RNA Mapping.

Supplemental Table 5. Summary of Results after Deconvolution and Filtering of the Whole-Genome Profiling Tags.

Supplemental Table 6. FPC Assembly Results for the High, Reduced, and Low-Stringency WGP Assemblies.

Supplemental Table 7. Final Contig and Scaffold Median Lengths after Superscaffold Construction through Integration with the Physical Map.

Supplemental Table 8. Summary Statistics of RNA-Seq Sequencing Data and the Alignments Mapping to Genes and Genome.

Supplemental Table 9. Mapping Statistics and Transcript Dynamics for Eight RNA Samples.

Supplemental Table 10. Summary Statistics of the Results from Different Gene Annotation Strategies.

Supplemental Table 11. Summary of Functional Annotations Derived from Interproscan.

Supplemental Table 12. Summary of smRNA Annotation.

Supplemental Table 13. Comparison of *T. hassleriana* Assembly with Nine Other Sequenced Crucifers.

Supplemental Table 14. Classification of Transposable Elements within the *T. hassleriana* Genome.

Supplemental Table 15. Statistics of the Ancestor Genome Construction within *A. thaliana*, *A. lyrata*, *B. rapa*, and *T. hassleriana*.

Supplemental Table 16. Syntenic and Collinearity Analysis between *Tarenaya* and Other Species.

Supplemental Table 17. Assembly Coverage of S-Locus-Like Regions in *T. hassleriana*.

Supplemental Table 18. Plant Genome Source Databases.

Supplemental Data Set 1. List of Syntenic Order of *Tarenaya* Genes in the Ancestor Block Configuration of the Genome.

Supplemental Data Set 2. MADS Box Gene Alignment as Used for the ML Phylogenetic Tree in Supplemental Figure 20.

ACKNOWLEDGMENTS

This work was supported by following funding sources to Beijing Genomics Institute, Shenzhen: State Key Laboratory of Agricultural Genomics, Guangdong Provincial Key Laboratory of core collection of crop genetic resources research and application (2011A091000047), Shenzhen Engineering Laboratory of Crop Molecular design breeding, and National Natural Science Funds for Distinguished Young Scholar (30725008). E.v.d.B., J.H., and M.E.S. were supported by the Netherlands Organization for Scientific Research (NWO VIDI Grant 864.10.001 and NWO Ecogenomics Grant 844.10.006). K.K. wishes to thank the Alexander-von-Humboldt Foundation and the BMBF for support. S.d.B. was supported by an NWO Experimental Plant Sciences graduate school "master talent" fellowship. A.P.M.W. thanks the Deutsche Forschungsgemeinschaft for support (Grants WE 2231/9-1 and EXC 1028). We also thank the co-principal investigators of the Brassicales Map Alignment Project for their support (Rod Wing, J. Chris Pires, Thomas Mitchell-Olds, Detlef Weigel, and S. Stephen Wright).

AUTHOR CONTRIBUTIONS

M.E.S., G.Z., J.M.H., and X.Zhu. designed the project. G.Z., M.E.S., P.Z. and S.C. led the sequencing and analysis. C.S., Z.Z., W.L., M.L., Y.T., J.W., X.X, H.Z. and Z.Q. assisted with sequencing and analysis. J.C. and G.F. did the SOAPdenovo genome assembly. J.C. and X.Zhong. did the annotation. P.Z., S.C., E.v.d.B., M.E.S., and C.B. did the evolutionary analysis. S.C., J.X., E.v.d.B., and J.H. constructed the physical map. M.E.S., S.C., and E.v.d.B. did the reproductive trait evolution analysis. K.K. and S.d.B. did the phylogenetic analysis of MADS box genes. C.K., A.Bräutigam., and A.P.M.W. conducted tissue-specific transcriptomic analyses of *T. hassleriana*. A.Bräutigam. and A.Becker. did the qRT-PCR analysis of MADS box genes. J.C.H. did analysis of the *TCP* gene family. M.E.S., E.v.d.B., and S.C. wrote the article.

Received May 7, 2013; revised July 6, 2013; accepted August 6, 2013; published August 27, 2013.

REFERENCES

Ackerman, C.M., Yu, Q., Kim, S., Paull, R.E., Moore, P.H., and Ming, R. (2008). B-class MADS-box genes in trioecious papaya: Two paleoAP3 paralogs, CpTM6-1 and CpTM6-2, and a PI ortholog CpPI. Planta **227:** 741–753.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815.

Ashburner, M., et al.; The Gene Ontology Consortium (2000). Gene ontology: Tool for the unification of biology. Nat. Genet. **25:** 25–29.

Barker, M.S., Vogel, H., and Schranz, M.E. (2009). Paleopolyploidy in the Brassicales: Analyses of the Cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. Genome Biol. Evol. **1:** 391–399.

Bartlett, M.E., and Specht, C.D. (2010). Evidence for the involvement of Globosa-like gene duplications and expression divergence in the evolution of floral morphology in the Zingiberales. New Phytol. **187:** 521–541.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S. R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **107:** 18724–18728.

Bennett, M.D., Leitch, I.J., Price, H.J., and Johnston, J.S. (2003). Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus ~25% larger than the *Arabidopsis* genome initiative estimate of ~125 Mb. Ann. Bot. (Lond.) **91:** 547–557.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. **14:** 988–995.

Blanc, G., Hokamp, K., and Wolfe, K.H. (2003). A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. **13:** 137–144.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003).

The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31:** 365–370.

Bowers, J.E., Chapman, B.A., Rong, J.K., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433–438.

Boyes, D.C., Nasrallah, M.E., Vrebalov, J., and Nasrallah, J.B. (1997). The self-incompatibility (S) haplotypes of *Brassica* contain highly divergent and rearranged sequences of ancient origin. Plant Cell **9:** 237–247.

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.

Busch, A., Horn, S., Mühlhausen, A., Mummenhoff, K., and Zachgo, S. (2012). *Corolla* monosymmetry: Evolution of a morphological novelty in the Brassicaceae family. Mol. Biol. Evol. **29:** 1241–1254.

Busch, A., and Zachgo, S. (2007). Control of corolla monosymmetry in the Brassicaceae *Iberis amara*. Proc. Natl. Acad. Sci. USA **104:** 16714–16719.

Busch, A., and Zachgo, S. (2009). Flower symmetry evolution: Towards understanding the abominable mystery of angiosperm radiation. BioEssays **31:** 1181–1190.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J., and Wittwer, C.T. (2009). The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. Clinical Chemistry **55:** 611–622.

Chen, K.G., Fan, B.F., Du, L.Q., and Chen, Z.X. (2004). Activation of hypersensitive cell death by pathogen-induced receptor-like protein kinases from *Arabidopsis*. Plant Mol. Biol. **56:** 271–283.

Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics **4:**10.

Colombo, M., Brambilla, V., Marcheselli, R., Caporali, E., Kater, M.M., and Colombo, L. (2010). A new role for the SHATTERPROOF genes during *Arabidopsis* gynoecium development. Dev. Biol. **337:** 294–302.

Couvreur, T.L.P., Franzke, A., Al-Shehbaz, I.A., Bakker, F.T., Koch, M.A., and Mummenhoff, K. (2010). Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol. Biol. Evol. **27:** 55–71.

Cruden, R.W., and Lloyd, R.M. (1995). Embryophytes have equivalent sexual phenotypes and breeding systems: Why not a common terminology to describe them? Am. J. Bot. **82:** 816–825.

Dassanayake, M., Oh, D.-H., Haas, J.S., Hernandez, A., Hong, H., Ali, S., Yun, D.-J., Bressan, R.A., Zhu, J.-K., Bohnert, H.J., and Cheeseman, J.M. (2011). The genome of the extremophile crucifer *Thellungiella parvula*. Nat. Genet. **43:** 913–918.

Dwyer, K.G., Kandasamy, M.K., Mahosky, D.I., Acciai, J., Kudish, B.I., Miller, J.E., Nasrallah, M.E., and Nasrallah, J.B. (1994). A superfamily of S locus-related sequences in *Arabidopsis*: Diverse structures and expression patterns. Plant Cell **6:** 1829–1843.

Edgar, R.C., and Myers, E.W. (2005). PILER: Identification and classification of genomic repeats. Bioinformatics **21** (suppl. 1)**:** i152–i158.

Edger, P.P., and Pires, J.C. (2009). Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. **17:** 699–717.

Endress, P.K. (1999). Symmetry in flowers: Diversity and evolution. Int. J. Plant Sci. **160** (S6)**:** S3–S23.

Engler, F.W., Hatfield, J., Nelson, W., and Soderlund, C.A. (2003). Locating sequence on FPC maps and selecting a minimal tiling path. Genome Res. **13:** 2152–2163.

Flagel, L.E., and Wendel, J.F. (2009). Gene duplication and evolutionary novelty in plants. New Phytol. **183:** 557–564.

Franzke, A., Lysak, M.A., Al-Shehbaz, I.A., Koch, M.A., and Mummenhoff, K. (2011). Cabbage family affairs: The evolutionary history of Brassicaceae. Trends Plant Sci. **16:** 108–116.

Guo, Y.L., Zhao, X., Lanz, C., and Weigel, D. (2011). Evolution of the S-locus region in Arabidopsis relatives. Plant Physiol. **157:** 937–946.

Hall, J.C., Sytsma, K.J., and Iltis, H.H. (2002). Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. Am. J. Bot. **89:** 1826–1842.

Haudry, A., et al. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. **45:** 891–898.

Hu, T.T., et al. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat. Genet. **43:** 476–481.

Hunter, S., et al. (2009). InterPro: The integrative protein signature database. Nucleic Acids Res. **37** (Database issue)**:** D211–D215.

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 Cleomaceae I. On the separate recognition of Capparaceae, Cleomaceae, and Brassicaceae. Ann. Mo. Bot. Gard. **98:** 28–36.

Jabbour, F., Nadot, S., and Damerval, C. (2009). Evolution of floral symmetry: A state of the art. C. R. Biol. **332:** 219–231.

Jaillon, O., et al; French-Italian Public Consortium for Grapevine Genome Characterization (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467.

Jiao, Y., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97–100.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28:** 27–30.

Kasahara, M. (2007). The 2R hypothesis: An update. Curr. Opin. Immunol. **19:** 547–552.

Katoh, K., Misawa, K., Kuma, K.i., and Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30:** 3059–3066.

Koornneef, M., and Meinke, D. (2010). The development of *Arabidopsis* as a model plant. Plant J. **61:** 909–921.

Kramer, E.M., Holappa, L., Gould, B., Jaramillo, M.A., Setnikov, D., and Santiago, P.M. (2007). Elaboration of B gene function to include the identity of novel floral organs in the lower eudicot *Aquilegia*. Plant Cell **19:** 750–766.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr. Protoc. Bioinformatics **11:** 7.

Li, L., and Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. **13:** 2178–2189.

Li, R.Q., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. **20:** 265–272.

Litt, A., and Kramer, E.M. (2010). The ABC model and the diversification of floral organ identity. Semin. Cell Dev. Biol. **21:** 129–137.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:** 955–964.

Machado, I., Cristina Lopes, A., Valentina Leite, A., and Virgíniade Brito Neves, C. (2006). *Cleome spinosa* (Capparaceae): Polygamodioecy and pollination by bats in urban and Caatinga areas, northeastern Brazil. Bot. Jahrb. Syst. Pflanzengesch. Pflanzengeogr. **127:** 69–82.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics **20:** 2878–2879.

Marquard, R.D., and Steinback, R. (2009). A model plant for a biology curriculum: Spider flower (*Cleome hasslerana* L.). Am. Biol. Teach. **71:** 235–244.

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.

Marshall, O.J. (2004). PerlPrimer: Cross-platform, graphical primer design for standard, bisulphite and real-time PCR. Bioinformatics **20:** 2471–2472.

McMillan, L.E.M., and Martin, A.C.R. (2008). Automatically extracting functionally equivalent proteins from SwissProt. BMC Bioinformatics **9:** 418.

Ming, R., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). Nature **452:** 991–996.

Mondragón-Palomino, M., and Theissen, G. (2009). Why are orchid flowers so diverse? Reduction of evolutionary constraints by paralogues of class B floral homeotic genes. Ann. Bot. (Lond.) **104:** 583–594.

Murat, F., Van de Peer, Y., and Salse, J. (2012). Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. Genome Biol. Evol. **4:** 917–928.

Nasrallah, M.E., Liu, P., Sherman-Broyles, S., Boggs, N.A., and Nasrallah, J.B. (2004). Natural variation in expression of self-incompatibility in *Arabidopsis thaliana*: Implications for the evolution of selfing. Proc. Natl. Acad. Sci. USA **101:** 16070–16074.

Navarro-Quezada, A.R. (2007). Molecular Evolution of Tropinone-Reductase-Like and Tau GST Genes Duplicated in Tandem in Brassicaceae. (Munich, Germany: LMU Munich).

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: Inference of RNA alignments. Bioinformatics **25:** 1335–1337.

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. Nucleic Acids Res. **37:** 289–297.

Pastuglia, M., Swarup, R., Rocher, A., Saindrenan, P., Roby, D., Dumas, C., and Cock, J.M. (2002). Comparison of the expression patterns of two small gene families of S gene family receptor kinase genes during the defence response in *Brassica oleracea* and *Arabidopsis thaliana*. Gene **282:** 215–225.

Patchell, M.J., Bolton, M.C., Mankowski, P., and Hall, J.C. (2011). Comparative floral development in Cleomaceae reveals two distinct pathways leading to monosymmetry. Int. J. Plant Sci. **172:** 352–365.

Preston, J.C., and Hileman, L.C. (2012). Parallel evolution of TCP and B-class genes in Commelinaceae flower bilateral symmetry. EvoDevo **3:** 6.

Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., and Vandepoele, K. (2012). i-ADHoRe 3.0—Fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Res. **40:** e11.

Punta, M., et al. (2012). The Pfam protein families database. Nucleic Acids Res. **40** (Database issue)**:** D290–D301.

Putnam, N.H., et al. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science **317:** 86–94.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. **61:** 539–542.

Rosin, F.M., and Kramer, E.M. (2009). Old dogs, new tricks: Regulatory evolution in conserved genetic modules leads to novel morphologies in plants. Dev. Biol. **332:** 25–35.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of *Arabidopsis thaliana* development. Nat. Genet. **37:** 501–506.

Schmittgen, T.D., and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative C(T) method. Nat. Protoc. **3:** 1101–1108.

Schnable, J.C., Wang, X., Pires, J.C., and Freeling, M. (2012). Escape from preferential retention following repeated whole genome duplication in plants. Front. Plant Sci. **3:** 94.

Schranz, M.E., Lysak, M.A., and Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: Building blocks of crucifer genomes. Trends Plant Sci. **11:** 535–542.

Schranz, M.E., and Mitchell-Olds, T. (2006). Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell **18:** 1152–1165.

Schranz, M.E., Mohammadin, S., and Edger, P.P. (2012). Ancient whole genome duplications, novelty and diversification: The WGD Radiation Lag-Time Model. Curr. Opin. Plant Biol. **15:** 147–153.

Shen, J.X., Fu, T.D., Yang, G.S., Ma, C.Z., and Tu, J.X. (2005). Genetic analysis of rapeseed self-incompatibility lines reveals significant heterosis of different patterns for yield and oil content traits. Plant Breed. **124:** 111–116.

Slotte, T., et al. (2013). The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat. Genet. **45:** 831–835.

Smaczniak, C., et al. (2012). Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. Proc. Natl. Acad. Sci. USA **109:** 1560–1565.

Soderlund, C., Longden, I., and Mott, R. (1997). FPC: A system for building contigs from restriction fingerprinted clones. Comput. Appl. Biosci. **13:** 523–535.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. **33** (Web Server issue)**:** W465–W467.

Stout, A.B. (1923). Alternation of sexes and intermittent production of fruit in the spider flower (*Cleome spinosa*). Am. J. Bot. **10:** 57–66.

Tang, H., Lyons, E., Pedersen, B., Schnable, J.C., Paterson, A.H., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics **12:** 102.

Tanksley, S.D. (2004). The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. Plant Cell **16** (suppl.)**:** S181–S189.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485:** 635–641.

Theissen, G. (2001). Development of floral organ identity: Stories from the MADS house. Curr. Opin. Plant Biol. **4:** 75–85.

Thomas, B.C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. **16:** 934–946.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics **25:** 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. **7:** 562–578.

Tsai, W.C., Kuoh, C.S., Chuang, M.H., Chen, W.H., and Chen, H.H. (2004). Four DEF-like MADS box genes displayed distinct floral morphogenetic roles in *Phalaenopsis* orchid. Plant Cell Physiol. **45:** 831–844.

Tsai, W.C., Pan, Z.J., Hsiao, Y.Y., Jeng, M.F., Wu, T.F., Chen, W.H., and Chen, H.H. (2008). Interactions of B-class complex proteins involved in tepal development in *Phalaenopsis* orchid. Plant Cell Physiol. **49:** 814–824.

Tuskan, G.A., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313:** 1596–1604.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. **10:** 725–732.

van Oeveren, J., de Ruiter, M., Jesse, T., van der Poel, H., Tang, J.F., Yalcin, F., Janssen, A., Volpin, H., Stormo, K.E., Bogden, R., van Eijk, M.J.T., and Prins, M. (2011). Sequence-based physical mapping of complex genomes by whole genome profiling. Genome Res. **21:** 618–625.

Vekemans, D., Proost, S., Vanneste, K., Coenen, H., Viaene, T., Ruelens, P., Maere, S., Van de Peer, Y., and Geuten, K. (2012). Gamma paleohexaploidy in the stem lineage of core eudicots: Significance for MADS-box gene and species diversification. Mol. Biol. Evol. **29:** 3793–3806.

Viaene, T., Vekemans, D., Irish, V.F., Geeraerts, A., Huysmans, S., Janssens, S., Smets, E., and Geuten, K. (2009). Pistillata—Duplications as a mode for floral diversification in (Basal) asterids. Mol. Biol. Evol. **26:** 2627–2645.

Wang, X.W., et al; Brassica rapa Genome Sequencing Project Consortium (2011). The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. **43:** 1035–1039.

Wing, R.A., Mitchell-Olds, T., Pires, J.C., Schranz, M.E., Weigel, D., and Wright, S. (2013). Brassicales Map Alignment Project (BMAP). http://www.brassica.info/resource/sequencing/bmap.php. Accessed August 22, 2013.

Wrzaczek, M., Brosché, M., Salojärvi, J., Kangasjärvi, S., Idänheimo, N., Mersmann, S., Robatzek, S., Karpiński, S., Karpińska, B., and Kangasjärvi, J. (2010). Transcriptional regulation of the CRK/DUF26 group of receptor-like protein kinases by ozone and plant hormones in *Arabidopsis*. BMC Plant Biol. **10:** 95.

Wu, H.-J., et al. (2012). Insights into salt tolerance from the genome of *Thellungiella salsuginea*. Proc. Natl. Acad. Sci. USA **109:** 12219–12224.

Xu, Z., and Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35** (Web Server issue)**:** W265–W268.

Yang, R., et al. (2013). The reference genome of the halophytic plant *Eutrema salsugineum*. Front. Plant Sci. **4:** 46.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan—An integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:** 847–848.

Zhang, X., Wang, L., Yuan, Y., Tian, D., and Yang, S. (2011). Rapid copy number expansion and recent recruitment of domains in S-receptor kinase-like genes contribute to the origin of self-incompatibility. FEBS J. **278:** 4323–4337.

VIII.2 CO-AUTHORED MANUSCRIPTS

Manuscript 5

Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae

*Erik van den Bergh, **Canan Külahoglu***, Andrea Bräutigam, Julian M. Hibberd, Andreas P.M. Weber, Xin-Guang Zhu, M. Eric Schranz*

Published in Current Plant Biology (2014), http://dx.doi.org/10.1016/j.cpb.2014.08.001

Impact Factor: 9.82

**Co-author*

Main findings:

This study investigated the role of whole genome duplication as a source of gene neofunctionalization within the evolution of C_4 photosynthesis. This was done by comparing the expression between paralog copies of *Gynandropsis gynandra* (C_4) with its C_3 relative *Tarenaya hassleriana*. The comparison of paralog synonymous substitution rate showed that *G. gynandra* and *T. hassleriana* share the paleohexaploidy. Gene copy number of photosynthetic gene families is similar between the C_3 and C_4 species. For the analyzed species polyploidy does not alter significantly the copy number of essential C_4 photosynthesis relevant genes.

The transcript abundance in the C_4 species is more strictly controlled compared to the C_3 species. Thus, the recruitment of existing genes through regulatory changes may have played a larger role in C_4 evolution than neofunctionalization of recently duplicated genes.

Contributions:

- Cleome transcriptome sequencing
- RNA-seq data processing
- Assembly and quantification
- Proof-reading of Manuscript

G Model
CPB-5; No. of Pages 8
ARTICLE IN PRESS

Current Plant Biology xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae

Erik van den Bergh[a], Canan Külahoglu[b], Andrea Bräutigam[b], Julian M. Hibberd[c], Andreas P.M. Weber[b], Xin-Guang Zhu[d], M. Eric Schranz[a,*]

[a] *Biosystematics, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands*
[b] *Institute of Plant Biochemistry, Center of Excellence on Plant Sciences(CEPLAS), Heinrich-Heine-University, D-40225 Düsseldorf, Germany*
[c] *Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom*
[d] *Plant Systems Biology Group, Partner Institute of Computational Biology, Chinese Academy of Sciences/Max Planck Society, Shanghai 200031, China*

ARTICLE INFO

Article history:
Received 15 May 2014
Received in revised form 19 August 2014
Accepted 23 August 2014

Keywords:
Plant genome evolution
Synteny
Cleomaceae
Brassicaceae
Bioinformatics
Whole genome duplication
Paleopolyploidy
C_4 photosynthesis

ABSTRACT

C_4 photosynthesis is a trait that has evolved in 66 independent plant lineages and increases the efficiency of carbon fixation. The shift from C_3 to C_4 photosynthesis requires substantial changes to genes and gene functions effecting phenotypic, physiological and enzymatic changes. We investigate the role of ancient whole genome duplications (WGD) as a source of new genes in the development of this trait and compare expression between paralog copies. We compare *Gynandropsis gynandra*, the closest relative of Arabidopsis that uses C_4 photosynthesis, with its C_3 relative *Tarenaya hassleriana* that underwent a WGD named Th-α. We establish through comparison of paralog synonymous substitution rate that both species share this paleohexaploidy. Homologous clusters of photosynthetic gene families show that gene copy numbers are similar to what would be expected given their duplication history and that no significant difference between the C_3 and C_4 species exists in terms of gene copy number. This is further confirmed by syntenic analysis of *T. hassleriana*, *Arabidopsis thaliana* and *Aethionema arabicum*, where syntenic region copy number ratios lie close to what could be theoretically expected. Expression levels of C_4 photosynthesis orthologs show that regulation of transcript abundance in *T. hassleriana* is much less strictly controlled than in *G. gynandra*, where orthologs have extremely similar expression patterns in different organs, seedlings and seeds. We conclude that the Th-α and older paleopolyploidy events have had a significant influence on the specific genetic makeup of Cleomaceae versus Brassicaceae. Because the copy number of various essential genes involved in C_4 photosynthesis is not significantly influenced by polyploidy combined with the fact that transcript abundance in *G. gynandra* is more strictly controlled, we also conclude that recruitment of existing genes through regulatory changes is more likely to have played a role in the shift to C_4 than the neofunctionalization of duplicated genes.
DATA: The data deposited at NCBI represents raw RNA reads for each data series mentioned: 5 leaf stages, root, stem, stamen, petal, carpel, sepal, 3 seedling stages and 3 seed stages of Tarenaya hassleriana and Gynandropsis gynandra. The assembled reads were used for all analyses of this paper where RNA was used. http://www.ncbi.nlm.nih.gov/Traces/sra/?study=SRP036637, http://www.ncbi.nlm.nih.gov/Traces/sra/?study=SRP036837

1. Introduction

Over sixty lineages of both monocot and eudicot angiosperms have evolved a remarkable solution to maximize photosynthesis efficiency under low CO_2 levels, high temperatures and/or drought: C_4 photosynthesis [1]. The evolution of this modified photosynthetic pathway represents a wonderful example of convergent evolution. While the changes necessary for the transition from C_3 to C_4 photosynthesis are numerous, the trait has a wide phylogenetic distribution across angiosperms, with 19 different plant

* Corresponding author. Tel.: +31 317483160.
E-mail addresses: erik.vandenbergh@wur.nl (E. van den Bergh), canan.kuelahoglu@uni-duesseldorf.de (C. Külahoglu), andrea.braeutigam@uni-duesseldorf.de (A. Bräutigam), jmh65@cam.ac.uk (J.M. Hibberd), andreas.weber@uni-duesseldorf.de (A.P.M. Weber), zhuxinguang@picb.ac.cn (X.-G. Zhu), eric.schranz@wur.nl (M. Eric Schranz).

http://dx.doi.org/10.1016/j.cpb.2014.08.001

G Model
CPB-5; No. of Pages 8
ARTICLE IN PRESS

families across the globe known to contain one or multiple members capable of C_4 photosynthesis [2]. Much research on eudicot C_4 has focused on *Flaveria* species (Asteraceae), which contains not only C_4 species but also a number of C_3/C_4 intermediates [3]. With the emergence of genomics and the choice of *Arabidopsis thaliana* as the genomics standard model organism, species in the Cleomaceae, a sister-family to the Brassicaceae (containing Arabidopsis and Brassica crops) have been proposed for genetic studies of C_4 [4,5].

C_4 plants spatially separate the fixation of carbon away from the RuBisCO active site by using phospho*enol*pyruvate carboxylase, an alternate carboxylase that does not react with oxygen. As a consequence they are more efficient under permissive conditions [6]. The typical C_4 system is characterized by a morphological change: so-called Kranz anatomy [7]. In this anatomy, specialized mesophyll (M) cells surround enlarged bundle sheath (BS) cells, with the leaf veins internal to the BS. Generally, the veination in C_4 leaves is increased [8]. This internal leaf architecture physically partitions the biochemical events of the C_4 pathway into two main phases. In the first phase, dissolved HCO_3^- is assimilated into C_4 acids by phospho*enol*pyruvate carboxylase (PEPC) in the mesophyll cells. In the second phase, these acids diffuse into the chloroplast loaded bundle sheath (BS) cells, where they are decarboxylated and the released CO_2 is fixed by RuBisCO. The increased CO_2 concentration in the BS cells allows carbon fixation by RuBisCO to be much more efficient by reducing photorespiration. Two subtypes of the C_4 biochemical pathway are defined, based on the most active C_4 acid decarboxylase that liberates CO_2 from C_4 acids in the bundle sheath: NADP-malic enzyme (NADP-ME), NAD-malic enzyme (NAD-ME); a facultative addition of phosphoenolpyruvate carboxykinase (PEPCK) activity can be present in either subtype [9]. The subtypes are used as a classification scheme for C_4.

The process of carboxylation and decarboxylation costs more energy than the simpler C_3 form of photosynthesis, but it diminishes photorespiration. In conditions of low atmospheric CO_2 pressure, photorespiration causes a major loss in photosynthetic output and the elaborate concentrating mechanisms of C_4 photosynthesis circumvent this [10].

All genes important for the C_4 pathway are expressed at relatively low levels in C_3 leaves [11]. The mechanism for recruitment of these genes into the C_4 pathway remains to be elucidated. For some ancestral C_3 genes changes in *cis*-regulatory elements, while in others changes in *trans* generate M and BS cell specificity [12–14], indicating variation in the mechanisms underlying gene recruitment into the C_4 pathway. It has been proposed that gene duplication and subsequent neofunctionalization of one gene copy has facilitated the alterations in gene expression that underlie the evolution of C_4 photosynthesis [15,16]. Gene duplication is proposed to be a (pre)condition for the evolution of C_4 because it allows the organism to maintain the original gene while a duplicate version can acquire beneficial changes. This can lead to significant changes in metabolism without the deleterious effect of modifications to essential genes. A recent study that compared convergent evolution of photosynthetic pathways with parallel evolution concluded that duplications are not essential for the development of C_4 biochemistry, but rather changes in expression and localization of specific genes [11,17]. However, this study highlighted just the number of C_4 genes and did not take into account the age and mechanism of gene duplications.

The modifications necessary for the anatomical changes from C_3 to C_4 photosynthesis are not well established. Recent work has shown that the SCARECROW (SCR) gene that is responsible for vein formation in roots, can produce proliferated bundle sheath cells as well as other changes that can be coupled to the shift to the Kranz anatomy [18]. Further work supports this relation by describing the role that the upstream interacting partner of SCR, SHORT-ROOT (SHR) plays in the variations in anatomy seen in various C_4 species [19,20].

Gene duplicates must be further refined by the mechanism by which they arise; either as single gene tandem duplication or whole genome duplication (WGD). Tandem duplications occur frequently, but the duplicates are often lost again resulting in a constant birth–death cycle of duplicate genes [21]. Second, there is whole genome duplication (WGD) or polyploidy, where all genes are simultaneously duplicated. After duplication there are often dramatic changes in the plant genomic structure, a process referred to as diploidization in which most genes return to single copy. However, the genes that are maintained in duplicate after WGD often have important functions in enzyme complexes (e.g. to maintain proper gene balance [22]) or can diversify and evolve new gene functions (e.g. neo-functionalization).

The contribution of WGD to photosynthesis-related genes has been studied in soybean, barrel-medic, Arabidopsis, and sorghum [23,24]. The polyploid and non-polyploid duplicated gene retention in *Glycine max*, *Medicago truncatula* and Arabidopsis for four classes of photosynthesis-related genes was compared: the Calvin–Benson–Bassham-cycle (CBBC), the light-harvesting complex (LHC), photosystem I (PSI) and photosystem II (PSII). It was found that photosystem genes were more dosage sensitive, with more duplicates derived only from WGD whereas CC gene families were often larger with more non-polyploid duplicates retained. In *Sorghum bicolor*, a recent WGD was reported to be an important origin of C_4 specific genes. Several key C_4 genes of this crop were found to be collinear with genes that function in C_3 photosynthesis when compared to maize and rice. Here, we combine the approaches of these two studies to examine the evolution of photosynthesis and C_4-related genes in C_3 and C_4 Cleomaceae species.

Gynandropsis gynandra (Fig. 1, blue clade) belongs to the NAD-ME C_4 photosynthesis sub-type [25,26] and is an important South-East Asian and African dry-season leafy vegetable (sometimes referred to as Phak-sian or African cabbage), and is closely related to horticultural C_3 species *Tarenaya hassleriana* (Fig. 1, pink clade). Both species are easily cultivated in the greenhouse, and a robust phylogenetic framework for Cleomaceae species is emerging [4,5,27]. There are two other independent origins of the C_4 within the Cleomaceae, *Cleome angustifolia* and *Cleome oxalidea* (Fig. 1, yellow clade), identified by carbon isotope discrimination [5,25]. Because of the economic importance and ease of growth, the C_4–C_3 contrast between *G. gynandra* and *T. hassleriana* makes this system most attractive and tractable. Both species also have relatively small genome sizes (*T. hassleriana* = 292 Mb and *G. gynandra* ≈ 1 Gb). *T. hassleriana* underwent a WGD named Th-α [28] but it is not yet known whether this event is shared with all or a subset of other Cleomaceae.

In this study we compare C_3 *T. hassleriana* of the Cleomaceae with C_4 *G. gynandra* of the same family. We use the knowledge of Brassicaceae gene functions to identify the important photosynthetic genes in both species and address the following questions: Does *G. gynandra* share the Th-α event? What is contribution of duplicate genes to photosynthesis and C_4-related gene families? And finally, what is the role of gene duplicates from WGD compared to continuous small-scale duplications?

2. Methods

2.1. Transcriptome sequencing and assembly

All transcriptome data was used directly from the Cleomaceae transcript atlas [17]. In the atlas, *T. hassleriana* genes were used as a reference to map transcripts from both species to Cleomaceae "unigenes" indicated by the gene name coined in the published *T.*

Please cite this article in press as: E. van den Bergh, et al., Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae, Curr. Plant Biol. (2014), http://dx.doi.org/10.1016/j.cpb.2014.08.001

G Model
CPB-5; No. of Pages 8
ARTICLE IN PRESS

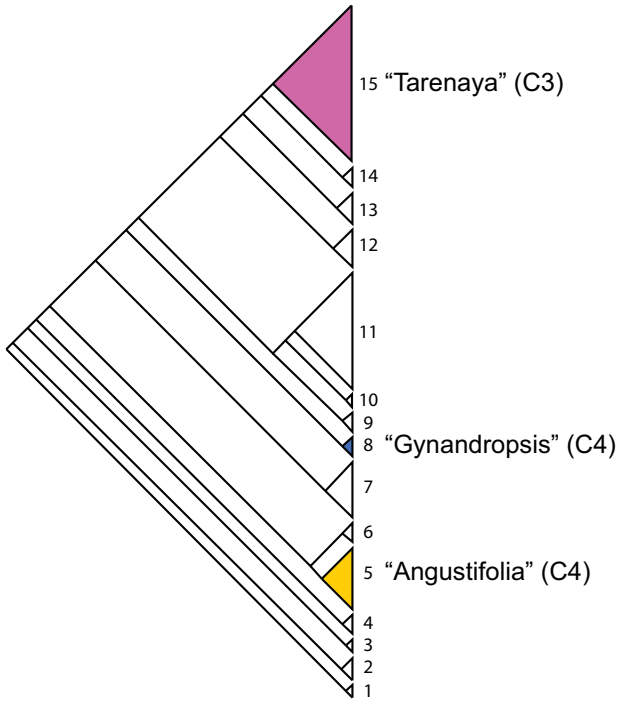


Fig. 1. Simplified phylogeny of Cleomaceae. Clades are numbered following the most recently published Maximum Likelihood phylogeny of Cleomaceae [25]. Clade 15 containing *T. hassleriana* is marked in pink. Clade 8 containing *G. gynandra* is marked in blue. Clade 5 (Yellow) contains the other origin of C_4 in Cleomaceae, with *C. angustifolia* and C_4/C_3 intermediate *C. paradoxa*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

hassleriana genome [29]. For gene quantification we used default BlatV35 parameters [30] in protein space for mapping, counting the best matched hit based on e-value for each read uniquely.

2.2. Homolog selection

A TBlastX [31] search of transcriptomes of *T. hassleriana* and *G. gynandra* was performed with default parameters (no evalue cutoff) to have a maximum number of hits for subsequent filtering. To filter paralogs and orthologs from these results, CIP/CALP filtering was used [32]. Cumulative Identity Percentage (CIP) is defined as the sum of the number of matching nucleotides for each high-scoring segment pair (HSP) of a pair of genes divided by the total lengths of those HSPs. Cumulative alignment length percentage (CALP) is defined as the sum of the alignment lengths of all HSPs of a matching gene pair divided by the total length of the query sequence. Both of these values give a reliable estimation of the similarity of two genes and is a more accurate method than evalue or bit score threshold filtering. A CIP/CALP threshold of 50/50 was chosen as a suitable cutoff point for orthology and/or paralogy.

2.3. Ks/4dtv calculation of paralog pairs

Paralogs identified with CIP/CALP filtering were aligned using Exonerate [33] with the coding2coding model parameter, using a custom output format through the "roll your own" parameter. The exact command line used was: "exonerate -m c2c seq1.fasta seq2.fasta –ryo \"%Pqs %Pts\\n" –showalignment false –verbose 0". The output from this command was fed into CodeML from the PAML package using standard parameters (Codonfreq = 2, kappa = 2, omega = 0.4). Output from PAML [34] was parsed using custom Perl scripts to read the synonymous substitution rate (*Ks*) and the fourfold transversion rate (4dtv). This workflow is identical to the established paralog identification pipeline Duppipe [35] using updated tools and more stringent selection using CIP/CALP.

2.4. Homolog clustering

Photosynthesis genes were selected from known functionally annotated Arabidopsis genes. Gene identifiers used for each family are listed hereafter and in Table 2. βCA: AT1G23730, AT1G58180, AT1G70410, AT3G01500, AT4G33580, AT5G14740. MDH (cytosolic): AT1G04410, AT5G43330, AT5G56720. MDH (mitochondrial): AT1G53240, AT2G22780, AT3G15020, AT3G47520, AT5G09660. MDH (peroxisomal): AT1G53240, AT2G22780, AT3G15020, AT3G47520, AT5G09660. MDH (plastidic): AT1G53240, AT2G22780, AT3G15020, AT3G47520, AT5G09660. NAD-ME: AT2G13560, AT4G00570. NADP-ME: AT1G79750, AT2G19900, AT5G11670, AT5G25880. PEPC: AT1G21440, AT1G53310, AT2G42600, AT3G14940. PPCK: AT1G08650, AT3G04530, AT3G04550, AT4G37870, AT5G28500, AT5G65690. These genes were then used as a BLAST database and queried with *T. hassleriana* and *G. gynandra* atlas unigenes. Hits were then filtered using a 50/50 CIP/CALP cutoff. Using custom Perl scripts, the hits of these hits were picked up, iterating recursively until convergence (no new hits found). All unique genes resulting from this process form a family cluster.

2.5. Synteny analyses

T. hassleriana genes were used as a query in the CoGe Synfind [36] program using the following parameters: Comparison algorithm: Last, Gene window size: 40, Minimum number of genes: 4, Scoring Function: Collinear, Syntenic depth: unlimited. As query genomes, the following were used: *A. arabicum* VEGI unmasked v2.5, *A. thaliana* Col-0 TAIR unmasked v10.02 and *T. hassleriana* BGI; Eric Scranz Lab; Weber lab unmasked v5.

3. Results

3.1. Evidence of WGD in both species confirming a shared event

Using the transcript sets of *G. gynandra* and *T. hassleriana*, paralogs were matched to each other by BLAST search and CIP/CALP filtering. In total, 55,014 paralogs were found: 26,883 in *T. hassleriana* covering 49% of transcript space and 28,131 in *G. gynandra* covering 48% of transcript space. Of all paralog pairs, *Ks* and fourfold transversion substitutions (4dtv) were determined and binned to establish an evolutionary time distribution (Fig. 2). In both species a large gene birth event has taken place around *Ks* = 0.4 (Fig. 2 between *Ks* = 0.25 and *Ks* = 0.5), which corresponds to the *Ks* window established earlier for the Th-α hexaploidy event [28]. The same analysis was performed using 4dtv values and results were extremely similar. Enumerating the paralogs that fall within the Th-α peak, we see that 15,785 gene pairs in *T. hassleriana* are retained from the Th-α paleohexaploidy, or ~29% of the total transcriptome. For *G. gynandra*, 16,096 gene pairs fall within the Th-α window, or around 27% of all transcripts.

3.2. Duplicate loss and retention in essential C_4 families

We examined six gene families that are essential in C_4 photosynthesis in detail: NAD malic enzyme (NAD-ME), NADP malic enzyme (NADP-ME), β carbonic anhydrase (βCA), malate dehydrogenase (MDH), phospho*enol*pyruvate carboxylase (PEPC) and phospho*enol*pyruvate carboxykinase (PPCK). Using Arabidopsis genes as a reference, homologous clusters were created using a

Please cite this article in press as: E. van den Bergh, et al., Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae, Curr. Plant Biol. (2014), http://dx.doi.org/10.1016/j.cpb.2014.08.001

G Model
CPB-5; No. of Pages 8
ARTICLE IN PRESS

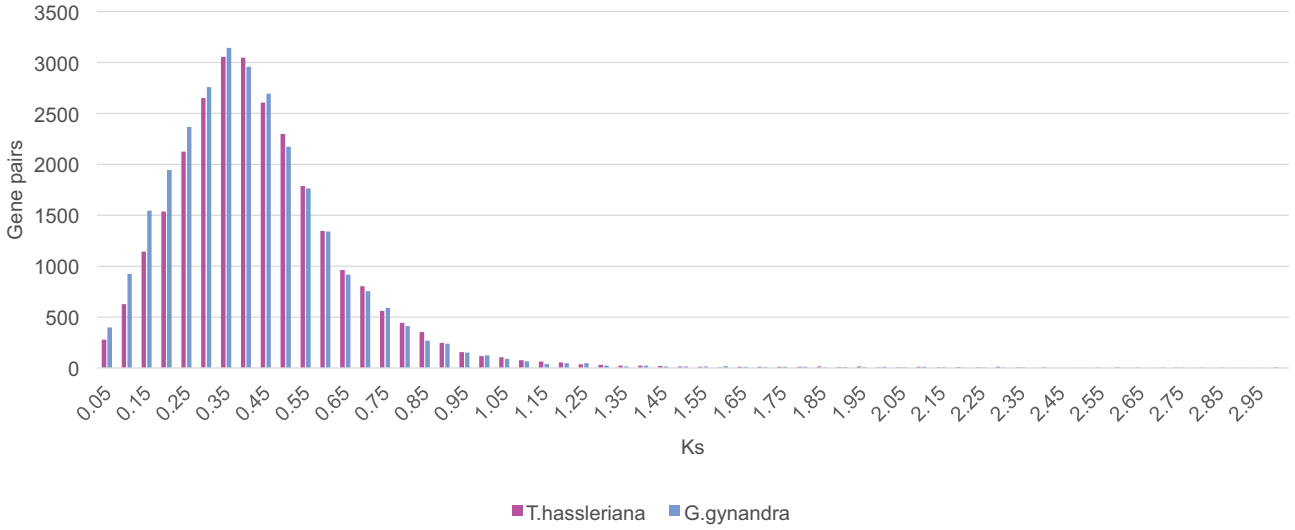


Fig. 2. Histogram showing the amount of gene pairs per Ks bin for *T. hassleriana* (pink) and *G. gynandra* (blue). The peak at around *Ks* = 0.45 is an indication of a massive gene birth event and is considered evidence of paleopolyploidy. Both species have an extremely similar peak, indicating that this is a shared polyploidy event. The *Ks* values of these peaks corresponds with *Ks* values found earlier for the Th-α hexaploidy event, indicating that this event has occurred before divergence of *T. hassleriana* and *G. gynandra*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

CIP/CALP cutoff of 50/50. 146 homologous pairs could be placed in a cluster across the three species comprising 105 unique genes (Table 1); 40 in *A. thaliana*, 57 in *T. hassleriana* and 49 in *G. gynandra*. In most cases both Cleomaceae species have around 1.5 times the number of genes of *A. thaliana* except, interestingly, the NADP-ME family where numbers are almost the same in all species. Also of note is that *T. hassleriana* has 16% more C4 related genes in total than *G. gynandra* (57 over 49).

All genes of one species in a cluster were then aligned to each other and the *Ks* value of each pairing was established and subsequently binned with a stepsize of *Ks* = 0.15 (Fig. 3). At the *Ks* corresponding to the Th-α hexaploidy, both *T. hassleriana* and *G. gynandra* show a relative increase of gene pairs with this amount of synonymous substitutions. *A. thaliana* at the *Ks* of its older At-α event shows a similar, if slightly lower increase. Even longer ago in

Table 1
C4 photosynthesis homolog cluster sizes in *A. thaliana*, *T. hassleriana* and *G. gynandra*. Both Cleomaceae species have around 1.5 times the number of genes of *A. thaliana* except the NADP-ME and NAD-ME families where numbers are lower than average in the Cleomaceae species resulting in a similar amount of homologs in each species for these two gene groups.

	A. thaliana	*T. Hassleriana*	*G. gynandra*
βCA	6	10	7
MDH (cyt.)	3	6	6
MDH (mit.)	5	6	6
MDH (per.)	5	8	6
MDH (plast.)	5	6	6
NAD-ME	2	3	3
NADP-ME	4	4	3
PEPC	4	8	6
PPCK	6	6	6
Total	40	57	49

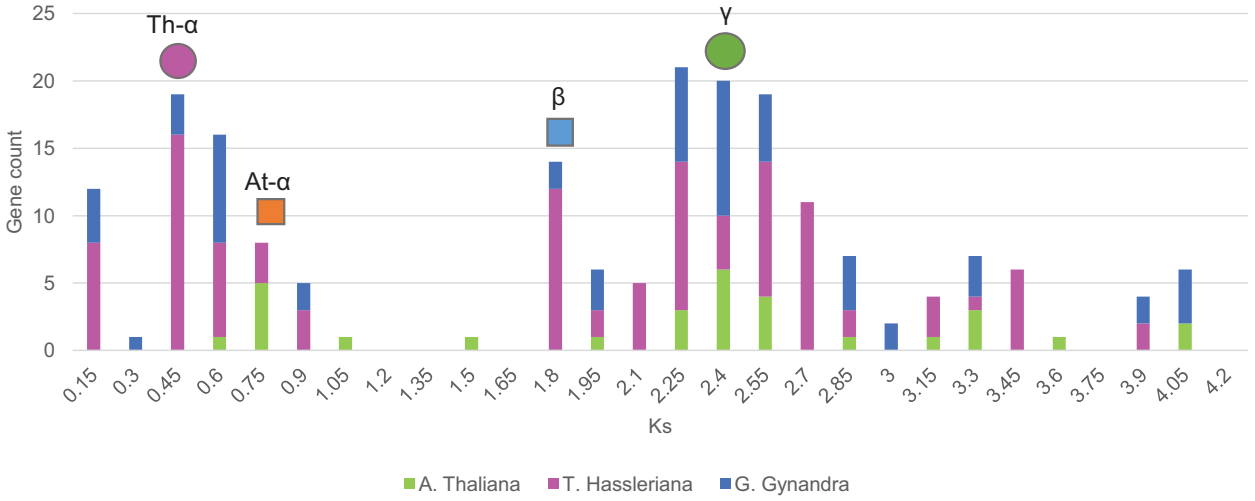


Fig. 3. Histogram showing *Ks* values of homolog gene clusters associated with C_4 photosynthesis: MDH, NAD-ME, NADP-ME, PEPC and βCA. Gene duplication events are marked at their associated *Ks* value and colored according to earlier publication [28]; a square indicates a duplication (tetraploidy), a circle indicates a triplication (hexaploidy). The contribution of the Th-α (pink circle) and the At-α (orange square) on photosynthesis related gene copy number can be seen at *Ks* = 0.45 and *Ks* = 0.6 respectively. The *β* event at *Ks* = 1.8 (blue square) has contributed substantially to the expansion of gene copy number in *T. hassleriana*. Further in evolutionary time, around *Ks* = 2.4, the γ event (green circle) that is also shared by all three species has contributed equally to the polyploid presence in photosyntenic orthologs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Please cite this article in press as: E. van den Bergh, et al., Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae, Curr. Plant Biol. (2014), http://dx.doi.org/10.1016/j.cpb.2014.08.001

G Model
CPB-5; No. of Pages 8
ARTICLE IN PRESS

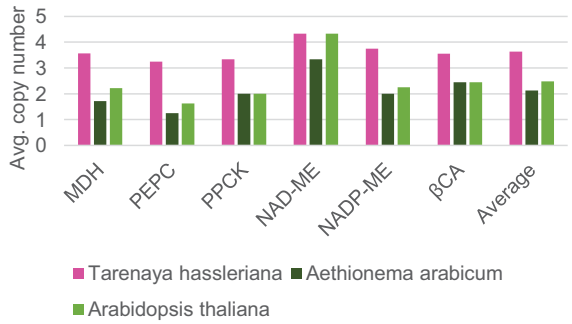


Fig. 4. Histogram showing average syntenic region copy number for *T. hassleriana*, *A. thaliana* and *Aethionema arabicum*. Because *A. arabicum* and *A. thaliana* both share a paleotetraploidy, the expected ratio of syntenic regions for *T. hassleriana*: *A. thaliana*: *Aethionema arabicum* is 3:2:2. In most cases, syntenic regions follow this distribution which is also reflected in the average ratio of all families being 3.6:2.1:2.5 (rightmost bars). The exception is NAD-ME, where the average region number in both *A. arabicum* and *A. thaliana* is as high as *T. hassleriana*.

evolutionary time at the *Ks* corresponding to the β event *T. hassleriana* has retained ~20% of C_4 related genes, where the other species show 2% and 0% retention for *G. gynandra* and *A. thaliana*, respectively. The final confirmed paleohexaploidy that all three species share, the ancient γ event at *Ks* = 2.4, has contributed substantially to the genetic makeup of all three species. In *A. thaliana* the number of relations that stem from the γ paleohexaploidy is 23%, with both Cleomaceae at 15% and 21% for *T. hassleriana and G. gynandra*, respectively.

3.3. Syntenic copy number variation

Syntenic analyses of the previously mentioned gene families was performed using CoGe Synfind [36]. Each *T. hassleriana* c4 related ortholog was used as a query with *T. hassleriana*, *Arabidopsis thaliana*, *A. arabicum* [37] as a basal representative of Brassicaceae. Thus for the *T. hassleriana*: *A. thaliana*: *A. arabicum* ortholog ratio we would theoretically expect 3 (Th-α):2 (At-α):2. Query results were enumerated and the average number of regions per family was determined (Fig. 4). For many families, the average is comparable to the 3:2:2 ratio, which is also represented by the average ratio (Fig. 4, rightmost set of bars) being 3.6:2.1:2.5. The exception is the NAD-ME family, which has seen more than expected retention with an orthologs ratio 4.3:3.3:4.3. The PEPC family also seems slightly under-retained in Brassicaceae, with a ratio of 3.3:1.3:1.6. Unfortunately, syntenic data is impossible to obtain without a sequenced genome so data syntenic regions of *G. gynandra* will have to be obtained in future work.

3.4. Regulation of photosynthetic homolog expression

Both Cleomaceae have substantially more copies of photosynthetic genes (Fig. 4). Using the Cleomaceae expression atlases [17], the expression of separate copies was compared in the C_3 and the C_4 species. In the expression atlas, the *T. hassleriana* coding sequence was used as a reference to map expression in both *T. hassleriana* and *G. gynandra* to a single Cleomaceae 'unigene'. Expression was quantified in nine different tissues including three developmental series: development from young to mature leaf (six stages), root, stem, stamen, petal, carpel, sepal, a seedling developmental series (three stages) and a seed time series (three stages).

For the photosynthetic gene families (NAD-ME, NADP-ME, PEPCK, PEPC, MDH, CA), homolog selection resulted in a data set of 43 unigenes with expression data for both Cleomaceae species.

Table 2
List of Arabidopsis genes used as representatives of C4 photosynthesis families. ATG identifiers correspond to identifier following the ATG system from the Arabidopsis Information Resource [43].

Gene family	ATG identifiers
βCA	AT1G23730 AT1G58180 AT1G70410 AT3G01500 AT4G33580 AT5G14740
MDH (cytosolic)	AT1G04410 AT5G43330 AT5G56720
MDH (mitochondrial)	AT1G53240 AT2G22780 AT3G15020 AT3G47520 AT5G09660
MDH (peroxisomal)	AT1G53240 AT2G22780 AT3G15020 AT3G47520 AT5G09660
MDH (plastidic)	AT1G53240 AT2G22780 AT3G15020 AT3G47520 AT5G09660
NAD-ME	AT2G13560 AT4G00570
NADP-ME	AT1G79750 AT2G19900 AT5G11670 AT5G25880
PEPC	AT1G21440 AT1G53310 AT2G42600 AT3G14940
PPCK	AT1G08650 AT3G04530 AT3G04550 AT4G37870 AT5G28500 AT5G65690

Expression levels were normalized and compared amongst photosynthetic gene families, examples of which are plotted for NAD-ME and βCA (Fig. 5). Immediately noticeable is the highly similar expression profiles of *G. gynandra* when compared to the more chaotic profiles of *T. hassleriana*. This is observed in all except one gene family. *G. gynandra* has 176 expressed unigenes with a highly correlated expression pattern (Pearson correlation > 0.95) whereas in *T. hassleriana* 87 unigenes share a highly correlated expression pattern (Pearson correlation > 0.95).

The expression pattern that is observed in *G. gynandra* in the β-CA family also correspond to their *A. thaliana* highest ranking match (Table 2). The cluster consisting of C.spinosa_00253, C.spinosa_13896, C.spinosa_18526 and C.spinosa_10164 for example all match highest to *A. thaliana* gene β carbonic anhydrase 4 (AT1G70410). The cluster consisting of C.spinosa_07642 and C.spinosa_13410 both map to carbonic anhydrase 1 (AT3G01500). A similar pattern is present in NAD-ME where the cluster of C.spinosa_03046 and C.spinosa_09126 both map to NAD-ME1 (AT2G13560) and the C.spinosa_12536 singleton maps to NAD-ME2 (AT4G00570).

Please cite this article in press as: E. van den Bergh, et al., Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae, Curr. Plant Biol. (2014), http://dx.doi.org/10.1016/j.cpb.2014.08.001

G Model
CPB-5; No. of Pages 8
ARTICLE IN PRESS

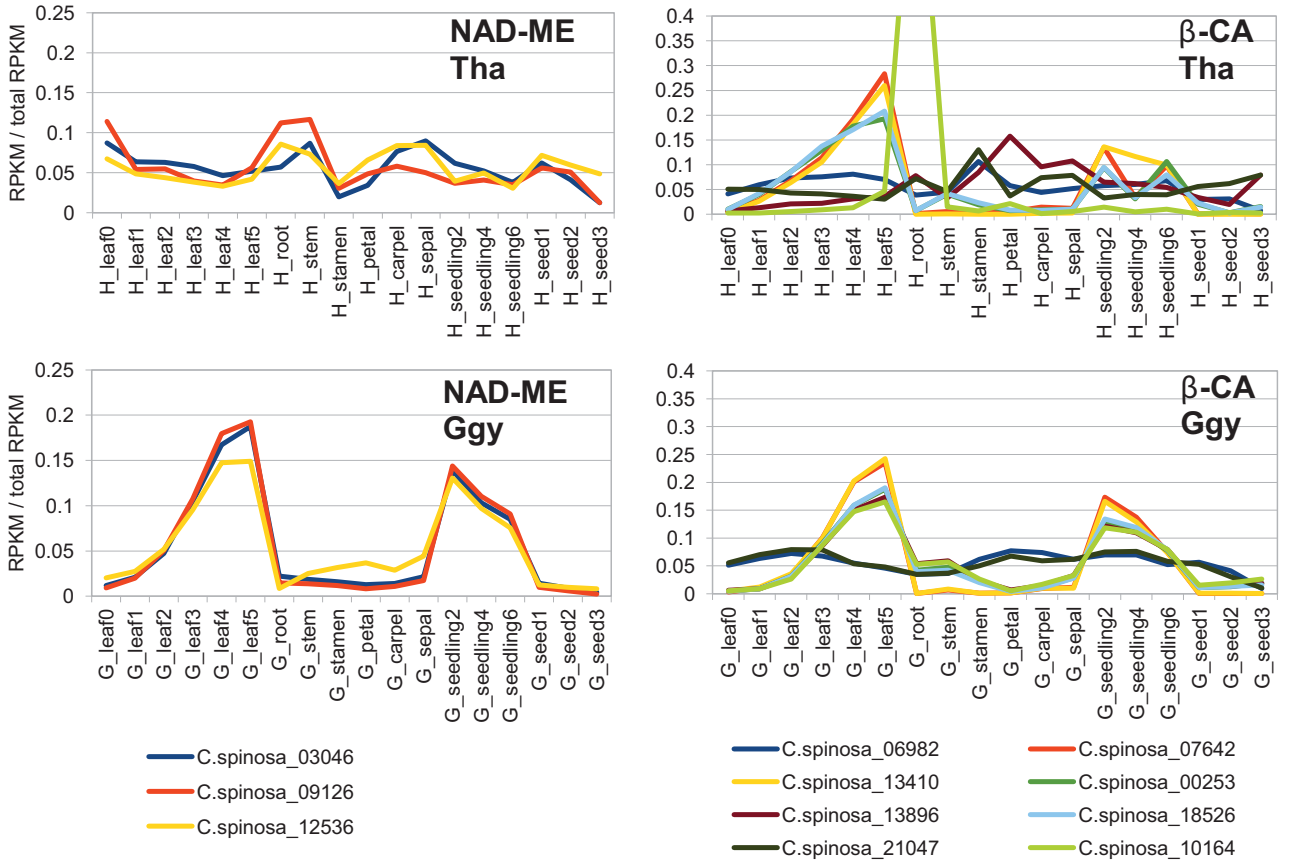


Fig. 5. Canalization in expression of NAD malic enzyme (top and bottom left) and β carbonic anhydrase (top and bottom right) homologs in *T. hassleriana* and *G. gynandra*. Top left: NAD-ME expression in *T. hassleriana*. Top right: βCA expression in *T. hassleriana*. Bottom left: NAD-ME expression in *G. gynandra*. Bottom right: βCA expression in *G. gynandra*. (Mapped) gene names and associated colors are displayed, see Materials and Methods for more details on the mapping of *G. gynandra* transcripts to *T. hassleriana* genes. Note that leaf0–leaf5 as well as seedling2–seedling6 and seed1–seed3 are time series of the same organ, with the leaf and seedling gradient being two days separated by stage. Transcription levels in *G. gynandra* (lower graphs) are more strictly regulated across organs, seeds and seedlings. The chaotic patterns in *T. hassleriana* (upper graphs) results in half the genes having a Pearson correlation > 0.95 compared to *G. gynandra*.

4. Discussion and conclusions

In this study, we have analyzed the transcriptomes of the C_3 *T. hassleriana* and C_4 *G. gynandra* to address the potential contribution of WGD and recent gene duplicates to the evolution of photosynthesis and C_4-pathway related genes. The initial comparison of *T. hassleriana* and *G. gynandra* was performed to identify the differential expression of key-genes involved in the NAD-ME C_4 biochemical pathway. However, it did not consider the role of gene duplicates. We show that very distinct patterns will occur when the duplication history is taken into account.

We could confirm the Th-α hexaploidy that has been found in *T. hassleriana* using an independent transcriptome dataset. We also find that *G. gynandra* shares this WGD with *T. hassleriana*, further establishing the occurrence of WGD in this lineage. Based on the phylogenetic position of both species in Cleomaceae, the Th-α duplication took place at least before the divergence of the two species which means that it is shared across Cleomaceae lineages 8–15 according to the latest phylogeny of the family [25]. Dating this polyploidy event in terms of absolute age is always a difficult task, however, here we find that the *Ks* rate of *G. gynandra* is extremely similar if not identical to *T. hassleriana.* Assuming then that mutation rates between these two species are the same, we can reaffirm the previous date estimation of Th-α at 13.7 mya [38].

The influence of the Th-α WGD event on photosynthetic gene composition is apparent, both in ortholog number as well as in syntenic region copy number for both species. From absolute orthologs numbers we can see that there is no increased retention between Cleomaceae species and even a slightly lower rate of retention in *G. gynandra.* This indicates that both species have experienced similar evolutionary constraints for a significant amount of time. Also we need to consider that genes sharing a similar sequence, do not necessarily have to share the same function. Even using strict CIP/CALP filtering which has been proved to be an accurate measure for the prediction of true orthologs [32], differential expression either in time, localization or regulation can substantially change the function of a gene. This is especially the case for genes in the core C_4 photosynthesis pathway, where many C_3 genes have been recruited into new functions [13,39].

When establishing *Ks* values of deeper ortholog nodes of photosynthesis genes, a large proportion of genes seem to have been retained from the γ duplication. For a trait that is likely to be highly dosage sensitive [23], we expect that gene loss will be rare and that remnants from this old paleohexaploidy are still present. However, considering the time that has passed since the γ paleohexaploidy event and on the basis of absolute gene copy numbers some gene loss has taken place predating the transition from C_3 to C_4.

The evolutionary importance of WGD events is made clear from the dominant presence of retained Th-α genes in both Cleomaceae species. However, certain questions remain: Can we couple this importance to the evolution of specific traits or in this case, C_4 photosynthesis? This is an old discussion, dating back to the works of

Please cite this article in press as: E. van den Bergh, et al., Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae, Curr. Plant Biol. (2014), http://dx.doi.org/10.1016/j.cpb.2014.08.001

G Model
CPB-5; No. of Pages 8
ARTICLE IN PRESS

Ohno who was the first to suggest that the massive radiation of vertebrates was caused by a whole genome duplication in the ancestor [40]. An earlier study on the evolution of photosynthesis in soybean, showed that the Calvin–Benson–Bassham cycle (CBBC) and the light harvesting complex (LHC) gene families show a greater expansion from single gene duplications than both photosystem groups. This is explained by the increased dosage sensitivity of photosystem genes: if some subunits are expressed differently due to duplications while others are not, this is deleterious for the system as a whole [23]. This acts as a conservation mechanism for gene copy number that does not affect the more loosely connected enzyme collection of the CBBC and LHC genes.

In *G. gynandra*, where the expression of C_4 genes is tightly linked in clusters we would expect a high retention of orthologs. However, this dependency on transcriptional regulation has not lead to an increased retention of photosynthetic genes, as evidenced by lower copy numbers for all C_4 gene families when compared to *T. hassleriana.* It is not likely that neofunctionalization of genes after polyploidy has played a major role in the shift to C_4 photosynthesis. The much more stringent transcriptional regulation of C4 cycle genes in *G. gynandra* when compared to *T. hassleriana* as evidenced in this study is in accordance with the alternative hypothesis, which states that this process was mainly due to recruitment of existing genes in transcriptional space as suggested by several authors [12,14,41,42].

We still have much to learn regarding the development of C_4 photosynthesis. When studying this exceptional trait, we must always consider the genetic history of the species in question. Here, we give evidence that duplications, on a large scale and small, contribute to trait evolution. The exact mechanisms behind the recruitment of these genes into new biochemical pathways however are still largely unknown. Current sequencing efforts for *G. gynandra* will significantly aid in finding the detailed mechanisms of gene and C_4 photosynthesis evolution. The *Cleome* genus provides an excellent model system for unraveling the evolutionary origin and workings of C_4 photosynthesis and hopefully will enable us to harvest the fruits of our knowledge on this remarkable form of plant energy conversion.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Cleome transcriptome sequencing, processing, assembly and quantification was done by CK. AB and APMW provided comments on handling highly expressed duplicates as well as proofreading the manuscript. EvdB performed the bioinformatic analyses. EvdB and MES prepared the manuscript. JMH and XZ proofread and edited the manuscript.

Acknowledgements

The work of EvB and MES was funded by NWO Vernieuwingsimpuls Vidi grant number 864.10.001. APMW acknowledges support by the Deutsche Forschungsgemeinschaft (SPP 1529; EXC 1028).

References

[1] R.F. Sage, P.-A. Christin, E.J. Edwards, The C4 plant lineages of planet Earth, J. Exp. Bot. 62 (2011) 3155–3169.
[2] R.F. Sage, The evolution of C4 photosynthesis, N. Phytol. 161 (2004) 341–370.
[3] M.S. Ku, J. Wu, Z. Dai, R.A. Scott, C. Chu, G.E. Edwards, Photosynthetic and photorespiratory characteristics of Flaveria species, Plant Physiol. 96 (1991) 518–528.
[4] N.J. Brown, K. Parsley, J.M. Hibberd, The future of C4 research – maize, Flaveria or Cleome? Trends Plant Sci. 10 (2005) 215–221.
[5] D.M. Marshall, R. Muhaidat, N.J. Brown, Z. Liu, S. Stanley, H. Griffiths, R.F. Sage, J.M. Hibberd, Cleome: a genus closely related to Arabidopsis, contains species spanning a developmental progression from C3 to C4 photosynthesis, Plant J. 51 (2007) 886–896.
[6] X.-G. Zhu, S.P. Long, D.R. Ort, Improving photosynthetic efficiency for greater yield, Annu. Rev. Plant Biol. 61 (2010) 235–261.
[7] G.E. Edwards, V.R. Franceschi, E.V. Voznesenskaya, Single-cell C4 photosynthesis versus the dual-cell (Kranz) paradigm, Annu. Rev. Plant Biol. 55 (2004) 173–196.
[8] A.D. McKown, N.G. Dengler, Vein patterning and evolution in C4 plants, Botany 88 (2010) 775–786.
[9] Y. Wang, A. Bräutigam, A.P.M. Weber, X.-G. Zhu, Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis, J. Exp. Bot. 65 (2014) 3567–3578.
[10] J.R. Ehleringer, T.E. Cerling, B.R. Helliker, C4 photosynthesis, atmospheric CO_2, and climate, Oecologia 112 (1997) 285–299.
[11] A. Bräutigam, K. Kajala, J. Wullenweber, M. Sommer, D. Gagneul, K.L. Weber, K.M. Carr, U. Gowik, J. Maß, M.J. Lercher, An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species, Plant Physiol. 155 (2011) 142–156.
[12] N.J. Brown, C.A. Newell, S. Stanley, J.E. Chen, A.J. Perrin, K. Kajala, J.M. Hibberd, Independent and parallel recruitment of preexisting mechanisms underlying C4 photosynthesis, Science 331 (2011) 1436–1439.
[13] J.M. Hibberd, S. Covshoff, The regulation of gene expression required for C4 photosynthesis, Annu. Rev. Plant Biol. 61 (2010) 181–207.
[14] K. Kajala, N.J. Brown, B.P. Williams, P. Borrill, L.E. Taylor, J.M. Hibberd, Multiple Arabidopsis genes primed for recruitment into C4 photosynthesis, Plant J. 69 (2012) 47–56.
[15] K. Monson Russell, Gene duplication, neofunctionalization, and the evolution of C_4 photosynthesis, Int. J. Plant Sci. 164 (2003) S43–S54.
[16] R.K.S. Monson, F. Rowan, The origins of C4 genes and evolutionary pattern in the C4 metabolic phenotype, in: Academic Press (Ed.), C4 Plant Biology, 1999, pp. 377–410, San Diego.
[17] C. Külahoglu, A.K. Denton, M. Sommer, J. Maß, S. Schliesky, T.J. Wrobel, B. Berckmans, E. Gongora-Castillo, C.R. Buell, R. Simon, et al., Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C3 and C4 Plant Species, The Plant Cell Online (2014), http://dx.doi.org/10.1105/tpc.114.123752, Advance online publication.
[18] T.L. Slewinski, A.A. Anderson, C. Zhang, R. Turgeon, Scarecrow plays a role in establishing Kranz anatomy in maize leaves, Plant Cell Physiol. 53 (2012) 2030–2037.
[19] T.L. Slewinski, A.A. Anderson, S. Price, J.R. Withee, K. Gallagher, R. Turgeon, Short-root1 plays a role in the development of vascular tissue and Kranz anatomy in maize leaves, Mol. Plant 7 (2014) 1388–1392.
[20] P. Wang, S. Kelly, J.P. Fouracre, J.A. Langdale, Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz anatomy, Plant J. 75 (2013) 656–670.
[21] S.B. Cannon, A. Mitra, A. Baumgarten, N.D. Young, G. May, The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*, BMC Plant Biol. 4 (2004) 10.
[22] P. Edger, J.C. Pires, Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes, Chromosome Res. 17 (2009) 699–717.
[23] J.E. Coate, J.A. Schlueter, A.M. Whaley, J.J. Doyle, Comparative evolution of photosynthetic genes in response to polyploid and nonpolyploid duplication, Plant Physiol. 155 (2011) 2081–2095.
[24] X. Wang, U. Gowik, H. Tang, J.E. Bowers, P. Westhoff, A.H. Paterson, Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses, Genome Biol. 10 (2009) R68.
[25] T.A. Feodorova, E.V. Voznesenskaya, G.E. Edwards, E.H. Roalson, Biogeographic patterns of diversification and the origins of C4 in Cleome (*Cleomaceae*), Syst. Bot. 35 (2010) 811–826.
[26] E.V. Voznesenskaya, N.K. Koteyeva, S.D. Chuong, A.N. Ivanova, J. Barroca, L.A. Craven, G.E. Edwards, Physiological, anatomical and biochemical characterisation of photosynthetic types in genus Cleome (*Cleomaceae*), Funct. Plant Biol. 34 (2007) 247–267.
[27] R.D. Marquard, R. Steinback, A model plant for a biology curriculum: spider flower (*Cleome hasslerana* L.), Am. Biol. Teach. 71 (2009) 235–244.
[28] M.S. Barker, H. Vogel, M.E. Schranz, Paleopolyploidy in the Brassicales: analyses of the cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales, Genome Biol. Evol. 1 (2009) 391–399.
[29] S. Cheng, E. van den Bergh, P. Zeng, X. Zhong, J. Xu, X. Liu, J. Hofberger, S. de Bruijn, A.S. Bhide, C. Kuelahoglu, The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers, Plant Cell Online 25 (2013) 2813–2830.
[30] W.J. Kent, BLAT – the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.
[31] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[32] F. Murat, Y. Van de Peer, J. Salse, Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes, Genome Biol. Evol. 4 (2012) 917–928.
[33] G.S. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison, BMC Bioinform. 6 (2005) 31.

Please cite this article in press as: E. van den Bergh, et al., Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae, Curr. Plant Biol. (2014), http://dx.doi.org/10.1016/j.cpb.2014.08.001

[34] Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood, Comput. Appl. Biosci. 13 (1997) 555–556.
[35] M.S. Barker, K.M. Dlugosch, L. Dinh, R.S. Challa, N.C. Kane, M.G. King, L.H. Rieseberg, EvoPipes.net. Bioinformatic tools for ecological and evolutionary genomics, Evol. Bioinform. 6 (2010) 143–149.
[36] E. Lyons, M. Freeling, How to usefully compare homologous plant genes and chromosomes as DNA sequences, Plant J. 53 (2008) 661–673.
[37] A. Haudry, A.E. Platts, E. Vello, D.R. Hoen, M. Leclercq, R.J. Williamson, E. Forczek, Z. Joly-Lopez, J.G. Steffen, K.M. Hazzouri, et al., An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions, Nat. Genet. 45 (2013) 891–898 (advance online publication).
[38] M.S. Barker, H. Vogel, M.E. Schranz, Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales, Genome Biol. Evol. 1 (2009) 391.
[39] U. Gowik, J. Burscheidt, M. Akyildiz, U. Schlue, M. Koczor, M. Streubel, P. Westhoff, cis-Regulatory elements for mesophyll-specific gene expression in the C4 plant Flaveria trinervia, the promoter of the C4 phosphoenolpyruvate carboxylase gene, Plant Cell Online 16 (2004) 1077–1090.
[40] S. Ohno, U. Wolf, N.B. Atkin, Evolution from fish to mammals by gene duplication, Hereditas 59 (1968) 169–187.
[41] U. Gowik, P. Westhoff, The path from C3 to C4 photosynthesis, Plant Physiol. 155 (2011) 56–63.
[42] B.P. Williams, S. Aubry, J.M. Hibberd, Molecular evolution of genes recruited into C4 photosynthesis, Trends Plant Sci. 17 (2012) 213–220.
[43] P. Lamesch, T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D.L. Alexander, M. Garcia-Hernandez, et al., The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools, Nucleic Acids Res. 36 (2011) D1009–D1014.

Please cite this article in press as: E. van den Bergh, et al., Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae, Curr. Plant Biol. (2014), http://dx.doi.org/10.1016/j.cpb.2014.08.001

VIII.3 CO-AUTHORED MANUSCRIPTS

Manuscript 6

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

Andrea Bräutigam, Simon Schliesky, ***Canan Külahoglu****, Colin P. Osborne and Andreas P.M. Weber*

Published in Journal of Experimental Botany (2014) 65: 13, pp. 3579–3593

Impact Factor: 5.79

**Co-author*

Main findings:

This study presents the leaf transcriptomes of the phosphoenolpyruvate carboxykinase (PEPCK) C_4 photosynthesis subtype grass *Megathyrsus maximus* and its sister species the C_3 grass *Dichantelium clandestinum*. These transcriptomes were compared with public available data of NAD-ME and NADP-ME C_4 subtype species to provide an integrative study of all C_4 photosynthesis subtypes. The core C_4 cycle of the PEPCK species *M. maximus* is closer related to the NAD-ME subtype than to the NADP-ME subtype plants. Differences between the PEPCK and NAD-ME subtype were the subcellular localization of transfer acid generation by differentially localized aspartate amino transferases, differences in decarboxylation biochemistry in form of alterations in PEPCK and NAD-ME transcript abundances. Within the scope of this study the flux of metabolites within the cell was estimated at least twice higher for C_4 plants compared to C_3 species. Based on the transcriptome results the PEPCK subtype C_4 cycle appears to be simpler than NAD-ME and NADP-ME plants, since it requires fewer alterations of intracellular transport processes and changes in electron transfer. This makes the PEPCK subtype an attractive target for C_4 engineering.

Contributions:

- Fresh culm cross sections
- Confocal microscopy
- Enzyme assays
- Manuscript editing

Journal of Experimental Botany, Vol. 65, No. 13, pp. 3579–3593, 2014
doi:10.1093/jxb/eru100 Advance Access publication 18 March, 2014
This paper is available online free of all access charges (see http://jxb.oxfordjournals.org/open_access.html for further details)

Journal of Experimental Botany
www.jxb.oxfordjournals.org

RESEARCH PAPER

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

Andrea Bräutigam[1,*,†], Simon Schliesky[1,†], Canan Külahoglu[1], Colin P. Osborne[2] and Andreas P.M. Weber[1,*]

[1] Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, Universitätsstrasse 1, D-40225 Düsseldorf, Germany
[2] Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

* To whom correspondence should be addressed. E-mail: andreas.weber@uni-duesseldorf.de
† These two authors contributed equally to this work.

Received 3 December 2013; Revised 4 February 2014; Accepted 10 February 2014

Abstract

C_4 photosynthesis affords higher photosynthetic carbon conversion efficiency than C_3 photosynthesis and it therefore represents an attractive target for engineering efforts aiming to improve crop productivity. To this end, blueprints are required that reflect C_4 metabolism as closely as possible. Such blueprints have been derived from comparative transcriptome analyses of C_3 species with related C_4 species belonging to the NAD-malic enzyme (NAD-ME) and NADP-ME subgroups of C_4 photosynthesis. However, a comparison between C_3 and the phospho*enol*pyruvate carboxykinase (PEP-CK) subtype of C_4 photosynthesis is still missing. An integrative analysis of all three C_4 subtypes has also not been possible to date, since no comparison has been available for closely related C_3 and PEP-CK C_4 species. To generate the data, the guinea grass *Megathyrsus maximus*, which represents a PEP-CK species, was analysed in comparison with a closely related C_3 sister species, *Dichanthelium clandestinum*, and with publicly available sets of RNA-Seq data from C_4 species belonging to the NAD-ME and NADP-ME subgroups. The data indicate that the core C_4 cycle of the PEP-CK grass *M. maximus* is quite similar to that of NAD-ME species with only a few exceptions, such as the subcellular location of transfer acid production and the degree and pattern of up-regulation of genes encoding C_4 enzymes. One additional mitochondrial transporter protein was associated with the core cycle. The broad comparison identified sucrose and starch synthesis, as well as the prevention of leakage of C_4 cycle intermediates to other metabolic pathways, as critical components of C_4 metabolism. Estimation of intercellular transport fluxes indicated that flux between cells is increased by at least two orders of magnitude in C_4 species compared with C_3 species. In contrast to NAD-ME and NADP-ME species, the transcription of photosynthetic electron transfer proteins was unchanged in PEP-CK. In summary, the PEP-CK blueprint of *M. maximus* appears to be simpler than those of NAD-ME and NADP-ME plants.

Key words: C_4 photosynthesis, *Dichanthelium clandestinum*, *Megathyrsus maximus*, PEP-CK, RNA-Seq, transcriptomics.

Introduction

Plants using C_4 photosynthesis display higher carbon conversion efficiency than C_3 plants (Amthor, 2010) and are thus among the most productive crop plants. C_4 plants also dominate many natural ecosystems because this trait enables efficient growth under water- and nitrogen-limited conditions at high temperatures. As the area of available arable land decreases and the human population increases, C_4 photosynthesis has become a trait of high potential for a second green revolution (Hibberd *et al.*, 2008; Maurino and Weber, 2013). To recreate this complex trait efficiently by synthetic

approaches, a mechanistic understanding of the genetic architecture controlling the biochemical, anatomical, and regulatory aspects of C_4 photosynthesis is required. Although the enzymes of the core cycle were discovered >50 years ago, knowledge about the metabolism underlying the C_4 trait remains incomplete. The engineering potential of C_4 metabolism was explored in the guinea grass *Megathyrsus maximus*.

C_4 photosynthesis increases photosynthetic efficiency by concentrating CO_2 at the site of Rubisco using a biochemical carbon-concentrating mechanism that is distributed between two compartments, the mesophyll cell (MC) and the bundle sheath cell (BSC), in most known C_4 species. The trait has convergently evolved at least 60 times (Sage *et al.*, 2011) and always employs phospho*enol*pyruvate carboxylase (PEPC) to incorporate bicarbonate into phospho*enol*pyruvate (PEP), yielding the four-carbon molecule oxaloacetate (OAA). For transfer to the site of Rubisco, OAA is converted to either malate by reduction or aspartate by transamination. Different evolutionary lineages, however, have different means to decarboxylate the now-organic carbon to release the CO_2 at the site of Rubisco: NADP-dependent malic enzyme (ME) decarboxylates malate to pyruvate in chloroplasts; NAD-ME decarboxylates malate to pyruvate in mitochondria; and phospho*enol*pyruvate carboxykinase (PEP-CK) decarboxylates OAA to PEP in the cytosol. The resulting C_3 acid is then transported back to the site of PEPC as PEP in the case of PEP-CK-based decarboxylation, or as pyruvate or alanine for NAD-ME and NADP-ME. In the chloroplasts, pyruvate is recycled to PEP by the action of pyruvate, phosphate dikinase (PPDK) with the reaction products pyrophosphate and AMP recycled by pyrophosphorylase (PPase) and AMP kinase (AMK). Historically, three different metabolic C_4 types were proposed based on the decarboxylation enzyme: the NADP-ME type, the NAD-ME type, and the PEP-CK type, of which the latter was considered the most complex (Hatch, 1987). An NADP-ME C_4-type leaf and an NAD-ME C_4-type leaf have been compared with closely related C_3 species globally at the transcriptome level which identified core C_4 cycle components and placed upper limits on the number of genes changed transcriptionally in C_4 metabolism (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011).

Among the C_4 plants with the highest contribution of PEP-CK activity to decarboxylation is the guinea grass *M. maximus*, one of the plant species in which the enzyme activity was originally described and therefore a prototypical PEP-CK plant (summarized in Hatch, 1987). *Megathyrsus maximus* has been taxonomically regrouped several times (Grass Phylogeny Working Group II, 2012), and has also been called *Panicum maximum* and *Urochloa maxima*. Other species with high PEP-CK activity in addition to NAD-ME activity are *Urochloa panicoides* (Ku *et al.*, 1980) and *Chloris gayana* (Hatch, 1987).

The biochemical characterization of PEP-CK-type C_4 plants identified carboxylation by PEPC as in all other C_4 plants (Ku *et al.*, 1980) and two decarboxylation enzymes, PEP-CK and NAD-ME (Ku *et al.*, 1980; Chapman and Hatch, 1983; Burnell and Hatch, 1988a, *b*; Agostino *et al.*, 1996). Exclusive decarboxylation by PEP-CK has not been reported to date. Carboxylation and decarboxylation are linked by the transfer acids malate, aspartate, alanine, pyruvate, and PEP (summarized in Hatch, 1987). In *C. gayana*, the distribution of transfer acids has been investigated by feeding labelled CO_2; both malate and aspartate became rapidly labelled, indicating that both are used as transfer acids. Furthermore, the labelling rate of aspartate was twice as high as that of malate, indicating an approximate flux ratio of 2:1 between aspartate and malate (Hatch, 1979). In *M. maximus*, the aminotransferase enzyme activities were localized to the cytosol (Chapman and Hatch, 1983) and the malate-producing malate dehydrogenases (MDHs) were present as both chloroplastidic NADP-MDH and cytosolic and mitochondrial NAD-MDH (Chapman and Hatch, 1983).

A high rate of PEP-CK decarboxylation is linked to malate decarboxylation in the bundle sheath and consumption of the resulting reducing equivalents (REs) either by reduction of OAA to malate or by the mitochondrial electron transport chain (Hatch, 1987; Burnell and Hatch, 1988a, *b*). The ATP produced is exported to the cytosol to fuel the PEP-CK reaction (Hatch *et al.*, 1988). It remains unresolved whether pyruvate kinase activity produces pyruvate from PEP (Chapman and Hatch, 1983) for transfer back to the mesophyll.

PEP-CK enzyme activity has also been reported for several NADP-ME and NAD-ME species: (Walker *et al.*, 1997; Wingler *et al.*, 1999; Bräutigam *et al.*, 2011; Pick *et al.*, 2011; Sommer *et al.*, 2012; Christin *et al.*, 2013; Muhaidat and McKown, 2013). Whether PEP-CK is an independent subtype or whether it is essentially similar to NAD-ME or NADP-ME species remains unresolved. Supplemental PEP-CK activity was apparently favoured during the evolution of C_4 plants, possibly because it lowers the concentrations and gradients of the transfer acids (Wang *et al.*, 2014), but it is unknown whether it is beneficial for engineering the trait.

Megathyrsus maximus displays a classical Kranz anatomy with large BSCs and few MCs between bundles (Yoshimura *et al.*, 2004). In this arrangement, the cell types are linked by plasmodesmata, which allow symplastic transport of the transfer acids along the concentration gradient (Evert *et al.*, 1977; Hatch, 1987; Botha, 1992; Bräutigam and Weber, 2011). However, this dependence upon symplastic transport has been questioned (Sowinski *et al.*, 2008) and the gradients measured between the cell types in maize do not quite reach the required steepness (Stitt and Heldt, 1985). In *M. maximus*, the photosynthetic rate is correlated with growth light intensity and with plasmodesmatal density (Sowinski *et al.*, 2007). The large BSCs have increased organelle number compared with C_3 BSCs and their chloroplasts have fully developed grana (Yoshimura *et al.*, 2004). As a consequence of linear electron transfer in the bundle sheath chloroplasts, oxygen is produced, leading to higher photorespiration compared with other C_4 plants (Furbank and Badger, 1982; Ohnishi and Kanai, 1983; Farineau *et al.*, 1984). However, the quantum yield for *M. maximus* is comparable with, or above, the quantum yield for *Zea mays* (NADP-ME+PEP-CK) and *Sorghum bicolor* (NADP-ME) (Ehleringer and Pearcy, 1983). Neither the intercellular transport rates of transfer acids nor

the global consequences of linear electron transfer in BSCs have been explored.

The recent sequencing of the model plant *Setaria italica* (Bennetzen *et al.*, 2012) and the detailed phylogenetic analysis of grasses (Grass Phylogeny Working Group II, 2012) enables RNA-Seq of the PEP-CK subtype of C_4 photosynthesis, by providing a mapping reference and the identification of suitable sister species, respectively. Although the phylogeny of the Paniceae tribe of grasses is not resolved with complete confidence (Grass Phylogeny Working Group II, 2012), the C_3 grass *Dichanthelium clandestinum* and the PEP-CK C_4 grass *M. maximus* are currently considered as monophyletic lineages that shared the last common ancestor 18 ± 4 Myr (million years) ago (Vicentini *et al.*, 2008; Grass Phylogeny Working Group II, 2012). *Dichanthelium clandestinum* is therefore among the closest living sister taxa to the PEP-CK-type model species *M. maximus* and was chosen for the comparison in the work reported here.

Two complementary strategies were chosen to extend the blueprint of C_4 photosynthesis to associated pathways and functions beyond the core cycle, which has already been described for the NAD-ME plant *C. gynandra* (Bräutigam *et al.*, 2011): (i) a broad analysis of C_4-related functions using comparative RNA-Seq data for PEP-CK (Paniceae, this study), NADP-ME (*Flaveria* species) (Gowik *et al.*, 2011*a*), and NAD-ME (*Cleome* species) (Bräutigam *et al.*, 2011), and leaf RNA-Seq data sets for *Z. mays* (Li *et al.*, 2011), *S. italica* (Bennetzen *et al.*, 2012), *S. bicolor*, *Oryza sativa*, and *Brachypodium distachyon* (Davidson *et al.*, 2012); and (ii) a detailed C_3 versus C_4 comparison between the PEP-CK species *M. maximus* and its C_3 sister species *D. clandestinum*.

Materials and methods

Plant growth and harvesting

Megathyrsus maximus (Collection of the Botanical Garden Düsseldorf) and *D. clandestinum* (grown from seed obtained from B&T World Seeds, Perpignan, France) plants were grown with 16 h of light at 24 °C. *Dichanthelium clandestinum* was maintained vegetatively. Harvesting was scheduled to the eight-leaf stage, which was 3–5 weeks after germination or tiller initiation. In the middle of the light period, the third leaf from the top—the third youngest—was sampled in three replicates for sequencing (one for 454 and two for Illumina sequencing) and five replicates for enzyme activities, and quenched in liquid nitrogen immediately after cutting. Pools of 20 plants per sample were harvested.

Enzyme activities

C_4 decarboxylation enzymes were extracted from frozen, ground leaves using 1 ml of buffer [25 mM TRIS-HCl (pH 7.5), 1 mM $MgSO_4$, 1 mM EDTA, 5 mM dithiothreitol (DTT), 0.2 mM phenylmethylsulphonyl fluoride (PMSF), and 10% (v/v) glycerol] per 10 mg of leaf powder. After desalting using NAP-5 size exclusion columns, enzyme activities of PEP-CK (Walker *et al.*, 1995), NAD-ME, and NADP-ME (Hatch and Mau, 1977) were measured photometrically based on the absorption change of NAD(P)H at 340 nm.

CO_2 assimilation rates and isotope discrimination

For three replicates of both species, the net leaf photosynthetic assimilation rate (A) was measured using a Li-Cor LI-6400XT infrared gas exchange analyser (LI-COR Inc., Lincoln, NE, USA). CO_2-dependent assimilation curves (A–C_i) were measured at 1500 µmol m^{-2} s^{-1} constant light. Light-dependent assimilation curves were measured at a constant external CO_2 concentration of 400 ppm.

For ^{13}C isotope discrimination, leaf powder was dried and analysed using the isotope ratio mass spectrometer IsoPrime 100 (IsoPrime Ltd, Cheadle, Manchester, UK). Results were expressed as relative values compared with the international standard (Vienna-PeeDee Belemnite).

RNA extraction and sequencing

Isolation of total RNA from ground tissue of *M. maximus* was performed using a guanidium thiocyanate extraction followed by an ethanol and a lithium chloride precipitation, as described by Chomczynski and Sacchi (1987). Extraction of total RNA from *D. clandestinum* was performed using a TRIS-borate buffer to cope with large amounts of polysaccharides, as described by Westhoff and Herrmann (1988). mRNA for 454 library preparation was enriched by using Qiagen Oligotex poly(A)-binding silicone beads and further prepared for sequencing as described in Weber *et al.* (2007). For Illumina sequencing two replicates of total RNA were used per sample. Library preparation and sequencing were carried out according to the manufacturer's suggestions by the local NGS facility (BMFZ, Biologisch-Medizinisches Forschungszentrum, Düsseldorf), using Roche Titanium chemicals for 454 and the TruSeq library kit for Illumina HiSeq 2000. Long and short read raw data were submitted to the short read archive (SUB440021, *D. clandestinum*; SUB439950, *M. maximus*).

Sequence assembly and expression statistics

De novo assembly was done using CAP3 (Huang and Madan, 1999) using default parameters on cleaned 454 reads. Reads were cleaned by trimming low quality ends, discarding reads of overall minor quality, and removal of exact duplicates using scripts of the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) as described in Schliesky *et al.* (2012). Contigs were annotated by BLAST best hit mapping to *S. italica* (v164) representative coding sequences. Quantitative expression was determined by mapping of all Illumina reads against *S. italica* representative coding sequences (v164) using BLAT (Kent, 2002) and counting the best hit for each read. Zero counts were treated as true 0. Expression was normalized to reads per mappable million and per kilobase (rpkm) *Setaria* CDS. Eight rpkm were chosen as the threshold of expression to discriminate background transcription. Differential expression was determined by DESeq (Anders and Huber, 2010), a negative binomial test, in R (R Development Core Team, 2012). A significance threshold of 0.05 was applied after Bonferroni correction for multiple hypothesis testing and is reported in Supplementary Table S3 available at *JXB* online. For all single genes mentioned in the text, changes in expression were confirmed using the 454 data set which was also mapped across species to *S. italica* as described in Bräutigam *et al.* (2011) (Supplementary Table S3). Pathway enrichment was determined by Benjamini–Hochberg correction (Benjamini and Hochberg, 1995). Fisher's exact test was used to test for over-/under-representation of MapMan categories.

Meta comparison of functional categories

Expression data for *B. dystachyon*, *S. bicolor*, and *O. sativa* were previously published by Davidson *et al.* (2012). Transcript sequences for mature *Z. mays* leaves (+4 cm sample) were obtained from the short read archive SRA012297 (Li *et al.*, 2010) and mapped to *S. italica* representative coding sequences. Expression data for five Flaveria species were taken from Gowik *et al.* (2011). Expression data for *Cleome gynandra* (C_4) and *Tarenaya hassleriana* (C_3) were taken from Bräutigam *et al.* (2011). The samples were produced in

different laboratories and with different sequencing technologies. Only the presence of C_4-related traits was interpreted, as absence calls may be due to inconsistent sampling with regard to leaf developmental state, time of day, and other variables.

EC (enzyme classifiers; Schomburg *et al.*, 2013) and Pfam (protein family; Sonnhammer *et al.*, 1997) annotations were added to the two reference transcriptomes, *S. italica* CDS (v164) and *Arabidopsis thaliana* CDS (TAIR10). Reduction of data complexity to functional classifiers was achieved by summing up all expression values mapping to the same EC or Pfam. Venn diagram sets were built through logical operators; that is, expression is higher/lower in all C_4 versus C_3 comparisons (see also Supplementary Table S2 at *JXB* online). Comparison pairs were chosen according to the sequencing method and experimenter: *M. maximus* versus *D. clandestinum* (this study), *S. bicolor* versus *O. sativa* and versus *B. dystachyon* (all from Davidson *et al.* 2012); *Z. mays* (Li *et al.* 2011) and *S. italica* (Bennetzen *et al.* 2012) were orphan data sets as no comparison partner was sequenced with the same technology and both were compared against *B. dystachion* as the C_3 reference. The dicots were compared as previously published (Bräutigam *et al.* 2011; Gowik *et al.*, 2011).

Leaf cross-sections for confocal microscopy

Fresh mature leaves (upper third of the leaf) of *M. maximus* and *D. clandestinum* were cut transversally and fixed in PBST [1× PBS buffer (137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.4 mM KH_2PO_4); 1% (v/v) Tween-20; 3% (v/v) glutaraldehyde] overnight at room temperature. Leaf cross-sections were stained with 0.1% 4',6-diamidino-2-phenylindole (DAPI) solution in phosphate-buffered saline (PBS) for 30 min. Subsequently, cross-sections were analysed with an LSM 780 (Zeiss) confocal microscope with a ×40 objective. Z-stack images were processed with LSM Zeiss software to produce maximum intensity overlay images.

Results

D. clandestinum *is well suited for a C_3 comparison with* M. maximus

The PACMAD clade of the grasses is exceptionally rich in C_4 plants (Christin *et al.*, 2013) to the point that it is difficult to identify and cultivate closely related C_3 species for comparative analyses. To confirm that *D. clandestinum* is a *bona fide* C_3 plant and to confirm the biochemical subtype of the C_4 plant *M. maximus*, different parameters were tested. The measured enzyme activities, stable isotopic carbon discrimination, A–C_i curves, and light curves indicated that *D. clandestinum* indeed represents a C_3 plant (Fig. 1). *Megathyrsus maximus* has high NAD-ME and PEP-CK enzyme activities as compared with *D. clandestinum*, but comparable activities of the NADP-ME decarboxylation enzyme (Fig. 1A). *Dichanthelium clandestinum* discriminates against ^{13}C at a $\delta^{13}C$ ratio of –30‰, while *M. maximus* shows C_4 typical

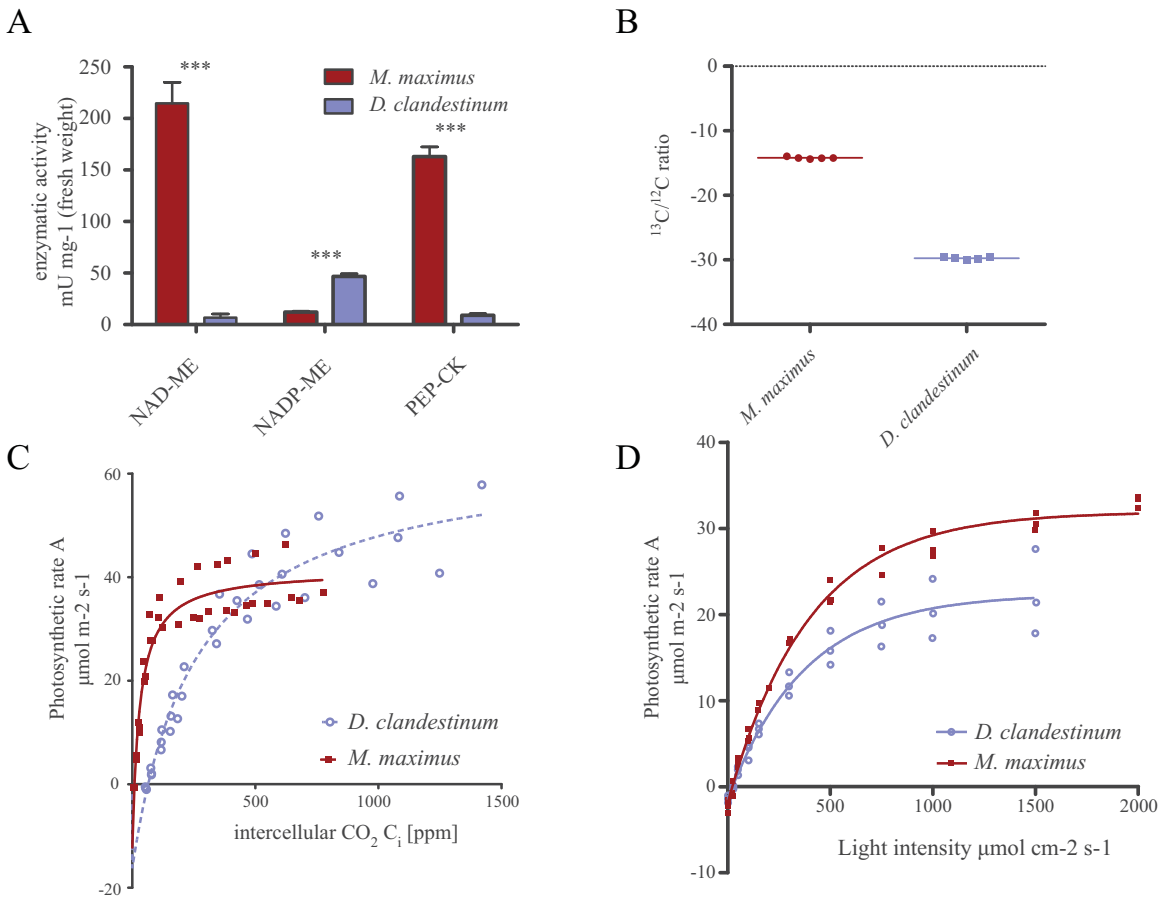


Fig. 1. Physiological characterization of *Megathyrsus maximus* and *Dicanthelium clandestinum*. Activity of the decarboxylation enzymes in *M maximus* and *D. clandestinum* (A); $^{13}C/^{12}C$ stable isotope ratio (B); A–C_i curves at 1500 µE (C); and light curves at 400 ppm CO_2 (D). ***P<0.001. (This figure is available in colour at *JXB* online.)

relaxation of carbon isotope discrimination with a $\delta^{13}C$ ratio of –13‰ (Fig. 1B). The *A*–C_i curve of *M. maximus* shows a low CO_2 compensation point of 9 ppm and saturation of the net carbon fixation rate at 41 μmol m^{-2} s^{-1}. The *A*–C_i curve of *D. clandestinum* plants grown alongside *M. maximus* indicates a CO_2 compensation point of 65 ppm and does not saturate even with high CO_2 concentrations, as is typical for a C_3 plant (Fig. 1C). The light response curves of CO_2 assimilation show similar rates for both types of plants at very low light intensities, with *M. maximus* continuously outgaining *D. clandestinum* as light increases. Thus, *M. maximus* has slightly higher quantum efficiency and saturates at a higher light intensity compared with *D. clandestinum* (Fig. 1D). In summary, the physiological data indicate that *D. clandestinum* is a suitable comparison partner for *M. maximus* due to its phylogenetic proximity and physiological characteristics typical of C_3 plants.

Quantitative and qualitative transcriptome information

The transcriptomes of both grass species were determined by RNA-Seq using two complementary technologies to gain quantitative gene expression information and provide a sequence resource optimized for C_4 unigene assembly. RNA-Seq libraries from two biological replicates of *M. maximus* and two biological replicates of *D. clandestinum* were sequenced with Illumina HiSeq2000 technology and yielded upwards of 53 million reads per replicate, of which >48 million reads were of high quality (Table 1). Reads were mapped cross-species to a closely related reference sequence database derived from the *S. italica* genome (Bennetzen *et al.*, 2012) and between 66% and 74% of reads matched the reference sequence database (Table 1). In the reference sequence database, 13 043 genes were matched with >8 rpkm, of which 792 were detected as differentially up-regulated in C_4 and 376 were detected as differentially down-regulated in C_4 (Table 1). In addition, 1.1 million and 0.9 million 454/Roche Titanium reads were generated and assembled for *M. maximus* and *D. clandestinum*, respectively, and mapped onto *S. italica* as a quality control for the Illumina mapping. The majority of gene expression differences followed similar trends in the 454 mapping or were not detected among the 454 reads; only 12 genes displayed inversely regulated patterns with the different sequencing technologies. Reads were filtered and trimmed based on a Phred score of 30 and assembled with CAP3 (Huang and Madan, 1999) to provide a reliable database of unigenes. C_4 cycle genes were covered by unigenes with full length (Supplementary Table S1 at *JXB* online). About 40 000 unigenes were generated for each species (Table 1).

Genes commonly up- or down-regulated in all C_4 decarboxylation types

Comparative RNA-Seq data for NADP-ME species versus C_3 sister species (Gowik *et al.*, 2011) and for NAD-ME species versus C_3 sister species (Bräutigam *et al.*, 2011), three RNA-Seq data sets for *S. bicolor*, *O. sativa*, and *B. dystachyon* from one comparative experiment (Davidson *et al.*, 2012), as well as orphan RNA-Seq data sets for two PACMAD NADP-ME grasses, *Z. mays* (Li *et al.*, 2011) and *S. italica* (Li *et al.*, 2011), are publicly available. By combining the public data with data from this study, the up- and down-regulated core C_4 genes altered in all C_4 species were identified.

Gene by gene comparisons may be limited between different C_3–C_4 species comparison pairs since for known C_4 genes, most notably PEPC, recruitment of paralogous genes has already been demonstrated (Westhoff and Gowik, 2004; Besnard *et al.*, 2009; Christin and Besnard, 2009). In addition, a function may be distributed among multiple genes, each of which singly does not appear changed. To overcome the inherent limitations of orthologous gene pair comparisons when analysing multiple species pairs, reads were summed to categories which represent a function rather than a particular gene. Enzymes were identified in the reference species *A. thaliana* and *S. italica* on the basis of EC numbers which cover ~5000 different enzymes (Schomburg *et al.*, 2013), of which 1073 are present in the references, and reads for each gene were summed based on the EC number. For example, reads mapping to different isogenes encoding PEPC are no longer represented by the gene identifier but they have been collapsed onto the EC number representing PEPC function (4.1.1.31). All proteins in both reference species were also assigned to their protein family on the basis of Pfam domains (Sonnhammer *et al.*, 1997), of which 4073 unique combinations are present in the references, and reads for each gene were summed based on the Pfam domain combination. Consequently, PEPC is no longer represented by a gene identifier but its function is represented by its Pfam domain combination pf00311. The functions up-regulated or down-regulated in all C_4 species compared with their related

Table 1. *Sequencing, mapping, and assembly statistics for* Megathyrus maximus *and* Dicanthelium clandestinum

Read mapping	*Megathyrus maximus* 1	*Megathyrus maximus* 2	*Dicanthelium clandestinum* 1	*Dicanthelium clandestinum* 2
No. of Illumina reads	61 703 536	56 780 148	53 079 709	56 765 538
No. of cleaned reads	56 470 008	52 282 627	48 160 148	50 328 269
Mappable reads (%)	41 570 126 (73.6%)	38 848 638 (74.3%)	34 151 633 (70.9%)	33 311 704 (66.2%)
No. of 454 reads	1 152 766		971 065	
No. of contigs in assembly	39 565		40 320	
Setaria CDS with >8 rpkm	13 043			
Differentially up-regulated	792			
Differentially down-regulated	376			

C_3 species and those limited to the two NAD-ME type based species were then analysed (Fig. 2A–D; Supplementary Table S2 at *JXB* online).

The functional analysis based on EC numbers indicated a consistent up-regulation of 16 functions in all C_4 comparisons. The C_4 enzymes with PPDK, PPase, AMK, PEPC, aspartate aminotransferase (AspAT), NADP-dependent malate dehydrogenase (NADP-MDH), and ME are up-regulated in all comparisons (Fig. 2A). In addition, one function related to starch synthesis, two functions related to sucrose synthesis, and six functions currently unlinked to C_4 were identified (Fig. 2A; Supplementary Table S2 at *JXB* online). Both NAD-ME species have 135 up-regulated functions in common, including PEP-CK, alanine aminotransferase (AlaAT), pyruvate dehydrogenase (PDH) kinase, and nine enzymes involved in purine synthesis and turnover (Fig. 2A). The 37 functions down-regulated in all C_4 comparisons include four of the Calvin–Benson (CBB) cycle and eight related to photorespiration (Fig. 2B). The down-regulated functions in both NAD-ME-type comparisons included aspartate kinase and aspartate oxidase, eight functions of pyrimidine synthesis, four of the CBB cycle, 11 of chlorophyll synthesis, and 16 of translation (Fig. 2B).

The functional analysis based on Pfam domain combinations showed 34 up-regulated functions in all C_4 species including PEPC, PPDK, phospho*enol*pyruvate phosphate translocator (PPT), and ME. Four photosystem-related functions, two functions related to starch synthesis, and one related to sucrose synthesis are also among those up-regulated (Fig. 2C). The 413 NAD-ME-type related up-regulated functions include PEP-CK, the pyruvate transporter (BASS2, Furumoto *et al.*, 2011), and the sodium:hydrogen antiporter (NHD; Furumoto *et al.*, 2011), all detected with high fold changes (Fig. 2C; Supplementary Table S2 at *JXB* online). Among the 38 down-regulated functions are the CBB cycle, photorespiration, and translation (Fig. 2D).

The analyses of C_4-related functions extend the known C_4 up-regulated traits to sucrose and starch synthesis and the C_4 down-regulated traits to the CBB cycle, photorespiratory functions, and translation. They also provide candidates for as yet unknown functions which may be C_4 related. The NAD-ME-type related functions include those that prevent the leakage of C_4 cycle metabolites into general metabolism.

The PEP-CK decarboxylation subtype is qualitatively similar to but quantitatively distinct from the NAD-ME

Given the blueprint of NAD-ME C_4 photosynthesis (Bräutigam *et al.*, 2011), it was tested whether the differentially regulated functions in the PEP-CK species are those already identified for the NAD-ME species. The occurrence of PEP-CK activity in species previously classified as NADP-ME and NAD-ME species and recent modelling efforts raised the question of whether the classification of PEP-CK as its own C_4 type is warranted (Wang *et al.*, 2014).

The C_4 genes were extracted from the complete data set (Supplementary Table S3 at *JXB* online) and compared with those of *C. gynandra* (Bräutigam *et al.*, 2011). *Megathyrsus maximus* and *C. gynandra* show elevated expression of enzymes and transporters known to be required for C_4 photosynthesis (Table 2). *Megathyrsus maximus* showed significantly increased transcripts encoding BASS2, NHD, PPDK, and PPT, which is similar to the dicotyledonous NAD-ME C_4 species *C. gynandra*. In comparison with a C_3 reference, the up-regulation of these transcripts was between 27-fold and 67.5-fold in *M. maximus* and between 15-fold and 226-fold

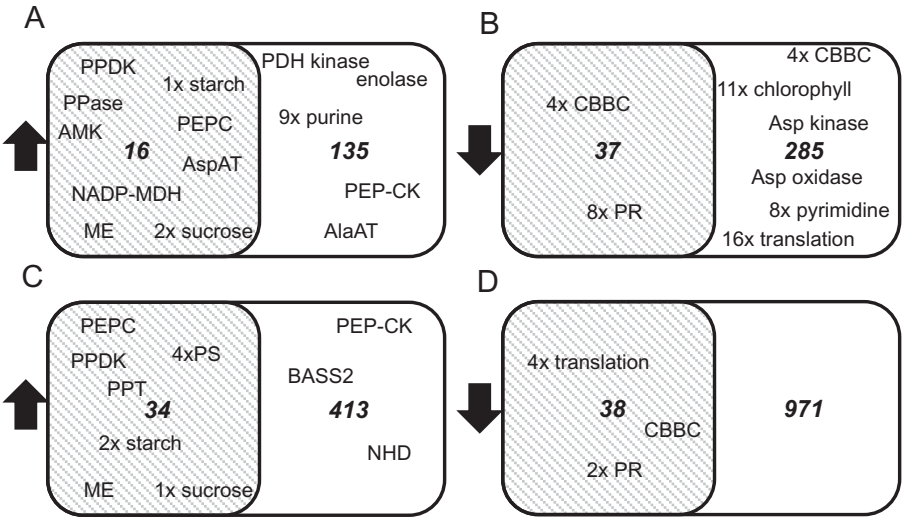


Fig. 2. Shared expression based on function in NAD-ME (white set) versus all C_4 species (grey set). Up- and down-regulated functions are based on expression of functions represented by enzyme classifiers (EC) (A, B) and by Pfam domain combinations (PDC) (C, D). PPDK, pyruvate phosphate dikinase; PPase, inorganic pyrosphate phosphorylase; AMK, adenosine monophosphate kinase; PEPC, phospho*enol*pyruvate carboxylase; AspAT, aspartate aminotransferase; MDH, malate dehydrogenase; ME, malic enzyme; PDH, pyruvate dehydrogenase; PEP-CK, phospho*enol*pyruvate carboxykinase; AlaAT, alanine aminotransferase; CBBC, Calvin–Benson–Bassham cycle; PR, photorespiration; Asp, aspartate; PPT, phospho*enol*pyruvate phosphate translocator; PS, photosynthesis; BASS2, pyruvate transporter; NHD sodium proton antiporter; all functions are listed in Supplementary Table S2 at *JXB* online.

Table 2. *The expression of C_4 cycle genes of* Megathyrsus maximus *in comparison with* Dicanthelium clandestinum *and* Cleome gynandra*, and their location in* M. maximus

Module	Gene name	*Setaria* ID	Function	Predicted location of translated protein	*M. maximus* expression (rpkm)	*D. clandestinum* expression (rpkm)	Fold- change	Significantly changed (DESeq, Bonferroni)	Fold change of function in *C. gynandra*
Regeneration	BASS2	Si001591m	Pyruvate sodium symport	Chloroplast	2797	69	40.5	Yes	87.3
	NHD	Si029362m	Sodium proton antiport	Chloroplast	838	31	27.0	Yes	**15.9**
	PPDK	Si021174m	Pyruvate→PEP	Chloroplast	13380	283	47.3	Yes	226.4
	PPa	Si017993m	Pyrophosphate→phosphate	Chloroplast	450.5	158.5	2.8	NS	3.2
	AMK	Si017707m	AMP→ADP	Chloroplast	985.5	114.5	8.6	NS	**8.9**
	PPT	Si013874m	PEP phosphate antiport	Chloroplast	405	6	67.5	Yes	15.0
Carboxylation	PEPC	Si005789m	PEP→OAA	Cytosol	18393	303.5	60.6	Yes	**77.6**
C_4 transfer acid	AspAT	Si001361m	Asp↔OAA	Cytosol	1273	79	16.1	Yes	2[a]
	GAP-DH	Si014034m	3-GPA→TP	Cytosol	4544	1538	3.0	NS	0.2
	MDH	Si036550m	Malate↔OAA	Cytosol	735	452	1.6	NS	0.44[a]
Decarboxylation	DIC	Si014081m	Malate phosphate antiport	mitochondrion	455	114	4.0	NS	519.0
NAD-ME	PIC	Si017569m	Phosphate proton symport	Mitochondrion	225	96	2.3	NS	2.5
	ME	Si000645m& Si034747m[b]	Malate→pyruvate	Mitochondrion	1299	230	5.6	NS	20.3
	Unknown/diffusion?		Pyruvate export						
Decarboxylation	PEP-CK	Si034404m	OAA→PEP	Cytosol	8819	99	89.5	Yes	8.6
PEP-CK	AAC	Si017474m	ATP ADP/P antiport	Mitochondrion	461	150	3.1	NS	0.4

Bold indicates use of a paralogous gene.
NS, non significant.
[a] A paralogue in a different compartment is up-regulated.
[b] Reads map to both malic enzymes

in *C. gynandra*. PPDK induction, however, was lower in *M. maximus* compared with *C. gynandra*, which might indicate increased regeneration of PEP by PEP-CK rather than PPDK. Both species also showed changes in AMK and PPase expression, but these were not expressed to a significantly higher extent in *M. maximus*. The NHD and AMK expressed at high levels are paralogous to the same proteins required for the C_4 cycle in the dicotyledonous plant (Table 2). The carboxylation enzyme PEPC was significantly up-regulated in both the dicot and the monocot, again using paralogues (Table 2). For the generation of the C_4 transfer acids malate and aspartate, only cytosolic AspAT was significantly up-regulated in *M. maximus*, while no up-regulation of the cytosolic isozyme was observed in *C. gynandra*. Cytosolic targeting was determined by localization prediction of the full-length protein of *M. maximus* (Supplementary Table S4 at *JXB* online). The most abundant transcript encoding MDH also encoded a cytosolic isozyme, suggesting use of the NAD-MDH form (Supplementary Table S4).

Two different decarboxylation modules using NAD-ME and PEP-CK, respectively, are active in the plants (Fig. 1A). In *M. maximus*, neither the transport protein DIC, responsible for antiport of malate into mitochondria against phosphate (Palmieri *et al.*, 2008), and PIC, responsible for symport of phosphate and protons (Pratt *et al.*, 1991; Hamel *et al.*, 2004), nor the decarboxylation enzyme NAD-ME were significantly changed, although all were up-regulated between 2.3- and 5.6-fold (Table 2). This is in stark contrast to the up-regulation detected for DIC and NAD-ME in *C. gynandra* which was between 20-and 519-fold. No candidate for pyruvate export from the mitochondria could be identified. The situation is reversed for the PEP-CK module where PEP-CK was significantly up-regulated 90-fold in *M. maximus* but only 8.6-fold in *C. gynandra*. The mitochondrial ATP-ADP translocase, AAC (Haferkamp *et al.*, 2002), is up-regulated in *M. maximus*, but not to a significant degree (Table 2). Orthologous AlaATs are significantly up-regulated by 37-fold in both species. Unlike the *C. gynandra* protein, which is predicted to be targeted to mitochondria, the *M. maximus* protein is predicted to be cytosolic (Supplementary Table S4 at *JXB* online). The *M. maximus* AlaAT protein showed a shortened N-terminus when aligned to the *S. italica* gene (Supplementary Table S4), hence *in silico* targeting predicted cytosolic localization. Finally, non-significant up-regulation of TPT and plastidic GAP-DH was detected in *M. maximus* to comparable levels as in *C. gynandra* (Table 2).

In addition to single gene analysis, differentially regulated genes were subjected to pathway enrichment analysis to detect changes in gene expression for whole pathways such as the CBB cycle, photorespiration, and photosynthesis. None of the pathways was significantly enriched among the differentially regulated genes (Supplementary Table S5 at *JXB* online).

The gene-by-gene and enrichment analyses revealed a similar but not identical blueprint for the PEP-CK species compared with the NAD-ME species. The core cycle blueprint was amended to include a companion transporter for the malate phosphate antiporter DIC, which couples it to the proton gradient with phosphate proton symport through PIC.

Energy requirements derived from the PEP-CK blueprint

The energy requirements of intracellular transport reactions were not considered when the energy balance of C_4 photosynthesis was originally calculated (i.e. Kanai and Edwards, 1999), although pyruvate transport was hypothesized to be active based on measurements of the metabolite concentration gradients in maize (Stitt and Heldt, 1985). To assess the energy requirements of the PEP-CK-based C_4 cycle, the

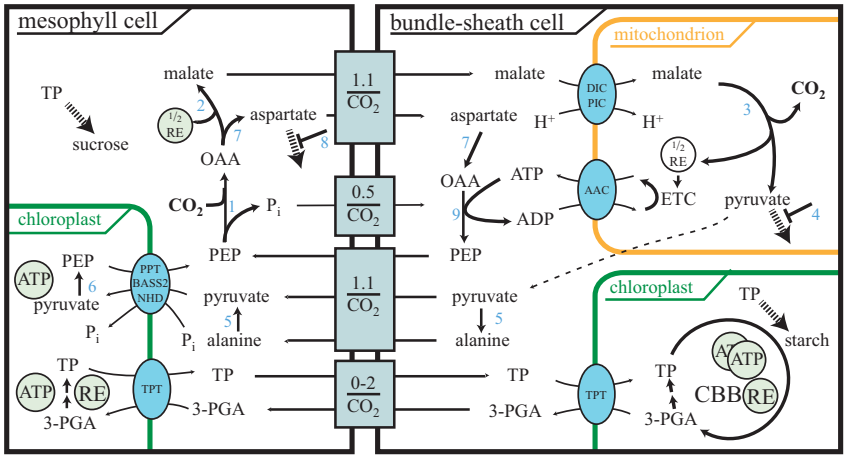


Fig. 3. Extended model for NAD-ME with high PEP-CK activity. Transport modules, consisting of one or more transporters, are shown together with the net transport through the module. Abbreviations: (1) Phospho*enol*pyruvate carboxylase; (2) malate dehydrogenase; (3) NAD-dependent malic enzyme (NAD-ME); (4) pyruvate dehydrogenase kinase; (5) alanine aminotransferase; (6) pyruvate, phosphate dikinase; (7) aspartate aminotransferase; (8) aspartate oxidase and aspartate kinase; (9) phospho*enol*pyruvate carboxykinase (PEP-CK); 3-PGA, 3-phosphoglyceric acid; TP, triose-phosphate; CBB, Calvin–Benson–Bassham cycle; OAA, oxaloacetic acid; RE, reducing equivalent; BASS2, pyruvate transporter; NHD, sodium proton antiporter; PPT, phospho*enol*pyruvate phosphate translocator; TPT, triose-phosphate phosphate translocator; ETC, electron transfer chain. Dashed arrows represent leakage to general metabolism. (This figure is available in colour at *JXB* online.)

amended blueprint was translated into a model of PEP-CK C_4 photosynthesis (Fig. 3).

Energy requirements are calculated following one turn of the cycle (Fig. 3): after PEP is carboxylated to OAA, half of the OAA is reduced to malate (Hatch *et al.*, 1988), requiring on average 0.5 REs derived from photosynthesis for each CO_2 (Fig. 3). The remaining OAA is transported as aspartate (Fig. 3). At the bundle sheath mitochondria, malate exchange for phosphate via DIC is coupled to the proton gradient via phosphate proton symport by PIC (Fig. 3). This process consumes the proton gradient of mitochondria. The proton gradient is also used to drive mitochondrial ATP synthesis for the PEP-CK reaction which decarboxylates OAA to PEP (Fig. 3) and is regenerated by oxidizing the NADH produced by malate decarboxylation (Fig. 3). The carboxylation, transfer, and decarboxylation thus consume on average 0.5 NADPH per CO_2 generated in photosynthetic electron transfer. During regeneration, the PPDK reactions require 2 ATP for the regeneration of each pyruvate but, since only half of the flux runs through malate decarboxylation and therefore pyruvate, only 1 ATP is required for each CO_2. The PPDK reaction is driven towards PEP regeneration by the PPase, which splits the energy-rich bond of pyrophosphate and makes the PPDK reaction irreversible *in vivo*. The production of PEP and its export through PPT creates the proton gradient required to import pyruvate and cycle sodium through the transport system (Fig. 3). Although the active transport of pyruvate is driven by the proton gradient, it requires no additional input of energy beyond that expended for the PPDK reaction (Furumoto *et al.*, 2011). The regeneration phase thus requires 1 ATP in total. The CBB cycle requires 3 ATP and 2 REs from photosynthesis, which may be consumed in the bundle sheath or mesophyll.

The total PEP-CK-based C_4 cycle, assuming no overcycling, thus requires 4 ATP and 2.5 NADPH from the photosynthetic electron transfer chain while solely NADP-ME-based C_4 photosynthesis requires 5 ATP and 2 NADPH and C_3 photosynthesis requires 3 ATP and 2 NADPH for each CO_2 (Kanai and Edwards, 1999). Engineering a PEP-CK-type C_4 cycle will thus avoid the adjustments required for the photosynthetic electron transfer chain since the demands in terms of the ATP and NADPH ratio are almost the same as in C_3 plants.

Intercellular transport derived from the PEP-CK blueprint

Engineering a C_4 cycle may require modifications to the symplastic transport interface (Weber and Bräutigam, 2013). To estimate the difference in intercellular transport for each MC, intercellular transport events between C_4 and C_3 were compared. Data from the scheme depicted in Fig. 3 were combined with anatomical data (Supplementary Fig. S1 at *JXB* online) and photosynthetic rates (Fig. 1C).

Since transport events are assessed per MC and not per leaf area, the number of MCs per leaf area was determined. In the C_4 plants, photosynthesis requires the MC and its adjacent BSC; in the C_3 plant, each MC is a self-contained unit. Microscopic imaging of leaf cross-sections revealed typical Kranz anatomy in *M. maximus* with large BSCs, each of which was connected to multiple MCs (Supplementary Fig. S1 at *JXB* online). The density of MCs was almost twice as high in the C_3 leaf compared with the C_4 leaf (Table 3). Since the photosynthetic rate per leaf area is also higher in *M. maximus* (Fig. 1C), almost twice as much CO_2 is fixed in each MC–BSC pair in *M. maximus* compared with an MC of *D. clandestinum* (5.4 versus 2.6 pmol CO_2 per unit and second). In *D. clandestinum*, only sucrose transport is required across the MC wall. Since each sucrose molecule carries 12 carbons, and since only half of the carbon is exported at any given time, with the remainder stored as starch, the assimilation of one molecule of CO_2 requires $1/12 \times 1/2 = 0.042$ transport events in the C_3 plant (Table 3). In contrast, the PEP-CK-based C_4 cycle requires between 2.75 and 4.75 transport events depending on the extent of RE shuttling because the C_4 acids, the C_3 acids, balancing phosphates, and REs are transported (Table 3). The total number of transport events is estimated by multiplying the number of CO_2 molecules assimilated with the number of transport events required for each CO_2 as 11.6–20.1 pmol s^{-1} in the C_4 species while for C_3 it is 0.1 pmol s^{-1}. C_4 photosynthesis requires between 100- and

Table 3. *Parameters for the calculation of transport requirements for the PEP-CK/NAD-ME C_4 cycle show that C4 photosynthesis requires 100–200 times more transport events*

Cell density was estimated from Supplementary Fig. S1 at JXB online and divided by photosynthetic parameters derived from Fig. 1 to yield the photosynthetic rate per cell (A). C_4 cycle transport requirements were derived from Fig. 3 and summed to calculate total transport events (B). Total transport events through plasmodesmata are calculated as A×B.

		M. maximus	*D. clandestinum*
Photosynthetic parameter	Photosynthetic cell density (Giga photosynthetic units m^{-2})	6.987	12.5
	Photosynthetic rate at 400 ppm (μmol m^{-2} s^{-1})	29.6	20.8
A	Photosynthetic rate CO_2 per cell (pmol CO_2 pu^{-1} s^{-1})	4.2	1.7
Metabolic parameter	C_4 acid (malate, aspartate)	1.1	
(transport events per CO_2)	C_3 acid (PEP, pyruvate, alanine)	1.1	
	Phosphate balance (P_i; 50% PEP assumed)	0.55	
	RE shuttle (triose-phosphate, 3-PGA)	0–2	
	Sucrose export		0.042
B	Total no. of transport events (transport events CO_2^{-1} pu^{-1})	2.75–4.75	0.042
A×B	No. of transport events per cell (pmol transport events s^{-1})	11.6–20.0	0.1

200-fold more transport events than C_3 photosynthesis, such that the intercellular transport capacity needs to be increased by approximately two orders of magnitude in C_4 (Table 3).

Engineering of the C_4 cycle will thus almost certainly require engineering of the BSC–MC interface, as it is highly unlikely that an existing C_3 MC could support the >100-fold increased symplastic flux.

Discussion

Assembly and mapping characteristics

This study was designed to compare two closely related C_3 and C_4 species to increase the probability of detecting C_4-related rather than species-related differences. While for several C_3 grass species, such as rice and *Brachypodium*, the genomes have already been sequenced and thus could serve as C_3 reference for comparative transcriptome sequencing, all of these belong to the BEP clade and have thus diverged 45–55 Myr ago from *M. maximus* (Grass Phylogeny Working Group II, 2012), which belongs to the PACMAD clade. *Dichanthelium clandestinum* was chosen as a C_3 species from within the PACMAD clade for the transcriptomic comparison presented here. Although the precise phylogenetic position of the *Dichanthelium* clade of Paniceae, which includes *D. clandestinum*, has not been determined, it was recently placed as sister to the group, which contains *M. maximus* (Grass Phylogeny Working Group II, 2012), with a divergence time of 14–22 Myr ago (Vicentini *et al.*, 2008). For quantification of steady-state transcript amounts, the RNA-Seq reads were mapped onto the coding sequences predicted from the *Setaria* genome. The closer phylogenetic proximity of *M. maximus* to *Setaria* is represented in the slightly higher mapping efficiency of its reads (Table 1). Overall, the mapping efficiency is above that of the *Flaveria* species on *Arabidopsis* (Gowik *et al.*, 2011*a*) but below that of the *Cleomaceae* on *Arabidopsis* (Bräutigam *et al.*, 2011). The disadvantage of a slightly uneven mapping efficiency was, however, outweighed by mapping reads from both species onto a common genome-based reference sequence, which enabled normalization to reads per kilobase per million reads. In addition, low abundance transcripts are frequently under-represented in contig assemblies, while high abundance transcripts were fragmented into multiple contigs per transcript. Establishing orthology, while possible with tools such as OrthoMCL, requires assumptions about similarities. Mapping onto a reference database as previously successfully established (Bräutigam *et al.*, 2011, Gowik *et al.*, 2011) was chosen to overcome this problem.

Contig assembly from Illumina reads results in fragmented contigs, especially for the high abundance contigs, as observed previously in other RNA-Seq projects (Bräutigam and Gowik, 2010; Franssen *et al.*, 2011; Schliesky *et al.*, 2012). The C_4 transcripts are among the most highly expressed transcripts in leaves of C_4 plants (Bräutigam *et al.*, 2011). To produce high confidence contigs, the transcriptome was sequenced by a long read technology, the reads cleaned with a high base quality threshold of Phred=30, and assembled with CAP3. Within the database, full-length contigs for all candidate C_4 genes were identified (Supplementary Tables S1, S4 at *JXB* online), validating a hybrid approach to quantification and database generation (Bräutigam and Gowik, 2010).

Are NAD-ME and the PEP-CK distinct subtypes of C_4 photosynthesis?

The three classical subtypes of C_4 photosynthesis, NADP-ME, NAD-ME, and PEP-CK, have been analysed by comparative transcriptome sequencing (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011; this study). If the two C_4 types NAD-ME and PEP-CK which both rely wholly or partially on NAD-ME-based decarboxylation were fundamentally different, major differences in the transcriptional profile would be expected. However, quantification of transcript abundance showed that the functions up-regulated in the NAD-ME plant *C. gynandra*, which shows some PEP-CK activity (Sommer *et al.*, 2012), and the PEP-CK plant *M. maximus*, which displays high PEP-CK activity, are quite similar.

The bicarbonate acceptor regeneration module is essentially identical. Both plant species belong to the sodium pyruvate transport group, as defined by Aoki *et al.* (1992), and show joint up-regulation of not only the sodium pyruvate symporter BASS2 (Furumoto *et al.*, 2011), but also the companion sodium:hydrogen antiporter NHD, and the PEP phosphate antiporter PPT (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011; Table 2). The generation of the transfer acids appears to be cytosolic as neither of the two plastidial dicarboxylate transporters, DiT1 (OAA/malate antiporter) (Weber *et al.*, 1995; Kinoshita *et al.*, 2011) and DiT2 (OAA/aspartate antiporter) (Renne *et al.*, 2003), was up-regulated (Supplementary Table S3 at *JXB* online) and the most abundant contigs encoding AspAT and MDH were predicted to be cytosolic (Table 2; Supplementary Table S4). The cytosolic localization relaxes the need to up-regulate organellar transporters, which are required to import substrates and export products. The two species use differentially localized AspATs, a mitochondrial isozyme in the case of *C. gynandra* (Sommer *et al.*, 2012) and a cytosolic one in the case of *M. maximus* (Table 2; Toledo-Silva *et al.*, 2013). For the decarboxylation process, both species use a combination of PEP-CK and NAD-ME and consequently have the same functions up-regulated. The degree of up-regulation, however, mirrors the enzyme activity differences, with PEP-CK transcripts being much more induced in *M. maximus* and NAD-ME and associated transporters much more induced in *C. gynandra* (Table 2). Hence the difference in decarboxylation biochemistry between both species rests in an altered balance between NAD-ME and PEP-CK activities, while the overall pathway is very similar.

At least part of the C_3 acid transport is accomplished through alanine to balance the amino groups between MCs and BSCs. The up-regulated AlaAT for both plants is an orthologous pair, which is targeted to organelles in *C. gynandra* (Bräutigam *et al.*, 2011; Sommer *et al.*, 2012) and *S. italica* (Supplementary Table S4 at *JXB* online). However, enzyme activity measurements placed high AlaAT activity in the

cytosol of *M. maximus* (Chapman and Hatch, 1983). The *in silico* translation of the *M. maximus* transcript revealed that it encodes a truncated version of AlaAT in comparison with the *Setaria* gene, in which a potential start ATG in-frame with the coding sequence is prefaced by a stop codon. The shortened protein is predicted to be cytosolic (Supplementary Table S4). Hence, the cytosolic AlaAT activity in *M. maximus* appears to have evolved by loss of the target peptide of an originally organellar-targeted protein. The simpler cycle suggests that the *M. maximus* blueprint is easier to engineer compared with the blueprints of NAD-ME (Bräutigam *et al.*, 2011; Sommer *et al.*, 2012) and NADP-ME species (Gowik *et al.*, 2011; Pick *et al.*, 2011; Denton *et al.*, 2013; Weber and Bräutigam, 2013).

Multiple species which had previously been grouped as NADP-ME or NAD-ME plants have different degrees of PEP-CK activity (Walker *et al.*, 1997; Pick *et al.*, 2011; Sommer *et al.*, 2012; Muhaidat and McKown, 2013) and modelling shows the advantages of supplemental PEP-CK activity in conferring environmental robustness to the pathway (Wang *et al.*, 2014), raising the question as to whether PEP-CK-type plants deserve their own group. While the functions up-regulated in *C. gynandra* and *M. maximus* are similar, there are differences with regard to localization of the enzymes generating the transfer acids. Whether the different enzyme localizations are tightly associated with the type and degree of use of the decarboxylation enzymes remains to be determined once additional transcriptomes are sequenced and a global view is enabled on more than just one prototypical species for each historical C_4 type. For engineering, it is probably advisable to follow the blueprint of a particular species since it is currently not clear whether differences in transfer acid generation are only species specific or are tied to other processes such as decarboxylation enzymes and therefore functionally relevant.

An extended model of C_4 photosynthesis with high PEP-CK activity

Understanding the evolution of C_4 metabolism and re-engineering a C_4 cycle in a C_3 plant requires a mechanistic understanding of the parts making up the system (Denton *et al.*, 2013). The global transcriptomics analysis of *M. maximus* compared with *D. clandestinum* enabled the extension of the C_4 metabolism model presented earlier for *M. maximus* (Hatch, 1987) and *C. gynandra* (Bräutigam *et al.*, 2011; Sommer *et al.*, 2012).

Transport processes and core cycle

The *M. maximus* analysis confirmed DIC as the mitochondrial malate importer (Table 2; Bräutigam *et al.*, 2011). The companion transporter, which couples malate transport to the proton gradient of the mitochondria and supplies mitochondria with inorganic phosphate for ATP production, is probably PIC (Hamel *et al.*, 2004; Table 2; Fig. 3). The only transporter which remains unknown at the molecular level is the mitochondrial pyruvate exporter. The candidate pyruvate transport protein, the human mitochondrial pyruvate carrier (MPC) (Bricker *et al.*, 2012; Herzig *et al.*, 2012), is not differentially expressed in *C. gynandra* and *M. maximus*. Potentially, pyruvate can traverse biomembranes in its protonated form by simple diffusion (Benning, 1986), although this is unlikely in a cellular context given that only one out of 10^5 molecules of pyruvic acid occurs in the protonated form at physiological pH values. Although early models did not take a reducing equivalent shuttle across both chloroplast envelopes into account for PEP-CK species (Hatch, 1987), possibly because *M. maximus* lacks chloroplast dimorphism (Yoshimura *et al.*, 2004), measurements of enzyme activity confirmed glyceraldehyde dehydrogenase in both MCs and BSCs of *U. panicoides* (Ku and Edwards, 1975), and RNA-Seq indicated modest up-regulation of the necessary transporters and enzymes (Table 2). Engineering a C_4 cycle will critically depend on correctly enabling the transport of substrates through transporters and companion transporters (Weber and von Caemmerer, 2010; Fig. 3). Balancing reducing power between MCs and BSCs via triose-phosphate/phosphate translocators in chloroplasts in both MCs and BSCs appears also to be required in species which lack chloroplast dimorphism (Table 2; Yoshimura *et al.*, 2004).

Knowledge about the intracellular transport proteins involved in C_4 photosynthesis has recently improved significantly (compare with Weber and von Caemmerer, 2010; Bräutigam and Weber, 2011; Denton *et al.*, 2013; Weber and Bräutigam, 2013), largely due to RNA-Seq-enabled identification and characterization of the chloroplast pyruvate transporter (Furumoto *et al.*, 2011), and the placement of several known transport proteins in the C_4 pathway (Taniguchi *et al.*, 2003; Bräutigam *et al.*, 2011; Gowik *et al.*, 2011*a*; Kinoshita *et al.*, 2011). However, information about the intercellular transport has not progressed since the discovery of sieve element-like plasmodesmata plates in the MC–BSC interface (Evert *et al.*, 1977; Botha, 1992).

The difference in total transport events between the C_3 and the C_4 species was estimated using the data provided by the model shown in Fig. 3, by images of the cellular architecture (Supplementary Fig. S1 at *JXB* online), and by photosynthetic rate measurements (Fig. 1C). The large difference in the requirement for intracellular transport between C_4 and C_3 pathways is not predominantly driven by the rather small differences in photosynthetic rates (Fig. 1C), but by two other factors: the number of MCs per leaf area and the number of transport events required for each CO_2 assimilated. The large BSCs, each of which borders several MCs, and the fact that *M. maximus* requires two cells in each photosynthetic unit means that the C_3 grass has almost twice as many photosynthetic units in the same leaf area. The net CO_2 assimilation capacity is thus not only higher by the ~20% higher photosynthetic rate per leaf area but—if normalized to the number of MCs—is almost twice as high for each unit. The second factor is the number of transport processes occurring over each interface. Intercellular transport for each C_3 cell is very low, 0.042 events per CO_2 assimilated for an MC. The transport events for the C_4 cycle are more difficult to estimate since, in addition to the comparatively fixed flux of C_4 and C_3 acids in the cycle, the PEP-balancing phosphate flux and the RE shuttle yield variable fluxes. However, even using

the lowest possible estimates, a >100-fold difference in transport events is predicted between the C_4 and C_3 plant interfaces. The interface itself is probably optimized for a balance of openness to enable the flux and closed-ness to enrich the CO_2 at the site of Rubisco, since different light intensities correlate with photosynthetic rates and plasmodesmatal density in *M. maximus* (Sowinski *et al.*, 2007). The fold change in transport events across the interface is in the range of the fold change expression changes for the C_4 genes (Tables 2, 23). The evolution and hence also re-engineering of the C_4 cycle must adapt the intercellular interface.

Accessory pathways to the core cycle

It is tempting to limit engineering efforts to the major transcriptional changes and therefore to the core cycle. However, accessory pathways to the core C_4 cycle may play a major role in adapting the underlying metabolism to the presence of the carbon-concentrating pump.

The comparison of multiple different C_3–C_4 pairs and therefore C_4 origins with each other provides a method to identify differentially regulated functions with biological significance, once the problem of paralogous genes carrying out the functions is overcome. By mapping RNA-Seq data to EC numbers and Pfam domains rather than individual genes, it has been possible to identify core C_4 genes (Fig. 2; Supplementary Table S2 at *JXB* online), which indicates that these methods are suitable to pick up additional C_4-related functions.

Both methods picked up functions involved in starch metabolism and sucrose synthesis (Fig. 2A, C). In the EC-based mapping, the sucrose synthesis pathway was present with two functions, the UDP-glucose pyrophosphorylase and the sucrose-phosphate synthase. Sucrose-phosphate synthase is the rate-limiting enzyme for sucrose synthesis in the C_3 plant *A. thaliana* (Häusler *et al.*, 2000; Strand *et al.*, 2000; Koch, 2004) and NDP sugar pyrophosphorylases are comparatively slow enzymes. The surplus of fixed carbon (Fig. 1C, D) leads to a surplus of triose-phosphates. In *Z. mays*, *Panicum miliaceum*, and *Brachiaria erucaeformis*, sucrose synthesis is localized to the mesophyll (Usuda and Edwards, 1980), which may also be the case in *M. maximus*. Both the localization of sucrose synthesis and the higher carbon assimilation rate contribute to more triose-phosphate at the site of sucrose synthesis and hence the need for greater sequestration (Fig. 3). Similarly, the higher rate of CO_2 assimilation (Fig. 2) and the localization of starch storage in the BSCs (Majeran and van Wijk, 2009; Majeran *et al.*, 2010) probably also require higher rates of starch synthesis to sequester the triose-phosphates efficiently (Figs 2, 3). When considering the engineering of C_4 photosynthesis, the sequestration of triose-phosphates is probably of low priority compared with the engineering of the enzymes and transport proteins, yet not adding these functions for triose-phosphate sequestration will probably limit the system to the capacity of C_3 photosynthetic plants, a 20% loss of potential productivity.

Insulating the C_4 cycle from other metabolic networks is also probably critical to avoid loss of cycle intermediates. No obvious proteins with functions in this context were identified in comparisons across all C_4 data sets (Fig. 2; Supplementary Table S2 at *JXB* online), although the uncharacterized functions may include such insulators (Supplementary Table S2). The analysis of only NAD-ME-based C_4 photosynthesis registered changes, which represent the overlap between the dicot *C. gynandra* and the grass *M. maximus*. Both species produce pyruvate in their mitochondria (Table 2; Bräutigam *et al.*, 2011) and use aspartate as a dominant transfer acid. Both NAD-ME species show higher PDH kinase and reduced aspartate kinase and aspartate oxidase transcript amounts (Fig. 2). These three enzymes control metabolite exit from the C_4 cycle as PDH kinase gates pyruvate decarboxylation for entry into the tricarboxylic acid (TCA) cycle, aspartate kinase controls entry into aspartate-derived amino acid metabolism, and aspartate oxidase controls entry into NAD synthesis. The leaking of cycle intermediates into other metabolism despite the insulation can be indirectly seen in the labelling pattern obtained by $^{14}CO_2$ feeding. If metabolites from the cycle are consumed, they need to be replaced from the CBB cycle and will thus carry label in C_2–C_4 of the four-carbon compounds and lead to label in the three-carbon compounds, which—if only cycling—should show no label at all. Indeed, labelling studies identified delayed labelling in both groups (e.g. Hatch, 1979), indicating that leaking of intermediates does occur. When engineering a C_4 cycle into a C_3 plant, limiting the leakage of cycle intermediates is probably required for all cycle metabolites to keep the cycle running robustly.

Two to three pathways are commonly down-regulated: the CBB cycle, photorespiration, and protein synthesis (Fig. 2). Reduced expression of these functions in C_4 species may not be required to engineer efficient CO_2 capture. However, reduced expression of CBB, photorespiration, and protein translation (Fig. 2) may be necessary to realize the nitrogen-saving benefits of C_4 photosynthesis which are common to C_4 plants (Sage, 2004).

NAD-ME species show an unusual pattern with regard to nucleotide metabolism; several functions of purine metabolism are up-regulated while several functions of pyrimidine synthesis are down-regulated. While one may speculate that the changes in purine metabolism are due to the altered ATP usage in these plants, the functional reason for these changes remains unknown.

Previous global transcriptome analyses found that genes encoding components of photosynthetic cyclic electron flow (CEF) were significantly up-regulated (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011), raising the question of whether such alterations to photosynthesis are required in all C_4 subtypes. The present analysis did not indicate differences in CEF in *M. maximus* compared with *D. clandestinum* (Supplementary Table S5 at *JXB* online). The reason lies in the high PEP-CK activity, which is fuelled by malate oxidation in mitochondria (Fig. 3). Malate is generated using photosynthetic REs leading to a 4:2.5 ATP:NADPH production ratio in photosynthesis which is very similar to that of C_3 photosynthesis at a 3:2 ratio and in contrast to the classical C_4 calculation of 5:2 (Kanai and Edwards, 1999). If considering engineering, a C_4 cycle with high PEP-CK activity together with malate decarboxylation in mitochondria removes the requirement

for dimorphic chloroplasts, which results in one less feature to be engineered.

It is tempting to think that the type of C_4 photosynthesis realized in *M. maximus* is less efficient because of higher energy input for the C_4 cycle (Fig. 3) and because of oxygen production in the bundle sheath, which increases the potential for photorespiration. Elevated photorespiration is indeed a feature of *M. maximus* (Furbank and Badger, 1982; Ohnishi and Kanai, 1983; Farineau *et al.*, 1984). However, the quantum efficiency of *M. maximus* is indistinguishable from that of *Z. mays* or *S. bicolor* (Ehleringer and Pearcy, 1983). It is surprising that the energy requirements derived from the model (Fig. 3) and the photorespiratory rate (Furbank and Badger, 1982; Ohnishi and Kanai, 1983; Farineau *et al.*, 1984) do not predict quantum efficiency.

The blueprint of C_4 metabolism in *M. maximus* is simpler compared with that of NAD-ME and NADP-ME plants, because the generation of transfer acids requires fewer adjustments in intracellular transport capacity and photosynthetic electron transfer, and at least some part of the insulators that prevent leakage of C_4 cycle intermediates into general metabolism are known. Thus, it represents an attractive target for engineering the C_4 cycle into a C_3 crop plant.

Supplementara data

Supplementary data are available at *JXB* online.
Figure S1. Cross-sections of *D. clandestinum* and *M. maximus.*
Table S1. *D. clandestinum* and *M. maximus* unigene fasta files.
Table S2. Excel table of Pfam and EC function analysis for all genes.
Table S3. Excel table of quantitative gene expression information including statistical analysis.
Table S4. Text document of selected full-length unigenes including alignment to *S. italica* genes and targeting prediction.
Table S5. Excel table of enrichment analysis for pathways.

Acknowledgements

The authors acknowledge excellent technical support for metabolite analysis by Katrin L. Weber and Elisabeth Klemp, and for RNA sequencing by the BMFZ, HHU Düsseldorf. The authors thank Alisandra Denton and especially Richard Leegood and Urte Schlüter for helpful comments on the manuscript. This work was supported by grants of the Deutsche Forschungsgemeinschaft to APMW (IRTG 1525 and EXC 1028 to APMW) and of the European Union Framework 7 Program (3to4 to APMW and CPO).

References

Agostino A, Heldt HW, Hatch MD. 1996. Mitochondrial respiration in relation to photosynthetic C_4 acid decarboxylation in C_4 species. *Australian Journal of Plant Physiology* **23,** 1–7.

Amthor JS. 2010. From sunlight to phytomass: on the potential efficiency of converting solar radiation to phyto-energy. *New Phytologist* **188,** 939–959.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106.

Aoki N, Ohnishi J, Kanai R. 1992. 2 Different mechanisms for transport of pyruvate into mesophyll chloroplasts of C_4 plants—a comparative-study. *Plant and Cell Physiology* **33,** 805–809.

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

Bennetzen JL, Schmutz J, Wang H, *et al*. 2012. Reference genome sequence of the model plant Setaria. *Nature Biotechnology* **30,** 555–559.

Benning C. 1986. Evidence supporting a model of voltage-dependent uptake of auxin into cucurbita vesicles. *Planta* **169,** 228–237.

Besnard G, Muasya AM, Russier F, Roalson EH, Salamin N, Christin PA. 2009. Phylogenomics of C_4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. *Molecular Biology and Evolution* **26,** 1909–1919.

Botha CEJ. 1992. Plasmodesmatal distribution, structure and frequency in relation to assimilation in C_3 and C_4 grasses in Southern Africa. *Planta* **187,** 348–358.

Bräutigam A, Gowik U. 2010. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biology* **12,** 831–841.

Bräutigam A, Kajala K, Wullenweber J, *et al*. 2011. 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, Weber APM. 2011. Transport processes—connecting the reactions of C_4 photosynthesis: In: Raghavendra AS, Sage RF, eds. *C_4 photosynthesis and related CO_2 concentrating mechanism. Advances in photosynthesis and respiration* , Vol. **32**. Dordrecht: Springer, 199–219

Bricker DK, Taylor EB, Schell JC, *et al*. 2012. A mitochondrial pyruvate carrier required for pyruvate uptake in yeast, *Drosophila*, and humans. *Science* **337,** 96–100.

Burnell JN, Hatch MD. 1988a. Photosynthesis in phospho*enol*pyruvate carboxykinase-type-C_4 plants—photosynthetic activities of isolated bundle sheath-cells from *Urochloa panicoides*. *Archives of Biochemistry and Biophysics* **260,** 177–186.

Burnell JN, Hatch MD. 1988b. Photosynthesis in phospho*enol*pyruvate carboxykinase-type-C_4 plants—pathways of C_4 acid decarboxylation in bundle sheath-cells of *Urochloa panicoides*. *Archives of Biochemistry and Biophysics* **260,** 187–199.

Chapman KSR, Hatch MD. 1983. Intracellular location of phospho*enol*pyruvate carboxykinase and other C_4 photosynthetic enzymes in mesophyll and bundle sheath protoplasts of *Panicum maximum*. *Plant Science Letters* **29,** 145–154.

Chomczynski P, Sacchi N. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Analytical Biochemistry* **162,** 156–159.

Christin PA, Besnard G. 2009. Two independent C_4 origins in Aristidoideae (Poaceae) revealed by the recruitment of distinct phospho*enol*pyruvate carboxylase genes. *American Journal of Botany* **96,** 2234–2239.

Christin PA, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ. 2013. Anatomical enablers and the evolution of C_4 photosynthesis in grasses. *Proceedings of the National Academy of Sciences, USA* **110,** 1381–1386.

Davidson RM, Gowda M, Moghe G, Lin HN, Vaillancourt B, Shiu SH, Jiang N, Buell CR. 2012. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *The Plant Journal* **71,** 492–502.

Denton AK, Simon R, Weber APM. 2013. C_4 photosynthesis: from evolutionary analyses to synthetic reconstruction of the trait. *Current Opinion in Plant Biology* **16,** 315–321.

Ehleringer J, Pearcy RW. 1983. Variation in quantum yield for CO_2 uptake among C_3 and C_4 plants. *Plant Physiology* **73,** 555–559.

Evert RF, Eschrich W, Heyser W. 1977. Distribution and structure of plasmodesmata in mesophyll and bundle-sheath cells of *Zea mays* L. *Planta* **136,** 77–89.

Farineau J, Lelandais M, Morot-Gaudry JF. 1984. Operation of the glycolate pathway in isolated bundle sheath strands of maize and *Panicum maximum*. *Physiologia Plantarum* **60,** 208–214.

Franssen SU, Shrestha RP, Brautigam A, Bornberg-Bauer E, Weber APM. 2011. Comprehensive transcriptome analysis of the highly

complex *Pisum sativum* genome using next generation sequencing. *BMC Bioinformatics* **12,** 227

Furbank RT, Badger MR. 1982. Photosynthetic oxygen exchange in attached leaves of C_4 monocotyledons. *Australian Journal of Plant Physiology* **9,** 553–558.

Furumoto T, Yamaguchi T, Ohshima-Ichie Y, *et al*. 2011. A plastidial sodium-dependent pyruvate transporter. *Nature* **476,** 472–473.

Gowik U, Brautigam A, Weber KL, Weber APM, 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.

Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins. *New Phytologist* **193,** 304–312.

Haferkamp I, Hackstein JHP, Voncken FGJ, Schmit G, Tjaden J. 2002. Functional integration of mitochondrial and hydrogenosomal ADP/ATP carriers in the *Escherichia coli* membrane reveals different biochemical characteristics for plants, mammals and anaerobic chytrids. *European Journal of Biochemistry* **269,** 3172–3181.

Hamel P, Saint-Georges Y, de Pinto B, Lachacinski N, Altamura N, Dujardin G. 2004. Redundancy in the function of mitochondrial phosphate transport in *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. *Molecular Microbiology* **51,** 307–317.

Hatch MD. 1979. Mechanism of C_4 photosynthesis in *Chloris gayana*—pool sizes and kinetics of CO_2-C14 incorporation into 4-carbon and 3-carbon intermediates. *Archives of Biochemistry and Biophysics* **194,** 117–127.

Hatch MD. 1987. C_4 photosynthesis—a unique blend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta* **895,** 81–106.

Hatch MD, Agostino A, Burnell JN. 1988. Photosynthesis in phospho*enol*pyruvate carboxykinase-type C_4 plants—activity and role of mitochondria in bundle sheath-cells. *Archives of Biochemistry and Biophysics* **261,** 357–367.

Hatch M, Mau S. 1977. Properties of phospho*enol*pyruvate carboxykinase operative in C_4 pathway photosynthesis. *Functional Plant Biology* **4,** 207–216.

Häusler RE, Schlieben NH, Nicolay P, Fischer K, Fischer KL, Flügge UI. 2000. Control of carbon partitioning and photosynthesis by the triose phosphate/phosphate translocator in transgenic tobacco plants (*Nicotiana tabacum* L.). I. Comparative physiological analysis of tobacco plants with antisense repression and overexpression of the triose phosphate/phosphate translocator. *Planta* **210,** 371–382.

Herzig S, Raemy E, Montessuit S, Veuthey JL, Zamboni N, Westermann B, Kunji ERS, Martinou JC. 2012. Identification and functional expression of the mitochondrial pyruvate carrier. *Science* **337,** 93–96.

Hibberd JM, Sheehy JE, Langdale JA. 2008. Using C_4 photosynthesis to increase the yield of rice—rationale and feasibility. *Current Opinion in Plant Biology* **11,** 228–231.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. *Genome Research* **9,** 868–877.

Kanai R, Edwards GE. 1999. The biochemistry of C_4 photosynthesis. In: Sage RF, Monson RK, eds. *C4 plant biology*. UK: Academic Press, 49–87.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Research* **12,** 656–664.

Kinoshita H, Nagasaki J, Yoshikawa N, Yamamoto A, Takito S, Kawasaki M, Sugiyama T, Miyake H, Weber APM, Taniguchi M. 2011. The chloroplastic 2-oxoglutarate/malate transporter has dual function as the malate valve and in carbon/nitrogen metabolism. *The Plant Journal* **65,** 15–26.

Koch K. 2004. Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. *Current Opinion in Plant Biology* **7,** 235–246.

Ku MSB, Edwards GE. 1975. Photosynthesis in mesophyll protoplasts and bundle sheath cells of various types of C_4 plants. 4. Enzymes of respiratory metabolism and energy utilizing enzymes of photosynthetic pathways. *Zeitschrift für Pflanzenphysiologie* **77,** 16–32.

Ku MSB, Spalding MH, Edwards GE. 1980. Intracellular localization of phosphoenolpyrvatue carboxykinase in leaves of C_4 and CAM plants. *Plant Science Letters* **19,** 1–8.

Li PH, Ponnala L, Gandotra N, *et al*. 2011. The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* **42,** 1060–1067.

Majeran W, Friso G, Ponnala L, *et al*. 2010. Structural and metabolic transitions of C_4 leaf development and differentiation defined by microscopy and quantitative proteomics in maize. *The Plant Cell* **22,** 3509–3542.

Majeran W, van Wijk KJ. 2009. Cell-type-specific differentiation of chloroplasts in C_4 plants. *Trends in Plant Science* **14,** 100–109.

Maurino VG, Weber APM. 2013. Engineering photosynthesis in plants and synthetic microorganisms. *Journal of Experimental Botany* **64,** 743–751.

Muhaidat R, McKown AD. 2013. Significant involvement of PEP-CK in carbon assimilation of C_4 eudicots. *Annals of Botany* **111,** 577–589.

Ohnishi J, Kanai R. 1983. Differentiation of photorespiratory activity between mesophyll and bundle sheath cells of C_4 plants 1. Glycine oxidation by mitochondria. *Plant and Cell Physiology* **24,** 1411–1420.

Palmieri L, Picault N, Arrigoni R, Besin E, Palmieri F, Hodges M. 2008. Molecular identification of three *Arabidopsis thaliana* mitochondrial dicarboxylate carrier isoforms: organ distribution, bacterial expression, reconstitution into liposomes and functional characterization. *Biochemical Journal* **410,** 621–629.

Pick TR, Bräutigam A, Schlüter U, *et al*. 2011. Systems analysis of a maize leaf developmental gradient redefines the current C_4 model and provides candidates for regulation. *The Plant Cell* **23,** 4208–4220.

Pratt RD, Ferreira GC, Pedersen PL. 1991. Mitochondrial phosphate transport—import of the H^+/P_i symporter and role of the presequence. *Journal of Biological Chemistry* **266,** 1276–1280.

R Development Core Team. 2012. *R: a language and environment for statistical computing*. Vienna, Austria.

Renne P, Dressen U, Hebbeker U, Hille D, Flugge UI, Westhoff P, Weber APM. 2003. The Arabidopsis mutant *dct* is deficient in the plastidic glutamate/malate translocator DiT2. *The Plant Journal* **35,** 316–331.

Sage RF. 2004. The evolution of C_4 photosynthesis. *New Phytologist* **161,** 341–370.

Sage RF, Christin PA, Edwards EJ. 2011. The C_4 plant lineages of planet Earth. *Journal of Experimental Botany* **62,** 3155–3169.

Schliesky S, Gowik U, Weber APM, Bräutigam A. 2012. RNA-seq assembly—are we there yet? *Frontiers in Plant Science* **3,** 220.

Schomburg I, Chang A, Placzek S, *et al*. 2013. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Research* **41,** D764–D772.

Sommer M, Bräutigam A, Weber APM. 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.

Sonnhammer ELL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28,** 405–420.

Sowinski P, Bilska A, Baranska K, Fronk J, Kobus P. 2007. Plasmodesmata density in vascular bundles in leaves of C_4 grasses grown at different light conditions in respect to photosynthesis and photosynthate export efficiency. *Environmental and Experimental Botany* **61,** 74–84.

Sowinski P, Szczepanik J, Minchin PEH. 2008. On the mechanism of C_4 photosynthesis intermediate exchange between Kranz mesophyll and bundle sheath cells in grasses. *Journal of Experimental Botany* **59,** 1137–1147.

Stitt M, Heldt HW. 1985. Generation and maintenance of concentration gradients between the mesophyll and bundle sheath in maize leaves. *Biochimica et Biophysica Acta* **808,** 400–414.

Strand A, Zrenner R, Trevanion S, Stitt M, Gustafsson P, Gardestrom P. 2000. Decreased expression of two key enzymes in the sucrose biosynthesis pathway, cytosolic fructose-1,6-bisphosphatase and sucrose phosphate synthase, has remarkably different consequences for photosynthetic carbon metabolism in transgenic Arabidopsis thaliana. *The Plant Journal* **23,** 759–770.

Taniguchi Y, Taniguchi M, Nagasaki J, Kawasaki M, Miyake H, Sugiyama T. 2003. Functional analysis of chloroplastic dicarboxylate transporters in maize. *Plant and Cell Physiology* **44,** S64–S64.

Toledo-Silva G, Cardoso-Silva CB, Jank L, Souza AP. 2013. De novo transcriptome assembly for the tropical grass *Panicum maximum* Jacq. *PLoS One* **8,** e70781.

Usuda H, Edwards GE. 1980. Localization of glycerate kinase and some enzymes for sucrose synthesis in C_3 and C_4 plants. *Plant Physiology* **65,** 1017–1022.

Vicentini A, Barber JC, Aliscioni SS, Giussani LM, Kellogg EA. 2008. The age of the grasses and clusters of origins of C_4 photosynthesis. *Global Change Biology* **14,** 2963–2977.

Walker R, Trevanion S, Leegood R. 1995. Phospho*enol*pyruvate 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.

Walker RP, Acheson RM, Tecsi LI, Leegood RC. 1997. Phospho*enol*pyruvate carboxykinase in C_4 plants: its role and regulation. *Australian Journal of Plant Physiology* **24,** 459–468.

Wang Y, Bräutigam A, Weber APM, Zhu XG. 2014. Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis. *Journal of Experimental Botany* **65** (in press).

Weber A, Menzlaff E, Arbinger B, Gutensohn M, Eckerskorn C, Flügge UI. 1995. The 2-oxoglutarate/malate translocator of chloroplast envelope membranes: molecular cloning of a transporter containing a 12-helix motif and expression of the functional protein in yeast cells. *Biochemistry* **34,** 2621–2627.

Weber APM, Bräutigam A. 2013. The role of membrane transport in metabolic engineering of plant primary metabolism. *Current Opinion in Biotechnology* **24,** 256–262.

Weber APM, von Caemmerer S. 2010. Plastid transport and metabolism of C_3 and C_4 plants—comparative analysis and possible biotechnological exploitation. *Current Opinion in Plant Biology* **13,** 257–265.

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. 2007. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiology* **144,** 32–42.

Westhoff P, Gowik U. 2004. Evolution of C_4 phospho*enol*pyruvate carboxylase. Genes and proteins: a case study with the genus Flaveria. *Annals of Botany* **93,** 13–23.

Westhoff P, Herrmann RG. 1988. Complex RNA maturation in chloroplasts. *European Journal of Biochemistry* **171,** 551–564.

Wingler A, Walker RP, Chen ZH, Leegood RC. 1999. Phospho*enol*pyruvate carboxykinase is involved in the decarboxylation of aspartate in the bundle sheath of maize. *Plant Physiology* **120,** 539–545.

Yoshimura Y, Kubota F, Ueno O. 2004. Structural and biochemical bases of photorespiration in C_4 plants: quantification of organelles and glycine decarboxylase. *Planta* **220,** 307–317.

VIII.4 CO-AUTHORED MANUSCRIPTS

Manuscript 7

Azolla domestication towards a biobased economy?

Paul Brouwer, Andrea Bräutigam, ***Canan Külahoglu****, Anne O. E. Tazelaar, Samantha Kurz, Klaas G. J. Nierop, Adrie van der Werf, Andreas P. M. Weber and Henriette Schluepmann*

Published in New Phytologist (2014) 202: pp.1069–1082
doi: 10.1111/nph.12708

Impact Factor: 6.54

**Co-author*

Main findings:

In this study explores methods and resources for advancing the domestication of mosquito fern (*Azolla filiculoides*). It features methods for *Azolla filiculoides* dissemination, cross-fertilizaion and cryopreservation/storage and transcriptome resources from RNA-sequencing of the microsporocarps, megasporocarps and sporophytes. A unigene database covering completely the central plant metabolism and most of the cellular processes and regulatory networks is provided. Furthermore, the transcriptome analysis showed, that A FLOWERING LOCUS T-like protein is involved in the transition to sexual repoduction in ferns.

Contributions:

- Handling of RNA samples
- Providing DNAse treatment protocols
- Quality checking RNA samples prior library preparation
- Manuscript proof-reading

New Phytologist

Research

Methods

Azolla domestication towards a biobased economy?

Paul Brouwer[1], Andrea Bräutigam[2], Canan Külahoglu[2], Anne O. E. Tazelaar[1], Samantha Kurz[2], Klaas G. J. Nierop[3], Adrie van der Werf[4], Andreas P. M. Weber[2] and Henriette Schluepmann[1]

[1]Molecular Plant Physiology Department, Utrecht University, Padualaan 8, 3584 CH, Utrecht, the Netherlands; [2]Institute for Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany; [3]Organic Geochemistry, Utrecht University, Budapestlaan 4, 3584 CD, Utrecht, the Netherlands; [4]Plant Research International, Droevendaalsesteeg 1, 6708 PB, Wageningen, the Netherlands

Author for correspondence:
Henriette Schluepmann
Tel: +31 30 253 3289
Email: h.schlupmann@uu.nl

Received: *31 October 2013*
Accepted: *22 December 2013*

New Phytologist (2014) **202:** 1069–1082
doi: 10.1111/nph.12708

Key words: *Azolla filiculoides*, biobased economy, cryopreservation, domestication, fern, RNAseq, sexual reproduction.

Summary

- Due to its phenomenal growth requiring neither nitrogen fertilizer nor arable land and its biomass composition, the mosquito fern *Azolla* is a candidate crop to yield food, fuels and chemicals sustainably. To advance *Azolla* domestication, we research its dissemination, storage and transcriptome.
- Methods for dissemination, cross-fertilization and cryopreservation of the symbiosis *Azolla filiculoides–Nostoc azollae* are tested based on the fern spores. To study molecular processes in *Azolla* including spore induction, a database of 37 649 unigenes from RNAseq of microsporocarps, megasporocarps and sporophytes was assembled, then validated.
- Spores obtained year-round germinated *in vitro* within 26 d. *In vitro* fertilization rates reached 25%. Cryopreservation permitted storage for at least 7 months. The unigene database entirely covered central metabolism and to a large degree covered cellular processes and regulatory networks. Analysis of genes engaged in transition to sexual reproduction revealed a FLOWERING LOCUS T-like protein in ferns with special features induced in sporulating *Azolla* fronds.
- Although domestication of a fern–cyanobacteria symbiosis may seem a daunting task, we conclude that the time is ripe and that results generated will serve to more widely access biochemicals in fern biomass for a biobased economy.

Introduction

In the coming decades we anticipate a rapid increase in world population that will greatly increase global demand for food, although this is constrained by the limited availability of arable land. With the depletion of fossil resources, plants will need to provide an increasingly large proportion of our requirements for energy and chemicals in addition to food (EPSO, 2005). Intensive agriculture using conventional crops is often associated with high inputs and negative climate impacts (Jensen *et al.*, 2012). For example, periodic application of excess nitrogen fertilizer leads to high nitrous oxide (N_2O) emissions (Smith *et al.*, 1997). N_2O has a global warming potential 310-fold higher than CO_2 and is, in terms of impact, the third most important greenhouse gas, after CO_2 and CH_4 (IPCC, 2007). To meet the challenges of lowering N_2O emissions and increasing production, novel crops that require less or no nitrogen fertilizer, use nonarable land with high biomass yields, and feed both food and chemical industries, are much sought after.

Azolla is one such potential crop: it is an aquatic fern that may be cultivated in closed systems on nonarable land as well as in freshwater basins. It is known for its high growth rates, doubling biomass in 2 d under favorable conditions (Wagner, 1997). *Azolla* thrives without the addition of nitrogen fertilizer to sustain its growth because it harbors symbiotic nitrogen-fixing cyanobacteria; this has led to its use as a nitrogen fertilizer in paddy fields of South-east Asia (Wagner, 1997). It is an accumulator of heavy metals and its use in waste water treatment has been demonstrated on a small scale (Costa *et al.*, 1999; Antunes *et al.*, 2001). *Azolla* is a high-protein animal feed, but its limited digestibility (Alalade & Iyayi, 2006; Abdel-Tawwab, 2008) may be the result of it containing tannins not lignin (Nierop *et al.*, 2011), together with other polyphenols (Fasakin, 1999). Compared to algae and free-living diazotrophic cyanobacteria, *Azolla* requires no mixing of the water body and is easier to harvest. To fully harness the potential of *Azolla*, however, its domestication is a prerequisite. Domestication requires protocols for the collection, storage and dissemination of reproductive structures (Willemse, 2009). In addition, breeding varieties adapted to specific uses and permitting containment of this otherwise invasive weed requires control over sexual reproduction. To accelerate the breeding process, genetic and sequence information is required. As with most ferns, *Azolla* is currently neither domesticated nor bred (Meyer *et al.*, 2012). Sequence information is lacking except for the genome

New Phytologist

sequence of its endosymbiotic cyanobacteria *Nostoc azollae* (Ran *et al.*, 2010).

Azolla is a member of the *Salviniaceae* and heterosporous: megaspores and microspores each form gametophytes bearing the gametes that mate to form the sporophyte, the dominant diploid phase of *Azolla*. A very small gametophyte yielding a single megaspore develops inside the megasporocarp (Peters & Perkins, 1993). By contrast many microsporangia with microspores packaged in several pseudocellular massulae develop in the microsporocarp (Herd *et al.*, 1985). *Azolla* spores inside the sporocarps generally exhibit strong resistance to external stresses, such as drought (Becking, 1987) and subzero temperatures (Janes, 1998). Using sporocarps to preserve biodiversity may seem a logical approach; however, sporocarps have not been generally available for *Azolla* species and the preservation methods will need to be improved to reach long-term and high viability. Long-term storage by preservation of whole sporophytes or only small parts of the meristems was not reported for *Azolla*. Instead, *Azolla* varieties are currently maintained by *in vitro* subculture in biodiversity collections such as the Biofertilizer Collection of the International Rice Research Institute (IRRI; Watanabe *et al.*, 1992). Continuous subculture, however, is laborious and thus prone to human error, and may promote somaclonal variation and adaptation of the specimens to the artificial environment under which they have been cultured during the past 25 yr. A more reliable method to preserve varieties needs to be developed that allows selection and breeding efforts.

Control over *Azolla* sexual reproduction will be of paramount importance for disseminating existing varieties or breed new varieties. Controlling the production of spores and fertilization will be most critical. A number of authors have described sexual reproduction in *Azolla* species (Becking & Donze, 1981; Becking, 1987; Peters & Meeks, 1989; Wagner, 1997; Zheng *et al.*, 2009; Carrapiço, 2010). These studies mainly focused on the vertical transfer of the cyanobacterial symbiont during the reproduction process. Methods were described that used sporocarps to raise new sporelings field plots, relying on the natural processes of fertilization on the floor bed (Quin-Yuan *et al.*, 1987; Shuying, 1987). None of the publications on sexual reproduction in *Azolla* described fertilization in a controlled laboratory environment, and none described the induction of spore formation *in vitro*. Knowledge of the environmental cues and molecular mechanisms controlling sporulation in *Azolla* is very scarce. Studies have reported outdoor conditions under which sporulation has occurred in different species, such as high population density (Becking, 1987; Janes, 1998), shorter days and colder nights (Kar *et al.*, 2002), high light intensity and high temperature (Becking, 1987). Herd *et al.* (1989) showed the effect of temperature regime on the sporulation of a large variety of *Azolla* species and strains, but could not establish a clear link with sporulation frequency. Also different growth-regulating substances could not induce sporulation in *A. pinnata* and *A. filiculoides* (Herd *et al.*, 1989). Kar *et al.* (2002) later showed that a combination of hormones could increase sporulation frequency and promote megasporocarp formation, but only in cultures that were already sporulating. As *Azolla* species adapted to differing environments over time, they likely evolved differential environmental cues triggering sporulation. To develop a reliable protocol to induce sporulation in several species, a strategy focusing on more downstream, molecular, processes controlling the transition to sexual reproduction will likely be more effective than studying environmental clues alone.

In flowering plants, the transition to sexual reproduction is controlled by multiple input pathways which measure day length, temperature, nutritional status and age of the plant. FLOWERING LOCUS T (FT) signals environmental cues perceived in leaves as it moves via the vasculature to the shoot meristems; there, FT activates LEAFY (LFY) and with it the transition to sexual reproduction. FT was not found in the genomes of the nonvascular lower plants *Physcomitrella* and *Selaginella* (Banks *et al.*, 2011) and has not previously been described in ferns, which were the first plants to evolve highly developed vasculature.

In lower plants LFY is thought to promote vegetative development of the gametophyte, and Floyd & Bowman (2007) proposed that LFY repression after fertilization would be required for development of the extended vegetative growth of the sporophytes in higher plants. LFY from the fern *Ceratopteris* was capable of partially suppressing the phenotype of the Arabidopsis *lfy* mutant but LFY from the moss *Physcomitrella* did not (Maizel *et al.*, 2005). Some targets of LFY are therefore conserved in ferns and higher plants. LFY activates the ABC genes in angiosperms and gymnosperms, thus promoting the transition to sexual reproduction. What induces the transition to sexual reproduction in ferns and other lower plants is mostly unknown.

Azolla belongs to an under-sampled group with regard to transcriptome or genome sequence resources. Studying molecular components that may control the transition to reproductive development in *Azolla* is therefore difficult. RNA sequencing of species without a sequenced genome provides a valuable resource. While the assemblies remain far from perfect (Schliesky *et al.*, 2012), both unigene databases (Brown *et al.*, 2011; Kajala *et al.*, 2012; Sommer *et al.*, 2012) and quantitative gene expression data (Brautigam *et al.*, 2011a; Gowik *et al.*, 2011) have successfully been used to explore the physiology and gene regulation in species without prior sequence resources.

In order to provide the basis needed for domestication of *Azolla*, we begin by describing a method to collect large amounts of *A. filiculoides* spores all year round. We define and illustrate key stages in the germination process then demonstrate *in vitro* fertilization and germination of *Azolla* spores. We further show that cryopreservation of fertilized megaspores using a drying pretreatment is effective for preserving *A. filiculoides* while also preserving the *N. azollae* symbiont, opening the way to genomic characterization of the cryopreserved variety. From sequencing reads of RNA from megasporocarps, microsporocarps and sporophytes we assemble a database of 37 649 unigenes which we annotate so as to provide a resource to molecular research. We then describe genes in *Azolla* possibly involved in inducing spore formation based on what is known from induction of the reproductive phase in flowering plants.

New Phytologist

Research 1071

Materials and Methods

Collecting *Azolla* sporocarps

Sporulating *A. filiculoides* Lam. was collected in mid-October 2012 from a ditch in Utrecht, the Netherlands (52°04′24″N; 5°08′53″E) and kept in demineralized water in a glasshouse at 5–15°C and 14 h days with a light intensity of at least 70 μmol m^{-2} s^{-1} Photosynthetic Photon Flux Density (PPFD).

Unfertilized megasporocarps were collected according to Toia *et al.* (1987) with modifications: sporulating plants were placed on top of a stack of sieves with 1000-, 500- and 200-μm mesh sizes and megasporocarps were detached from the stems using a strong water jet. Residue recovered on the 200-μm mesh size sieve was then layered on a 3 M sorbitol: water step gradient and centrifuged at 400 ***g*** for 10 min, resulting in a clear layer of pure megasporocarps. Megasporocarps were washed three times by centrifugation with 50 ml water. To collect mostly fertilized megasporocarps, sediment accumulating at the bottom of the containers in which sporulating cultures were kept was used in the above procedure, except that the sorbitol: water step gradient was layered on top of the residue of the 200-μm mesh size sieve to obtain a clear layer of fertilized megasporocarps. Mature microsporocarps were plucked manually from the plants.

Documenting megaspore germination

A mixture of megasporocarps and microsporangia was germinated in 400 μl of either *Azolla* growth medium at pH 5.5 (Watanabe *et al.*, 1992) or demineralized water in a growth cabinet set at 25°C day : 15°C nights with 12 h light (40–70 μmol s^{-1} m^{-2} PPFD). Spores were scored for germination over a period of 6 wk. Morphological changes to the megasporocarps were tracked under a binocular Leica Axioskop light microscope, using either × 10 or × 5 objectives, and key events documented by digital imaging in dark field using a Leica PFC 420C Camera (Carl Zeiss BV, Sliedrecht, Netherlands). Images of sporulating *Azolla* plants and *Azolla* sporelings were made using a Nikon DXM12000 camera (Nikon Instruments Europe BV, Amsterdam, Netherlands) on either a Zeiss Axiovert 35M reversed microscope or a Zeiss Stemi SV11 stereo microscope (Carl Zeiss). Mature *Azolla* plants were photographed using a Nikon D300S DSLR camera with a 60 mm macro objective.

Documenting *Nostoc azollae*

Nostoc azollae phycobilisomes differ in their spectral emissions from plant pigments (Rigbi *et al.*, 1980), therefore, a Leica SP2 confocal fluorescence microscope, equipped with either × 16 or × 40 objectives and a helium–neon laser with excitation wavelength of 543 nm, was used to visualize *N. azollae* fluorescence in the range 630–670 nm. Plant tissue fluorescence was visualized in the range 560–630 nm and/or 680–750 nm. During capture, images were frame-averaged over 16 frames. To determine the presence of cyanobacteria in sporelings, two glass plates were pressed against each other, thereby squeezing the cyanobacteria cells out of the leaf pockets.

Fertilizing *A. filiculoides* spores *in vitro*

Fifty megasporocarp batches were mixed during 20 s with microsporangia at various ratios in 1 ml water, then incubated in darkness and room temperature (RT) for 2–13 d (fertilization periods), before megasporocarps (without free microsporangia and massulae) were transferred to *Azolla* growth medium (8 ml) and left to develop in 25°C day : 15°C nights with 12 h light (40–70 μmol m^{-2} s^{-1} PPFD). After 2 wk development germination was scored at regular intervals. All conditions were tested in triplicate.

Testing and optimizing cryopreservation protocols for long-term storage of *A. filiculoides*

The cryopreservation protocols summarized in Table 1(a) were tested on batches of 20 megasporocarps. Each condition was evaluated in duplicate. For cryopreservation involving cryoprotectant pretreatment, 1 ml of the cryoprotective solution was added to the megasporocarps immediately before snap-freezing in liquid nitrogen (LN). Batches were kept for 5 min in LN, then thawed on a heating plate set at 30°C for 1.5 min and washed thrice in 1 ml medium. To test cryopreservation using a drying pretreatment, megasporocarps were dried for 1, 4 and 8 d in the fume hood at RT and then snap frozen in LN without added fluid. Batches were kept for 5 min in LN then thawed for 1.5 min at 30°C before adding 1 ml medium. For each cryopreservation protocol, a treatment control was included that was not frozen but still exposed to the cryopreservation pre-treatment; additionally two controls were included that received no treatment. Germination was as in the Documenting megaspore germination section.

In order to optimize cryopreservation involving drying as pretreatment, megasporocarps with attached massulae were collected from sediment and the number of megasporocarps in each batch was increased to 250–600 megasporocarps. Drying conditions tested included 1, 4, 7, 16 and 32 d drying at RT in the fume hood as above. Additional batches were dried for 1, 4 and 7 d at constant temperature (CT) of 26°C. For each cryopreservation pre-treatment a control batch was included that was exposed to the cryopreservation pre-treatment only. Two batches were included that were not subjected to any cryopreservation treatment. These batches served as controls and allowed comparison of germination rates between megaspores collected from plant material and spores collected from growth container sediments. The freezing, thawing and germination of the batches was as above.

In order to test long-term storage, batches of spores collected from sediment and dried for 7 d at RT were frozen in LN, then transferred to −80°C and stored for 1 d, 1, 2, 5 and 7 months. Another set of dried batches was frozen at −20°C and then stored for only 1 and 2 months. A final set of batches was not subjected to a drying treatment, but directly frozen in the −20°C freezer,

New Phytologist

Table 1 Cryopreservation protocols tested on Azolla filiculoides spores. (a) Cryopreservation pre-treatment definitions[1] and (b) Germination rate and viability for unfrozen and frozen megasporocarps after cryopreservation pre-treatment[2]

Cryopreservation pre-treatments	Concentration/treatment time	Previously reported use
(a)		
DMSO + Glycerol	2 M DMSO, 3.2 M Glycerol	Tree fern, *Dicksonia Sellowiana,* Rogge *et al.* (2000)
DMSO + EG + PVP	2 M DMSO, 4.8 M EG, 2% PVP	Zebrafish embryo, Riesco *et al.* (2012)
Sucrose	0.4 M Sucrose	Fern, *Pteris adscensionis*, Barnicoat *et al.* (2011)
Trehalose	0.4 M Trehalose	Fern, *Pteris adscensionis*, Barnicoat *et al.* (2011)
DMSO + EG + Glucose + Trehalose	2 M DMSO, 2.4 M EG, 3.2 M Glucose, 0.4 M Trehalose	Duckweed, *Lemna Minor*, Parsons & Wingate (2012)
Drying	1 d, 4 d, 8 d	Tree fern, *Dicksonia Sellowiana*, Rogge *et al.* (2000)

	No freezing			Freezing		
Pre-treatment	*N*	Viability (%)	Germination (%)	*N*	Viability (%)	Germination (%)
(b)						
None (control)	32	67	23	42	0	0
DMSO + Glycerol	29	93	7	34	0	0
DMSO + EG + PVP	36	79	23	41	0	0
Sucrose	37	92	29	49	0	0
Trehalose	50	82	57	70	0	0
DMSO + EG + Glucose + Trehalose	35	71	48	71	0	0
1 d drying	31	51	9.40	36	19	0
4 d drying	24	24	21	27	30	0
8 d drying	31	38	10	36	25	6

[1]Dimethyl sulfoxide (DMSO), ethylene glycol (EG) and Polyvinyl pyrrolidone (PVP).
[2]Viability is the percentage of megasporocarps with floats. Freezing was in liquid nitrogen. Each condition was tested in duplicate. *N*, number of megasporocarps tested.

but these repeatedly had zero germination. Zero and 7 months storage were tested in duplicate and triplicate, respectively.

RNA sequencing and quantitative RT-PCR (qRT-PCR)

Ferns, in general, and heterosporous ferns like *Azolla*, in particular, represent a particularly under-sampled group with regard to sequence information. Tissues from three different developmental phases of the complex lifecycle were therefore chosen for RNA sequencing (RNA-seq) to capture the transcriptome of *Azolla*. Microsporocarps and megasporocarps were plucked from sporulating *Azolla* grown on demineralized water in the glasshouse. Sporophytes were sterile, nonsporulating, grown in medium with and without nitrogen, and collected at 6 h intervals over 24 h starting 1 h before dawn. Total RNA was extracted (Spectrum Plant Total RNA Kit; Sigma-Aldrich) from megasporocarps, microsporocarps and sporophytes, then digested with DNase I. Two replicate extractions were pooled for each of the reproductive tissues and eight replicate extractions were pooled for the sporophyte tissue (corresponding to four time points and growth with/without nitrogen). Poly(A+) RNA was enriched using oligo (dT) Dynabeads (Ambion). To enrich capped transcripts cDNA was synthesized using the Clontech SMARTer kit (Takara Bio Europe SAS, Saint-Germain-en-Laye, France). Libraries were made from 100 ng template independently for each of the three tissues using the Ion Plus Fragment Library Kit with Ion Xpress™ Barcode Adapters (BC12, BC13 and BC01; Life Technologies). PCR amplifications and emulsions were generated using the Ion PGM™ Template OT2 400 Kit and Ion OneTouch™ 2 System (Life Technologies) with emulsions at 8 pM. Sequencing was completed with the Ion 316™ Chip v2 on a Ion PGM™ sequencer.

The resulting sequencing reads were inspected with the FastQC tools (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Using fastx tools (http://hannonlab.cshl.edu/fastx_toolkit/), reads were trimmed by removing bases with a Phred score < 20; reads which were pruned to < 50% of their original length by this step were discarded. The remaining reads were filtered for those which had < 90% of bases with a Phred score above 20 and reads shorter than 50 bases were discarded. The trimmed and filtered reads were assembled using CLC Genomics Workbench software (CLC Bio, Aarhus, Denmark) with default parameters (Brautigam *et al.*, 2011b). Quality was assessed based on read length distribution and unigene annotation (Schliesky *et al.*, 2012). The unigene database was annotated using BLAT (Kent, 2002) against proteins from *Selaginella moellendorffii*, *Arabidopsis thaliana* and *Nostoc azollae* to identify the proportion of genes similar to those in plant and symbiont genomes (Supporting Information Notes S1).

The unigene database was uploaded to the KEGG Automated Annotation Server (KAAS, http://www.genome.jp/tools/kaas/) to test the coverage of common pathways (Moriya *et al.*, 2007). The resulting maps were exported from the server, curated for pathways not present in plants, and sorted according to content (proteins present in the unigene database in green colour; see Notes S2–S4).

New Phytologist

Research 1073

All reads were then mapped onto the unigene database with default parameters using CLC Genomics Workbench (Table S1). The relative read counts for each unigene were normalized by the total read counts in each tissue (reads per million mapped reads; rpm). To find signature genes for the microsporocarp, megasporocarp and sporophyte, a small group of unigenes (1%, 3%, and 3%, respectively) with read coverage above 100 rpm were identified for each tissue. Annotation of the highly read unigenes against *A. thaliana* served to assign ontology terms in the ontology list derived from MapMan (Usadel *et al.*, 2006, 2009) in Excel. Fisher's Exact test was then applied to evaluate whether the highly read unigenes were enriched in any one ontology term of pathways (Fisher, 1922); *P*-values were corrected for multiple hypothesis testing by the Bonferroni method.

In order to validate the unigene sequences and quantify gene-expression, RNA was extracted from *A. filiculoides* nonsporulating sporophytes grown in the cabinet, sporulating sporophytes collected from the wild (September 2013) and microsporocarps, and then cDNA synthesized. The primers for qRT-PCR were for the references *AfTUBULIN* (AfTUBF: CCTCCGAAAACT CTCCTTCC; AfTUBR: GGGGGTGATCTAGCCAAAGT) and *AfADENINE PHOSPHORIBOSYLTRANSFERASE* (AfAPTF: TAGAGATGCATGTGGGTGCAGT; AfAPTR: AAAA GCGGTTTACCACCCAGTT) (Salmi & Roux, 2008). Further qRT-PCR primers were for *AfFT* (AfFTF: AAGAGATTTG GCAAGCTGGA; AfFTR: TAGCAACCACCAACAGCATC), *AfSOC1* (AfSOCF: ATGGGATCGTAAGGCTTCAAAA; AfSOCR: AGCAGAGCACACAGGTCTCAAC), *AfLFY* (AfLFYF: GCGGCAAGAGGAAGAGATAGA; AfLFYR: AGT GGATGTGCTCTTGCTGAA) and *AfCAL* (AfCALF: TTTG CATCTTTCGCTCTCTCA; AfCALR: CCAAGCTGCACAA TGTAAGGA). Data was from three biological replicates, significance was assessed by *t*-test with $P < 0.05$.

Results

Azolla filiculoides spores can be collected year round

A difficulty for researching sexual reproduction in *Azolla* is access to sporulating populations. Storage of sporulating *A. filiculoides* in rainwater at temperatures varying from 5 to 15°C and 14 h light with minimum intensity of 40 μmol m^{-2} s^{-1} PPFD d^{-1} resulted in a constantly sporulating population. Megasporocarps were collected by subsequent sieving and purification. Yields from plants varied from 2890 (December 2012) to 55 190 (June 2013) megasporocarps purified from 1 m^2 of *Azolla* mat (standing crop density 2.1–2.8 kg FW m^{-2}). By contrast, yields from the sediment were much higher, ranging from 157 600 (December 2012) to 343 740 (August 2013) purified from 1 m^2 of *Azolla* mat. Sediment FW increased as organic material accumulated, reaching an average of 14 kg m^{-2} in August 2013. Year-round access to large amounts of spores allowed testing of *in vitro* germination, *in vitro* fertilization and preservation methods, and allowed extraction of RNA from sporocarps.

Cotyledon emergence, not float emergence characterizes megaspore fertilization

Key stages in the sexual reproduction of *Azolla* were documented by three dimensional reconstruction of digital images from dark-field microscopy. Sporocarps developed under the sporophyte (Fig. 1a) in pairs, first as microsporocarp pairs, then as the sporophytes reached maturity as micro and megasporocarp pairs (Fig. 1b). The megasporocarp (Fig. 1c), containing the female megaspore and cyanobacteria akinetes, quickly sunk. Detached microsporocarps generally burst open, releasing the microsporangia (Fig. 1d); once these burst, massulae were released and

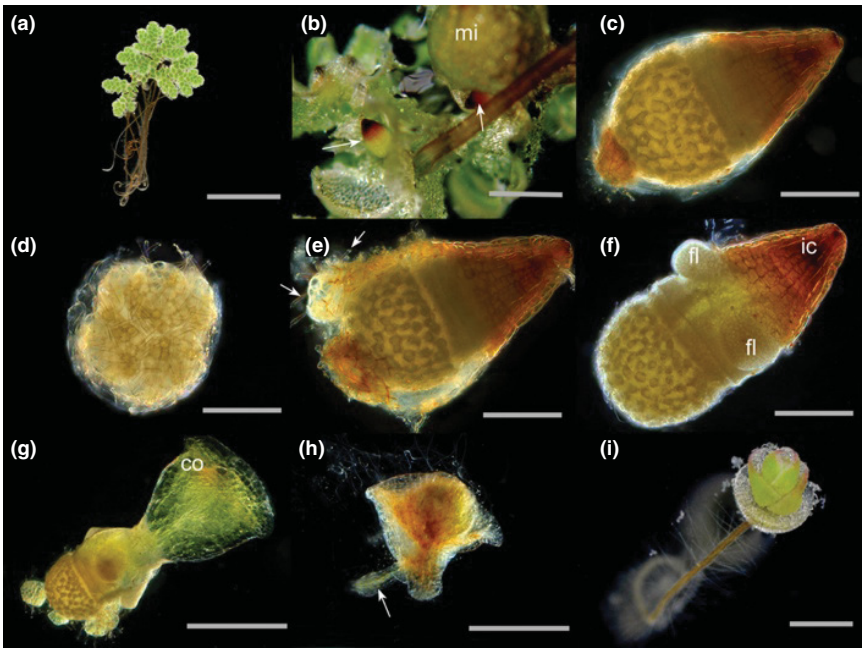


Fig. 1 Key stages in the sexual reproduction of *Azolla filiculoides*. (a) Sporophyte; bar, 10 mm. (b) Megasporocarps (arrows) and microsporocarp (mi) at the underside of a sporulating plant; bar, 1 mm. (c) Detached megasporocarp; bar, 0.2 mm. (d) Microsporangium containing four massulae; bar, 0.2 mm. (e) The massulae's glochidia (arrows) allow it to attach to the megasporocarp; bar, 0.2 mm. (f) Megasporocarp floats (fl) emerge from beneath the indusium cap (ic); bar, 0.2 mm. (g) Cotyledons (co) push away indusium cap and emerge from the megasporocarp; bar, 0.5 mm. (h) Sporeling with root (arrow) detaches from the megasporocarp; bar, 0.5 mm. (i) *Azolla* sporeling floating; bar, 1 mm.

New Phytologist

entangled in filamentous appendages of the epispore wall of the megaspore (Fig. 1e). To study spore fertilization and preservation, morphological changes denoting successful fertilization and germination were characterized.

The first morphological change to the megasporocarp was elongation of the megasporocarp and emergence of floats from underneath the indusium cap (Fig. 1f). Nonfertilized megaspores also developed floats, especially when nutrients were present in the medium. Float development could be the result of independent development of the female gametophytes, or rapid proliferation of the cyanobacteria pushing the floats outward. Hence, megasporocarps with floats did not automatically imply that fertilization had taken place but instead revealed whether or not the megaspores were viable. Fertilized viable megaspores developed further into sporelings with a cotyledon (Fig. 1g). Root development was visible shortly after (Fig. 1h). Sporelings became sufficiently buoyant to float at the surface when first leaves appeared (Fig. 1i).

Germination rates up to 30% were achieved after *in vitro* fertilization, of which 95% occurred within 26 d (Fig. 2a). Nutrient availability promoted germination, but was not essential: when a 1 : 1.5 mixture of megasporocarps and microsporangia was placed in medium or demineralized water, germination rates reached 30.1% and 8.1%, respectively. In both cases no nitrogen was present, indicating that sufficient nitrogen reserves were present to allow germination.

To conclude, observations on spore germination revealed that the emergence of cotyledons but not of floats was a reliable measure of megaspore fertilization.

Azolla megasprores can be fertilized *in vitro*

Controlled fertilization is a prerequisite for breeding. We therefore set out to test whether *Azolla* spores could be fertilized *in vitro* and how much time would be required for this process. Megasporocarps collected from sporophytes were incubated for various durations with microsporangia, from burst microsporocarps, then transferred to fresh medium in the absence of massulae and left to germinate. Germination frequencies obtained varied from 7% to 27% (Fig. 2b). Megasporocarps without added microsporangia did not develop sporelings, confirming that megaspores were not fertilized when attached to the sporophytes nor during collection. Incubation for fertilization beyond 6 d from 9 to 13 d increased germination frequencies above 20%. Useful *A. filiculoides* spore fertilization rates were thus obtained *in vitro*, within a practicable time span.

Drying rather than cryoprotectant pre-treatment permits spore survival to cryopreservation

In order to preserve the diversity of natural varieties or varieties developed for breeding and further dissemination, various stages of *Azolla* will need to be preserved over long periods of time without loss of viability or genetic alterations. Since spores are the natural dissemination stages of *Azolla*, we tested whether these could be simply preserved by drying, but spores in dried macrosporocarps were never viable when extending storage to 4 wk.

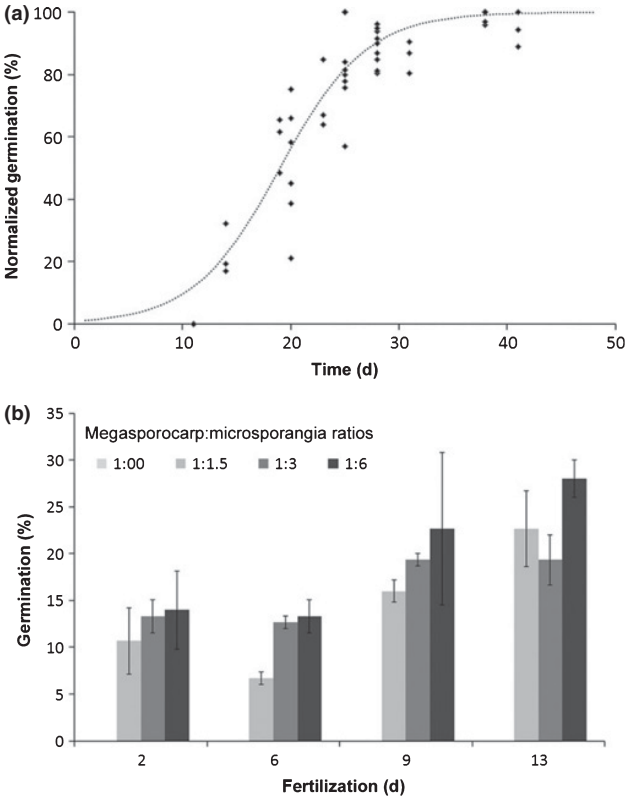


Fig. 2 *Azolla filiculoides* germination and *in vitro* fertilization. (a) Germination frequencies over a 5-wk period: germination was scored when cotyledons had emerged from megasporocarps, data from several experiments, frequencies were normalized by dividing the number of germinated megasporocarps at given time points by the number of germinated megasporocarps after 40 d for each experiment. The trend-line is the best-fit sigmoid function with $\beta = 0.4$ and $T0.5 = 18$ d. (b) Germination rate as a function of fertilization period, for different megasporocarp: microsporangia ratios. Spores were mixed, incubated in darkness at room temperature (RT) for 2, 6, 10 and 13 d for fertilization, then megasporocarps only transferred to growth medium for germination. Germination was scored after 40 d. Error bars, ± SE.

Alternatively spores were frozen directly in −20°C or liquid nitrogen (LN), either in medium or without, but neither gave viable spores.

We then tested cryopreservation protocols (Table 1a) employing cryoprotectants. Cryoprotectant mixtures did not generally affect the viability of megaspores or germination frequencies (Table 1b, *No freezing*). Cryoprotectant mixtures, however, did not permit survival to the freeze/thaw cycle (Table 1b, *Freezing*). Sporelings were solely recovered when spores had been dried in a fume hood for 8 d at RT before freezing. Drying pre-treatments moreover allowed 10–30% of the megasporocarps to develop floats, an indication for viability, whilst all other pre-treatments did not (Table 1b, *No freezing*- viability). The low germination of 23% in the untreated control megasporocarps revealed that only about a quarter of the megaspores used for this screening experiment were fertilized. We concluded that drying pre-treatment changed the physiology of fertilized *A. filiculoides* megaspores so as to resist freeze/thaw cycles.

New Phytologist

Table 2 Survival of *Azolla filiculoides* spores to the freeze/thaw cycle after various drying pre-treatments

Pre-treatment	Freezing	*N*	Germination (%)
None (control)	No	281	50.96
RT 1 d	LN	391	0.00
RT 4 d	LN	382	0.79
RT 7 d	LN	247	5.66
RT 16 d	LN	283	1.06
RT 32 d	LN	305	2.62
CT 26°C 1 d	LN	371	4.86
CT 26°C 4 d	LN	340	26.50
CT 26°C 7 d	LN	279	50.62

Megasporocarps collected from sediment were dried 1–32 d at either fluctuating room temperature (RT) of constant temperature (CT) of 26°C. *N*, number of megasporocarp tested. Megasporocarps were either not frozen (No) or frozen in liquid nitrogen (LN), thawed and then germinated as described in the Materials and Methods section.

We extended the drying pre-treatment at RT to 32 d and tested drying at higher temperature of 26°C, simulating natural drought conditions. To improve the percentage of fertilized megaspores, megasporocarps from sediment were used which also allowed using greater numbers of megasporocarps per condition to increase the sensitivity of our test. A germination frequency of 51% in the untreated control revealed that half of the megasporocarps from sediment were fertilized and viable (Table 2). The optimal duration of drying at RT was 7 d, after which the frequency of survival decreased. Constant temperature (CT) at 26°C during the drying improved survival over the lower and fluctuating RT tremendously: the batch dried at CT 26°C for 7 d before the freeze/thaw cycle reached 50% germination, which was nearly equal to that of the untreated control (51%) and almost 10 times higher compared to 7 d drying at RT (6%). Spores dried for 7 d at RT and 26°C had a water content of 18.0% and 13.1%, respectively. Hence, drying at 26°C was more efficient and improved survival rates of cryopreserved spores up to a level nearly equal to that of untreated spores.

Nostoc azollae symbionts survive and cryopreservation of the symbiosis is possible without loss of viability for at least 7 months

Survival of the *N. azollae* symbionts residing within the megasporocarp cone tip is especially important as the symbionts fix nitrogen, which is of agronomic importance. Filamentous cyanobacteria were present in *Azolla* fronds recovered from cryopreserved megasporocarps (Fig. 3a) and exhibited heterocysts at equal frequency to untreated controls (Fig. 3b). In addition, all cryopreserved batches from drying optimization experiments yielded sporelings that grew on nitrogen-free medium except those batches subjected to < 4 d drying pre-treatment at RT.

In order to test whether spore viability persisted when frozen over longer periods, batches of fertilized megasporocarps were dried for 7 d at RT, then snap frozen in LN and stored at −80°C (or frozen and stored at −20°C) for up to 7 months (Fig. 3c).

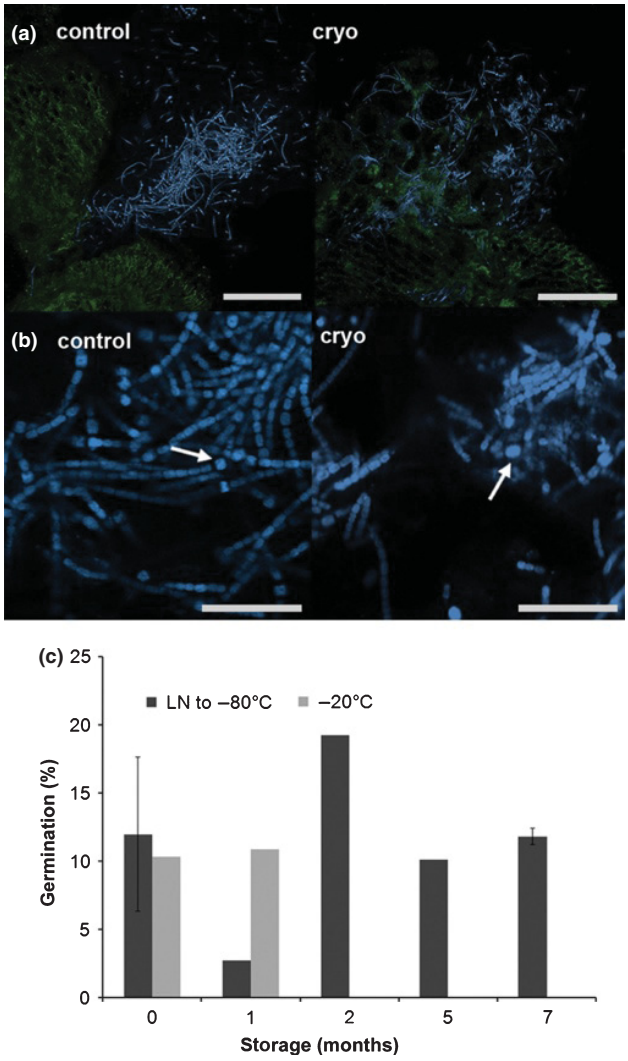


Fig. 3 Survival of the symbiont *Nostoc azollae* and long-term cryopreservation of the *Azolla* symbiosis. (a) *N. azollae* filaments in 7 wk *Azolla* fronds developing from a fresh megasporocarp (control) and a megasporocarp cryopreserved for 2 months at −80°C (cryo); bar, 0.15 mm. Confocal fluorescence microscopy with excitation 534 nm and fluorescence recorded at 680–750 nm (green) and 630–670 nm (blue). (b) Close-up of the filaments as in (a) depicting heterocysts (arrows); bar, 37.5 μm. (c) Germination of cryopreserved spores stored up to 7 months. Megasporocarps with massulae attached were dried for 7 d at room temperature (RT), batches of *c*. 200 megasporocarps were frozen in liquid nitrogen then stored at −80°C or frozen and stored at −20°C. Thawed megasporocarps were scored for germination after 5 wk. Batches stored at −20°C were not viable beyond the first month storage with zero germination after 2 and 5 months. Error bars, ± SD, $n = 3$ batches for 7 months storage.

Germination rates varied between 2.72% and 19.20% and after 7 months averaged 13%. Although a large amount of variation was observed there was no loss in viability related to the storage period over at least 7 months.

In our hands cryopreservation of neither sporophytes nor small explants was successful and therefore if the cryopreservation of *Azolla* varieties were to be contingent on the availability of spores,

New Phytologist

a method to reliably induce sporulation will need to be developed. Given the absence of reports on successful induction of sporulation, we chose to investigate molecular pathways that may control transition to sexual reproduction.

A 37 649 unigene database of *A. filiculoides* covers metabolism, cellular processes and regulatory networks extensively

A. filiculoides belongs to a particularly neglected phylogenetic group of heterosporous ferns, the *Salviniaceae*, for which no sequence resource exists. To provide for phylogenetic and molecular studies, sequencing reads were obtained independently from three differing stages of the complex lifecycle of *Azolla*: microsporocarps, sporophytes and megasporocarps. Microsporocarps did not contain contaminating *N. azollae* but were not aseptic. Sporophytes were aseptically grown but contained contaminating *N. azollae*. Megasporocarps, central for *Azolla* reproductive biology, contained contaminating *N. azollae* and were not aseptic. Sequencing reads were cleaned and assembled into a unigene database. Because the assembly could not be benchmarked against closely related species, reads were aggressively cleaned to preclude erroneous assemblies removing two-thirds of all reads (Table S1, Read cleaning). The resulting reads were then assembled into a database comparable with previous assemblies (Notes S1; Brautigam & Gowik, 2010). Annotation against *A. thaliana*, *S. moellendorffii* and *N. azollae* matched two-thirds of the unigene database against the plants but less than one-fifth against the symbiont (Table S1). All but 41 unigenes matched by the symbiont also matched the plants, in most cases with better Expectation values (e-values; Table S2). mRNA purification and cDNA library synthesis thus efficiently discriminated against bacterial transcripts.

Without a close relative with a sequenced genome, the unigene database was benchmarked against core plant pathways as represented in KEGG. Unigenes in the database covered all nuclear encoded genes of both the light and dark reaction of photosynthesis, all but two genes required to synthesize all 20 amino acids, all genes required to synthesize both purine and pyrimidine nucleotides, and the genes for sulfur and ammonia assimilation (Notes S2). Starch and sucrose synthesis, the TCA cycle and glycolysis were completely covered. The synthesis and modification of fatty acids and lipids were largely represented; phenylpropanoid metabolism including lignin precursors, terpenoid synthesis including carotenoids, porphyrin and chlorophyll synthesis were completely covered. Gene groups involved in cellular maintenance such as peroxisomes, the proteasome, the ER including ER trafficking, DNA replication and repair, RNA synthesis and processing were well represented (Notes S3). The regulatory pathways of circadian rhythm, hormone perception and pathogen perception were essentially complete (Notes S4). In summary, the unigene database of *A. filiculoides* containing 37k unigenes covers central metabolism entirely, and cellular processes and regulatory networks to a large degree.

All reads were then mapped on the unigene database: 77%, 76% and 74% of reads from for megasporocarps, microsporocarps and sporophytes, respectively, matched unigenes of the database. Unigenes were labeled 'expressed' if at least one read could be detected. The majority of unigenes – 20 598 out of 37 649 – were expressed in all tissues, while only 7% at most were specific to sporophyte tissue (Fig. 4a). mRNA extracted from the differing tissues were thus reproducibly from the fern *A. filiculoides*.

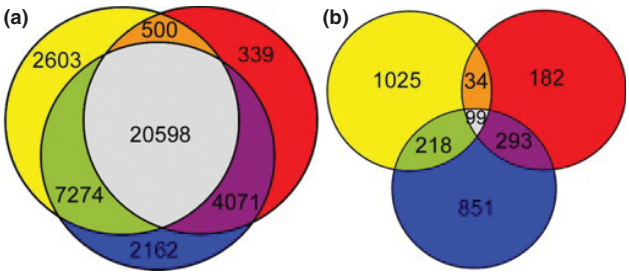


Fig. 4 Distribution of unigenes over megasporocarps, microsporocarps and sporophytes. mRNA was extracted from megasporocarps, microsporocarps and nonsporulating sporophytes, sequenced, then sequences assembled in a unigene database as described in the Materials and Methods section. 77%, 76% and 74% of reads matched unigenes and were used to assign unigene distribution over the differing tissues: microsporocarp (blue), megasporocarp (red), sporophyte (yellow). (a) Distribution taking all unigenes into account. (b) Distribution of highly expressed unigenes counting at least 100 reads per million reads.

The reproducibly high proportion of mapped reads in all tissues allowed us to test whether highly abundant reads from each tissue reflect tissue biology: when only unigenes with rpm > 100 were considered, the different tissues revealed little intersection and thus expression was characteristic (Fig. 4b). Sporophytes exhibited highly read unigenes enriched for Calvin cycle, core nitrogen metabolism, photorespiration, photosystem I and chlorophyll synthesis over other pathways (Table 3); a read signature typical of leaf tissues (Brautigam *et al.*, 2011a,b; Gowik *et al.*, 2011). By contrast, highly read megasporocarp unigenes were enriched in storage protein synthesis, in mitochondrial electron transfer, and ATPase, and in the syntheses of sterols and derivatives (Table 3). Highly read unigenes from microsporocarps were similarly enriched in mitochondrial electron transfer including ATPase and synthesis of sterol and sterol derivatives, and in addition, in proteasome, in cytosolic ribosomes as well as lipid synthesis and transfer proteins. Signatures of highly read unigenes therefore confirmed that both reproductive tissues engaged in storage compound synthesis fueled by catabolic metabolism, which was consistent with collection of the reproductive organ when still on the sporophyte to minimize microbial contamination.

Pathways leading to the onset of reproductive organ formation in Arabidopsis are present in *Azolla*

Genes controlling the onset of flowering in Arabidopsis may possibly also be involved in the transition to the reproductive phase in ferns: 165 unigenes from the *A. filiculoides* database were most similar to Arabidopsis genes involved in flowering and correspond to 64 different Arabidopsis genes (Table S2). Table S3

New Phytologist

Research 1077

Table 3 Pathways enriched in *Azolla* megasporocarps (mega), microsporocarps (micro) and sporophytes

Pathway	Unigenes present in all	Highly expressed in			Percentage highly expressed in			Fishers Exact test for enrichment in[3]		
		mega (475)[1]	micro (1144)[1]	sporophyte (1025)[1]	mega (1%)[2]	micro (3%)[2]	sporophyte (3%)[2]	mega	micro	sporophyte
Calvin cycle	37	0	2	7	0	5	19	1	3.13E-01	**5.65E-05**
Nitrogen – core assimilation	27	0	0	13	0	0	48	1	1	**6.30E-14**
Photorespiration	46	2	3	13	4	7	28	1.15E-01	1.65E-01	**1.97E-10**
Photosynthesis PS I	27	0	1	8	0	4	30	1	5.67E-01	**4.27E-07**
Chlorophyll	53	0	1	7	0	2	13	1	1	**5.79E-04**
Storage protein	26	7	12	1	27	46	4	**2.69E-08**	**4.05E-12**	5.14E-01
Lipids – sterol and derivatives	70	8	12	1	11	17	1	**2.99E-06**	**1.29E-06**	1
Mitochondrial electron transfer/ATPase	132	9	16	3	7	12	2	**5.03E-05**	**2.95E-06**	1
Lipids – general	43	2	9	0	5	21	0	1.03E-01	**4.96E-06**	6.33E-01
LTP	21	3	6	2	14	29	10	2.27E-03	**2.93E-05**	1.12E-01
Proteasome	87	2	15	5	2	17	6	3.02E-01	**5.78E-08**	9.04E-02
Ribosome cytosol	197	2	34	6	1	17	3	1	**2.78E-16**	6.63E-01

[1]Number of highly expressed unigenes in this tissue.
[2]Expected % if distribution was even.
[3]Statistically significant *P*-values are indicated in bold text. *P*-values were corrected for multiple hypothesis testing by the Bonferroni method.

scores the presence of proteins associated with control of flowering in genomes of higher and lower plants and in the transcriptomes of the ferns *Ceratopteris richardii* and *A. filiculoides*. All major proteins associated with circadian rhythm and the photoperiod flowering pathways had homologs in *Azolla* tissues. Furthermore a large proportion of proteins associated with the vernalization or autonomous pathway in Arabidopsis had homologs in *Azolla*. LFY was identified in *Azolla*, and a number of SQUAMOSA BINDING PROTEIN-like (SPL) proteins were identified in both *Azolla* and *Ceratopteris*, including SPL1, SPL2 and SPL9-like in *Azolla*. A total of nine MADS-box like proteins in *Azolla* include potential homologs of SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1) and CAULIFLOWER (CAL1).

FT was not found in lower plant genomes and possibly, the integrating role of FT might have evolved along with (lignified) vasculature in ferns. An FT-like protein is present in both *Azolla* and *Ceratopteris* (Fig. 5a). The *Azolla* FT-like protein contains the motifs D-P-D-x-P-S-P-S (Fig. 5a, Box I) and G-x-H-R (Fig. 5a, Box IV) conserved for PEBP-like proteins (Hedman *et al.*, 2009). The amino acid associated with FT/TFL switching differs in the fern FT-like proteins: fern FT-like proteins have F, while Y and H are characteristic for FT and TFL1, respectively (Fig. 5a, Box II). Furthermore MFT has a characteristic P at box VI, whereas fern proteins have a corresponding D instead. In the conserved region indicated by box I both *Azolla* and *Ceratopteris* have an A at the fourth position, similar to MFT and BROTHER OF FT (BFT), instead of a V found in FT and TFL. The *Azolla* FT-like protein was repeatedly detected in microsporocarps and megasporocarps but not in the sporophyte. qRT-PCR further confirmed high expression of *Azolla FT*-like in microsporocarps compared to sporophytes (not shown) and revealed induced expression in sporulating sporophytes compared to nonsporulating sporophytes (Fig. 5b) along with *SOC1* and *LFY* (Fig. 5c). We conclude that an FT-like protein, neither characteristically FT nor MFT, is induced as sporophytes undergo reproductive development.

Mapping the *Azolla* proteins to Arabidopsis flowering pathways reveals that whole Arabidopsis pathways towards flowering are not obviously induced in reproductive tissues of *Azolla* (Fig. 6). Nonetheless *Azolla* proteins like FT, SPL1 and CAL1 are commonly detected in the reproductive organs whilst they were not detected in the sporophyte. By contrast, *Azolla* proteins like SPL9 and SPL2 seem to be restricted to the sporophyte (Fig. 6) as is the *Azolla* protein like Class II KNOX reported to repress gametophytic development in *Physcomitrella* (Table S2; Sakakibara *et al.*, 2013).

Discussion

Methods that may be invaluable for the domestication and breeding of *Azolla* are presented using the species *A. filiculoides* as a starting point. Collecting large amounts of clean fertilized megaspores will be critical for dissemination. Fertilizing spores *in vitro* will be important for breeding. Cryopreserving *Azolla* will be essential to preserve biodiversity and store varieties particularly suited for production. Importantly, cryopreservation of an *A. filiculoides* variety now opens the way to genomic investigations with the safety of cryopreserving the plant genotype sequenced. Our first RNAseq experiment showed that *Azolla* mRNA was mostly devoid of contaminating *N. azollae* RNA and generated a 37 649 unigene database that extensively covered plant metabolism, cellular processes and regulatory networks. Networks controlling the transition to reproductive development

New Phytologist

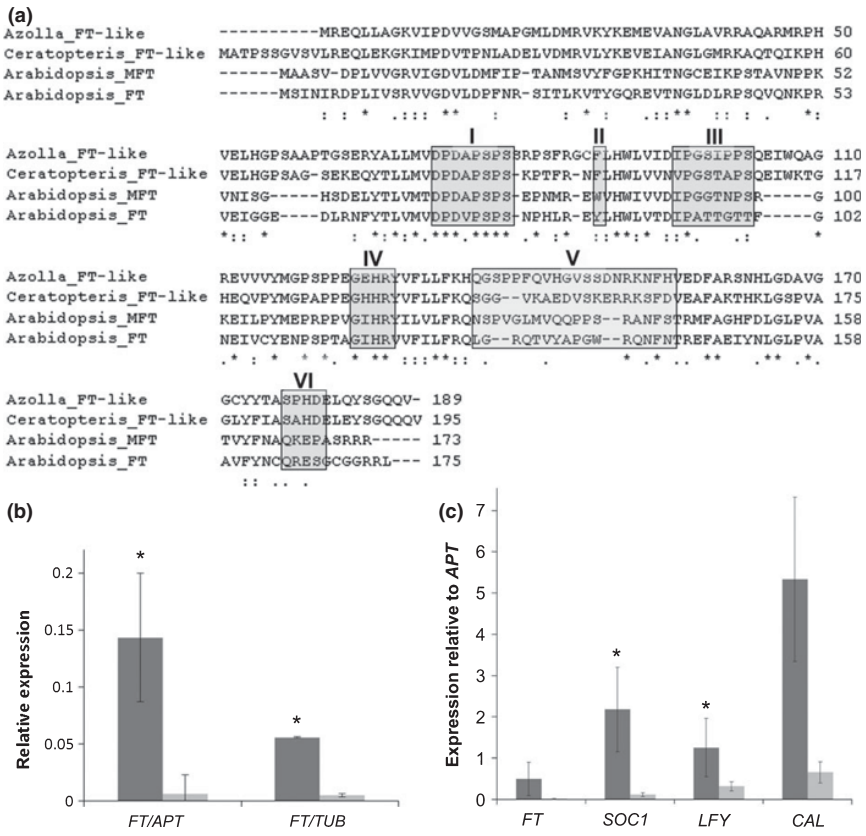


Fig. 5 *Azolla* FT-like. (a) Alignment of fern FT-like proteins with FT and MFT from Arabidopsis. Read IDs were extracted from the *A. filiculoides* database matching a theoretical DNA encoding the FT from *Ceratopteris* using BLAST then assembled using CAP3. The longest resulting contig covers the query FT for 94% of the sequence. Features are boxed: I, PEBP conserved D-P-D-x-P-S-P-S motiv; II, His/Tyr residues involved in FT/TFL switching; III, conserved MFT/FT region; IV, PEBP conserved G-x-H-R motif; V, B-region; VI, box with P characteristic for MFT (Hedman *et al.*, 2009). (b) *AfFT* expression relative to *TUBULIN* or *APT* in sporophytes sporulating (dark gray) from the wild or nonsporulating (light gray) from the growth cabinet. qRT-PCR was as described in the Materials and Methods section, $n = 3$, ± SD. (c) Expression of *AfFT*, *AfSOC1*, *AfLFY* and *AfCAL* relative to *APT* in sporophytes as in (b), $n = 3$, ± SD. *, $P < 0.05$.

in higher plants were present in *Azolla* and included an FT-like protein induced in sporulating *Azolla*.

In order to establish the above methods, we supplemented existing documentation on *Azolla* germination from Becking (1987), Peters & Meeks (1989), Wagner (1997) and Carrapiço (2010). Our conditions for *in vitro* fertilization and germination of the *A. filiculoides* megaspores resembled those described for *Marsilea vestita*, another heterosporous fern from the *Marsileaceae*, a family related to the *Salviniaceae* (Mahlberg & Yarus, 1977), although the time span required for *Azolla* germination is much greater.

Cryopreservation in ferns was achieved previously and is generally based on the desiccation and frost tolerance of the spores (Pence, 2008). Janes (1998) reported successful storage of *A. filiculoides* spores by freezing, in tap water, at −10°C, for at least 19 d. We were unable to reproduce this with material collected in the Netherlands, but this may be due to our material from plants being grown in the relatively protected environment of the glasshouse, where they were not exposed to stressful conditions. A different method for preserving *Azolla* varieties, not relying on the spores, consists of keeping a stem-tip culture of an *Azolla* frond under sterile conditions at low temperature (0–10°C), allowing preservation up to 12 months (Xu *et al.*, 2011). Preserving stem-tip cultures is laborious and precludes scaling up of the procedure. Cryopreservation is potentially more reliable and our results for spores subjected to a controlled drying pre-treatment open the way to long-term high efficiency preservation of varieties from *A. filiculoides*. Whether cryopreservation of spores from other species of *Azolla* will be similarly successful will require further investigation.

Because spores are the natural dissemination form of ferns, the control of spore induction for reliable and mass scale production of dissemination stages of any fern variety is a prerequisite to its domestication. However, very little knowledge on fern spore induction was available as research on ferns lags far behind that of angiosperms and even lycophytes and bryophytes (Muthukumar *et al.*, 2013). There is as yet no single fern genome sequenced and transcriptomes have been published from only two ferns, *Ceratopteris richardii* (Bushart *et al.*, 2013) and *Pteridium aquilinum* (Der *et al.*, 2011). Insight from these was limited, however, due to the evolutionary distance and specific ecological niche of the floating fern family of *Salviniaceae*.

RNA-seq from three different tissues of *Azolla*, the sporophyte and both micro- and megasporocarps, yielded a unigene database. Analysis of the database showed that *Azolla* shares core metabolism and regulation with genomic model plants (Notes S2). The Unigene database completely covers primary metabolism and mostly covers cellular processes and regulatory pathways. The unigene length distribution and annotation was comparable to that produced in other sequencing efforts (Brautigam & Gowik, 2010). To test whether the unigene database is in principle suitable for future quantitative RNA-seq, the reads from the three different tissues were mapped onto the assembly. Two thirds of reads could be mapped. Considering that during

New Phytologist

Research 1079

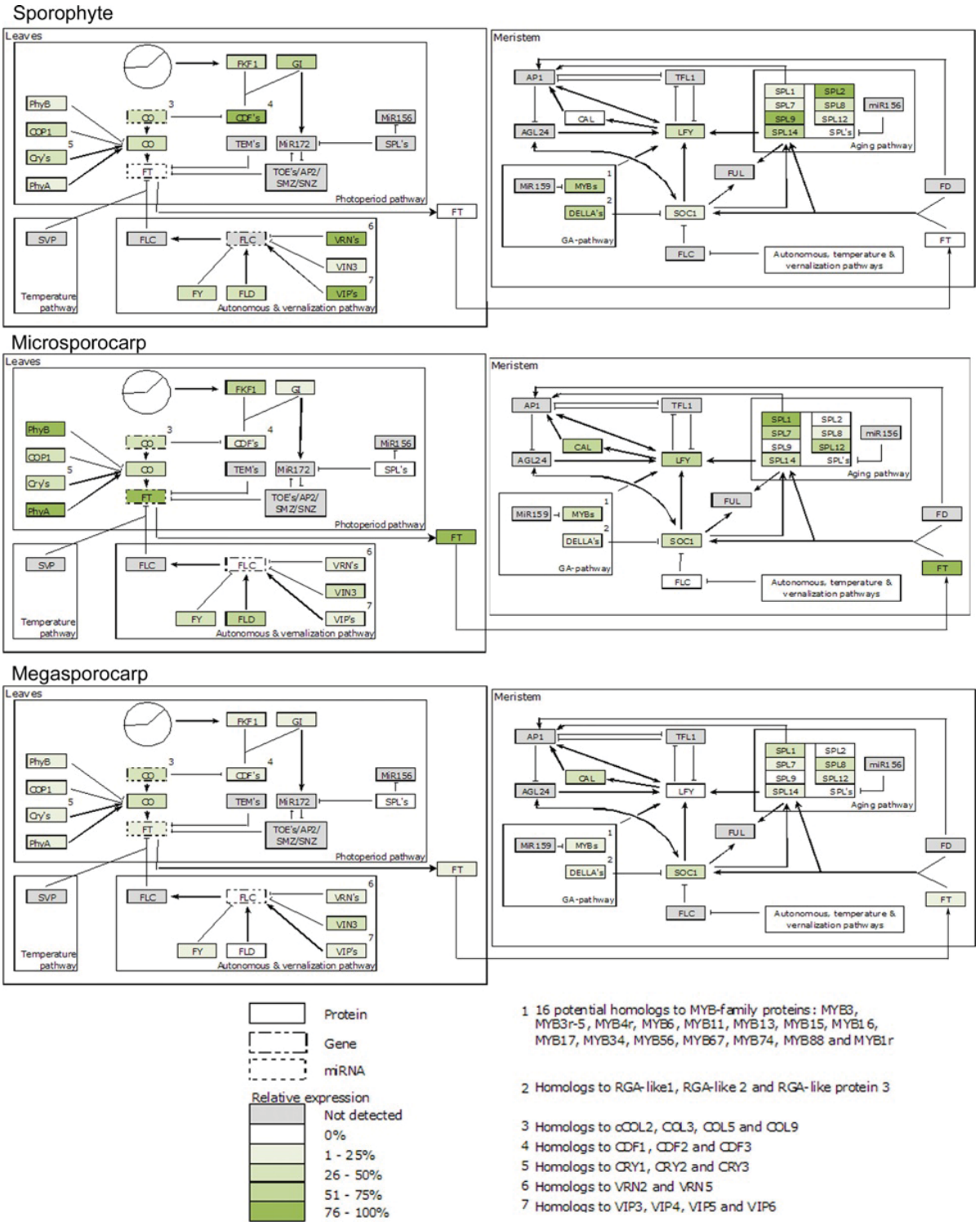


Fig. 6 Proteins from the floral induction pathways of Arabidopsis and their potential homologs in the differing tissues of Azolla: nonsporulating sporophyte, microsporocarp and megasporocarp. Layout is based on the onset of flowering in Arabidopsis with resulting transformation of the shoot apical meristem into an inflorescence meristem after Srikanth & Schmid (2011) and Jung *et al.* (2012). Presence of the unigenes in *Azolla* was extracted from Supporting Information Table S2 and normalized counts given in % over the three tissues visualized in shades of green. Gray boxes, genes not identified in the unigene database.

mapping of sequenced reads onto the corresponding genome *c.* 90% of the reads map (Hamisch *et al.*, 2012) and that during cross-species mapping between 60% (Gowik *et al.*, 2011) and 80% (Brautigam *et al.*, 2011b) of reads map on the reference sequence database, the mapping was efficient. The long read technology of Ion Torrent sequencing employed in this study thus proved to be a cost-effective way to produce a unigene database for a species without sequence resources. The low annotation frequency of the unigene database of *Azolla* (Notes S1) compared to the annotation frequencies of flowering plant RNA-seq efforts (Brautigam *et al.*, 2011b) demonstrates the evolutionary distance of *Azolla* from genomic models and indicates a large potential for new gene discovery, particularly in the light of *Azolla's* unusual secondary metabolism (Nierop *et al.*, 2011).

Assigning ontology terms of the differing pathways to the highly read unigenes indicated that read counts for the pathways reflected the biology of the tissues. Both reproductive tissues are catabolic in nature: the energy-consuming mitochondrial electron transfer chain and its ATPase were overrepresented among the highly read unigenes. The microsporocarp is very rich in lipids and in protein as judged by microscopic stains; the megasporocarp is also very rich in protein but only moderately rich in lipids (Lucas & Duckett, 1980). Microsporocarps are still in the process of maturing when attached to the plant and meiotic development has not been completed (Lucas & Duckett, 1980). Hence, the prevalence of proteasome and cytosolic ribosome components among the highly read unigenes in microsporocarps may reflect the maturation process. Meiosis as a category was not tested but selected genes involved in meiotic processes could be detected in microsporocarp mRNA, but not in the other two (Table S2).

Similar to other RNA-seq efforts, the *Azolla* transcriptome database will serve to elucidate *Azolla* biology, for example the transition to sexual reproduction and sporocarp formation. The *Azolla* unigene database contained candidate homologs of many genes controlling sexual transition in Arabidopsis. Whether these actually have a function in regulating sporulation in *Azolla*, however, is uncertain as many of the flowering-related genes have been associated with multiple developmental functions in seed plants. Several flowering-related genes were only identified because multiple tissues were included in the initial sequencing: *Azolla* proteins like FT, CAL and SPL1 were reproducibly read in the *Azolla* reproductive tissues but not in the sporophytes. By contrast, *Azolla* proteins like SPL9 and SPL2 were reproduciby read in the sporophyte whilst absent in the reproductive organs. *Azolla* FT-like expression was confirmed by qRT-PCR to be induced in sporulating as opposed to nonsporulating sporophytes. A crucial future step will be to attempt transformation of *Azolla* to test the function of genes identified in the unigene database. Only very recently, Muthukumar *et al.* (2013) reported successful transformation of the ferns *P. vittata* and *C. thalictroides* using spores as the transformation targets.

In conclusion, we present for the first time methods for storage and dissemination, as well as an annotated database of genes that may contribute to the domestication of *Azolla*, a candidate crop that is highly productive, reaching 40 tons ha^{-1} yr^{-1} dry weight yield (AzoFaSt Performance Report, 2013), requiring no nitrogen fertilizer and growing in areas not previously used as arable land. The domestication of a fern/cyanobacteria symbiosis may seem like a daunting task, but we feel that the time is ripe. In addition, results generated will serve more widely to access the wealth of biochemicals in biomass hidden within the pteridophytes for the bio-economy of the future.

Acknowledgements

We thank Ronald Leito and Frits Kindt of the Biology Imaging Center, Utrecht University, for help with microscopy. We acknowledge expert technical support by BMFZ, Heinrich Heine University, Düsseldorf. Funding was from the AzoFaSt Pathfinder, Climate-KIC, European Institute of Technology, and the Deutsche Forschungsgemeinschaft (EXC 1028). We further thank anonymous reviewers for helpful comments on the manuscript.

References

Abdel-Tawwab M. 2008. The preference of the omnivorous-macrophagous, *Tilapia zillii* (Gervais), to consume a natural free-floating Fern, *Azolla pinnata*. *Journal of the World Aquaculture Society* **39**: 104–112.

Alalade OA, Iyayi EA. 2006. Chemical composition and the feeding value of Azolla (*Azolla pinnata*) meal for egg-type chicks. *International Journal of Poultry Science* **5**: 137–141.

Antunes APM, Watkins GM, Duncan JR. 2001. Batch studies on the removal of gold(III) from aqueous solution by *Azolla filiculoides*. *Biotechnology Letters* **23**: 249–251.

AzoFaSt Performance Report. 2013. Van der Werf A, ed. *Project performance report pathfinder project "AzoFaSt".* [WWW document] URL http://www.uu.nl/faculty/science/EN/contact/Researchinstitutes/IOEB/research/groups/mpp/research/Pages/Trehalose.aspx [accessed 1 September 2013].

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA *et al.* 2011. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**: 960–963.

Barnicoat H, Cripps R, Kendon J, Sarasan V. 2011. Conservation *in vitro* of rare and threatened ferns - case studies of biodiversity hotspot and island species. *In Vitro Cellular and Developmental Biology – Plant* **47**: 37–45.

Becking JH. 1987. Endophyte transmission and activity in the Anabaena-Azolla association. *Plant and Soil* **100**: 183–212.

Becking JH, Donze M. 1981. Pigment distribution and nitrogen fixation in *Anabaena azollae*. *Plant and Soil* **61**: 203–226.

Brautigam A, Gowik U. 2010. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biology (Stuttgart, Germany)* **12**: 831–841.

Brautigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Mass J, Lercher MJ *et al.* 2011a. An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiology* **155**: 142–156.

Brautigam A, Mullick T, Schliesky S, Weber AP. 2011b. Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C(3) and C(4) species. *Journal of Experimental Botany* **62**: 3093–3102.

Brown NJ, Newell CA, Stanley S, Chen JE, Perrin AJ, Kajala K, Hibberd JM. 2011. Independent and parallel recruitment of preexisting mechanisms underlying C(4) photosynthesis. *Science* **331**: 1436–1439.

Bushart TJ, Cannon AE, ul Haque A, San Miguel P, Mostajeran K, Clark GB, Marshall Porterfield D, Roux SJ. 2013. RNA-Seq analysis identifies potential modulators of gravity response in spores of *Ceratopteris* (*Parkeriaceae*): evidence

for modulation by calcium pumps and apyrase activity. *American Journal of Botany* **100**: 161–174.

Carrapiço F. 2010. *Azolla* as a superorganism. Its implications in symbiotic studies. In: Seckbach J, Grube M, eds. *Symbioses and stress: joint ventures in biology. Cellular origin, life in extreme habitats and astrobiology*, **17**. Dordrecht, the Netherlands: Springer, 227–241.

Costa ML, Santos MC, Carrapiço F. 1999. Biomass characterization of *Azolla filiculoides* grown in natural ecosystems and wastewater. *Hydrobiologia* **415**: 323–327.

Der JP, Barker MS, Wickett NJ, dePamphilis CW, Wolf PG. 2011. *De novo* characterization of the gametophyte transcriptome in bracken fern. Pteridium aquilinum. *BMC Genomics* **12**: 99.

EPSO. 2005. European plant science: a field of opportunities. *Journal of Experimental Botany* **56**: 1699–1709.

Fasakin EA. 1999. Nutrient quality of leaf protein concentrates produced from water fern (*Azolla africana* Desv) and duckweed (*Spirodela polyrrhiza* L. Schleiden). *Bioresource Technology* **69**: 185–187.

Fisher RA. 1922. On the interpretation of χ^2 from contingency tables, and the calculation of *P. Journal of the Royal Statistical Society* **85**: 87–94.

Floyd KS, Bowman JL. 2007. The ancestral developmental tool kit of land plants. *International Journal of Plant Sciences* **168**: 1–35.

Gowik U, Brautigam A, Weber KL, Weber AP, Westhoff P. 2011. Evolution of C_4 photosynthesis in the genus *Flaveria*: how many and which genes does it take to make C4? *Plant Cell* **23**: 2087–2105.

Hamisch D, Randewig D, Schliesky S, Brautigam A, Weber APM, Geffers R, Herschbach C, Rennenberg H, Mendel RR, Hansch R. 2012. Impact of SO_2 on *Arabidopsis thaliana* transcriptome in wildtype and sulfite oxidase knockout plants analyzed by RNA deep sequencing. *New Phytologist* **196**: 1074–1085.

Hedman H, Källman T, Lagercrantz U. 2009. Early evolution of the *MFT*-like gene family in plants. *Plant Molecular Biology* **70**: 359–369.

Herd YR, Cutter EG, Watanabe I. 1985. A light and electron microscopic study of microsporogenesis in *Azolla microphylla*. *Proceedings of the Royal Society of Edinburgh* **86B**: 53–58.

Herd YR, Cutter EG, Watanabe I. 1989. The effects of temperature and selected growth regulating substances on sporulation in the aquatic fern *Azolla*. *American Fern Journal* **79**: 136–142.

IPCC. 2007. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL, eds. *Climate change 2007: the physical science basis. Contribution of Working Group I to the fourth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge, UK/New York, NY, USA: Cambridge University Press.

Janes R. 1998. Growth and survival of *Azolla filiculoides* in Britain: II. Sexual reproduction. *New Phytologist* **138**: 377–384.

Jensen ES, Peoples MB, Boddey RM, Gresshoff PM, Hauggaard-Nielsen H, Alves BJR, Morrison MJ. 2012. Legumes for mitigation of climate change and the provision of feedstock for biofuels and biorefineries – a review. *Agronomy for Sustainable Development* **32**: 329–364.

Jung JH, Ju Y, Seo PJ, Lee JH, Park CM. 2012. The SOC1-SPL module integrates photoperiod and gibberellic acid signals to control flowering time in Arabidopsis. *Plant Journal* **69**: 577–588.

Kajala K, Brown NJ, Williams BP, Borrill P, Taylor LE, Hibberd JM. 2012. Multiple Arabidopsis genes primed for recruitment into C(4) photosynthesis. *Plant Journal* **69**: 47–56.

Kar PP, Mishra S, Singh DP. 2002. Azolla sporulation in response to application of some selected auxins and their combination with gibberellic acid. *Biology and Fertility of Soils* **35**: 314–319.

Kent WJ. 2002. BLAT – The BLAST-like alignment tool. *Genome Research* **12**: 656–664.

Lucas RC, Duckett JG. 1980. A cytological study of the male and female sporocarps of the heterosporous fern *Azolla filiculoides* Lam. *New Phytologist* **85**: 409–418.

Mahlberg PG, Yarus S. 1977. Effects of light, pH, temperature, crowding on megaspore germination and sporophyte formation in *Marsilea*. *Journal of Experimental Botany* **28**: 1137–1146.

Maizel A, Busch MA, Tanahashi T, Perkovic J, Kato M, Hasebe M, Weigel D. 2005. The floral regulator LEAFY evolves by substitutions in the DNA binding domain. *Science* **308**: 260–263.

Meyer RS, DuVal AE, Jensen HR. 2012. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytologist* **196**: 29–48.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35**: 182–185.

Muthukumar B, Joyce B, Elless M, Stewart N. 2013. Stable transformation of ferns using spores as targets: *Pteris vittata* (Chinese brake fern) and *Ceratopteris thalictroides* (C-fern 'Express'). *Plant Physiology* **163**: 648–658.

Nierop KGJ, Speelman EN, de Leeuw JW, Reichart G-J. 2011. The omnipresent water fern *Azolla caroliniana* does not contain lignin. *Organic Geochemistry* **42**: 846–850.

Parsons JL, Wingate V. 2012. Methods and compositions for the cryopreservation of duckweed. Patent Application Publication US 2012/0190004 A1.

Pence VC. 2008. Cryopreservation of Bryophytes and Ferns. In: Reed BM, ed. *Plant cryopreservation: a practical guide*. New York, NY, USA: Springer, 117–140.

Peters GA, Meeks JC. 1989. The Azolla-Anabaena symbiosis: basic biology. *Annual Review of Plant Physiology and Plant Molecular Biology* **40**: 193–210.

Peters GA, Perkins SK. 1993. The Azolla-Anabaena symbiosis: endophyte continuity in the Azolla life-cycle is facilitated by epidermal trichomes. *New Phytologist* **123**: 65–75.

Quin-Yuan X, Yan-Ru S, Guang-Li Y, Ke-Lin P. 1987. Germination of *Azolla filiculoides* Lam. sporocarps and factors affecting their growth. In: Smith WH, Cervantes EP, eds. *Azolla utilization*, Proceedings of the Workshop on AzoIla Use Fuzhou, Fuijan, China. Los Baños, Philippines: International Rice Research Institute, 33–38.

Ran L, Larsson J, Vigil-Stenman T, Nylander JAA, Ininbergs K, Zheng WW, Lapidus A, Lowry S, Haselkorn R, Bergman B. 2010. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS ONE* **5**: 329–493.

Riesco MF, Martínez-Pastor F, Chereguini O, Robles V. 2012. Evaluation of zebrafish (*Danio rerio*) PGCs viability and DNA damage using different cryopreservation protocols. *Theriogenology* **77**: 122–130.

Rigbi M, Rosinski J, Siegelman HW, Sutherland JC. 1980. Cyanobacterial phycobilisomes: selective dissociation monitored by fluorescence and circular dichroism. *Proceedings of the National Academy of Sciences, USA* **77**: 1961–1965.

Rogge GD, Viana AM, Randi AM. 2000. Cryopreservation of spores of *Dicksonia sellowiana*: an endangered tree fern indigenous to South and Central America. *Cryo-Letters* **21**: 223–230.

Sakakibara K, Ando S, Yip HK, Tamada Y, Hiwatashi Y, Murata T, Deguchi H, Hasebe M, Bowman JL. 2013. *KNOX2* genes regulate the haploid-to-diploid morphological transition in land plants. *Science* **339**: 1067–1070.

Salmi ML, Roux SJ. 2008. Gene expression changes induces by spaceflight in single-cells of the fern *Ceratopteris richardii*. *Planta* **229**: 151–159.

Schliesky S, Gowik U, Weber AP, Brautigam A. 2012. RNA-Seq assembly – are we there yet? *Frontiers in Plant Science* **3**: 220.

Shuying L. 1987. Method for using *Azolla filiculoides* sporocarps to culture sporophytes in the field. In: Smith WH, Cervantes EP, eds. *Azolla utilization*. Proceedings of the Workshop on AzoIla Use Fuzhou, Fuijan, China. Los Baños, Philippines: International Rice Research Institute, 27–32.

Smith KA, McTaggart IP, Tsuruta H. 1997. Emissions of N_2O and NO associated with nitrogen fertilization in intensive agriculture, and the potential for mitigation. *Soil Use and Management* **13**: 296–304.

Sommer M, Brautigam A, Weber AP. 2012. The dicotyledonous NAD malic enzyme C_4 plant *Cleome gynandra* displays age-dependent plasticity of C_4 decarboxylation biochemistry. *Plant Biology (Stuttgart, Germany)* **14**: 621–629.

Srikanth A, Schmid M. 2011. Regulation of flowering time: all roads lead to Rome. *Cellular and Molecular Life Sciences* **68**: 2013–2037.

Toia RE Jr, Buzby KM, Peters GA. 1987. Sporocarps of *Azolla mexicana* Presl. I. isolation and purification of megasporocarps and microsporangia. *New Phytologist* **106**: 271–279.

Usadel B, Nagel A, Steinhauser D, Gibon Y, Blasing OE, Redestig H, Sreenivasulu N, Krall L, Hannah MA, Poree F *et al.* 2006. PageMan: an

interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. *BMC Bioinformatics* 7: 8.

Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M. 2009. A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant, Cell & Environment* **32**: 1211–1229.

Wagner GM. 1997. *Azolla*: a review of its biology and utilization. *Botanical Review* **63**: 1–26.

Watanabe I, Roger PA, Ladha JK, van Hove C. 1992. *Biofertilizer germplasm collections at IRRI*. Los Baños, Philippines: International Rice Research Institute.

Willemse MTM. 2009. Evolution of plant reproduction: from fusion and dispersal to interaction and communication. *Chinese Science Bulletin* **54**: 2390–2403.

Xu G, Zheng X, Ye H, Wang J, Lin Y. 2011. Method for preserving *Azolla* variety. Patent CN102119659 A.

Zheng W, Bergman B, Chen B, Zheng S, Xiang G, Rasmussen U. 2009. Cellular responses in the cyanobacterial symbiont during its vertical transfer between plant generations in the *Azolla microphylla* symbiosis. *New Phytologist* **181**: 53–61.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Summaries RNAseq and assembly

Table S2 *Azolla filiculoides* unigene database annotation

Table S3 Genes associated with reproductive phase transition in genomes of higher and lower plants and in transcriptomes of ferns

Notes S1 *A. filiculoides* unigene database.

Notes S2 *A. filiculoides* unigene database proteins in metabolism.

Notes S3 *A. filiculoides* unigene database proteins in cellular processes.

Notes S4 *A. filiculoides* unigene database proteins in regulatory pathways.

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

VIII.5 CO-AUTHORED MANUSCRIPTS

Manuscript 8

Co-expression analysis as tool for the discovery of transport proteins in photorespiration

Christian Bordych, Marion Eisenhut, Thea R. Pick, ***Canan Külahoglu**** *and Andreas P.M. Weber*

Published in Plant Biology (2013) 15, pp. 686–693
doi:10.1111/plb.12027

Impact Factor: 2.40

**Co-author*

Main findings:

This method review is evaluating the power of co-expression analysis for revealing unknown components using the example of the photorespiratory pathway in plants. The article explains the principle of co-expression analysis and shows the application of different tools for identifying new candidate genes of photorespiratory pathway. A comparison of the results of ATTED based co-expression analysis with the method of Weighted Gene Correlation Network Analysis (WGCNA) revealed that both methods yield too many putative candidates, which should be limited by additional filtering. Different strategies for data set filtering are described in the review.

Contributions:

- Performed Weighted Gene Co-Expression Analyses

plant biology

Plant Biology ISSN 1435-8603

REVIEW ARTICLE

Co-expression analysis as tool for the discovery of transport proteins in photorespiration

C. Bordych, M. Eisenhut, T. R. Pick, C. Kuelahoglu & A. P. M. Weber

Institute of Plant Biochemistry, Center of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, Düsseldorf, Germany

Keywords
Arabidopsis; co-expression; photorespiration; transporters.

Correspondence
A. P. M. Weber, Institute of Plant Biochemistry, Center of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, Universitätsstraße 1, D-40225 Düsseldorf, Germany.
E-mail: andreas.weber@uni-duesseldorf.de

Editor
H. Rennenberg

Received: 17 December 2012; Accepted: 25 February 2013

doi:10.1111/plb.12027

ABSTRACT

Shedding light on yet uncharacterised components of photorespiration, such as transport processes required for the function of this pathway, is a prerequisite for manipulating photorespiratory fluxes and hence for decreasing photorespiratory energy loss. The ability of forward genetic screens to identify missing links is apparently limited, as indicated by the fact that little progress has been made with this approach during the past decade. The availability of large amounts of gene expression data and the growing power of bioinformatics, paired with availability of computational resources, opens new avenues to discover proteins involved in transport of photorespiratory intermediates. Co-expression analysis is a tool that compares gene expression data under hundreds of different conditions, trying to find groups of genes that show similar expression patterns across many different conditions. Genes encoding proteins that are involved in the same process are expected to be simultaneously expressed in time and space. Thus, co-expression data can aid in the discovery of novel players in a pathway, such as the transport proteins required for facilitating the transfer of intermediates between compartments during photorespiration. We here review the principles of co-expression analysis and show how this tool can be used for identification of candidate genes encoding photorespiratory transporters.

INTRODUCTION

Starting in 1978, several key enzymes of photorespiratory metabolism (reviewed in Timm & Bauwe, this issue) were identified by Chris and Shauna Somerville and Bill Ogren, utilising a forward genetic approach (reviewed in Somerville 2001). Shifting mutagenised *Arabidopsis thaliana* populations from elevated CO_2 concentrations (1%) to atmospheric conditions (0.03%) revealed mutants, which were not able to cope with ambient CO_2 concentrations. By labelling photosynthetic products with $^{14}CO_2$ and subsequent enzymatic tests, five individual players in the photorespiratory C_2 cycle were discovered, namely phosphoglycolate phosphatase (Somerville & Ogren 1979), serine:glyoxylate aminotransferase (Somerville & Ogren 1980b), glutamate synthase (Somerville & Ogren 1980a), serine hydroxymethyltransferase (Somerville & Ogren 1981) and glycine decarboxylase (Somerville & Ogren 1982). Based on this work, a significant part of the photorespiratory cycle was deciphered at the biochemical level. Since the detection of glycerate kinase (Boldt *et al.* 2005), all key enzymes of the core C_2 cycle have now been identified at the genetic level in the model plant *A. thaliana* and have been characterised biochemically.

The C_2 cycle is a highly compartmentalised pathway localised to chloroplasts, peroxisomes, mitochondria and the cytosol. Thus, some intermediates of the cycle need to be transported out of a compartment and re-imported into another one *via* transport proteins for further reactions to occur (for review see Eisenhut *et al.* 2012). A large number of these transport processes in photorespiration between chloroplasts, peroxisomes and mitochondria remain uncharacterised, even more than 30 years after discovery of the first enzyme involved in photorespiration. Forward genetic screens have, to date, contributed little to closing the gaps in knowledge about the transport of photorespiratory intermediates between compartments.

In this article, we review co-expression analysis and explore it as a novel approach to discover these missing players in photorespiration. We demonstrate that co-expression analysis can serve as a candidate discovery tool for reverse genetic approaches to identify as yet unknown proteins involved in transport process during photorespiration.

IDENTIFICATION OF CANDIDATE GENES THROUGH CO-EXPRESSION ANALYSIS

In order to discover unknown transport proteins involved in photorespiratory metabolism, we chose a novel contemporary approach, the combined use of co-expression analysis and reverse genetic studies of deduced candidate genes. The subsequent workflow is shown in Fig. 1 and is explained below.

The principle idea of co-expression analysis

Within the last 15 years it became possible to measure the expression level of every single gene of an organism under various conditions in a high-throughput manner, utilising microarrays, and thereby creating vast amounts of data (Brazma & Vilo 2000). It was a major challenge to visualise, to order and to integrate these data into public available databases, which had to be developed to make efficient use

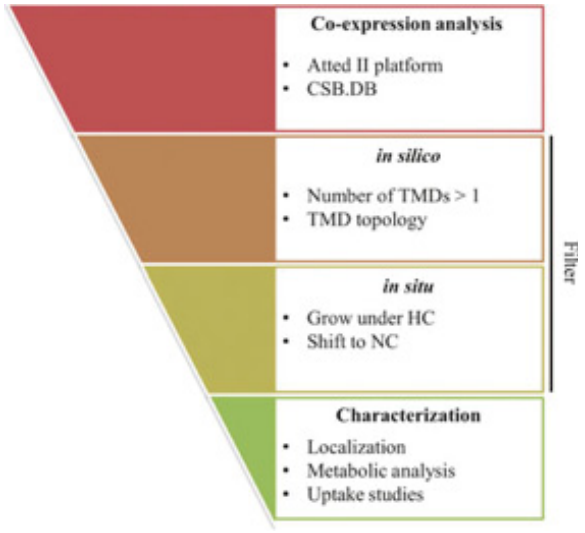


Fig. 1. Discovery of new genes involved in photorespiratory transport can be divided into four major parts; each part severely reduces the number of possible candidate genes. Co-expression analysis is carried out on the Atted-II platform. All hits are filtered afterwards for transmembrane proteins with more than one predicted transmembrane domain (TMD). Distribution of predicted TMDs can be compared with known transport proteins in the transport database (TransportDB; www.membranetransport.org). Finally, T-DNA lines are screened for a photorespiratory phenotype; mutants showing a positive result are subsequently characterised for the gene function.

of these resources. Exploratory multivariate statistical tools, such as hierarchical clustering, became the technique of choice for analysing these large datasets to finally let the raw data speak: converting 'data into information and then information into knowledge' (Eisen *et al.* 1998). Basically, in these approaches a similarity score is assigned to each possible gene pair, comparing their expression states under many different conditions. On the basis of these scores a distance matrix is generated, and in the case of clustering approaches, a dendrogram is produced in which the genes showing the most similar expression patterns are located next to each other (Eisen *et al.* 1998).

Clustered genes within a dendrogram are thought likely to participate in the same process (*e.g.,* a metabolic pathway). As proof of principle, Persson *et al.* (2005) provided evidence for a functional relationship between co-expressed genes involved in cellulose synthesis during secondary cell wall formation. At least three distinct cellulose synthase genes (*CESA4*, *CESA7* and *CESA8*) are required for cellulose synthesis during secondary cell wall formation. Knocking-out each one of these genes caused, in each case, a defect in cell wall formation, indicating that these genes code for subunits that build up a cellulose synthase complex. All three proteins have been shown to be present at the same time in the same tissue in *Arabidopsis*, pointing to orchestrated gene expression of corresponding genes (Taylor *et al.* 2003). Based on this observation Persson *et al.* (2005) hypothesised that additional partners involved in secondary cell wall synthesis are also co-expressed with these known genes. To identify some of these genes, microarray data were screened for further co-expressed genes. The study was based solely on expression data from Affymetrix ATH1 microarrays to avoid distortions caused by data from different microarray platforms. To exclude those chips representing potential outlier data, quality control was performed prior to clustering. Basically, data from each chip were tested to fit a model derived from data of all remaining chips in the set ('Deleted Residuals'). Using this 'quality control', 95 out of 503 available Affymetrix chips were rejected for future studies.

Results from subsequent clustering showed that *CESA* genes, which are involved in primary and secondary cell wall formation, 'exhibit a high degree of co-expression based on correlation coefficients' (Persson *et al.* 2005). This observation is consistent with the finding that all corresponding proteins are expressed simultaneously in a time- and space-dependent manner (Taylor *et al.* 2003), and serves as a prerequisite for further analyses. *CESA4, 7* and *8* were top of the list of co-regulated genes, with the three *CESA* genes 'confirming tight co-regulation of the three CESA subunits' (Persson *et al.* 2005). Four different genes with no functional annotation from the list of the 40 most co-regulated genes of *CESA4, 7* and *8* were investigated for their function in secondary cell wall synthesis. *Arabidopsis* knockout mutants for two of these genes showed a phenotype resembling a defect in secondary cellulose synthesis, indicating a functional relationship between the proteins encoded by the co-expressed genes (Persson *et al.* 2005).

Yonekura-Sakakibara *et al.* (2007) took advantage of co-expression analysis to identify a gene involved in specific glycosylation of flavonol in *Arabidopsis* out of 107 candidate genes from a multigene family. Investigation of glycosylation patterns of flavonoid structures demanded five additional genes be involved in glycosylation in *Arabidopsis*, besides the four already known genes. Correlation between known structural and regulatory genes involved in flavonoid biosynthesis in *Arabidopsis* and all 107 members of the multigene family were analysed. Showing that the already identified and functionally characterised genes of flavonoid biosynthesis cluster together, these authors proved that the functional relationship of these genes is mirrored by their expression state. For five out of the initial 107 candidate genes a significant positive correlation with the bait genes was observed. Two independent knockout mutants for one of these genes were analysed, demonstrating a deficiency in synthesis of a specific flavonol. This phenotype could be rescued by re-introduction of the intact gene (Yonekura-Sakakibara *et al.* 2007). The combination of co-expression analysis, metabolic profiling and reverse genetics was shown to 'be a versatile tool for functional identification of genes that belong to a multigene family and to complete the model of a particular metabolic pathway in *Arabidopsis*' (Yonekura-Sakakibara *et al.* 2007).

These and a number of other studies (*e.g.,* Stuart *et al.* 2003; Lee *et al.* 2004) demonstrate the power of co-expression analysis as a candidate discovery tool, which encouraged us to explore this approach for the identification of yet unknown genes putatively involved in the transport of photorespiratory intermediates between the compartments. This idea gained further support from a study that demonstrated that genes for peroxisomal enzymes involved in photorespiration are co-expressed (Reumann & Weber 2006), which was later confirmed through mutant analyses, demonstrating a role of peroxisomal malate dehydrogenases in photorespiration (Cousins *et al.* 2008).

Application of co-expression analysis tools

During the last few years several tools have been developed to make the results of co-expression analyses publicly available (for a detailed review on popular web tools and databases see Aoki *et al.* 2007). However, for efficiently using such web tools to connect yet unknown genes with functionally annotated ones, one should be aware of all features offered by the web tool. The following paragraphs focus on different features and their consequences for co-expression results.

To measure co-expression between two genes, one has to apply a measure for co-expression for each gene pair. Most online tools use Pearson's correlation coefficient (PCC). PCC values range between −1 (anti-correlated, gene A is up-regulated when gene B is down-regulated in many different experiments/conditions) and +1 (correlated, increased expression of gene A goes together with increased expression of gene B over many samples). PCC has been shown to be prone to outliers, rendering gene pairs correlated instead of non-correlated. To circumvent this effect, several tools utilise Spearman's correlation coefficient (SCC), which does not use the expression values of each gene directly for calculation, but the ranks which are assigned to each gene under each condition beforehand (Usadel *et al.* 2009). In the case of the web tool Atted, mutual rank (MR) is used as a measure of co-expression, which resembles the geometric mean of the correlation ranks (which are based on PCC) between two genes (Obayashi & Kinoshita 2009). These authors state that a functional relationship between genes is not always mirrored by a high PCC, thus true relationships sometimes cannot be resolved. To circumvent this problem, a new correlation measurement was introduced. MR puts emphasis on the fact that the correlation rank for the relationship of gene A to gene B differs from the rank for gene B to gene A, and represents the geometric mean of both rankings (Obayashi & Kinoshita 2009).

Several co-expression web tools (*e.g.,* Atted) accept more than one query gene. Co-expression analysis with groups of bait genes results in a list of genes that are co-expressed with the entire group of queries. This option turned out to be useful if it is not known which query serves as the best, leading to more reliable candidate lists of co-expressed genes (Usadel *et al.* 2009). Further, a multi-query run can confirm results for runs with a single query to provide additional support. Since co-expression analysis using web tools with pre-calculated co-expression scores is a very rapid method, this additional step with groups of query genes can easily be included in the analysis, when offered by the platform.

The most available online tools include the option to preselect conditions (condition-dependent) under which co-expression should be investigated. Thereby, co-expression analysis can be restricted to a defined set of microarrays (thus conditions or experiments) (Obayashi *et al.* 2011). This feature is very useful, because one can limit the search to single tissues or special conditions to further improve the quality of the results. Some popular online tools add the option of individualising conditions by allowing a user-specified set or even upload data that are not publicly available. Nonetheless, a condition-independent approach (using all available microarray data) can still prove very fruitful in finding overall gene relationships or obtaining an initial overview (Usadel *et al.* 2009). Furthermore, a condition-independent approach combined with a condition-dependent approach has been reported to be effective in discovering new interaction partners (Hirai *et al.* 2007).

The number of online tools for co-expression analysis is still increasing, so that only a limited overview of their features can be discussed here. However, running co-expression analysis *via* an online platform requires profound knowledge about all given features in order to select the right platform, subsequently the appropriate features. Usadel *et al.* (2009) have published a clear and detailed article about different online tools dealing with co-expression.

Co-expression results

Once the right parameters for co-expression analysis are set up, one can obtain most co-expressed genes within a few minutes. However, the main work is still to probe the list of genes for promising candidate genes, thus further manual filtering using *a priori* knowledge is required. As an example for such a filtering approach, a common workflow from our lab to identify candidate genes coding for photorespiratory transporters is described here.

Initial co-expression analysis was carried out on the Atted-II platform (atted.jp; Obayashi *et al.* 2007) using 13 genes coding for soluble proteins involved in photorespiratory metabolism (Table 1), resulting in a list of 300 ranked co-expressed genes per query gene (in total 3,900 single hits). As previously shown,

Table 1. Co-expression of photorespiratory genes. Co-expression analysis was carried out on the Atted platform (www.atted.jp) using each AGI of column 1. The resulting TOP300 co-expressed genes were screened for photorespiratory genes from column 1. Column 'TOP300' shows how often each photorespiratory gene is present, while 'TOP50' presents how frequent the gene is among the first 50 ranked genes.

locus	description	TOP300	TOP50
At1 g11860	Glycine decarboxylase T protein (GDT1)	7	3
At1 g23310	Glutamate:Glyoxylate aminotransferase 2 (GGT2)	7	5
At1 g32470	Glycine decarboxylase H protein 3 (GDH3)	9	4
At1 g48030	Glycine decarboxylase L protein (mLPD1)	8	5
At1 g68010	Hydroxypyruvate reductase 1 (HPR1)	7	4
At2 g26080	Glycine decarboxylase P protein 2 (GDP2)	6	2
At2 g35370	Glycine decarboxylase H protein 1 (GDH1)	8	4
At4 g33010	Glycine decarboxylase P protein 1 (GDP1)	10	5
At4 g37930	Serine hydroxymethyltransferase (SHM1)	8	4
At5 g04140	Fd-dependent glutamate synthase (Fd-GOGAT)	12	6
At5 g12860	Dicarboxylate transporter 1 (DiT1)	5	3
At5 g35630	Glutamine synthase 2 (GS2)	11	5
At5 g64290	Dicarboxylate transporter 2.1 (DiT2.1)	4	2

genes for photorespiratory enzymes in peroxisomes are simultaneously expressed in a variety of experiments and form a cluster (Reumann & Weber 2006). This finding was already reflected in the initial results obtained from the Atted-II platform, as many of the genes coding for these enzymes were frequently placed among the TOP300 genes co-expressed for each query, while most of these hits were ranked within the TOP50 of all hits (Table 1).

Filtering

Subsequent filtering of co-expression analysis results strongly depends on the initial question asked. In this example, we were looking for additional players in photorespiration, specifically for transport proteins. In order to restrict the list of genes co-expressed with known genes of photorespiration, we applied different filters to increase the probability of identifying genes coding for transporters. The filtering procedure did not aim at generating a complete list of all photorespiratory transporters already present in the original list generated through co-expression analysis, but rather an essential list of the most promising candidate genes that should be analysed in detail through follow-up experiments, such as reverse genetic analyses.

Criterion 1: Prediction of transmembrane domains (in silico)
First, a list of membrane proteins was generated through discarding all genes for which the encoded protein had been experimentally proven to be soluble. The online platform ARAMEMNON (http://aramemnon.botanik.uni-koeln.de; Schwacke *et al.* 2007) served as the basis for this analysis. ARAMEMNON merges data from up to 17 individual prediction programs, and generates a consensus prediction for transmembrane domains (Schwacke *et al.* 2003). Further support is added from experimental data (if available) *via* linking to corresponding publications. Second, genes encoding proteins with less than two predicted transmembrane domains (TMDs) are discarded. Single TMDs often render proteins membrane-bound, rather than enabling them to transport metabolites or cofactors. It cannot be ruled out that some genuine transport proteins were lost in this step, but the list of putative candidates was considerably shortened, which therefore made this step reasonable. Third, focus centred on all proteins showing typical distributions of TMDs, as frequently found in known transport proteins. While there is no single common pattern known for the distribution of TMDs along the entire peptide chain, the transporter database (www.
membranetransport.org; Ren *et al.* 2004, 2007) provides a good overview of how TMDs are distributed in proteins that have been characterised as transporters. ARAMEMNON visualises data from transmembrane helice prediction programs and generates a consensus prediction (TmConsens), including the number of predicted TMDs and the amino acid sequences that contribute to the estimated transmembrane helices. For additional support, the subcellular proteomic database (SUBA; http://suba.plantenergy.uwd.edu.au) is linked and gives hydropathy plots (Heazlewood *et al.* 2007) for comparison with the obtained TmConsens data.

ARAMEMNON is not only helpful in terms of identification of transmembrane proteins, but also in discovering genes that are members of multi-gene families. We decided to exclude genes that are members of larger families because of possible redundant functions of the gene family members, which would complicate reverse genetic analyses. Clearly, this filter criterion might exclude genes that play a role in the transport of photorespiratory intermediates; however, it helps focus on single-copy genes whose effect on photorespiration is easier to study using reverse genetic screens. That is, multi-gene family members were given lower priority for experimental testing, but were not excluded from the screen.

Criterion 2: Photorespiratory phenotype (in vivo)
The second step of filtering was to check public seed stocks for availability of corresponding T-DNA lines. Today, *Arabidopsis* T-DNA insertion lines are available for most known *Arabidopsis* genes, and can be ordered from the stock centres: NASC (Nottingham Arabidopsis Stock Centre, Nottingham University, UK) and ABRC (Arabidopsis Biological Resource Center, Ohio State University, USA). Whenever possible, at least two individual T-DNA lines were ordered to check if the knockout of a candidate gene results in a photorespiratory phenotype, as for example observed for plants deficient in serine hydroxymethyltransferase activity (Voll *et al.* 2006; Jamai *et al.* 2009). Several photorespiratory phenotypes are reviewed in this Special Issue in the paper of Timm & Bauwe (2012).

T-DNA lines were grown under high CO_2 conditions (HC; 0.3% CO_2) and selected for homozygous plants. For shifting experiments, homozygous seeds were spotted on solid MS medium, together with wild-type plants as a negative control, and grown for approximately 2 weeks under HC conditions. One half of these plant sets was shifted to normal CO_2 conditions (NC; 0.038% CO_2), while the other half remained under HC conditions as a control. Plants were checked continuously for a photorespiratory phenotype, *e.g.,* leaf chlorosis or stunted growth, for a maximum of 10 days. To investigate effects in later developmental stages, plants grown under HC conditions were transferred to soil and further grown at HC and used for shifts to NC at later stages.

In some rare cases, a homozygous knockout line was not available for analysis due to embryo lethality. However, in such cases it is possible to check the candidate by applying the miRNA (Bartel 2004) approach, *e.g.,* through generation of artificial miRNA (amiRNA) lines (Schwab *et al.* 2006). A common approach in our lab is to repress the transcript amount of the corresponding genes with amiRNA under control of either a constitutive promoter or an inducible one. Silencing in the constitutive lines can reveal a phenotype in photorespiratory shifting experiments as long as plants are viable. Controlling the amiRNA expression level using an inducible promoter is particularly helpful in cases where repression from a constitutive promoter fails, *e.g.,* if constitutive down-regulation of the candidate gene prevents establishment of viable seedlings.

SCREENING OF CANDIDATE GENES FOR POSSIBLE FUNCTION IN PHOTORESPIRATION

Arabidopsis T-DNA lines adversely affected in NC conditions, having pale green or yellow leaves, were further characterised. The first step in the detailed characterisation of these selected candidates was localisation of the corresponding protein inside the plant cell. It was necessary to verify that this protein is indeed targeted to the membrane of the predicted organelle for it to function there as transporter.

In a second step, the metabolic response of mutants in the candidate genes to a shift from HC to NC conditions was studied. Levels of metabolites associated with photorespiration were measured using gas chromatography–mass spectrometry (GC/MS) to test for altered levels of specific metabolites between mutant and wild-type plants. Accumulation of a specific metabolite might indicate a deficiency in transport of this metabolite across an organellar membrane. This hypothesis could then be tested through recombinant expression of the candidate protein in *Escherichia coli*, *Saccharomyces cerevisiae* or a cell-free expression system (Nozawa *et al.* 2007; Haferkamp & Linka 2012), followed by reconstitution into proteoliposomes and radiolabelled flux assay (for review see Haferkamp & Linka 2012). In some cases, heterologous expression of proteins and reconstitution in artificial liposome systems fail, so that the transported compound or corresponding transport kinetics remain elusive. In such cases, transport assays with isolated organelles (Werdan & Heldt 1972) has afforded an alternative approach to assess possible impaired transport processes, through comparison of uptake rates of organelles from mutant lines with those of wild-type organelles.

Further, a search for homologue proteins in *E. coli*, *S. cerevisiae* or cyanobacteria could reveal closely related genes that facilitate transport of the same compound. Because the deletion of a single gene in these organisms is relatively easy and straightforward, mutant lines for homologous genes can be generated to investigate the impact of the mutation. Growth studies in different media supplemented with various nitrogen or carbon sources can provide information about the transported substrate.

IDENTIFICATION OF CANDIDATE GENES INVOLVED IN PHOTORESPIRATORY TRANSPORT

We employed the above protocol in the search for photorespiratory transporters and generated ten promising candidates according to our defined criteria (Table 2). We focused on these genes, which (i) are co-regulated with several genes of the photorespiratory pathway; and (ii) show more than one predicted transmembrane domain and transporter-like distribution of transmembrane domains, according to the transporter database (Ren *et al.* 2004, 2007). T-DNA lines and amiRNA lines for these candidates are currently under investigation to select for those with a photorespiratory phenotype (*in vivo* filtering).

Table 2. Genes co-expressed with soluble photorespiratory enzymes. Putative candidates after *in silico* analysis are shown. Co-expression analyses were carried out for all identified genes of photorespiration. Most abundant genes were manually filtered for the existence of transmembrane domains and transporter-like distribution of transmembrane domains. For all noted genes, a protein with more than one transmembrane domain is predicted. Description of the locus is according to the ARAMEMNON platform (http://aramemnon.botanik.uni-koeln.de/).

locus	description
At1 g32080	LrgB-like membrane protein, required for chloroplast development (AtLrgB)
At5 g59250	putative VGT subfamily sugar transporter
At3 g61870	putative membrane protein of unknown function
At1 g44920	putative membrane protein of unknown function
At2 g21960	protein of unknown function
At2 g42770	putative peroxisomal PMP22-type protein of unknown function
At4 g35760	VKOR-type thiol oxidoreductase (AtLTO1/AtVKOR-DsbA)
At1 g22850	putative vesicle transport SNARE-associated protein
At1 g49380	putative cytochrome c biogenesis factor
At2 g35260	putative CAAX-type-II prenyl protease

COMPARISON OF ATTED WITH WEIGHTED GENE CORRELATION NETWORK ANALYSIS

The chances of identifying a true positive hit in the search for photorespiratory transporters will clearly increase when focusing on those genes that encode proteins containing predicted TMDs. Indeed, this strategy has proved to be a powerful tool for the identification of new transport proteins involved in photorespiration (see section 'Proof of principle'). To evaluate robustness of the Atted-based approach in possibly identifying additional candidate genes, we have also employed weighted gene correlation network analysis (WGCNA; Langfelder & Horvath 2008). This approach was based on data from a subset of Affymetrix ATH1 arrays (Schmid *et al.* 2005) with *Arabidopsis* wild-type expression data from 64 different experiments. Expression values were normalised (arithmetic mean of three replicates calculated, then log2-transformed and finally median-centred) and WGCNA performed on the R platform, with a calculated soft threshold of seven and a static tree cut height of 0.8 for module detection. Without exception, the photorespiratory key enzymes (Table 1) were assigned to one distinct module, along with all of the TOP10 candidates listed in Table 2, and selected after co-expression analysis. The probability that ten randomly drawn genes (Table 2) would all be members of the WGCNA module that represents photorespiratory key genes is $P = 8.1 \times 10^{-11}$, further strengthening our selection. The size of this module (2241 genes) exceeds the list of results from Atted before application of the filter (1246 genes, excluding all duplicates). Interestingly, 78% of the genes obtained from Atted are present in the WGCNA module containing the photorespiratory key enzymes (Fig. 2). This indicates that both methods are largely consistent with each other. However, the number of putative candidates from both systems requires additional filters to reduce the list of genes. Since the module generated with WGCNA analysis is rather large, further analysis of this module might add additional can-

Fig. 2. Comparison between Atted-II and WGCNA results. A total of 972 genes are present in results from both analyses, constituting roughly 80% of all genes obtained from Atted-II.

didates to our list of interest. This work is currently ongoing in our lab.

Both co-expression and WGCNA analysis, performed individually, are not sufficient to yield lists of candidate genes that are short enough to investigate *in vivo*. Nevertheless, they are reasonable first steps to provide a first suggestion of putative candidates. In combination with the right setup of filters, such *in silico* experiments provide a powerful but still simple tool for the initial phase of our studies.

PROOF OF PRINCIPLE

We have demonstrated that co-expression analysis can actually function as a viable tool for the discovery of candidate genes involved in photorespiratory transport. We were able to identify three new genes coding for photorespiratory transporters using the workflow described above. These genes had remained hidden in previous attempts using forward genetic approaches.

Identification of *AtBOU* as photorespiratory transporter

The A BOUT DE SOUFFLE (BOU) protein was successfully identified as a transporter involved in shuttling intermediates in the photorespiratory C_2 cycle (Eisenhut *et al.* 2013). Since *BOU* was already under investigation, it has been excluded from the list of TOP10 candidates (Table 1). Nevertheless, it reached one of the highest ranks before the filter was applied to co-expression results. Out of 13 possible hits, *BOU* was found ten times in the list of co-expressed genes.

The idea of screening *Arabidopsis bou* mutants for a photorespiratory phenotype originated from initial co-expression results showing that *BOU* mRNA is simultaneously expressed with genes for photorespiratory key enzymes. *BOU* gene expression is apparently particularly coordinated with the expression of genes coding for the GDC multi-enzyme complex, which is central in the photorespiratory glycine-to-serine conversion. *BOU* promoter activity is triggered by light and restricted to leaf tissues and cotyledons, supporting a pivotal role in photosynthesis and photorespiration (Eisenhut *et al.* 2013).

Indeed, *bou* knockout plants suffer in ambient air, having chlorotic leaves and growth retardation, but growing much like the wild type when kept under HC. Despite these morphological alterations, further data support the role of BOU in photorespiration. Gas exchange measurements demonstrated that knockout plants have strongly reduced CO_2 fixation rates, and therefore have elevated CO_2 compensation points in comparison to the wild type under HC conditions, which increase upon a shift to NC conditions. Moreover, photorespiratory metabolism is disturbed in the *bou* mutant, as indicated in metabolic analysis. Most remarkably, the glycine level in the mutant is greatly increased in comparison to that of wild-type plants. Together with the finding that mitochondrial glycine degradation is strongly reduced in *bou*, it is anticipated that the function of BOU will be connected with GDC activity. We discriminate between direct transport of glycine into mitochondria by BOU from shuttling of a GDC cofactor (Eisenhut *et al.* 2013). Studies with spinach mitochondria revealed the ability of glycine to diffuse through the mitochondrial membrane if the glycine concentration exceeds 0.5 mM, which is the case in photosynthetic tissues (Yu *et al.* 1983). In the *bou* mutant, the concentration of glycine is even higher, which makes the cofactor hypothesis more likely (Eisenhut *et al.* 2013). To date, the specific substrate transported *via* BOU has not been identified, but ongoing research is focusing on GDC cofactors.

Identification of *DiT1* as photorespiratory transporter

The plastidial 2-oxoglutarate (2-OG)/malate transporter (*DiT1*, *AtDiT1*, *OMT1*) is an interesting candidate, due to its clustering with the already described *DiT2.1* (Taniguchi *et al.* 2002; Renne *et al.* 2003) in a co-expression analysis approach and its sequence homology with *DiT2.1* (*AtpDCT1*). *dit1* mutant plants were shown to suffer under NC conditions, resulting in retarded development, small leaf size, frequently emerging shoots and a decrease in chlorophyll content; however, the mutant still survives under NC conditions, while *DiT2.1*-deficient plants are only viable under HC conditions (Kinoshita *et al.* 2011). DiT1 supplies the chloroplast with 2-OG, which is utilised by FD-GOGAT in the chloroplast and forms a double transporter system together with DiT2.1, further participating in the export of synthesised glutamate and re-fixation of ammonium ions originating from the photorespiratory cycle (Schneidereit *et al.* 2006; Kinoshita *et al.* 2011).

Identification of *AtPLGG1* as photorespiratory transporter

The gene *PLGG1* (*At1 g32080*) was ranked as a most promising candidate according to our co-expression analysis (Table 1). Analysed *plgg1-1* knockout plants develop chlorotic regions along the leaf lamina when grown under ambient air (NC) conditions. This phenotype can be complemented to almost wild type-like when *plgg1-1* knockout plants are kept under HC. We very recently demonstrated that PLGG1 functions as the long-sought plastidic glycerate/glycolate transporter (Pick *et al.* 2013).

Experiments showed that PLGG1 is the translocator mediating glycolate export and glycerate import over the chloroplast membrane, linking chloroplasts to peroxisomes during photorespiration. Metabolite analysis revealed that glycolate and glycerate accumulate in *plgg1-1* plants in response to a shift from HC to NC. Labelling experiments demonstrated that the export of glycolate is blocked in *plgg1-1* plants because, after an $^{18}O_2$ pulse, the level of labelled glycolate increases rapidly compared to the wild type. In contrast to this observation, a slow accumulation of labelled glycerate in *plgg1-1* plants compared to the wild type was measured, pointing to a block in glycerate import. Further experiments linking glycerate import to O_2 evolution at photosystem II revealed reduced O_2 generation in *plgg1-1* chloroplasts compared to wild-type chloroplasts, pointing again to impaired import of glycerate into PLGG1-deficient chloroplasts. These data mirror and complete previous observations demonstrating that glycolate and glycerate are transported by the same protein without facilitating strict counter-exchange (Howitz & McCarty 1986), as needed in photorespiration.

PERSPECTIVE

In this review we describe a four-step protocol that integrates *in silico* and *in vivo* experiments to identify new transport proteins for the photorespiratory cycle. The effectiveness of this

approach is demonstrated through the identification of BOU, DiT1 and PLGG1 as photorespiratory transporters. Thus, the co-expression screen offers a promising method to shed light on yet unknown transport processes between chloroplasts, peroxisomes and mitochondria. Complete understanding of these processes is essential in future attempts to remodel photorespiratory carbon flux in the effort to decrease photorespiratory carbon loss, which can account for up to 20% of net CO_2 assimilation (Cegelski & Schaefer 2006).

The workflow outlined here for the discovery of candidate genes involved in photorespiration can be used, with minor individual adjustments, to develop hypotheses about gene functions not related to photorespiration (Reumann & Weber 2006). To find missing links in other pathways, the initial query genes used for co-expression analysis must be adjusted, together with filtering criteria in the second step, to match the needs of the particular query. Since co-expression analysis through publicly available online resources is now fast and simple, it provides a valuable first step in the search for currently unknown gene functions and for annotating orphan steps in biological pathways and processes.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (PROMICS Research Group, WE 2231/8-1).

REFERENCES

Aoki K., Ogata Y., Shibata D. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology*, **48**, 381–390.

Bartel D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Boldt R., Edner C., Kolukisaoglu U., Hagemann M., Weckwerth W., Wienkoop S., Morgenthal K., Bauwe H. (2005) D-GLYCERATE 3-KINASE, the last unknown enzyme in the photorespiratory cycle in Arabidopsis, belongs to a novel kinase family. *The Plant Cell*, **17**, 2413–2420.

Brazma A., Vilo J., (2000) Gene expression data analysis. *FEBS Letters*, **480**, 17–24.

Cegelski L., Schaefer J. (2006) NMR determination of photorespiration in intact leaves using *in vivo* $^{13}CO_2$ labeling. *Journal of Magnetic Resonance*, **178**, 1–10.

Cousins A.B., Pracharoenwattana I., Zhou W., Smith S.M., Badger M.R. (2008) Peroxisomal malate dehydrogenase is not essential for photorespiration in Arabidopsis but its absence causes an increase in the stoichiometry of photorespiratory CO_2 release. *Plant Physiology*, **148**, 786–795.

Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA*, **95**, 14863–14868.

Eisenhut M., Pick T., Bordych C., Weber A. (2012) Towards closing the remaining gaps in photorespiration – the essential but unexplored role of transport proteins. *Plant Biology*, **15**, 676–685.

Eisenhut M., Planchais S., Cabassa C., Guivarc'h A., Justin A.M., Taconnat L., Renou J.P., Linka M., Gagneul D., Timm S., Bauwe H., Carol P., Weber A.P. (2013) Arabidopsis A BOUT DE SOUFFLE is a putative mitochondrial transporter involved in photorespiratory metabolism and is required for meristem growth at ambient CO_2 levels. *The Plant Journal*, **73**, 836–849.

Haferkamp I., Linka N. (2012) Functional expression and characterisation of membrane transport proteins. *Plant Biology*, **14**, 675–690.

Heazlewood J.L., Verboom R.E., Tonti-Filippini J., Small I., Millar A.H. (2007) SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Research*, **35** (Database issue), D213–D218.

Hirai M.Y., Sugiyama K., Sawada Y., Tohge T., Obayashi T., Suzuki A., Araki R., Sakurai N., Suzuki H., Aoki K., Goda H., Nishizawa O.I., Shibata D., Saito K. (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proceedings of the National Academy of Sciences USA*, **104**, 6478–6483.

Howitz K.T., McCarty R.E. (1986) d-glycerate transport by the pea chloroplast glycolate carrier: studies on [1-C]d-glycerate uptake and d-glycerate dependent O_2 evolution. *Plant Physiology*, **80**, 390–395.

Jamai A., Salome P.A., Schilling S.H., Weber A.P., McClung C.R. (2009) Arabidopsis photorespiratory serine hydroxymethyltransferase activity requires the mitochondrial accumulation of ferredoxin-dependent glutamate synthase. *The Plant Cell*, **21**, 595–606.

Kinoshita H., Nagasaki J., Yoshikawa N., Yamamoto A., Takito S., Kawasaki M., Sugiyama T., Miyake H., Weber A.P., Taniguchi M. (2011) The chloroplastic 2-oxoglutarate/malate transporter has dual function as the malate valve and in carbon/nitrogen metabolism. *The Plant Journal*, **65**, 15–26.

Langfelder P., Horvath S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Lee H.K., Hsu A.K., Sajdak J., Qin J., Pavlidis P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Research*, **14**, 1085–1094.

Nozawa A., Nanamiya H., Miyata T., Linka N., Endo Y., Weber A.P., Tozawa Y. (2007) A cell-free translation and proteoliposome reconstitution system for functional analysis of plant solute transporters. *Plant and Cell Physiology*, **48**, 1815–1820.

Obayashi T., Kinoshita K. (2009) Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Research*, **16**, 249–260.

Obayashi T., Kinoshita K., Nakai K., Shibaoka M., Hayashi S., Saeki M., Shibata D., Saito K., Ohta H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Research*, **35** (Database issue), D863–D869.

Obayashi T., Nishida K., Kasahara K., Kinoshita K. (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant and Cell Physiology*, **52**, 213–219.

Persson S., Wei H., Milne J., Page G.P., Somerville C.R. (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences USA*, **102**, 8633–8638.

Pick T.R., Bräutigam A., Schulz M.A., Obata T., Fernie A.R., Weber A.P.M. (2013) PLGG1, a plastidic glycolate glycerate transporter, is required for photorespiration and defines a unique class of metabolite transporters. *Proceedings of the National Academy of Sciences USA*, doi: 10.1073/pnas.1215142110.

Ren Q., Kang K.H., Paulsen I.T. (2004) TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Research*, **32**(Database issue), D284–D288.

Ren Q.H., Chen K.X., Paulsen I.T. (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Research*, **35**, D274–D279.

Renne P., Dressen U., Hebbeker U., Hille D., Flugge U.I., Westhoff P., Weber A.P. (2003) The Arabidopsis mutant *dct* is deficient in the plastidic glutamate/malate translocator DiT2. *The Plant Journal*, **35**, 316–331.

Reumann S., Weber A.P. (2006) Plant peroxisomes respire in the light: some gaps of the photorespiratory C2 cycle have become filled – others remain. *Biochimica et Biophysica Acta*, **1763**, 1496–1510.

Schmid M., Davison T.S., Henz S.R., Pape U.J., Demar M., Vingron M., Scholkopf B., Weigel D., Lohmann J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nature Genetics*, **37**, 501–506.

Schneidereit J., Hausler R.E., Fiene G., Kaiser W.M., Weber A.P. (2006) Antisense repression reveals a crucial role of the plastidic 2-oxoglutarate/malate translocator DiT1 at the interface between carbon and nitrogen metabolism. *The Plant Journal*, **45**, 206–224.

Schwab R., Ossowski S., Riester M., Warthmann N., Weigel D. (2006) Highly specific gene silencing by artificial microRNAs in Arabidopsis. *The Plant Cell*, **18**, 1121–1133.

Schwacke R., Schneider A., van der Graaff E., Fischer K., Catoni E., Desimone M., Frommer W.B., Flugge U.I., Kunze R. (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. *Plant Physiology*, **131**, 16–26.

Schwacke R., Fischer K., Ketelsen B., Krupinska K., Krause K. (2007) Comparative survey of plastid and mitochondrial targeting properties of transcription factors in Arabidopsis and rice. *Molecular Genetics and Genomics*, **277**, 631–646.

Somerville C.R. (2001) An early Arabidopsis demonstration. Resolving a few issues concerning photorespiration. *Plant Physiology*, **125**, 20–24.

Somerville C.R., Ogren W.L. (1979) A phosphoglycolate phosphatase-deficient mutant of Arabidopsis. *Nature*, **280**, 833–836.

Somerville C.R., Ogren W.L. (1980a) Inhibition of photosynthesis in Arabidopsis mutants lacking leaf glutamate synthase activity. *Nature*, **286**, 257–259.

Somerville C.R., Ogren W.L. (1980b) Photorespiration mutants of *Arabidopsis thaliana* deficient in serine-glyoxylate aminotransferase activity. *Proceedings of the National Academy of Sciences USA*, **77**, 2684–2687.

Somerville C.R., Ogren W.L. (1981) Photorespiration-deficient mutants of *Arabidopsis thaliana* lacking mitochondrial serine transhydroxymethylase activity. *Plant Physiology*, **67**, 666–671.

Somerville C.R., Ogren W.L. (1982) Mutants of the cruciferous plant *Arabidopsis thaliana* lacking glycine decarboxylase activity. *The Biochemical Journal*, **202**, 373–380.

Stuart J.M., Segal E., Koller D., Kim S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Taniguchi M., Taniguchi Y., Kawasaki M., Takeda S., Kato T., Sato S., Tabata S., Miyake H., Sugiyama T. (2002) Identifying and characterizing plastidic 2-oxoglutarate/malate and dicarboxylate transporters in *Arabidopsis thaliana*. *Plant and Cell Physiology*, **43**, 706–717.

Taylor N.G., Howells R.M., Huttly A.K., Vickers K., Turner S.R. (2003) Interactions among three distinct CesA proteins essential for cellulose synthesis. *Proceedings of the National Academy of Sciences USA*, **100**, 1450–1455.

Timm S., Bauwe H. (2012) The variety of photorespiratory phenotypes – employing the current status for future research directions on photorespiration. *Plant Biology*, **15**, 737–747.

Usadel B., Obayashi T., Mutwil M., Giorgi F.M., Bassel G.W., Tanimoto M., Chow A., Steinhauser D., Persson S., Provart N.J. (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell & Environment*, **32**, 1633–1651.

Voll L.M., Jamai A., Renne P., Voll H., McClung C.R., Weber A.P. (2006) The photorespiratory Arabidopsis *shm1* mutant is deficient in SHM1. *Plant Physiology*, **140**, 59–66.

Werdan K., Heldt H.W. (1972) Accumulation of bicarbonate in intact chloroplasts following a pH gradient. *Biochimica et Biophysica Acta*, **283**, 430–441.

Yonekura-Sakakibara K., Tohge T., Niida R., Saito K. (2007) Identification of a flavonol 7-O-rhamnosyltransferase gene determining flavonoid pattern in Arabidopsis by transcriptome coexpression analysis and reverse genetics. *Journal of Biological Chemistry*, **282**, 14932–14941.

Yu C., Claybrook D.L., Huang A.H.C. (1983) Transport of glycine, serine, and proline into spinach leaf mitochondria. *Archives of Biochemistry and Biophysics*, **227**, 180–187.

IX. ACKNOWLEDGEMENTS

This PhD thesis would not have been possible without the guidance and support of many people, whose contribution and assistance helped me during the time of experimental work and the process of writing.

I would like to thank wholeheartedly,

Prof. Dr. Andreas Weber, for letting me write my dissertation in your laboratory, for your constant support, great ideas and inspiration. I am very grateful for the knowledge I gained in your laboratory and for the great opportunity of going to Michigan State University.

Prof. Dr. Rüdiger Simon, for your valuable ideas, suggestions and especially critical questions during the committee meetings and for being the co-reviewer of my thesis.

Dr. Andrea Bräutigam, for your support and practical advice and for being the Post-Doc in charge during my doctorate time.

Prof. Dr. C. Robin Buell and **the Buell lab**, thank you so much for letting me stay in your laboratory for my MSU research stint and looking after me so well. The stay in your lab was a great experience. I would like to express my gratitude to all the members of the **Buell lab**, who always helped me kindly making me feel at home. Especially, I would like to thank **Dr. Elsa Gongorra-Castillo** for teaching me RNA-seq data analysis so patiently and **Dr. Kevin Childs** for sharing with me his knowledge and scripts about WGCNA analysis.

Dr. Sigrun Wegener-Feldbrügge, for you great support as iGRAD-plant Coordinator. Your wonderful organization and coordination of the program made my life so much easier.

all **iGRAD-plant members** for the enjoyable retreats and lively discussion during our meetings

Prof. Dr. Barb Sears, for your enormous support during my stay in East Lansing and letting me borrow the nostalgic blue truck everyone called “George”.

Prof. Dr. Martin Lercher und **Prof. Dr. Peter Westhoff** for your support during my laboratory rotations. I would also like to thank **Dr. Christian Esser** for teaching me very patiently Perl programming, during my laboratory rotation, which helped me a great deal in the subsequent data analyses. Thank you **Dr. Udo Gowik**, for teaching me how to prepare hand cross-sections of fresh leaves during my Westhoff lab rotation. This method was one of the most helpful skills for my leaf anatomical analyses.

all Weber lab members for their help and being a great team. Thank you so much for making my time in the Weber laboratory so enjoyable. I will remember all the funny occasions and terrace BBQs fondly.

the **“C_4 group”** (**Alisandra K. Denton, Manuel Sommer, Thomas J. Wrobel, Dominik Brilhaus, Simon Schliesky, Dr. Thea R. Pick, Dr. Urte Schlüter** and **Dr. Andrea Bräutigam**) for helpful and insightful discussions and valuable ideas.

Alisandra K. Denton and **Dr. Jan Wiese** for proofreading my dissertation.

everyone, who was involved in the presented manuscripts. I would like to thank all those I could not mention here individually.

my family, especially my **parents**, for your constant encouragement, love and support throughout my life.

Ferdi, for having you in my life.

The **Deutsche Forschungsgemeinschaft** and **iGRAD-plant** for funding.