Heinrich Heine
Universität
Düsseldorf

Phylogenetic Analyses on the Evolution of C_4 Photosynthesis

Inaugural-Dissertation

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

vorgelegt von
Janina Maß
aus Duisburg

Düsseldorf, März 2020

Aus dem Institut für Computergestützte Zellbiologie
der Heinrich-Heine-Universität Düsseldorf

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

Berichterstatter:

1. Prof. Dr. Martin J. Lercher
2. Prof. Dr. Andreas P. M. Weber

Tag der mündlichen Prüfung: 27.10.2020

Eidesstattliche Versicherung und Selbstständigkeitserklärung

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

Ort, Datum Janina Maß

「痛みは避けられない。苦しみはオプションです。」

村上春樹

HEINRICH-HEINE-UNIVERSITÄT DÜSSELDORF

Zusammenfassung

Mathematisch-Naturwisselschaftliche Fakultät
Institut für Informatik

Phylogenetic Analyses on the Evolution of C_4 Photosynthesis

Janina Maß

Der C_4-Stoffwechselweg, der seinen Namen von der initialen Fixierung von Kohlenstoff als 4-Kohlenstoff-Verbindung erhält, kann als Erweitung zum evolutionär älteren C_3-Stoffwechselweg der Photosynthese beschrieben werden. C_4-Photosynthese bringt Pflanzen unter heißen und trockenen Bedingungen Vorteile dadurch, dass der nachteilige Prozess der Photorespiration unterdrückt wird. In C_4-Pflanzen findet die abschlieißende Kohlenstofffixierung statt nachdem Kohlenstoffdioxid in der Nähe von Rubisco angereichert wird, was durch verschiedene biochemische und anatomische Änderungen ermöglicht wird. Trotz der Notwendigkeit komplexer Veränderungen hat sich die C_4-Photosynthese mehrfach konvergent entwickelt. C_4 -Photosynthese entstand unanbhängig in Clustern, z.B. in der Familie der Cleomaceae, die in den *Manuskripten 4* and *5* untersucht wurde; einige Cluster enthalten noch vorhandene intermediäre Spezies. Die effiziente Wasser- und Stickstoffverwertung machen es erstrebenswert, den C_4- Stoffwechselweg in C_3-Pflanzen einzubringen um die landwirtschaftliche Produktion zu verbessern. Um letztlich das Ziel der Einbringung des C_4-Weges zu erreichen, ist es notwendig seine Komplexität zu verstehen.

Die Manuskripte in dieser Arbeit zeigen Methoden und Analysen auf, deren Zielsetzung ein besseres Verständnis der Komplexität der C_4-Photosynthese ist. Dabei werden Genexpression und genomische Sequenzen betrachtet. Die Möglichkeiten, die solch große Datensätze, die aus verschiedenen Quellen stammen, bringen, führen allerdings auch zu hoher Heterogenität der Daten und machen sorgfältige Prozessierung nötig. Eine Komponente davon – die Maximierung nutzbarer Alignmentinformation durch Outlier-Filterung – wird in *Manuskript 1* beschrieben.

In den Manuskripten *4* und *5* werden Genexpressionsdaten zweier nah verwandter Spezies aus der Familie der Cleomaceae verglichen, die den C_4 bzw. C_3 Stoffwechselweg nutzen. In den Manuskripten *2* und *3* untersuchen wir die speziellen C_4-Gewebe Mesophyll (M)- und Bündelscheidenzellen (BS) während der Blattentwicklung der wichtigen C_4 -Nutzpflanze *Zea mays*. *Manuskript 2* ist eine Meta-Studie, in dem wir unseren eigens generierten Datensatz mit denen anderer Studien, die BS- und M-Gewebe untersuchen, vergleichen. Dabei wird die Nützlichkeit der Datenintegration und Querreferenzierung mehrerer Studien deutlich. Der Datensatz wurde zum Download und zur Visualisierung

als Ressource zur Verfügung gestellt. Die Daten wurden darüberhinaus in *Manuskript 3* verwendet, in dem wir die Beziehung von Expressionsdivergenz zum Selektionsdruck untersuchen. Dies beinhaltete die Durchführung automatisierter phylogenetischer Analysen, einschließlich des Herausfilterns von Sequenzen mit dem in *Manuskript 1* beschriebenen Tool, um das Auftreten positiver Selektion zu untersuchen. Dabei fanden wir qualitative und quantitative Belege für die präkonditionierende Rolle von Genduplikation in der Evolution der C_4-Photosynthese.

HEINRICH HEINE UNIVERSITY

Abstract

Faculty of Mathematics and Natural Sciences
Institute for Computer Science

Phylogenetic Analyses on the Evolution of C_4 Photosynthesis

by Janina Maß

The C_4 pathway, which derives its name from the initial fixation of carbon as a four-carbon compound, can be described as an *add-on* to the evolutionarily older C_3 pathway of photosynthesis. It provides advantages for plants under hot, arid conditions as it suppresses the detrimental process of photorespiration. In C_4 plants final carbon fixation takes place after concentrating carbon dioxide near Rubisco, which is made possible by various biochemical and anatomical alterations. Despite the complex adaptions required, C_4 photosynthesis evolved many times convergently. Independent C_4 origins occur in clusters, e.g. in the Cleomaceae family investigated in *Manuscripts 4* and *5*; some of the clusters include extant intermediary species. The high water and nitrogen use efficiency of the C_4-trait make it a desirable target for introduction into non-C_4 crops to improve agricultural production. To ultimately achieve the goal of engineering C_4, it is necessary to understand the complexity of the trait.

Manuscripts included in this thesis present methods or analyses targeted at understanding the complexity of the C_4-trait via examination of gene expression and genomic sequence. The opportunities opened up by sourcing large-scale data from different pools come at the price of high heterogeneity and necessitate careful processing. One aspect of this – maximizing usable alignment information via outlier filtering – is described in *Manuscript 1*.

Manuscript 4 and Manuscript 5 both focus on comparing gene expression between closely related Cleomaceae species that utilize the respective photosynthesis pathways. In *Manuscript 2 and Manuscript 3* we take a look at the key C_4 tissues, mesophyll (M) and bundle sheath (BS) during leaf development in the important C_4 crop species *Zea mays*. *Manuscript 2* is a meta-study comparing our own generated data set with those of other studies targeting BS and M tissues; it highlights the usefulness of integrating and cross-referencing multiple studies. The data was made available for download or visualization as a community resource. The data was further used in *Manuscript 3*, where we compared expression divergence to phylogenetic signal. This included performing an automated phylogenetic analysis, including sequence filtering with the tool from *Manuscript 1*, to infer the presence of positive selection. We found qualitative and

quantitative evidence for the preconditioning role of gene duplication in the evolution of C_4 photosynthesis.

Acknowledgements

This thesis is only possible due to the efforts of my supervisors, my colleagues and, of course, my co-authors.
A big thank you goes to everyone involved in the projects and to everyone who was part of the iGRADplant graduate school.
In particular, I'd like to thank the following people:
Prof. Dr. Martin Lercher for the opportunity to work on this project, for his support and his patience.
Prof. Dr. Andreas Weber and his lab for the provided expertise and assistance.
Dr. Christian Esser for all his support and also being the funniest office mate in the world.
Dr. Gabriel Gelius-Dietrich who was a great inspiration for me when I saw him hacking away regex-littered sed and awk commands and I thought that was so cool.
Dr. Anna Kersting for being super helpful, friendly and pragmatic (and for bringing her plant Emily Ficus into our office).
Everyone in the Bioinformatics group, and especially Dr. David Heckmann, Dr. Thomas Laubach, Dr. Sabine Thuß, Rafael Dellen, and Anja Walge who were so great to be around.
I'm grateful for Dr. Sigrun Wegener-Feldbrügge who took good care of all of us grad students and made managing the iGRADplant graduate school look effortless.
I thank Prof. Dr. Shinhan Shiu and his lab for the great time at Michigan State, as well as Prof. Dr. Barb Sears for all the help we received from her during our stay in Michigan.
Thanks also go to the DFG for funding.
Further thanks also go to the Pythonfoo crowd at the Chaosdorf, especially bison, who I've learned a lot from.
Marie Bolger and Dr. MaPi Cendrero for sharing the pain and encouraging me not to give up on this thesis during my time in Jülich.
Last, and decidedly not least, Ali for so much – professionally for helping me with statistics, paper writing, understanding biology, and personally, for keeping me sane :) .

Contents

Abbreviations

BS	**B**undle **s**heath
CBBC	**C**alvin-**B**enson-**B**assham **c**ycle
CO_2	**C**arbon **dio**xide
Gbp	**G**iga**b**ase **p**airs
kbp	**k**ilo**b**ase **p**airs
Mbp	**M**ega**b**ase **p**airs
M	**M**esophyll
O_2	**O**xygen
PS	**P**hoto**s**ystem
Rubisco	**R**ib**u**lose-1,5-**bis**phosphate **c**arboxylase/**o**xygenase

Chapter 1

Introduction

1.1 The Importance of Understanding Evolution

Evolution describes the gradual accumulation of changes of heritable traits in populations over time often as fitness-increasing adaptions driven by natural selection (Darwin, 1859). For evolution to take place, it needs genetic variability, through mutations, that can be passed on to the next generation and may lead to new traits.

The impact of mutations can be neutral or detrimental, but also beneficial under certain circumstances. An interesting example is the sickle cell trait with its protection against malaria (Aidoo et al., 2002). Sickle cell anemia, an inherited disorder for which both alleles need to be mutated, leads to a shorter life expectancy. Therefore, it is expected that the mutated allele would be uncommon. However in regions with endemic malaria, it is highly prevalent (Piel et al., 2010) due to sickle cell hemoglobin having a survival advantage against malaria.

Research on evolution is not only figuring out how the apparent current biodiversity could have come into existence, e.g. how birds evolved from small carnivorous dinosaurs (Xu et al., 2003), but also impacts our everyday lives in various ways, be it medical or agricultural advancements. Medical examples of how understanding evolution affects our everyday lives include continuous development of influenza vaccines to keep pace with ongoing changes, and managing antibiotic treatment regimes to minimize development of antibiotic resistance in bacteria. In agriculture, understanding the principles behind evolution is important for breeders who are interested in optimizing traits in crop plants. This varies from trying to cross in traits, such as disease resistance, from wild close relatives to very ambitious goals such as trying to engineer a highly beneficial complex trait such as C_4 photosynthesis in distantly related key crop species. The C_4 engineering community looks for hints on how to engineer C_4 photosynthesis in the natural occurrence of the trait. With evolutionary intermediates representing stable states with any necessary preconditions established in order.

Simple and Complex Traits

Some traits follow simple, Mendelian genetics; the presence or absence of this trait is entirely determined by the genotype at one genetic locus. Of these, some are recessive like certain types of albinism (Dessinioti et al., 2009), requiring an individual to have two copies of the allele before the phenotype becomes apparent. In other cases, as e.g. with Huntington's disease (Walker, 2007), the mutant allele is dominant, and the phenotype manifests as soon as there is a single copy of the allele present. Finally, many simple traits, like the sickle cell trait discussed above, have an intermediate phenotype in heterozygous individuals. In plants or animals, such simple trait-related alleles can be screened for and possibly utilized for breeding with relative ease.

In contrast to simple traits, complex traits like the yield or flavor of a crop plant are influenced by many different genetic loci. Naturally, many complex traits are of key interest for human prosperity and therefore our breeding and engineering efforts. However, generally speaking, attempts to influence complex traits are limited as the understanding of the entire complexity is hard to achieve. In particular, quantitative modifications to traits such as yield may be an ongoing challenge; but bringing a complex trait into a species where it was previously absent, has been historically intractable.

One reason to be hopeful about future efforts to engineer complex traits is that despite their complexity, many evolve in a convergent, repeated manner. Well known examples of this include eyesight (Gehring, 2005) and flight. A particularly striking example of repeated evolution is C_4 photosynthesis, a trait which increases photosynthetic efficiency in hot, arid, and high light conditions, which has evolved at least 66 times (Sage et al., 2012). While its highly convergent evolution gives researchers optimism that it may be possible to establish the C_4 photosynthetic trait in non-C_4 crop plants, doing so will require an extraordinary understanding of the overall complexity of the trait. Obtaining such an understanding will require large scale, cross species analyses. These analyses, in turn, must be reliable enough to ultimately help generate a high-precision, high-recall list of necessary genetic changes.

1.2 Advancements and Challenges in Comparative Studies

Advances in technology combined with accumulating public knowledge on genetic sequences of many species open new possibilities in performing the sort of large-scale comparative analyses that are necessary to understand complex traits. However, the entirety of the data generation and analysis pipeline, from the cultivation of each species prior to sequencing through to the last phylogenetic comparison, has the potential to influence the final conclusions. Re-generating or even just re-analyzing all the data going into a pipeline is generally not feasible, yet awareness of the various strengths and

weaknesses of early data analyses can help keep any potentially introduced errors from propagating through a pipeline, hindering analysis or even potentially leading to false conclusions.

1.2.1 Genomes and Transcriptomes, Sequencing, Assembly, and Annotation

The world of DNA and RNA sequencing has changed rapidly over the years. In the 1990s the first fully sequenced genomes were published, starting off with the bacterium *Haemophilus influenzae* in 1995 (Fleischmann et al., 1995), followed by model organisms *S. cerevisiae* (Goffeau et al., 1996), a unicellular eukaryote, and the bacterium *E. coli* (Blattner et al., 1997). Starting in 1990, the human genome was sequenced as a large consortium project with the help of researchers from across 20 institutions and with a cost of about $2.7 billion, and an initial draft was published in 2001 (International Human Genome Sequencing Consortium, 2001). In contrast, the similarly sized and recently released American cockroach genome has 19 contributors on the authors list (Li et al., 2018). Differences like this have been made possible by advances in sequencing technologies.

Revolutionary at the time of its invention, Sanger sequencing (Sanger et al., 1977) can produce reads with a very low error rate of about one kbp in length (Shendure and Ji, 2008), however, the technology produces one read per reaction tube, making obtaining coverage of larger genomes highly work and cost intensive. With many genomes measuring multiple Gbp in length, and containing repetitive regions, Sanger technology also had to be complemented with other tools to capture the larger structures. Historical methods to capture long range info involved time consuming protocols to create, on the smaller end, jumping libraries (such as mate-pairs or fosmids) that rely on the circularization of DNA, and the capture and sequencing of the join-point to identify regions from several to tens of kbps of each other. On the longer end, artificial chromosomes can be transformed, maintained and amplified in species such as *E.coli* and then sequenced individually (as reviewed in Ekblom and Wolf (2014)).

Beginning in 2005, a wave of second generation sequencing (2GS) technologies became available which broke up the one-reaction-one-sequence paradigm by allowing for the simultaneous sequencing of thousands or even millions of sequences from one library that have been – by methods specific to individual technologies – distributed and fixed across a surface . Incorporation of fluorescent nucleotides can be measured at each position with a laser and camera. These methods rely on PCR to amplify the DNA prior to sequencing, which has some consequences for the resulting data. The PCR ultimately biases the technologies towards sequencing of shorter fragments, for example in the range of a few hundred bp. On the upside, these methods maintain respectable accuracy (∼99.9% for

Illumina sequencing (Fox et al., 2014)), as not individual molecules, but rather clusters of molecules are sequenced.

Second generation sequencing already started a revolution in genomics with the basic coverage required for sequencing a genome becoming obtainable in individual labs. In this time frame, not only key model, but also generally high-interest species such as hot pepper (*Capsicum annuum*) (Kim et al., 2014) or hemp (*Cannabis sativa*) (Van Bakel et al., 2011) were sequenced. However, obtaining long range contiguity information remained challenging, and many genomes released in this time frame were highly fragmented (e.g. the barley genome (International Barley Genome Sequencing Consortium, 2012)).

Besides genomics, second generation sequencing was, and is, used heavily for transcriptomes. RNAseq is frequently used complementary to genome sequencing. For instance RNAseq can be provided as additional information for gene annotation. Additionally however, de novo transcriptome assembly allows for the reconstruction of a species' putative transcriptome from RNAseq data alone. While frequently much better than assembling with no or a very distant reference, de novo transcriptome assembly is complicated by the dynamic range of RNAseq data making e.g. a rare transcript essentially indistinguishable from a sequencing error. Thus, generated sequences can be biased and difficult to work with.

Around 2010, several third generation sequencing technologies started being developed, tested, and used (Pacific Biosciences (English et al., 2012), Oxford Nanopore (Mikheyev and Tin, 2014)). These technologies focus on sequencing individual molecules of DNA (or even directly RNA), removing PCR from the picture and obtaining impressive lengths. Sequences can start from several kbp, and now frequently reach 100s of kbps. Sequencing of individual molecules comes at the cost of a lower signal to noise ratio, which results in lower accuracy, with error rates ranging from nowadays 2% up to 38% in an early assessment (Laver et al., 2015). These long range technologies led again to an increase in the amount of genomes being sequenced, but also to a general improvement in contiguity. Complementing 3GS data with existing 2GS methods has made challenges posed by the high error rate largely surmountable. More genomes are released with these technologies on a daily basis.

Finally, several new technologies have provided ways to obtain Mbp scale contiguity information in a high throughput manner (BioNano Genomics, Hi-C, 10x Genomics), which can be used to scaffold the anyways improving genome assemblies and rapidly and cheaply achieve draft genomes of seemingly comparable quality to reference assemblies. As these different approaches have their strengths and weaknesses, multiple technologies are typically used in tandem (e.g. Dudchenko et al. (2017), Shi et al. (2016), Zimin et al. (2017)).

All this development has several implications to working with public data today. First, the data is heterogeneous: Some assemblies are more or less fragmented; some assemblies have been annotated with the help of more or less RNAseq information or with RNAseq information derived from very different sets of tissues. Some assemblies have been created by genome specialists, others by the lab most interested in the particular species. All these technologies come with their own specific error models, reflected again in the final assembly. Second, the ratio of person-time to data production has become very low. While there are many cases where more data can compensate for human time investment in curation, there are still plenty of cases where this changed ratio can lead to error. Steps such as selfing a species, to sequence a homozygous individual are often omitted. Several infamous genomes were published with the inclusion of adapter and technical sequences, e.g. the carp genome (Xu et al., 2014) where Illumina adapter sequences had not been removed. Current genome papers often use all their data for assembly, and reserve none for verification, leading to no reliable estimate of final quality. Third, the sheer amount of data available – challenging to work with or not – opens up new opportunities for cross species comparisons and understanding evolution.

1.2.2 Aims of Phylogenetic Analyses

Phylogenetic analyses is an umbrella term that encompasses a huge variety of distinct analyses – from comparing the incidence of particular SNPs within populations to tracing the signature of ancient whole genome duplications across whole kingdoms. However, these analyses retain some commonalities. They all ultimately center around comparing biological sequence (DNA, RNA, or protein) to determine the relation of the sequenced entities. Further, once the basis for comparison and the relationships have been established, phylogenetic analyses are the basis for investigating further questions, such as: What occurred during evolution to produce the sequences found today. This may mean checking for evidence of selective pressure, examining the possibility of gene duplication or loss, investigating hints for horizontal gene transfer, or more. Below discusses in more detail the various steps and aims that occur during a typical cross-species phylogenetic analysis.

Gene Clustering Often cross-species comparative studies focus not on comparing the entirety of the genomic sequence, but rather, to make the problem more tractable, exclude intergenic regions and focus on comparing only the genes. Generally, the first thing to determine is which genes are likely to have descended from a common ancestor. This can be done simply by sequence-based clustering. First a preliminary graph with links between homologous genes is constructed. This is commonly based on homology identified by the Basic Local Alignment Search Tool (Altschul et al., 1997). Then, clusters are identified within the graph using an algorithm such as Markov chain clustering

(MCL) (Van Dongen, 2000). Such sequence-based methods can be further augmented with the addition of information such as expected phylogenetic distance (when clustering individual genes) (Emms and Kelly, 2015), or inclusion of additional information such as the ordering of the genes (synteny) (Lechner et al., 2014). Ultimately gene clustering produces a list of orthogroups and the genes within each orthogroup, which can be used for further analyses.

Alignment Nucleotide sequences can under several types of mutations including nucleotide substitution, insertions and deletions. Such sequences need to be aligned before they can be properly compared. For two sequences an optimal global alignment (for a given scoring scheme) can be obtained with the dynamic programming algorithm Needleman-Wunsch (Needleman and Wunsch, 1970). In practice, many tools are available for multiple sequence alignment, and range from tools such as ClustalW (Thompson et al., 2002), which very aggressively tries to find any matching positions it can between sequences, to tools such as PRANK (Löytynoja, 2014), which uses bootstrapping to repeatedly perform alignments with some data omitted and when finished only reports robustly aligned regions as aligned. Multiple sequence alignment can be performed in either protein space, which is suitable for aligning distantly related sequences, or in nucleotide space when there is a need to differentiate very similar sequences.

Tree reconstruction After multiple sequence alignment, the next step in many comparative phylogenetic analyses is reconstructing the phylogenetic tree of the given sequences. Phylogenetic trees are often reconstructed based on maximum parsimony (e.g. (Plotree and Plotgram, 1989)), maximum likelihood (e.g. Guindon et al. (2010) Stamatakis (2006)), or Baysian inference (e.g. Ronquist and Huelsenbeck (2003), Drummond and Rambaut (2007)) methods. Challenges include propagation of earlier errors (tree reconstruction can be sensitive to misalignments (Ogden and Rosenberg, 2006)); and reconstruction specific issues such as long branch attraction (Kolaczkowski and Thornton, 2009). Accurate reconstruction benefits from densely packed data with only relatively small phylogenetic distances between individual sequences. Methods such as repeated bootstrapping with some data omitted can be used to assign confidence to specific branches of the consensus tree. The resulting phylogenetic trees can be used to answer questions ranging from which species or genes are most closely related, to the relative timings of speciation and duplication events. Further, trees can be used to identify horizontal gene transfer events.

Selective pressure A frequent question about evolution is which genes were evolving under negative selection, where the sequences are conserved, neutral selection where the

sequences may drift over time, or positive selection where the sequences are under pressure to change rapidly or in a specific manner. Some portions of the genomic sequence are generally more conserved than others. For instance, protein-coding regions are generally more conserved than non-coding regions, and within coding regions, non-synonymous substitutions that change the amino acid (often in the first and second codon positions) are rarer than synonymous substitutions that do not result in an amino acid change (often in the third codon position) (reviewed in Booker et al. (2017)). Synonymous substitutions are generally assumed to be under neutral selection and occurring at a clade-specific but fairly constant rate across time; they can be used in combination with dated fossil evidence for dating speciation and duplication events. The ratio of non-synonymous substitution rate (dN) to synonymous substitution rate (dS) can be used to estimate whether genes are under positive or negative selection in genes diverging 10s of millions of years ago (Obbard et al., 2012). Positive selection testing with dN/dS is sensitive to the accuracy of the phylogenetic tree and further to the multiple sequence alignment, with sequencing errors or misalignments potentially causing a false signal of positive selection (Mallick et al., 2009).

Gene duplication Another frequent question about evolution is when and how genes were duplicated. The relative timing of a gene duplication can be determined directly from an accurate phylogenetic tree. Additional information can be used to understand how a gene duplication occurred. For instance, the duplication of many genes in the same species at a specific time is evidence for a whole genome duplication event. Other cases can also indicate how duplication occurred, for instance when other genes in the same region were duplicated, when the intergenic sequence around the gene was duplicated or when the intron structure was conserved (Qiao et al., 2018). Gene duplication is often thought to reduce the selective pressure on the duplicated paralogs and thereby allow them more flexibility to evolve new or specialized functions (Lawton-Rauh, 2003).

Challenges

Successful cross species comparison relies heavily on both the quality and comparability of the input data. Generally, differences in protein abundance or protein structure are of interest. However, an accurate proteome is dependent on extended prior steps, including sequencing the genome, the transcriptome, and combining both of the above with de novo gene prediction and homology mapping to predict gene models.

Quality and Consistency in Data Generation Ideally, methods of data generation and analysis would be held constant, however, the sheer cost and effort that go into sequencing a genome make re-sequencing or even re-annotating a comparative rarity. Further, sometimes differences between species, e.g. genome size, necessitate different

sequencing or analysis methods. Therefore researchers must be able to work with the potentially heterogeneous data. Many cross-species analyses theoretically have more statistical or conclusive power the more data is included. This often leads researchers to take in as many related sequences as can be acquired, even when some of the additional sequences are of lower quality. One reason sequences of comparable quality may not be available is limitations on the input material. For instance, the first draft of the Neanderthal genome was sequenced largely from 40,000-year-old femur bone fossils, which produced a very low-quality draft (Green et al., 2010) assembly compared to a subsequent sequencing project that started with a 130,000-year-old fossilized toe bone (Prüfer et al., 2014). When projects seek to sequence RNA instead of DNA, the within-organism variance can make obtaining a specific tissue incredibly difficult. For instance, there is a deal of interest in understanding the differentiation of neurons, yet separating and sorting individual neurons is a difficult task, often resulting in working with tiny amounts of material, degraded material, or both (Lacar et al., 2016).

In both of the above examples, small amounts of input material can be amplified by PCR with generic primers. However, this will also amplify any DNA present, including trace amounts of contaminants. Contamination can also occur for a variety of other reasons from simple mistakes, the presence of pestilent species, to unavoidable cases, such as, e.g. sequencing species in strict symbiotic relationships where one species cannot be cultivated without the presence of the another. Contamination, depending upon amount, can result in the need for additional filtering, can turn a single genome project into a meta-genome project, or in extreme cases result in reads from the target genome being rare amongst all reads obtained. Unsurprisingly, read-level contaminations can find their way into the final assembly. Such contaminations can then cause trouble for downstream analyses, potentially appearing as outlying sequences or masquerading as horizontal gene transfer (Lercher and Pál, 2007).

In addition to material limitations, continuously developing sequencing technologies mean that publicly available sequences for different species are frequently based on entirely different technologies such as short or long read sequencing, with lower or higher error rates, with random error or consistent bias (as discussed in more detail above).

The heterogeneity introduced during data generation should be taken into consideration during further analyses.

Quality and consistency in data processing Sequencing DNA or RNA may form the basis, yet it is only the start of how subsequent analyses may differ. The choice of tool for sequence assembly depends upon the sequencing technology and characteristics (e.g. repeat content and genome size) of the target.

The method used to determine the location and structure of genes in a genome assembly is a further potential source of heterogeneity in the data. There are a variety of

tools for the base *de novo* gene calling program, and many must be specifically trained on the target or a closely related species (Korf, 2004, Stanke and Waack, 2003). This is further compounded by differing availability of extrinsic data (homology, RNASeq). Researchers face additional choices such as whether to call alternative splicing events or allow partial gene models. Each genome project is followed by a near-unique gene calling process of collecting and using extrinsic data to best define genes on the given genome. However, some large-scale attempts at consistency exist, as large databases such as NCBI may run a comparable gene calling pipeline fed with large amounts of extrinsic homology data on uploaded genomic sequences (Souvorov et al., 2010).

In summary, genomes will vary in their completeness, continuity, error rate, and error types. Similarly, the associated provided gene sequences will have both omissions and mis-called genes; and the frequency of these will be very nearly genome specific. A researcher hoping to use publicly available genomes and associated gene sequences must be able to perform their analyses in a way that is robust to such differences.

Specific challenges in cross-species analysis The afore mentioned differences in availability, generation, and processing of sequencing data can lead to a variety of challenges during cross-species phylogenetic comparisons. Finding *evidence of absence* is a conundrum, which inherently makes any inferences of gene loss challenging. While a few target gene loss cases can be back-checked with further wet lab analyses, evidence of large-scale deletions must be interpreted with care so as not to mistake a relatively incomplete genome assembly for one that has undergone extensive gene loss.

Not only can assembly errors masquerade as gene loss, but multiple alleles may be assembled separately, masquerading as gene duplication. Inclusion of *de novo* transcriptome assemblies greatly exacerbates this problem as different alleles, sequencing errors or simply splice isoforms can masquerade as paralogs originating from gene duplication.

The density and phylogenetic distribution of available sequences affect the reconstruction of phylogenetic trees. Phylogenetic tree reconstruction often suffers from long branch attraction, whereby distantly related sequences are incorrectly grouped together in the resulting phylogenetic tree. This can result from the introduction of outlying sequences, or simply from a lack of phylogenetic resolution. This causes a trade-off for the researcher, who risks inclusion of more outliers but gains resolution with the incorporation of more species.

In addition to any issues caused by long branch attraction, outlying sequences can cause misalignments. This can happen for sequences that are completely mis-assigned, but can also be caused by outliers such as extremely distant sequences, chimeric sequences, or sequences with minor structural errors such as inclusion of some UTR or intron sequences into the final CDS prediction. Generally, global alignment tools are intended for use on truly homologous sequences, and many will force the best alignment

possible, even when this alignment is improbable. Misaligned sequences have low identity with the other sequences. Further misaligned sequences lack typical characteristics of aligned homologous sequences, such as more conserved non-synonymous than synonymous sites. Misalignments can harm or lead to erroneous conclusions in almost any analysis that requires an alignment as input.

One approach to avoid further problems caused by outlying sequences is to filter them out. This is however a non trivial task as comparative analyses frequently work with 10s of thousands of gene families (or orthogroups). Generation of automated rules for detection of outlying sequences in orthogroups is challenging.

1.2.3 Contributions of *Manuscript 1* towards Facilitating Cross-Species Comparative Studies

The first manuscript in this thesis focusses on implementing an early step in a cross-species comparison to improve the robustness and reliability of the down stream analyses. Specifically, *Manuscript 1* describes a tool, seqSieve, which filters outlying sequences from a multiple sequence alignment. This tool helps maximize the useful information that can be obtained from the data, while avoiding inclusion of data that is too divergent to contribute.

Manuscript 1

This manuscript describes seqSieve, a high-throughput, customizable software for removing outlying sequences from multiple sequence alignments for the purpose of quality control in large-scale studies. Such outliers can cause gaps or misaligned regions in a multiple sequence alignment which can have detrimental effects, such as information loss or bias introduction, on downstream analyses. Outlying sequences are detected by a customizable scoring system that factors in (unique) gaps, (unique) insertions, mismatches, or a weighted combination thereof.

We then compare seqSieve to several other filtering tools, of which it performs respectably in terms of maximizing ungapped sum of pairs. Analysis of differences between the tools indicated that seqSieve's more inclusive and robust scoring system is more successful in detecting outlying sequences.

This tools is particularly suited for use in a pipeline for detecting positive selection on a genome wide scale. Starting from primary transcripts for each species of interest, such a positive selection detecting pipeline could look like the following: Orthogrouping, multiple sequence alignment, seqSieve, phylogeny reconstruction, and finally testing for positive selection per site. As multiple sequence alignment occurs early in a pipeline, any errors are propagated and amplified during later steps frequently leading to false

positive signals of positive selection (Fletcher and Yang, 2010). Further, the widely used PAML codeml tool does not properly handle gapped regions (Yang, 2007).

1.3 C_4 Photosynthesis as a Complex Trait

The majority of organic carbon on planet earth was fixed in a light-driven reaction known as photosynthesis. While the majority of species in the Plant Kingdom, Viridiplantae, use the "classic" photosynthetic pathway, known as C_3 photosynthesis, a disproportionate amount of the earth's net primary productivity (up to 23%)(Still et al., 2003) comes from the approximately 3% of Viridiplantae species that have evolved an add-on to the classic C_3 pathway (Kellogg, 2013), namely a complex trait known as C_4 photosynthesis (Hatch and Slack, 1968). C_4 photosynthesis is more efficient in certain (hot, arid) environments than the C_3 pathway (Amthor, 2010).

The most well known part of the C_4 photosynthetic trait is a biochemical pump that increases the amount of CO_2 in the vicinity of the central carbon fixing enzyme Rubisco. That said, efficient C_4 photosynthesis requires more extensive adjustments in anatomy and metabolism of the plant.

The efficiency advantages of C_4 photosynthesis come from the suppression of the wasteful photorespiration process. All plants ultimately fix CO_2 via Rubisco, which results in the immediate production of 3-phosphoglycerate (3-PGA). Every sixth molecule of 3-PGA produced can be used in sugar and ultimately biomass production, while the rest must be recycled to form the precursors in the Calvin-Benson-Bassham-Cycle (CBBC). However, Rubisco cannot always discriminate effectively between CO_2 and O_2. When O_2 is fixed in the place of CO_2, a molecule of 2-phosphoglycolate (2-PG) is produced which is toxic for the plant. Recycling 2-PG results in a loss of both CO_2 and energy (reviewed in Hagemann et al. (2016)), and in some conditions photorespiration can reduce overall photosynthetic efficiency by 30% (Sharkey, 1988, Zhu et al., 2004).

By concentrating pre-fixed CO_2 around Rubisco, C_4 photosynthesis suppresses the wasteful fixation of O_2 and thereby reduces photorespiration and increases photosynthetic efficiency.

1.3.1 C_4 Biochemistry

In C_4 plants, Rubisco and the CBBC are found in an interior compartment, most frequently, bundle sheath tissue. The biochemical pump transports carbon from the exterior mesophyll tissue into the bundle sheath tissue.

The biochemical pump fixes carbon onto a 3-carbon scaffold in the mesophyll, producing a 4-carbon organic acid, which travels to the bundle sheath, where it is decarboxylated, releasing a molecule of CO_2. The 3-carbon scaffold then travels back to the mesophyll to start the cycle again (Hatch and Slack, 1968).

For the biochemical pump to work efficiently, it requires anatomical alterations, for instance, to keep mesophyll and bundle sheath tissues in close proximity.

1.3.2 Anatomical Modifications in C_4 Plants

Anatomical modifications in C_4 plants include reduced distance between mesophyll and bundle sheath tissues, more space for Rubisco and the CBBC in the bundle sheath.

One of the most obvious anatomical changes is the increased vein density found in many C_4 species. Specifically, the veins become spaced tightly enough that the cells occur in a specific ratio of vein:BS:M:M:BS:vein, so that every mesophyll cell is directly bordering a bundle sheath cell. This minimizes the distance metabolites are required to diffuse between cells (reviewed in Sage (2004)).

Another anatomical change is the enlarged size and increased organellar content of the bundle sheath cells. The cell layer in C_3 species, which is recruited to become the bundle sheath in C_4 species, is not always the same, but is generally derived from a small, vein-accompanying tissue without photosynthetic function (Edwards and Voznesenskaya, 2010). Additional changes may include modifications of thylakoid structure, lignification of the bundle sheath cell wall and plasmodesmata (Sage, 2004).

1.3.3 Convergent Evolution of C_4 Photosynthesis

Despite its complexity, the C_4 pathway has evolved at least 66 times (Sage et al., 2012). Some of what makes this possible is likely the presence of all the genes in the C_4 cycle in all plants, relying on co-option from an already existing pool of genes and not novel gene evolution. However, this still doesn't explain how the sheer complexity can be repeatedly obtained. This has often led to the hypothesis that C_4 evolution may require fewer changes than is immediately apparent via the recruitment of master regulator(s) or pathway(s) (Westhoff and Gowik, 2010). Many of the C_4 origins occur in clusters and some include intermediate species that are not yet C_4, but already have a partial carbon concentrating mechanism in place. Such clades provide an opportunity to test hypothesis in the evolutionary origin of C_4 photosynthesis (Gowik et al., 2011). These clades with C_3-C_4 intermediate species show evidence for a gradual evolution of C_4 photosynthesis, in which an early step is the establishment of a cycle for scavenging the CO_2 lost during photorespiration, now known as C_2 photosynthesis (Sage et al., 2014). After this, the gradual up-regulation of C_4 enzymes and gradual localization to the bundle sheath appears to provide continuous fitness benefits Heckmann et al. (2013). Despite the smooth evolutionary slope once initiated, strong clustering in C_4 origins hints at preconditioning steps that make it more probable in some lineages than others. These are thought to include e.g. gene duplication and increased vein spacing.

Using cross-species comparisons to understand the differences between C_3, C_4, and intermediate species remains an active area of research.

1.3.4 Contributions of Manuscripts *2*, *3*, *4*, and *5* towards Understanding the Complexity of C_4 Photosynthesis

Despite our extensive ability to describe the complexity of C_4 plants, we still don't fully grasp the molecular basis of anatomical and metabolic changes, which remains a major hurdle in the way towards engineering C_4 photosynthesis into C_3 plants.

Three manuscripts in this thesis try to address this knowledge gap by leveraging large data sets to better understand the molecular basis of some of the complexity of C_4 photosynthesis. Specifically, they employ high-throughput sequencing to simultaneously measure the abundance of all mRNA transcripts in each target sample. All three manuscripts are comparative (between species or tissues) and together expand our knowledge of C_4 photosynthesis on the transcriptional level.

In *Manuscript 4* (see 3.1) mature leaf tissue of closely related C_3 and C_4 Cleomaceae species is compared. In *Manuscript 5* (see 3.2), the comparison of these Cleomaceae species is expanded to look at a broad sampling of plant tissues with a focus on leaf development. Finally, *Manuscript 2* and *Manuscript 3* (see 2.2, 2.3) compare expression between mesophyll and bundle sheath tissues in the developing leaf of a C_4 maize plant.

Manuscript 4

In *Manuscript 4*, high-throughput 454 mRNA sequencing was used to quantitatively compare two closely related C_3 and C_4 Cleomaceae species, to compile and contrast transcription profiles and identify candidate genes related to the C_4 pathway. Specifically, this manuscript looks at gene expression in mature leaf tissue of *Gynandropsis gynandra* (previously referred to as *Cleome gynandra*), a C_4 plant, and *Tarenaya hassleriana* (previously referred to as *Cleome spinosa*), a C_3 plant. As anatomical differences are setup early in development, prior to this manuscript it was unknown whether extensive differential regulation between C_3 and C_4 tissues was to be expected beyond the core biochemical pump. Comparative analysis revealed over 600 differentially expressed genes between the species. Such differentially expressed candidates found in this study include genes associated with transport, chloroplast movement and expansion, plasmodesmatal connectivity, cell wall modification and transcription factors.

As a pioneering work in C_4 cross-species transcriptome comparison, this manuscript also had to establish that its results were robust to the methods used. The Cleomaceae species proved to be phylogenetically close enough to the model plant species *Arabidopsis thaliana* to allow for cross-species read mapping. This simplified the analysis in some ways by avoiding some of the troubles of *de novo* transcriptome assemblies. Similarly it

circumvented the issue of establishing orthologous relationships between the Cleomaceae species as mapping both species to *A. thaliana* was sufficient. However, much effort was put into validation and optimization of the method and to establish it was robust and allowed for biologically meaningful comparison. Two read mapping tools BLAST and BLAT (BLAST Like Alignment Tool) were compared throughout, and output and thresholds for both tools were optimized.

Manuscript 5

In *Manuscript 5*, the comparison of *Gynandropsis gynandra* (C_4) and *Tarenaya hassleriana* (C_3) from *Manuscript 4* is expanded to include a more holistic view of the plant and in particular to capture early leaf development, so as to understand the setup of the anatomy that is critical to C_4 photosynthesis. Indeed, a delay in tissue differentiation could be linked to denser venation in the C_4 species. Examination of transcripts near the center of the delayed expression module, led to identification of candidate regulatory genes. Similarly, differential expression indicated increased endoreduplication in the C_4 species. Endoreduplication is the additional duplication of DNA without cell division, and is frequently found in enlarged cell types. Here, further analyses could confirm the increased nuclear size in a subset of cells, and microscopy could identify enlarged nuclei in the bundle sheath cells that are enlarged in the C_4 species. Further, this work provided the community with a compendium of differences between the C_3 and C_4 species during development and in a variety of tissues which will continue to support future research.

As above, coming to the biological conclusions first required a clean bioinformatics analysis, including the summarizing and extraction of precise information from 108 RNAseq samples totalling over 2.2 billion reads. As much more RNAseq data and the *T. hassleriana* genome were now available, further comparison between the bias introduced in working with a cross-species mapping vs a *de novo* assembly was performed. Cross species mapping obtained a lower specificity in mapping reads between closely-related paralogs, however, expression by gene family remained very comparable. In contrast, the *de novo* assembly suffered from frequent missing, fragmented or otherwise inaccurate transcripts. Therefore, the central analysis was performed based on cross-species mapping of reads from both species to the newly available *T. hassleriana* genome. Even after quantification and basic differential expression was complete, more in depth bioinformatic analyses were required to interpret the large dataset. This included three different clustering methods to e.g. confirm the clustering of replicates or check that tissues were well-matched between species. Clustering and summarizing based upon functional annotation were used to obtain an overview and description of the dataset, which led to further, more-targeted tests. In particular, the transcriptional modules that

showed a delay in the C_4 developmental series could be identified in both the k-means and the hierarchical clustering.

Manuscript 2

In this manuscript, we enriched the two most prominently C_4 -specialized tissues, BS and M, along a developing maize leaf and sequenced their transcriptome. We compared our results with a variety of other BS and M separation studies in maize or related species. Inter-study comparison identified both weaknesses and strengths in particular methods, and highlighted the strength of the meta-comparison. Specifically, comparison allowed for the identification of chronic method-related bias in separation techniques leading to different levels of 3' bias (an indicator of RNA degradation). Some studies showed severe, yet constant bias, but more worryingly some other studies showed minimal but uneven bias. Here, higher degradation in the M led to erroneous conclusions. Besides identifying weaknesses in studies, the meta-comparison allowed us to draw conclusions that would have been hard to support based on any individual study. For instance, both maize developmental studies indicated that protein synthesis related transcripts switched from generally BS-specific early in development to generally M-specific in mature maize tissue, which likely supports the high expression of the M-specific PSII. There was more information compiled here than could be readily analyzed in a single paper. However, this could be a useful resource for any researcher interested in BS- and M-specific transcripts in grasses, so we developed and provided a website for visualizing and summarizing the dataset. The meta-comparison relied heavily on bioinformatics analyses. Reads were mapped with the fast, spliced short read mapper TopHat to the respective genome of each species *Zea mays*, *Panicum virgatum* and *Setaria viridis*. Clustering was used to gain an overview and determine how to best compare the studies. Differential expression analysis was performed for each study, including an R package, contamDE, to provide a more robust analysis with the partial enrichment achieved in this study. Homologous relationships between species were established with BLAST. Further, non-standard analyses were necessary, for instance, it was necessary to modify the Picard Tools software to compute 3' bias on a gene-by-gene basis. The website visualization and community tool was developed as part of this study.

Manuscript 3

The manuscript deals with the question of how changes in gene expression following gene duplication have contributed to the evolution of C_4 photosynthesis. Gene duplication has long been hypothesized to facilitate the evolution of complex traits, but most of the evidence for this is based on timing and anecdotal studies. Here we perform a

genome-wide phylogenetic and gene expression analysis to elucidate whether gene duplication in the grass lineage appears to be linked to the evolution of the complex trait C_4 photosynthesis.

We performed large scale pairwise and gene family-wise transcriptional and phylogenetic analyses. The gene family-wise analyses involved performing many multiple sequence alignments, filtering these with seqSieve (as described in *Manuscript 1*), reconstructing phylogenetic trees, and testing for positive selections via PAML/codeml.

Overall we could link gene duplication level to increased expression divergence, tissue specificity, and gain or loss of a photosynthetic expression pattern. We found tentative evidence linking expression divergence to positive selection. In particular, the known core C_4 paralogs showed unusually high expression divergence, tissue specificity, correlation to the photosynthetic expression pattern and positive selection. Beyond the core C_4 genes, functional categories related to C_4 photosynthesis were enriched in paralogs with key divergence patterns. For instance, transcripts of ATP-consuming photosynthetic paralogs diverge in expression between mature M and BS tissue.

Bibliography

Aidoo, M., Terlouw, D. J., Kolczak, M. S., McElroy, P. D., ter Kuile, F. O., Kariuki, S., Nahlen, B. L., Lal, A. A., and Udhayakumar, V. (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. *The Lancet*, 359(9314):1311–1312.

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 research*, 25(17):3389–3402.

Amthor, J. S. (2010). From sunlight to phytomass: on the potential efficiency of converting solar radiation to phyto-energy. *New Phytologist*, 188(4):939–959.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., et al. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462.

Booker, T. R., Jackson, B. C., and Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC biology*, 15(1):98.

Darwin, C. (1859). On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. *London*, 1:859.

Dessinioti, C., Stratigos, A. J., Rigopoulos, D., and Katsambas, A. D. (2009). A review of genetic disorders of hypopigmentation: lessons learned from the biology of melanocytes. *Experimental dermatology*, 18(9):741–749.

Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1):214.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95.

Edwards, G. E. and Voznesenskaya, E. V. (2010). C4 photosynthesis: Kranz forms and single-cell C4 in terrestrial plants. In *C4 photosynthesis and related CO2 concentrating mechanisms*, pages 29–61. Springer.

Ekblom, R. and Wolf, J. B. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications*, 7(9):1026–1042.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology*, 16(1):157.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., et al. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one*, 7(11):e47768.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512.

Fletcher, W. and Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular biology and evolution*, 27(10):2257–2267.

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., and Loeb, L. A. (2014). Accuracy of next generation sequencing platforms. *Next generation, sequencing & applications*, 1.

Gehring, W. J. (2005). New perspectives on eye development and the evolution of eyes and photoreceptors. *Journal of Heredity*, 96(3):171–184.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. *Science*, 274(5287):546–567.

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(6):2087–2105.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., et al. (2010). A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3):307–321.

Hagemann, M., Weber, A. P., and Eisenhut, M. (2016). Photorespiration: origins and metabolic integration in interacting compartments. *Journal of experimental botany*, 67(10):2915.

Hatch, M. D. and Slack, C. R. (1968). A new enzyme for the interconversion of pyruvate and phosphopyruvate and its role in the C4 dicarboxylic acid pathway of photosynthesis. *Biochemical Journal*, 106(1):141–146.

Heckmann, D., Schulze, S., Denton, A., Gowik, U., Westhoff, P., Weber, A., and Lercher, M. (2013). Predicting C4 Photosynthesis Evolution: Modular, Individually Adaptive Steps on a Mount Fuji Fitness Landscape. *Cell*, 153(7):1579 – 1588.

International Barley Genome Sequencing Consortium (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 491(7426):711.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860.

Kellogg, E. A. (2013). C4 photosynthesis. *Current Biology*, 23(14):R594–R599.

Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J. M., Lee, H.-A., Seo, E., Choi, J., Cheong, K., Kim, K.-T., et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature genetics*, 46(3):270.

Kolaczkowski, B. and Thornton, J. W. (2009). Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PLoS One*, 4(12):e7891.

Korf, I. (2004). Gene finding in novel genomes. *BMC bioinformatics*, 5(1):59.

Lacar, B., Linker, S. B., Jaeger, B. N., Krishnaswami, S. R., Barron, J. J., Kelder, M. J., Parylak, S. L., Paquola, A. C., Venepally, P., Novotny, M., et al. (2016). Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nature communications*, 7:11022.

Laver, T., Harrison, J., O'neill, P., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D. J. (2015). Assessing the performance of the oxford nanopore technologies minion. *Biomolecular detection and quantification*, 3:1–8.

Lawton-Rauh, A. (2003). Evolutionary dynamics of duplicated genes in plants. *Molecular phylogenetics and evolution*, 29(3):396–409.

Lechner, M., Hernandez-Rosales, M., Doerr, D., Wieseke, N., Thévenin, A., Stoye, J., Hartmann, R. K., Prohaska, S. J., and Stadler, P. F. (2014). Orthology detection combining clustering and synteny for very large datasets. *PLoS One*, 9(8):e105015.

Lercher, M. J. and Pál, C. (2007). Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Molecular biology and evolution*, 25(3):559–567.

Li, S., Zhu, S., Jia, Q., Yuan, D., Ren, C., Li, K., Liu, S., Cui, Y., Zhao, H., Cao, Y., et al. (2018). The genomic and functional landscapes of developmental plasticity in the American cockroach. *Nature Communications*, 9(1):1008.

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. In *Multiple sequence alignment methods*, pages 155–170. Springer.

Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome research*, 19(5):922–933.

Mikheyev, A. S. and Tin, M. M. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular ecology resources*, 14(6):1097–1102.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Obbard, D. J., Maclennan, J., Kim, K.-W., Rambaut, A., O'Grady, P. M., and Jiggins, F. M. (2012). Estimating divergence dates and substitution rates in the Drosophila phylogeny. *Molecular Biology and Evolution*, 29(11):3459–3473.

Ogden, T. H. and Rosenberg, M. S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Systematic biology*, 55(2):314–328.

Piel, F. B., Patil, A. P., Howes, R. E., Nyangiri, O. A., Gething, P. W., Williams, T. N., Weatherall, D. J., and Hay, S. I. (2010). Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature communications*, 1:104.

Plotree, D. and Plotgram, D. (1989). PHYLIP-phylogeny inference package (version 3.2). *cladistics*, 5(163):6.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43.

Qiao, X., Yin, H., Li, L., Wang, R., Wu, J., Wu, J., and Zhang, S. (2018). Different modes of gene duplication show divergent evolutionary patterns and contribute differently to the expansion of gene families involved in important fruit traits in pear (Pyrus bretschneideri). *Frontiers in plant science*, 9:161.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Sage, R. F. (2004). The evolution of C4 photosynthesis. *New phytologist*, 161(2):341–370.

Sage, R. F., Khoshravesh, R., and Sage, T. L. (2014). From proto-Kranz to C4 Kranz: building the bridge to C4 photosynthesis. *Journal of experimental botany*, 65(13):3341–3356.

Sage, R. F., Sage, T. L., and Kocacinar, F. (2012). Photorespiration and the evolution of C4 photosynthesis. *Annual review of plant biology*, 63:19–47.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467.

Sharkey, T. D. (1988). Estimating the rate of photorespiration in leaves. *Physiologia Plantarum*, 73(1):147–152.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135.

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., et al. (2016). Long-read sequencing and de novo assembly of a Chinese genome. *Nature communications*, 7:12065.

Souvorov, A., Kapustin, Y., Kiryutin, B., Chetvernin, V., Tatusova, T., and Lipman, D. (2010). Gnomon–NCBI eukaryotic gene prediction tool. *National Center for Biotechnology Information*, pages 1–24.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Stanke, M. and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(suppl_2):ii215–ii225.

Still, C. J., Berry, J. A., Collatz, G. J., and DeFries, R. S. (2003). Global distribution of C3 and C4 vegetation: Carbon cycle implications. *Global Biogeochemical Cycles*, 17(1):6–1–6–14. 1006.

Thompson, J. D., Gibson, T., Higgins, D. G., et al. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics*, pages 2–3.

Van Bakel, H., Stout, J. M., Cote, A. G., Tallon, C. M., Sharpe, A. G., Hughes, T. R., and Page, J. E. (2011). The draft genome and transcriptome of Cannabis sativa. *Genome biology*, 12(10):R102.

Van Dongen, S. M. (2000). *Graph clustering by flow simulation*. PhD thesis.

Walker, F. O. (2007). Huntington's disease. *The Lancet*, 369(9557):218–228.

Westhoff, P. and Gowik, U. (2010). Evolution of C4 photosynthesis—looking for the master switch. *Plant Physiology*, 154(2):598–601.

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., Xu, J., Zheng, X., Ren, L., Wang, G., et al. (2014). Genome sequence and genetic diversity of the common carp, Cyprinus carpio. *Nature Genetics*, 46(11):1212.

Xu, X., Zhou, Z., Wang, X., Kuang, X., Zhang, F., and Du, X. (2003). Four-winged dinosaurs from China. *Nature*, 421(6921):335.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591.

Zhu, X.-G., Portis, A., and Long, S. (2004). Would transformation of C3 crop plants with foreign Rubisco increase productivity? A computational analysis extrapolating from kinetic properties to canopy photosynthesis. *Plant, Cell & Environment*, 27(2):155–165.

Zimin, A. V., Puiu, D., Hall, R., Kingan, S., Clavijo, B. J., and Salzberg, S. L. (2017). The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *Gigascience*, 6(11):gix097.

Chapter 2

First Author Manuscripts

2.1 Manuscript 1: seqSieve – Removing Outliers from Multiple Sequence Alignments

Overview

Title: seqSieve – Removing Outliers from Multiple Sequence Alignments
Authors: Janina Maß, Alisandra K. Denton, and Martin J. Lercher
Submitted to PeerJ

First authorship

Contributions

- Project design
- Python implementation
- Software testing
- Data interpretation
- Writing manuscript

seqSieve – Removing gap-inducing sequences from multiple sequence alignments

Janina Maß[1], Alisandra K. Denton[2], Martin J. Lercher[3]

February 2018

[1]Institute for Theoretical and Quantitative Biology, Heinrich Heine University, 40225 Düsseldorf, Germany

[2]Institute for Plant Biochemistry, Heinrich Heine University, 40225 Düsseldorf, Germany

[3]Institute for Computer Science and Department of Biology, Heinrich Heine University, 40225 Düsseldorf, Germany

Abstract

The inadvertent inclusion of poorly matching sequences can cause gaps or misaligned regions in a multiple sequence alignment and detrimentally affect downstream analyses. Few methods exist to remove such disruptive sequences. Here, we present seqSieve, a high-throughput, customizable, and user-friendly Python application that iteratively removes the most outlying sequences in an alignment, realigning the remaining sequences in each iteration. seqSieve detects outlying sequences by a user-adjustable assessment of mismatches and gaps introduced by the sequence. Compared to alternative filtering tools, seqSieve achieves a higher alignment quality, measured as the number of pairwise nucleotide or amino acid matches in ungapped regions. seqSieve is freely available from http://pypi.python.org/pypi/seqSieve.

1 Introduction

As the availability of sequence information and the scope of sequencing projects increases, so does the potential to introduce errors through the inclusion of non-fitting or partial information. During selection of homologous sequences for phylogenetic or comparative analyses, sequences may be wrongly included due to the homology of a small region; the non-matching rest of the sequence may then induce alignment errors.

More sequence data and higher phylogenetic resolution may sometimes lead to more complete and confident results [1, 2]; however, adding sequences that are only partially homologous can ruin an analysis by creating a poor alignment, typically signified by many gaps. Corresponding misalignments can lead to less accurate phylogenies [3] and false discoveries in positive selection testing [4].

While a true insertion or deletion is part of the phylogenetic signal, excessive gaps are frequently a sign of misalignment. Gaps introduced by non-matching sequences can bias topologies, as gaps are frequently treated simply as ambiguous data [5, 6], and can lead to loss of information in tools that exclude gapped columns completely [7, 8]. Removing confounding sequences or portions of an alignment can greatly improve accuracy in phylogeny reconstruction or selection testing [9–13].

While various tools facilitate filtering poorly aligned columns in an alignment, we found few tools designed to filter out non-matching sequences.

These included GUIDANCE [14], which best controlled error-rate in a comparison of three filtering tools [15]; OD-seq [16], which removes sequences with an anomalously large distance to the remaining sequences; and MaxAlign [13], which removes gap-causing sequences after a multiple sequence alignment (MSA) and optimizes the number of nucleotides or amino acids in ungapped portions of the alignment (ungapped alignment area, or UAA; eq 1.). Katoh and Toh [17] suggest that combining the iterative removal of sequences performed by MaxAlign with iterative realignment may improve results, and several studies have opted for a final realignment after using MaxAlign, or even manual iteration [18–20]. The tool presented here, seqSieve, optimizes the UAA while iteratively removing outlying sequences from an MSA and realigning the remaining sequences.

2 Methods

The first criterion used in this study was the ungapped alignment area (UAA), defined as

$$UAA := r \cdot c' \quad , \tag{1}$$

where r is the number of alignment rows (nucleotide or amino acid sequences) and c' is the number of alignment columns devoid of gaps. The second criterion used in this study was the ungapped sum of pairs (USoP), defined as the number of identical pairs of nucleotides or amino acids found in ungapped columns,

$$USoP := \sum_{p=1}^{c'} \sum_{i=1}^{r-1} \sum_{j=i+1}^{r} s(a_{ip}, a_{jp}), \tag{2}$$

$$s(a_{ip}, a_{jp}) = \begin{cases} 1, & a_{ip} = a_{jp} \\ 0, & a_{ip} \neq a_{jp} \end{cases}, \tag{3}$$

where a_{ip} is the nucleotide or amino acid in row i and column p of the alignment matrix [21].

seqSieve is implemented in Python and can use either of the alignment programs MAFFT [22] or PRANK [23] for MSA. seqSieve is designed for high-throughput use and thus runs from a command line interface, accepting either single alignments or a folder containing alignments as input. As iterative realignment is time consuming, seqSieve is implemented to run multiple MSA optimizations on multiple cores in parallel. seqSieve is available from the Python Package Index (http://pypi.python.org/pypi/seqSieve) and has been tested under Linux and OS X. seqSieve iteratively removes the most outlying sequence(s), realigns the remaining sequences, and calculates the changed UAA. The final alignment returned by seqSieve is the one for which the UAA can not be increased further through the removal of another sequence.

seqSieve has a customizable scoring system, allowing outlying sequences to be detected by (unique) gaps, (unique) insertions, mismatches, or a weighted combination of these. Whether a region has gaps or insertions is determined by majority rule. By default, sequences are penalized for gaps and insertions by an amount proportional to the percentage of ungapped

and gapped sequences, respectively. For example, in an alignment of 10 sequences, if only sequence A had a gap in column 1, it would be penalized by 0.9; if sequences A, B, and C had a gap in column 2 while the remaining sequences did not, they would each be penalized by 0.7. The sequence(s) with the maximum penalty are removed, and the remaining sequences are realigned and scored again. seqSieve records and reports statistics of the initial alignment and following iterations, and provides a summary graph; for an example, see Figure 1 A.

We downloaded a test dataset from Ensembl v75 [24], consisting of all 1,484 orthologous groups for which the genome of each of eight animal species contained exactly one ortholog; the animals contributing to this dataset were *Bos taurus*, *Drosophila melanogaster*, *Gallus gallus*, *Mus musculus*, *Takifugu rubripes*, *Caenorhabditis elegans*, *Equus caballus*, and *Homo sapiens*. Prior to filtering, the data was aligned with MAFFT (Supplemental Dataset 1). To compare the performance of seqSieve to alternative approaches, we executed GUIDANCE, MaxAlign v1.1, OD-seq, and seqSieve with default parameters.

3 Results

Compared to the original (raw) MAFFT alignments, seqSieve was able to increase the UAA in 54% of the orthologous groups, with an average improvement of 673 letters (13% of the average initial UAA). seqSieve increased UAA more than GUIDANCE and OD-seq in most cases, but slightly less than Max-

Align (Figure 2A). The cases where MaxAlign improved UAA substantially more than seqSieve were manually checked to understand the differences with the aim of improving the algorithm. However, visual examination indicated the alignment quality may often be higher in the seqSieve filtered sequences; for examples, see Figure 1B). To quantify the alignment quality, the number of matching nucleotide or amino acid pairs at each ungapped position in the alignment was summed to yield the ungapped sum of pairs (USoP; Eq. (2)). MaxAlign, GUIDANCE, and seqSieve each increased USoP relative to the original alignment, with seqSieve having the greatest effect. With these modes, seqSieve improved USoP significantly more often than either MaxAlign or GUIDANCE ($p < 0.001$, Fisher's Exact Test; Figure 2B).

In addition to how they score outlying sequences, MaxAlign and seqSieve differ in the iterative re-alignment performed by seqSieve. To evaluate the effect of iterative realignment, the test set was run through seqSieve with re-alignment disabled, and this result compared with a final re-alignment thereof as well as with full iterative realignment. In a small number of cases (32 of 1484), the sequence selection differed with iterative re-alignment disabled. All of these cases had higher UAA, and 22 had higher USoP with iterative re-alignment.

4 Discussion

With real data, seqSieve was able to substantially increase UAA, and outperformed other tools in terms of improving alignment quality as measured

by USoP. Both UAA and USoP are good measures of how much information can be utilized by gap-excluding downstream programs such as PAML [8]. Further, the sum-of-pairs scoring function [21], of which USoP is a simple subset, is a standard quality measure used, for instance, by ClustalW [25], and, in a weighted form, by Mafft [22]. For applications where maximizing information in the ungapped columns of an alignment is beneficial, the two programs specialized at this, MaxAlign and seqSieve, are top performers. While MaxAlign has a more agressive algorithm at optimizing UAA, this occassionally came at the cost of favoring long, poorly aligned sequences over shorter, but more reliably aligned sequences. By penalizing both gaps and insertions, seqSieve was able to outperform MaxAlign, Guidance, and OD-seq in optimizing USoP. GUIDANCE and OD-seq are both heuristics that lack an optimization criterion. While both seqSieve and MaxAlign aim at optimizing a scoring function, the way in which this is achieved differs between the two algorithms. In seqSieve, outlying sequence are identified by scoring a dynamic (or parameterized) combination of gap and insertion penalties, while the UAA is only used to decide if the alignment is improved by the removal of these sequences. In contrast, MaxAlign removes the sequence(s) resulting in the greatest increase in UAA per sequence removed. As poor alignments cause not just gaps, but also insertions and mismatches, the more inclusive outlier identification used by seqSieve was more successful in improving alignment quality. Another difference between seqSieve and MaxAlign is that seqSieve employs iterative realignment, which cleans the alignments from any artifacts resulting from previously identified outliers. While iterative realignment only rarely affected sequence selection, it im-

proved UAA and USoP in these rare cases.

5 Conclusion

While more and more sequencing data becomes available, it is important to realize that more data alone may fail to benefit phylogenetic analyses, or may even have detrimental effects [1]. Problems are of course amplified when (partially) non-homologous sequences are added to an alignment. Automated and self-optimized filtering by seqSieve can aid in avoiding the pitfalls of such non-matching additions, especially in high-throughput analyses.

Availability and Requirements

seqSieve is a platform-independent Python script, and is freely available from http://pypi.python.org/pypi/seqSieve. seqSieve requires two other python packages: matplotlib and numpy, as well as an external alignment program: MAFFT and/or PRANK.

References

[1] Sievers F, Dineen D, Wilm A, Higgins DG. Making automated multiple alignments of very large numbers of protein sequences. Bioinformatics. 2013;29(8):989–995.

[2] Prosdocimi F, Linard B, Pontarotti P, Poch O, Thompson JD. Controversies in modern evolutionary biology: the imperative for error detection and quality control. BMC Genomics. 2012;13:5.

[3] Ogden TH, Rosenberg MS. Multiple sequence alignment accuracy and phylogenetic inference. Systematic Biology. 2006;55(2):314–328.

[4] Fletcher W, Yang Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Molecular Biology and Evolution. 2010;27(10):2257–2267.

[5] Simmons MP. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. Molecular Phylogenetics and Evolution. 2012;62(1):472–484.

[6] Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biology. 2009;58(1):130–145.

[7] Creevey C, McInerney JO. CRANN: detecting adaptive evolution in protein-coding DNA sequences. Bioinformatics. 2003;19(13):1726.

[8] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007;24(8):1586–1591.

[9] Talavera G, Castresana J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Systematic Biology. 2007;56(4):564–577.

[10] Hartmann S, Vision TJ. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evolutionary Biology. 2008;8(1):95.

[11] Misof B, Misof K. A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments : A More Objective Means of Data Exclusion. Systematic Biology. 2009;58(1):21–34.

[12] Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Molecular Biology and Evolution. 2012;29(1):1–5.

[13] Gouveia-Oliveira R, Sackett PW, Pedersen AG. MaxAlign: maximizing usable data in an alignment. BMC Bioinformatics. 2007;8(1):312.

[14] Penn O, Privman E, Landan G, Graur D, Pupko T. An alignment confidence score capturing robustness to guide tree uncertainty. Molecular Biology and Evolution. 2010;27(8):1759–1767.

[15] Jordan G, Goldman N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Molecular biology and evolution. 2011;29(4):1125–1139.

[16] Jehl P, Sievers F, Higgins DG. OD-seq: outlier detection in multiple sequence alignments. BMC Bioinformatics. 2015;16(1):1–11.

[17] Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics. 2008;9(4):286–298.

[18] Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. Patterns of positive selection in seven ant genomes. Molecular biology and evolution. 2014;31(7):1661–1685.

[19] Scott KA, Porter SL, Bagg EA, Hamer R, Hill JL, Wilkinson DA, et al. Specificity of localization and phosphotransfer in the CheA proteins of Rhodobacter sphaeroides. Molecular Microbiology. 2010;76(2):318–330.

[20] Audelin AM, Cowan SA, Obel N, Nielsen C, Jørgensen LB, Gerstoft J. Phylogenetics of the Danish HIV epidemic: the role of very late presenters in sustaining the epidemic. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2013;62(1):102–108.

[21] Altschul SF. Gap costs for multiple sequence alignment. Journal of theoretical biology. 1989;138(3):297–309.

[22] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 2013;30(4):772–780.

[23] Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(30):10557–10562.

[24] Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Research. 2014;42(D1):D749–D755.

[25] Thompson JD, Gibson T, Higgins DG, et al. Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics. 2002;00:2.3:2.3.1–2.3.22.

Additional Files

Additional file — Test dataset

Figures

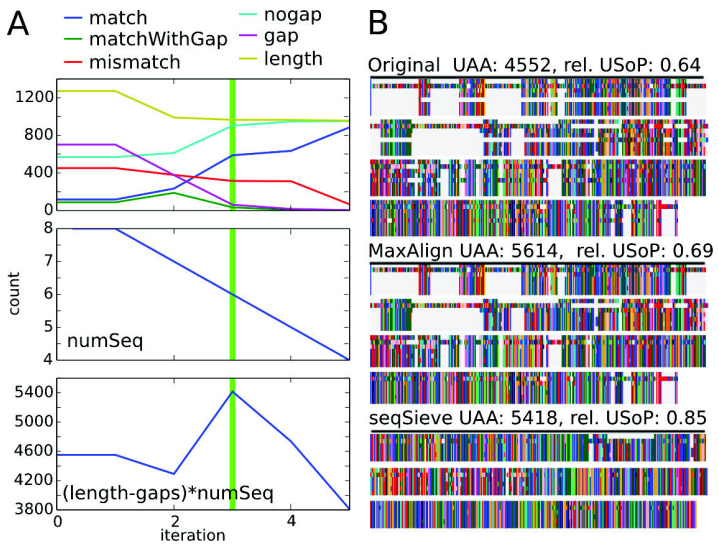


Figure 1: Filtering of an example orthogroup, where filtering with MaxAlign showed a greater improvement of UAA, but seqSieve showed a greater improvement of the relative USoP (values shown above alignments). (A) seqSieve graphical report showing the change in sequence statistics, number, and the optimization criterion, UAA, over the iterations. The green bar marks the optimal sequence set selected by seqSieve. (B) The MAFFT alignment of the original sequences, sequences after filtering with MaxAlign, and sequences after filtering with seqSieve.

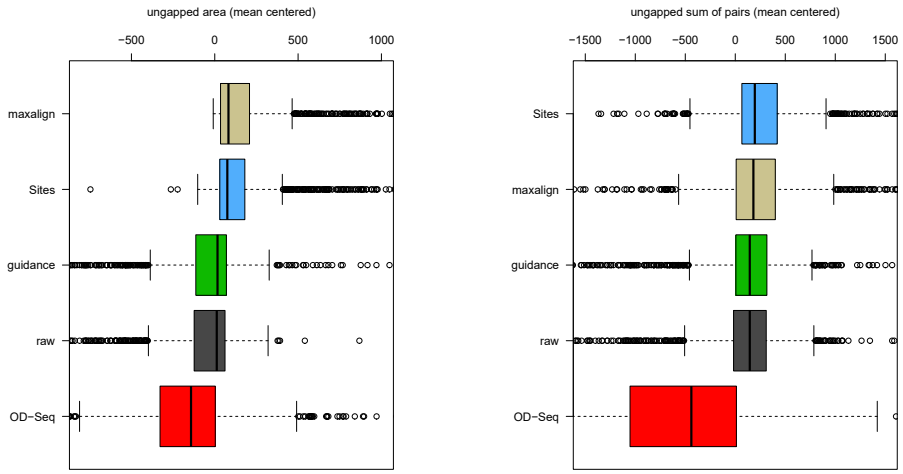


Figure 2: Alignment improvement by different algorithms. (A) seqSieve and MaxAlign typically showed a greater improvement of UAA than either MaxAlign or OD-seq. (B) seqSieve achieved a higher ungapped sum of pairs (USoP) in most cases. "Sites" is the default mode of SeqSieve. "raw" is the original MAFFT alignment.

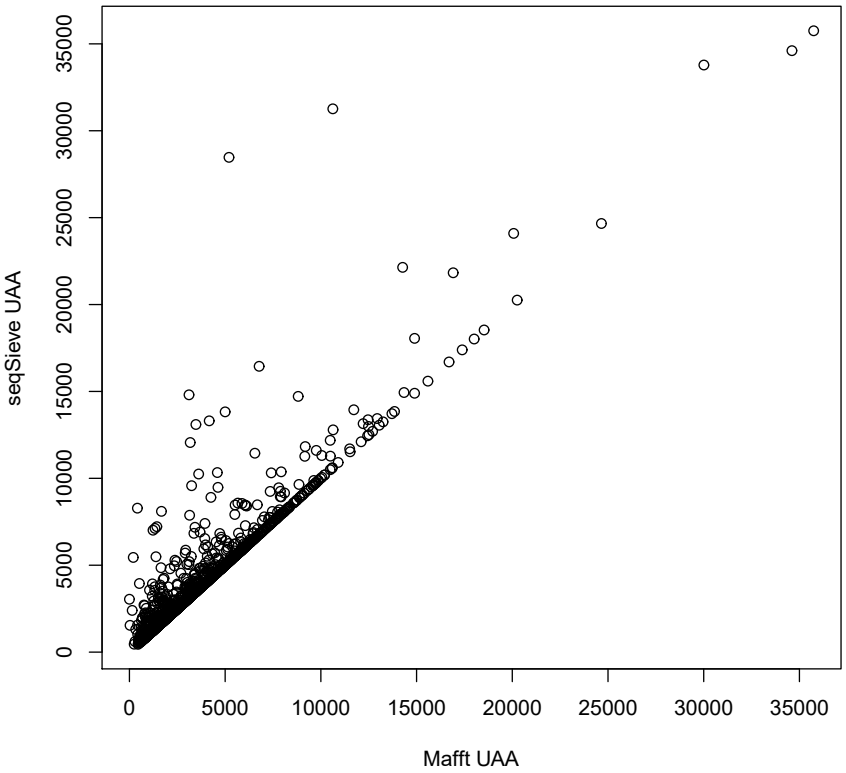


Figure S 1: Relationship between UAA for the original (raw) MAFFT alignment and the alignment improved with seqSieve.

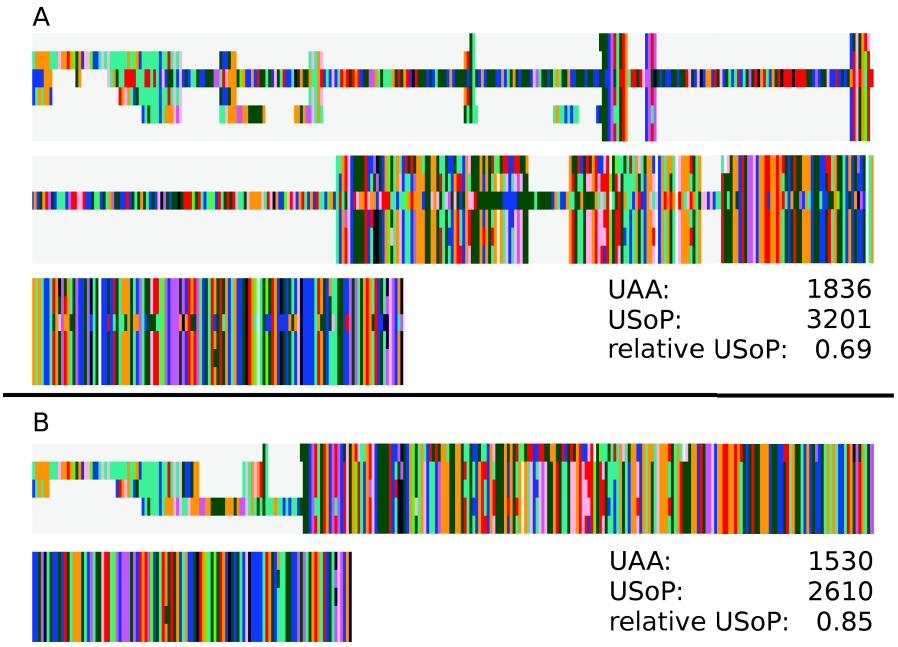


Figure S 2: The quality of alignments was visually evaluated where UAA was most different after filtering between MaxAlign and seqSieve. Example orthogroup where running MaxAlign (A) increased UAA, but not relative USoP, more than seqSieve (B). Alignments were plotted with vizqes (http://pypi.python.org/pypi/vizqes).

2.2 Manuscript 2: Freeze-quenched maize mesophyll and bundle sheath separation uncovers bias in previous tissue-specific RNA-Seq data.

Overview

Title: Freeze-quenched maize mesophyll and bundle sheath separation uncovers bias in previous tissue-specific RNA-Seq data

Authors: Alisandra K. Denton*, Janina Maß*, Canan Külahoglu, Martin J. Lercher, Andrea Bräutigam, Andreas P. M. Weber

* Co-first author.

Published in Journal of Experimental Botany, Vol. 68, No. 2 pp. 147–160, 2017
Co-first authorship

Contributions

- Assistance in wet lab work (harvest for full-gradient and metabolite analysis, metabolite extraction, RNA library preparation), together with AKD
- Discussion and interpretation of data
- Contribution and guidance in Bioinformatics analysis
- Modification of Picard Tools to calculate gene-wise 3' bias
- Editing of manuscript
- Interactive web data visualization and web app deployment

Journal of Experimental Botany, Vol. 68, No. 2 pp. 147–160, 2017
doi:10.1093/jxb/erw463 Advance Access publication 2 January 2017
This paper is available online free of all access charges (see http://jxb.oxfordjournals.org/open_access.html for further details)

RESEARCH PAPER

Freeze-quenched maize mesophyll and bundle sheath separation uncovers bias in previous tissue-specific RNA-Seq data

Alisandra K. Denton[1,*], Janina Maß[2,*], Canan Külahoglu[1], Martin J. Lercher[2], Andrea Bräutigam[1,3] and Andreas P. M. Weber[1,†]

[1] Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS), iGRAD-Plant Program, Heinrich-Heine-University, 40225 Düsseldorf, Germany
[2] Institute of Informatics, Cluster of Excellence on Plant Sciences (CEPLAS), iGRAD-Plant Program, Heinrich-Heine University, 40225 Düsseldorf, Germany
[3] Network Analysis and Modeling Group, IPK Gatersleben, Corrensstrasse 3, D-06466 Stadt Seeland, Germany

* Co-first author.
† Correspondence: andreas.weber@uni-duesseldorf.de

Received 3 August 2016; Editorial decision 18 November 2016; Accepted 18 November 2016

Editor: Robert Sharwood, Australian National University

Abstract

The high efficiency of C_4 photosynthesis relies on spatial division of labor, classically with initial carbon fixation in the mesophyll and carbon reduction in the bundle sheath. By employing grinding and serial filtration over liquid nitrogen, we enriched C_4 tissues along a developing leaf gradient. This method treats both C_4 tissues in an integrity-preserving and consistent manner, while allowing complementary measurements of metabolite abundance and enzyme activity, thus providing a comprehensive data set. Meta-analysis of this and the previous studies highlights the strengths and weaknesses of different C_4 tissue separation techniques. While the method reported here achieves the least enrichment, it is the only one that shows neither strong 3′ (degradation) bias, nor different severity of 3′ bias between samples. The meta-analysis highlighted previously unappreciated observations, such as an accumulation of evidence that aspartate aminotransferase is more mesophyll specific than expected from the current NADP-ME C_4 cycle model, and a shift in enrichment of protein synthesis genes from bundle sheath to mesophyll during development. The full comparative dataset is available for download, and a web visualization tool (available at http://www.plant-biochemistry.hhu.de/resources.html) facilitates comparison of the the *Z. mays* bundle sheath and mesophyll studies, their consistencies and their conflicts.

Key words: C_4, cell separation, maize, meta-analysis, transcriptomics.

Introduction

Specialization and coordination between two cell types improves photosynthetic efficiency in most C_4 photosynthetic plants. Specifically, most C_4 plants shuttle carbon from a surrounding mesophyll (M) tissue into a surrounded bundle sheath (BS) tissue (Hatch, 1987). The shuttling concentrates CO_2 around the carbon fixing enzyme, Rubisco, thereby suppressing photorespiration and increasing photosynthetic efficiency. This lends selective advantage to

C_4 plants in photorespiration-inducing (e.g. hot and arid) environments (Schulze *et al.*, 1996). The high photosynthetic efficiency and stress tolerance of C_4 species has led to interest in engineering the trait. However, the complexity of the trait—with many changes to anatomy and metabolism beyond the core biochemical pump—makes this an ambitious goal, which will require a full systems-level understanding of both the mature C_4 trait and its development to be achieved (Sage and Zhu, 2011).

BS and M cells show extensive specialization in metabolism and anatomy in C_4 plants. In the classic C_4 arrangement—Kranz anatomy—enlarged BS cells form a ring around the vascular bundle and are in turn surrounded by M cells (Hatch, 1987). Narrow vein spacing means each M cell borders a BS cell, allowing direct transfer of metabolites between them. Compared with a C_3 leaf, there is a massive increase in the relative amount of BS tissue, allowing for a division of labor between cell types that includes both photosynthesis and major facets of other metabolism (Majeran *et al.*, 2010; Friso *et al.*, 2010). Following Rubisco, most enzymes in the Calvin–Benson–Bassham Cycle (CBBC) and the linked photorespiratory cycle are restricted to the BS (Broglie *et al.*, 1984; Rawsthorne *et al.*, 1988; Döring *et al.*, 2016). In *Z. mays*, distribution of photosystem II and therefore linear electron transport and reducing equivalent regeneration are restricted to the M, while the BS relies on ATP from cyclic electron transport around photosystem I and biochemical shuttles that transfer reducing equivalents to the BS for energy (Romanowska *et al.*, 2008; Wang *et al.*, 2014; Bellasio and Griffiths, 2014). Subsets of metabolism are divided up between the two cell types with, for instance, amino acid, nucleotide, and isoprenoid synthesis in the M, and sulfur metabolism and starch synthesis in the BS (Majeran *et al.*, 2005; Friso *et al.*, 2010).

Information on anatomical and metabolic changes has been gained through comparative proteomic and transcriptomic studies both between C_3 and C_4 species (e.g. Bräutigam *et al.*, 2011, 2014; Gowik *et al.*, 2011; Wang *et al.*, 2014; Covshoff *et al.*, 2016), and between isolated tissue types (Majeran *et al.*, 2005; Friso *et al.*, 2010; Li *et al.*, 2010; Chang *et al.*, 2012; Tausta *et al.*, 2014; John *et al.*, 2014; Aubry *et al.*, 2014). Many of the differences between cell types are set up early in development, and tissue maturation studies have obtained mechanistic insights. For instance, comparison of C_4 and C_3 Cleomaceae species linked delayed photosynthetic differentiation to extended vein proliferation and ultimately closer vein spacing in the C_4 species (Külahoglu *et al.*, 2014). In *Z. mays* carefully comparing the primordia of Kranz leaf tissue with non-Kranz husk tissue implicated the recruitment of the ScareCrow regulatory module from the root epidermis to BS cells (Wang *et al.*, 2013). Potentially due to the difficulties of isolating cell types, to date there has only been one transcriptomics (Tausta *et al.*, 2014) and one proteomics (Majeran *et al.*, 2010) study that have looked at immature M and BS tissue. These studies have shown the early establishment of tissue specificity of major C_4 enzymes and the roles of M and BS cells in sink *vs* source tissue to logically reflect the broader changes between source and sink tissue. As neither of the above studies could look at metabolites, and interstudy comparisons have produced distinct results on cell specificity—particularly of transcription factors (Tausta *et al.*, 2014)—we judged further analysis to be warranted.

Here we successfully perform an 'omics'-scale analysis on developmental tissue separated by a method developed by Stitt and Heldt (1985), and thus simultaneously capture changes in the transcriptome, enzymatic activities, and the metabolome. A subsequent meta-analysis of this and other BS and M separation studies highlights the strengths and weaknesses of each of the various separation methods, and the advantages of using complementary techniques. The comparative dataset has been made available for visual exploration or download, and can assist both in experimental design both for BS/M related studies and for studies in the broader category of tissue separation.

Materials and methods

Plant genome data

Genome and gene-model data was downloaded for *Setaria viridis* (v1.1/v311; Bennetzen *et al.*, 2012) and *Panicum virgatum* (v1.1/v273; DOE-JGI, 2016) from Phytozome 11.0 (Goodstein *et al.*, 2012). The AGPv3.22 release of the *Zea mays* genome with the 5b+ filtered gene set was obtained from ensemble plants (Kersey *et al.*, 2016) and Gramene (Tello-Ruiz *et al.*, 2016), respectively. Orthologs were identified by best BLAST (Altschul *et al.*, 1997) hit from *Z. mays* to *S. viridis* or *P. virgatum*.

External RNAseq data

Complementary RNAseq data were downloaded from the sequence read archives (Kodama *et al.*, 2012) and European nucleotide archives (Leinonen *et al.*, 2010). We included two additional *Z. mays* BS and M separation studies (Chang *et al.*, 2012: SRP009063; Tausta *et al.*, 2014: SRP035577); corresponding whole developmental leaf sections (Li *et al.*, 2010; SRP002265); *Z. mays* tissue atlas (Sekhon *et al.*, 2013; SRP010680); and primordial leaf and husk tissue (Wang *et al.*, 2013; SRP028231). The non-*Z. mays* studies were separation of BS and M cells in *S. viridis* (John *et al.*, 2014; ERA275647) and *P. virgatum* (Rao *et al.*, 2016; SRP062667).

Note that as the original authors included the same precise set of sequences for BS and M tissues in section 14 (Li *et al.*, 2010; Tausta *et al.*, 2014), and reported the same plant growth conditions, we've considered these studies broadly comparable. However, to avoid redundancy, the BS and M samples for section 14 are only included with Tausta *et al.* (2014).

Plant growth conditions and harvest

Z. mays B73 was grown in the summer of 2012 under conditions previously described (Pick *et al.*, 2011). The third leaf was harvested when it measured 18 cm from the second ligule to the leaf tip. Two different harvesting methods were performed. In the first, a leaf gradient with five sequential developmental slices (4 cm each) was harvested with the 'leaf guillotine' (see Fig. S1A available at Dryad Digital Repository http://dx.doi.org/10.5061/dryad.tf6q6; Pick *et al.*, 2011). This method required 10 s to extract the third leaf and properly align it, which does not allow for reliable estimates of the high-turnover photosynthetic metabolite distributions. Therefore, a second harvesting method was performed, in which the plants were positioned above two liquid nitrogen containers and two 8 cm

slices were cut with connected scissors (see Fig. S1A, B at Dryad) achieving a delay of less than 1 s between slicing and flash-freezing. Metabolite abundance and enzyme activity were measured from both harvest sets; the full five-slice gradient was used for RNAseq.

Tissue enrichment

M and BS tissues were enriched using a method modified from Stitt and Heldt (1985). Ground material was filtered through 250, 80, and 41 µm meshes on liquid nitrogen. Three fractions were selected for further analysis. The 'BS-e' fraction showed the most enrichment of BS tissue (it did not pass through 80 µm mesh); the 'M-e' fraction showed most enrichment in M tissue (it passed through 41 µm mesh); and the 'I-e' fraction showed intermediate, but consistent, proportions of tissues (it did not pass through 41 µm mesh).

Extraction and abundance measurements metabolites and enzymes

Enzymes were extracted and desalted as described in Bräutigam *et al.* (2014), and the enzyme activity was measured through colorimetric assays as described in Hatch and Mau (1977) and Walker *et al.* (1995). Metabolites were extracted and quantified via gas chromatography–electron-impact time-of-flight mass spectrometry as described in Rudolf *et al.* (2013). Both low-signal metabolites and individual replicates with a percentage abundance in BS more than 3 standard deviations from the mean were excluded. The integrated peaks were divided by the area of the ribitol (internal standard) peak and the fresh weight, and to further reduce noise and compensate for FW/DW differences between the cell types by the mean abundance for the replicate. Therefore, normalized differences between metabolites represent not absolute distribution, but distribution relative to the other metabolites, particularly sucrose and the other highly abundant metabolites.

Sequencing and estimating transcriptional abundances

RNA was extracted with QIAGEN RNeasy Plant kits, according to the manufacturer's instructions except for an extra wash step in 80% ethanol after the standard wash steps. Libraries were prepped from RNA with an RNA integrity number >8 and sequenced with the Illumina HiSeq 2000 platform. The quality was checked with FastQC (Andrews, 2010). Quality and adapter trimming was performed with Trimmomatic (Bolger *et al.*, 2014). Trimmed reads were mapped to their respective genomes with Tophat2 (Kim *et al.*, 2013) and the unique counts per locus were quantified with HTSeq (Anders *et al.*, 2015); transcripts per million (TPM) was calculated from the unique counts and gene length. Coverage metrics including 3′ bias were calculated with PicardTools 2.4.1: CollectRnaSeqMetrics (Wysoker *et al.*, 2012). Non-default parameters used for bioinformatics programs are provided (see Table S1 at Dryad). The same pipeline was used for all studies except as necessitated by experimental differences (e.g. paired *vs* single end reads), or otherwise noted.

Differential expression and tissue specificity normalization

Differential expression *P*-values and $\log_2$ fold changes were calculated with EdgeR (Robinson *et al.*, 2009). Where no replicates were available (Chang *et al.*, 2012), the mean common dispersion from the remaining studies was used. Additionally, due to the low level of enrichment achieved in this study, ContamDE (Shen *et al.*, 2016), a cross-contamination tolerant package for RNAseq statistics, was employed for the data generated here. As necessary for interstudy comparisons in *Z. mays*, $\log_2$ fold changes from edgeR (Chang *et al.*, 2012; Tausta *et al.*, 2014) and ContamDE (this study) were quantile normalized, and the fully normalized TPM back calculated from the quantile normalized $\log_2$ fold change and mean TPM.

Estimation of initial tissue specificity by 'deconvolution'

The distribution of metabolites and enzyme activities was compared with the distribution of markers to estimate the original tissue specificity in a method modified from Stitt and Heldt (1985). First, all data were converted into fraction of total by developmental slice. Second, marker enzyme activities were used as proxies for the amount of M (phosphoenolpyruvate carboxylase (PEPC) activity) and BS (NADP-malic enzyme (ME) activity) tissue in each enrichment fraction. The slope of a regression line between the ln(target/M) against ln(BS/M) estimated the fraction of the target found in pure BS (see Fig. S1C at Dryad). *P*-values were calculated with a null hypothesis of slope=0.5 (50% M, 50% BS). This was automated with a linear regression in R and calculated for every metabolite and non-marker enzyme. To estimate the 'pure' abundance values, the estimated fraction in BS and M (1–fraction BS) were multiplied by 2× the average abundance value for the developmental slice.

Functional category enrichment testing

Functional categories were assigned with Mercator (Lohse *et al.*, 2014). Enrichment was tested with Fisher's exact test, and the false discovery rate calculated according to (Benjamini and Yekutieli, 2001).

Statistics

Unless otherwise noted, all statistical analysis was performed in the R statistical environment (R Development Core Team, 2011) and whenever a test was performed more than 20 times, the false discovery rate (Benjamini and Hochberg, 1995) was calculated from the resulting *P*-values.

Accession numbers

The reads related to this article have been deposited in the Sequence Read Archives under the accession number SRP052802.

Results

Validation of separation method

Here, we enriched BS and M cells along a developing *Z. mays* leaf by grinding and serial filtration (Stitt and Heldt, 1985). Two harvesting methods were used, the first using a 'guillotine' (Pick *et al.*, 2011) to sample five contiguous 4 cm slices from tissue just emerging from the ligule (slice 5) to the leaf tip (slice 1). In the second, targeted at capturing unadulterated metabolite levels, two 8 cm slices were harvested in full illumination and quenched in liquid nitrogen within a second of cutting. M and BS tissues were enriched using a method modified from Stitt and Heldt (1985) that capitalizes on the distinct physical properties of M and BS cells to enrich them in different separation fractions as ground tissue is filtered through serially smaller meshes over liquid nitrogen. The activity of C_4 enzymes and the metabolite levels were measured from both harvests, and RNAseq was performed on material from the five-slice gradient.

The distribution of tissue specific markers indicated BS and M tissue were successfully enriched (see Fig. S1D and Dataset S1 at Dryad). The classic BS marker is NADP-ME, the enzyme responsible for releasing the carbon from C_4 acids in the BS. NADP-ME activity and transcripts were both higher in the coarsest (from here on, BS-e for bundle

sheath enriched) separation fraction; in between in the middle (from here on, I-e for intermediate enrichment) separation fraction; and lowest in the finest (from here on, M-e, for mesophyll enriched) fraction (see 'Materials and methods' for details). The classic M marker, PEPC, the C_4 fixing enzyme, showed the opposite pattern, with highest activity and transcript abundance in the fine, M-e fraction. While the enrichment was strongest in mature tissue, it was also apparent in the youngest tissue (slice 5). For non-marker enzymes and metabolites, the original distribution was estimated based on the marker enzymes (see 'Materials and methods'; Fig S1C at Dryad).

This enrichment method was chosen over other separation methods both for sample integrity and to obtain data on metabolite abundance, enzyme activity, and transcript abundance from the same material. However, in the rapid harvest (with less than 1 s between cutting and quenching in liquid nitrogen), very few significant differences were found between metabolite levels in M and BS ((iso)-citric acid and malonic acid were both enriched in BS slice 3–4; FDR<0.05; Fig. S2C at Dryad). In contrast, many metabolites showed significant differences based on leaf age (10 metabolites with FDR<0.05 between slice 3–4 and slice 1–2 in the sub-1 s harvest, and 20 with FDR<0.05 between at least one of the neighboring slices in the 10 s harvest; Fig. S2E and Dataset S1 at Dryad). The observed developmental changes were very similar between the sub-1 s and 10 s harvest; however, there were a few exceptions. One example is phenylalanine, which increased in abundance with leaf age in the fast harvest, but decreased in the slow harvest (Fig. S2C, D at Dryad). Although not statistically significant, the BS *vs* M trend of several metabolites corresponded with expectations. Notably, serine and the other photorespiratory metabolites were higher in the BS, where they are expected to be produced, both in the faster (Fig. 1B) and, to a lesser extent, also in the slower (Fig.S2B at Dryad) harvest. Malate, which presumably moves from M to BS entirely based on a diffusion gradient, tended towards enrichment in the mature M (slices 1–2, 1–3; Fig. 1A and Fig. S2A at Dryad). Further, there is a modest consistency between previous studies measuring distribution of metabolites and that measured here (Fig. 1C). All measured core C_4 metabolites shift from putative BS towards putative M enrichment between slice 3–4 and 1–2 (Fig. 1A). Such synchronized changes could relate to increasing flux (or changing rate-limiting steps) in the C_4 cycle. The differences between harvest speeds highlights how labile these metabolites can be, and discrepancies between studies or low enrichment values may simply reflect response to conditions and the readiness with which they pass the plasmodesmata, respectively. Higher confidence in metabolite distribution will require more replicates, and, potentially, more defined conditions.

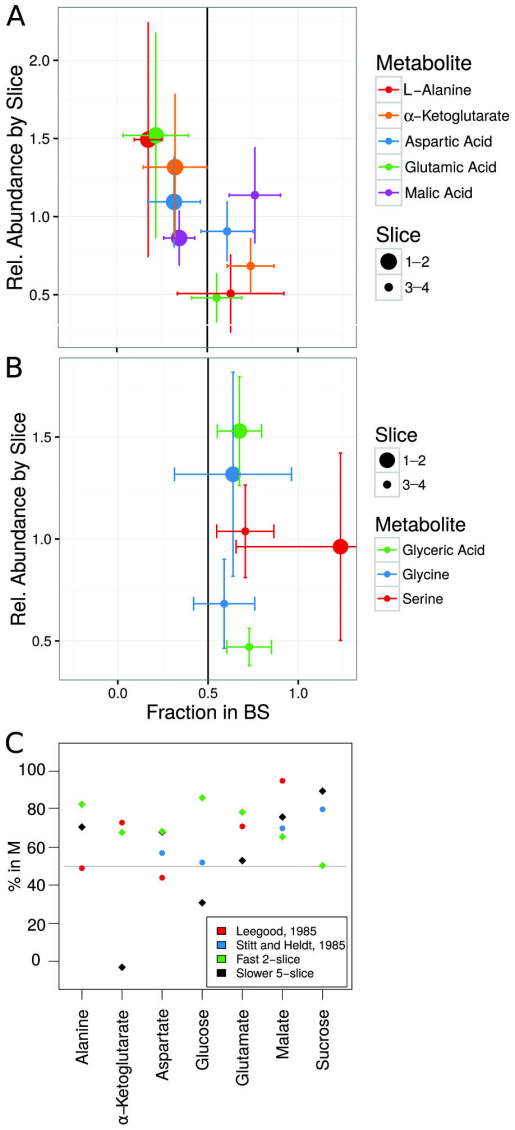


Fig. 1. Metabolites. (A, B) The estimated tissue enrichment and abundance of measurable metabolites associated with the photorespiratory cycle (A) and the C_4 cycle (B). Error bars indicate standard error. (C) Comparison of metabolite tissue enrichment measured by Leegood (1985) and Stitt and Heldt (1985) with the average of slice 1 and 2 in the slower five-slice harvest and slice 1–2 in the faster two-slice harvest.

Comparison with other separated transcriptomes

Quantitative study comparison

While this separation method provides high integrity and allowed us to simultaneously measure transcripts, metabolites, and enzyme activities, it comes with its own caveats due to the limited enrichment. As separation studies will likely continue, either in new species or with variations such as separating the husk (Huang and Brutnell, 2016), we evaluated the advantages and disadvantages of different separation methods and their effect on biological results. We compiled a comparative dataset from all existing M/BS specific full RNAseq experiments in monocots. These covered mechanical and enzymatic separation in *Z. mays* (Chang *et al.*, 2012); mechanical separation in *S. viridis* (John *et al.*, 2014); laser micro-dissection in *Z. mays* (Li *et al.*, 2010; Tausta *et al.*, 2014);

mechanical micro-dissection in *P. virgatum* (Rao *et al.*, 2016); and the serial filtration performed here (referred to as 'Denton 2016' in figures). While the data encompass three origins and two subtypes of C_4 photosynthesis, and BS and M cell specificity is not expected to match perfectly, previous studies have found substantial conservation even between monocots and dicots (Aubry *et al.*, 2014). Overall, the combination of mechanical BS preparation and enzymatic (Chang *et al.*, 2012) or leaf rolling (John *et al.*, 2014) M separation achieved the highest marker enrichment, followed by the micro-dissection studies (Li *et al.*, 2010; Tausta *et al.*, 2014; Rao *et al.*, 2016), while the method used here, as anticipated from the original report (Stitt and Heldt, 1985), showed the least enrichment (Fig. 2A). Consistent with the lower enrichment, this study showed the lowest statistical power of the various methods with an average of 2100 discoveries (FDR<0.05) per slice, compared with 4030–12 777 discoveries for the other (biological-replicate-including) studies when computed with edgeR. Therefore, a cross contamination aware R-package, contamDE, which includes a factor for the relative tissue enrichment of each replicate, was employed. With contamDE an average of 4479 discoveries (BS-e *vs* M-e FDR<0.05) were made per slice, and this was used for further analysis (see Table S2 at Dryad).

Tissues were matched to achieve a more in-depth comparison between the *Z. mays* studies. For mature tissues, the sample from Chang *et al.* (2012) was most similar to section 14 from Tausta *et al.* (2014) and to slice 2, here, while the youngest section in Tausta *et al.* (2014) was most similar to slice 4, here (Spearman correlation, Fig. S3A at Dryad). The Tausta *et al.* (2014) study was able to detect genes with a lower log fold change (relative to the total log fold change distribution) than either Chang *et al.* (2012), with just one replicate, or this study, with low enrichment. However, examining log fold change indicated the differences between studies ran deeper than statistical power, with many genes significant in one study not enriched or even significantly enriched in the opposite direction in another study (Fig 2B–D).

Qualitative study comparison

For a more qualitative look at the differences between studies we performed a hierarchical clustering of samples from this study, those from Chang *et al.* (2012) and Tausta *et al.* (2014), and the unseparated sections from Li *et al.* (2010) that corresponded to Tausta *et al.* (2014). The samples clustered primarily by study, followed by leaf age and then M and BS, with some mixing (Fig. 2E). Between-study differences could in theory come from growth conditions and plant age, from differences in separation method or from a combination thereof, and all studies but Li *et al.* (2010) and Tausta *et al.* (2014) used distinct growth and harvest conditions (see Table S3 at Dryad). Notably, the unseparated sections from Li *et al.* (2010), which were grown comparably to those from Tausta *et al.* (2014) clustered not with the associated leaf sections of Tausta *et al.* (2014), but with the respective older or younger serial filtration data here, indicating a substantial role of separation method in clustering. Indeed, one of the gene clusters (3) was primarily expressed at a lower level across the laser micro-dissection (Tausta *et al.*, 2014) samples compared with all the other samples (including Li *et al.*, 2010). RNA is known for its degradability under procedures like laser micro-dissection, and Li *et al.* (2010) clearly reported the 3′ bias in the laser micro-dissection section 14, but did not at that time have the comparative studies to evaluate how this would globally affect the results. A list of genes most dramatically affected by laser micro-dissection was obtained by looking for genes with significantly different abundance between unseparated (Li *et al.*, 2010) and the laser micro-dissection separated section 14 (Li *et al.*, 2010; Tausta *et al.*, 2014). The majority (3298 of 3362) of the differentially regulated genes were downregulated in the laser micro-dissection samples. These laser micro-dissection 'downregulated' genes were depleted in BS *vs* M, differences shared with this study (Fig. 3A; Fisher's exact test, $P<0.001$). Further, these genes showed several functional enrichments (MapMan categories), including major categories such as transport and signaling; and minor categories such as minor CHO metabolism.callose, GARP G2-like transcription factor family and Class XI Myosin (Dataset S2 at Dryad). Finally, the strong 3′ bias resulted in a low diversity library compared with the other studies (see Fig. S3B, C at Dryad).

Considering the effect degraded RNA can have, we evaluated the 3′ bias across studies to see how the other separation methods compared. The three prime bias was highest in the laser (Tausta *et al.*, 2014) and mechanical (Rao *et al.*, 2016) micro-dissection studies; however, it was present to various degrees in at least some samples of the other separation studies and in multiple other *Z. mays* studies without separation (Li *et al.*, 2010; Sekhon *et al.*, 2013; Wang *et al.*, 2013; Fig. 3C). Notably, both studies that used distinct methods for isolation of BS strands and M cells (John *et al.*, 2014; Chang *et al.*, 2012) showed minor 3′ bias, but each M sample showed more than its corresponding BS sample (Fig. 3C). The 3′ bias was not spread evenly across all genes, but was higher in the 199 genes where Chang *et al.* (2012) and slice 2 (this study) were significantly, but oppositely, enriched in the BS and M, respectively (see Fig. S4 at Dryad). Overall mild increases in 3′ bias between samples are prominent in these 199 genes and their orthologs, notably including the M samples in Chang *et al.* (2012) and Tausta *et al.* (2014) and one BS replicate from this study. The orthologs of these 199 genes, measured by John *et al.* (2014) mostly (138 of 184; 75%) were enriched in the same direction as Chang *et al.* (2012), while those measured by Rao *et al.* (2016) mostly (101 of 152; 66%) agreed with this study (Fig. 3C). In contrast, neither cross-species comparison showed a notable BS or M bias in orthologs of the opposite gene set—the 14 genes where Chang *et al.* (2012) and slice 2 (this study) were significantly enriched in the M and BS, respectively (Fig. 3D). In summary, despite evolutionary distance between *Z. mays* and *S. viridis*, the studies with higher, degradation-marking 3′ bias in the M than BS (Chang *et al.*, 2012; John *et al.*, 2014) share a set of 'BS enriched' genes that conflict with the M enrichment seen in *Z. mays* (this study) and *P. virgatum* (Rao *et al.*, 2016).

To determine if different RNA quality and 3′ bias relate to some of the discrepancies between the *Z. mays* studies,

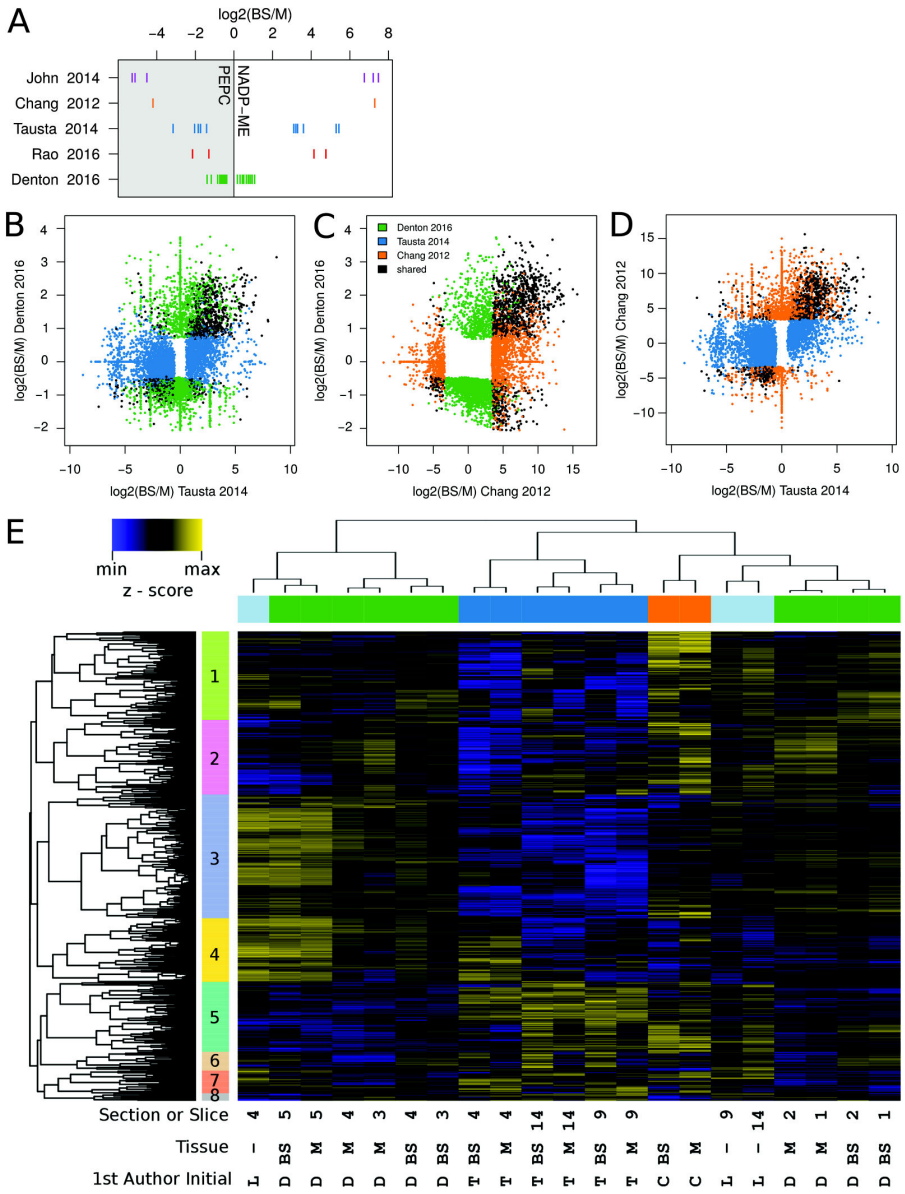


Fig. 2. Interstudy comparison. (A) Enrichment of the classic BS (NADP-ME) and M (PEPC) marker genes in each study. (B–D) Log_2 fold change of genes that were significantly enriched in BS or M in at least one of the paired *Z. mays* studies. (E) Hierarchical clustering of fully normalized log2 (TPM) for *Z. mays* samples, with Pearson and Spearman correlation-based distance for genes and samples, respectively. Genes filtered to those with TPM min>0, max>50. Side colors included to help delineate studies on the *x*-axis and major clusters on the *y*-axis. C, Chang *et al.* (2012); D, this study; L, Li *et al.* (2010); T, Tausta *et al.* (2014).

we quantified the level of 3′ bias on genes in two different conflict sets—conflict set 1: BS specific in Chang *et al.* (2012) or Tausta *et al.* (2014) and M specific in the comparable tissue here, or BS (Chang *et al.*, 2012) and M (Tausta *et al.*, 2014, section 14); conflict set 2: as conflict set 1 but with BS and M switched). This showed that conflict set 1 had the most 3′ bias across all studies while conflict set 2 had the same or even less bias than the whole gene set (Fig. 3B). One of the genes in 'conflict set 1' is related to the C_4 cycle, namely phosphoenolpyruvate carboxylase kinase (PPCK; GRMZM2G178074), which regulates PEPC in the M (Vidal and Chollet, 1997). The coverage across the PPCK locus shows a mild 3′ bias in unseparated studies and in both BS and M samples here, with higher coverage in the M (Fig. 4). In the laser micro-dissection study, there is a strong 3′ bias in both samples, with more remaining coverage in the M sample, while in the Chang *et al.* (2012) sample, there is a mild 3′ bias in the BS sample, but a strong 3′ bias in the M sample,

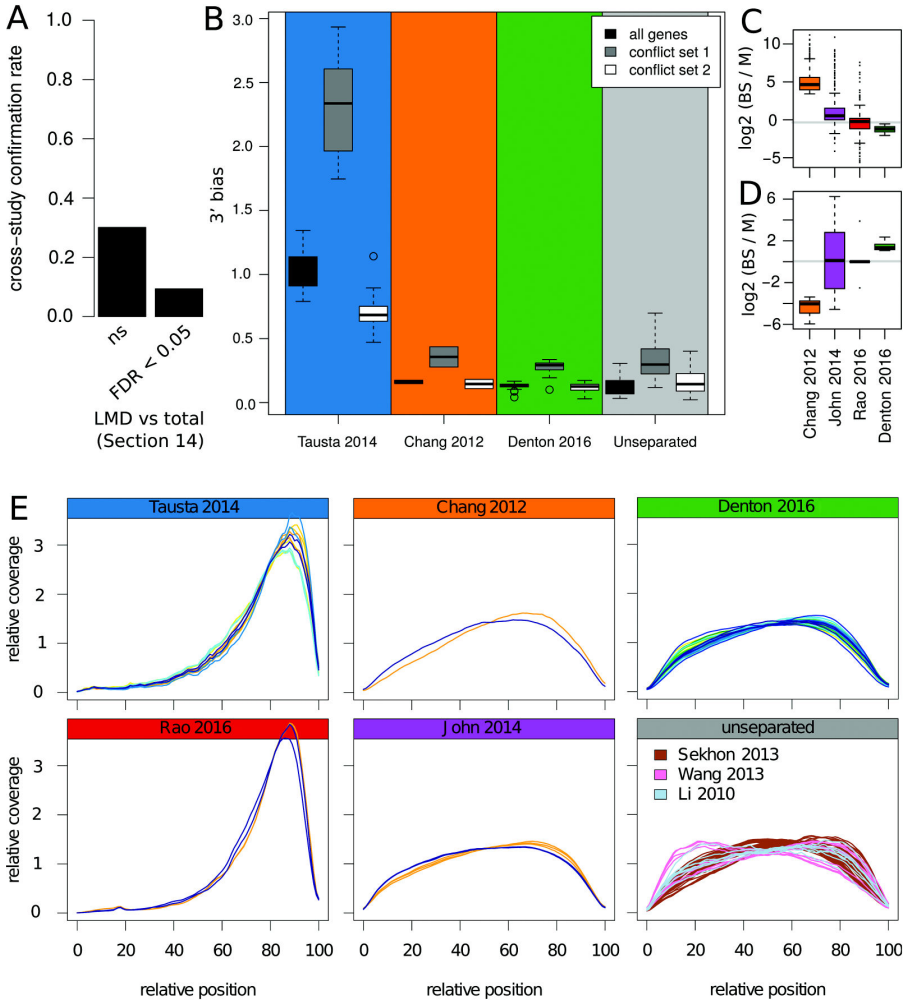


Fig. 3. Technical bias. (A) The fraction of significant differences discovered here (slice 2) that were shared with the Tausta *et al.* (2014; section 14) study broken up based on whether these genes were of significantly lower abundance in the laser micro-dissected section 14 compared with whole section 14 (Li *et al.*, 2010; Tausta *et al.*, 2014). (B) The 3′ bias observed in the coverage for the genomic background and the two conflict sets in each *Z. mays* separation study, and all the unseparated samples of Li *et al.* (2010), Wang *et al.* (2013), and Sekhon *et al.* (2013). (C, D) The tissue enrichment of the *Z. mays* genes and the *S. viridis* (John *et al.*, 2014) and *P. virgatum* (Rao *et al.*, 2016) orthologs where the *Z. mays* gene was significantly more abundant in the BS in Chang *et al.* (2012), and the M in slice 2 (this study) (C), or vice versa (D). (E) Transcript coverage by study. For all BS and M separation studies, blue represents BS, yellow represents M, and tissue maturity increases from light to dark. Green represents I-e in this study (Denton 2016).

causing PPCK to appear higher in the BS based on total read count. While not all genes in 'conflict set 1' looked like this (e.g. many had a very strong 3′ bias across every sample and study; not shown), other similar examples were not hard to find (see Fig. S5 at Dryad). Further, components of 'conflict set 1' were enriched in several MapMan categories. These included three transcription factor sub-categories (PHD finger, pseudo ARR, and putative), and minor CHO metabolism.callose in genes BS specific in Chang *et al.* (2012), and M specific in slice 2 of this study (Dataset S2 at Dryad).

Another likely artifact of the separation method is the residual contamination with non-M and non-BS tissue types. The mechanical separation methods are expected to co-purify the vascular bundle with the BS cells, while the serial grinding and filtration used here presumably includes all cell types in at least one of the enrichment fractions. To confirm and quantify these expectations would require unambiguous markers that were known to, for instance, be highly specific to the vascular tissues and absent from M or BS. In the absence of fully characterized markers in *Z. mays*, we tested a variety of candidates, largely known from other species.

Putative vascular markers were initially selected from the literature based on functions expected to be highly vascular specific. Enzymes associated with lignification of protoxylem elements (LAC17) were more abundant in the BS base sample (Fig. S6A at Dryad; FDR<0.05 for three of the four expressed). Similarly, homologs to Arabidopsis XYLEM CYSTEIN PROTEASE (XCP) 1 (GRMZM2G066326) and

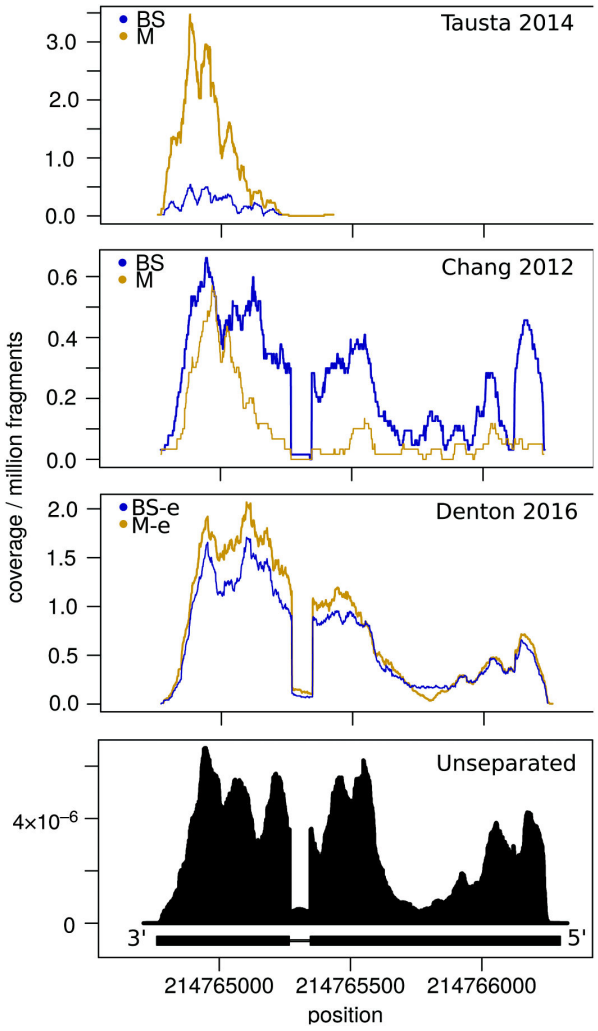


Fig. 4. Coverage of example gene PPCK. Read depth across genomic region of PPCK (GRMZM2G178074; which is in conflict set 1) in the various *Z. mays* separation studies, and in the unseparated samples of Li *et al.* (2010), Wang *et al.* (2013), and Sekhon *et al.* (2013).

2 (GRMZM2G367701), involved in programmed cell death in the xylem, were higher in the BS base sample (Fig. S6B at Dryad; FDR<0.001 for all three). These markers, however, were not expressed in older tissue and thus could not be used for interstudy comparisons. SUCROSE TRANSPORTER 2 (SUT2), frequently used as a companion cell marker in Arabidopsis (AT2G02860; Meyer *et al.*, 2000), has five homologs in *Z. mays*, for which the cumulative expression was enriched in the BS across all *Z. mays* studies (see Fig. S6C at Dryad). A study on phloem transported RNAs in Arabidopsis (Deeken *et al.*, 2008) provided a larger list of potential vascular markers; however, the cumulative expression was again higher in the BS across studies (Fig. S6D at Dryad). We further examined sets of genes that included 'phloem' (Fig. 5A), 'xylem' (Fig. S6E at Dryad), or 'vascular' (Fig. S6F at Dryad) in their descriptions. Cumulative expression of these keyword gene sets was largely higher in the BS across studies; however, for 'phloem' and 'vascular' genes, BS enrichment in the laser micro-dissected samples was less than BS enrichment of these genes in the mechanical separation studies.

We further evaluated the distribution of putative epidermal markers. A previous study using laser micro-dissection to separate epidermal and M tissues identified two epidermal specific genes in *Z. mays* (Javelle *et al.*, 2010). The more highly expressed of these, GRMZM2G345700, was consistently higher in the M samples (Fig. S6G at Dryad; FDR<0.05 in six of nine comparisons), while the less highly expressed GRMZM2G387360 was not significantly enriched. A broader look at all genes including the words 'epidermal' in their descriptions (Fig. 5B) showed higher cumulative expression in the M in most comparisons, while the most substantial M enrichment appeared to be in this study (Fig. 5B). It is hard to draw a firm conclusion in the absence of unambiguous markers, as expression patterns in epidermal cells may be more similar to M than BS, and vice versa for vascular expression. However, both expectation and a view on the broader patterns support co-purification of vascular tissues with the mechanical BS purification methods, co-purification of epidermal tissue with the M in the serial filtration method used here, and generally less co-purification using laser micro-dissection.

The strengths of interstudy comparison

Multiple study comparisons allow for confidence in results that would seem dubious alone. In this study, aspartate aminotransferase stood out as having transcript enrichment in M cells (Fig. 6A) that was contrary to the expected even distribution between cell types in the current *Z. mays* C_4 model (Furbank, 2011; Pick *et al.*, 2011). Comparison with the other datasets confirmed the same pattern in all NADP-ME studies (*Z. mays* and *S. viridis*). Previous studies (Chang *et al.*, 2012; Tausta *et al.*, 2014) have mentioned a low-expression BS specific AspAT paralog, or the detection of AspAT in both BS and M proteomic studies as balancing explanations. However, in both transcriptomics and proteomics (Friso *et al.*, 2010; Majeran *et al.*, 2010) the total abundance is much higher in the M. This is further supported by the high M specificity of the AspAT enzyme activity (Fig. 6A). A similar but less pronounced pattern in transcripts could be found for alanine aminotransferase (see Fig. S7A at Dryad).

Consistent BS or M enrichment as the leaf develops helps increase confidence, both as a repeat observation and as a simple explanation consistent with the gradual nature of changes in transcript abundance during leaf development (Pick *et al.*, 2011). On the flip side, however, it seems less likely that a gene changed from BS specific to M specific or vice versa during development. We used the interstudy comparison to evaluate the reliability of observed switches in enrichment across leaf development. As expected, genes that were significantly enriched in M-then-BS or BS-then-M in sections 4 and 14 of Tausta *et al.* (2014) were much less likely to find cross-study support (same enrichment direction in slice 4 and 2 of this

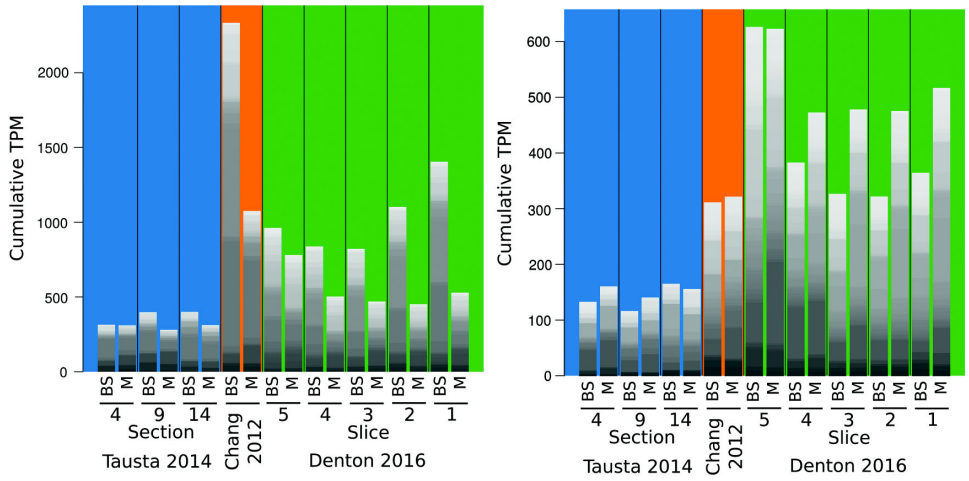


Fig. 5. Co-purification of additional tissues. Fully normalized abundance of genes that included the word 'phloem' (A) or 'epidermal' (B) in their MapMan description.

study) than their M-then-M or BS-then-BS enriched counterparts (19% *vs* 78%, Fisher's exact test $P<0.001$). However, the 48 genes that were significantly enriched in the BS in section 4 (Tausta *et al.*, 2014) and in the M in section 14 (Tausta *et al.*, 2014) with support from this study showed enrichment in the functional category 'protein.synthesis.ribosomal protein.eukaryotic.60S subunit' and all parental categories there of (Dataset S2 at Dryad). Further investigation showed that both the 60S and 40S ribosomal subunits have a clear pattern with strong BS enrichment in young but entirely unsheathed tissue (section 4, Tausta *et al.*, 2014; slice 4, here). As the leaf develops the strong BS enrichment fades, and even switches to a mild M enrichment (Fig. 6B and Fig. S7B at Dryad). To determine if the mature M enrichment could be related to supporting the high turnover of photosystem II components, we included the data for *S. viridis* (John *et al.*, 2014) and *P. virgatum* (Rao *et al.*, 2016) in the analysis. Notably the 60S and 40S ribosomal subunits showed M enrichment in *S. viridis* (Fig. 6B and Fig. S7B at Dryad), in which photosystem II, like in *Z. mays*, is primarily localized to the M (Fig. 6C). In contrast, these subunits showed BS enrichment in *P. virgatum* (Fig. 6B and Fig. S7B at Dryad), in which photosystem II is not primarily localized to the M (Fig. 6C).

Data accessibility and visualization

To facilitate public comparison of these transcriptomes, we are providing (i) a *Z. mays* gene browser with gene-specific or gene-group visualization of the data from BS/M separation studies in *Z. mays*; and (ii) all the data analysed in this study (including non-*Z. mays* BS *vs* M comparisons, and unseparated *Z. mays* studies) in tabular format (Dataset S3 and S4 at Dryad). The *Z. mays* gene browser aims to facilitate comparison and critical evaluation of the similarities and differences between these studies. To this end, the graphics include the separation method in the display and necessary contextual data (e.g. unseparated samples from Li *et al.* (2010) corresponding to the laser micro-dissection samples from Tausta *et al.* (2014), and 3′ bias (Fig. 7). Further, the browser includes several pre-loaded gene sets to help users compare studies (Fig. 7B). These sets include, for example 'conflict set 1' described above. Further gene sets include three gradations of highly supported M or BS specific genes across studies (735, 365, and 126 significant differences; shared between 7+, 8+, or all 9 of the comparisons, respectively), and highly supported M or BS specific transcription factors (52 significant differences shared between 7+ comparisons), and transcription factors of special interest in immature tissue (36 significant differences in two of the three youngest comparisons (Tausta *et al.*, 2014 section 4, and slice 4 and 5, here) and higher in foliar than husk primordia in Wang *et al.* (2013). Full lists and descriptions are provided in Dataset S5 at Dryad and with the visualization tool at http://www.plant-biochemistry.hhu.de/resources.html.

Discussion

Despite the variety of BS and M separation methods used and increasing number of studies, no method presents itself as a clear best option. Rather, the various methods come with advantages and disadvantages, which should be considered both when planning the experiment and evaluating the data.

The only fast methods for metabolite extraction are leaf rolling (Leegood, 1985) for M compared with whole tissue, and grinding and serial filtration on liquid nitrogen as performed here (Stitt and Heldt, 1985). Only a handful of metabolites measured in these studies overlap, and the correlation between studies is modest. Elements of this study might help clarify why and plan the next experiment. First, there were very substantial differences between the <1 s harvest and the 10 s harvest and between the different leaf slices. This highlights that the dynamics of abundance of metabolites make them extremely sensitive to both conditions and harvest methods. Considering the dominance of age, conditions and

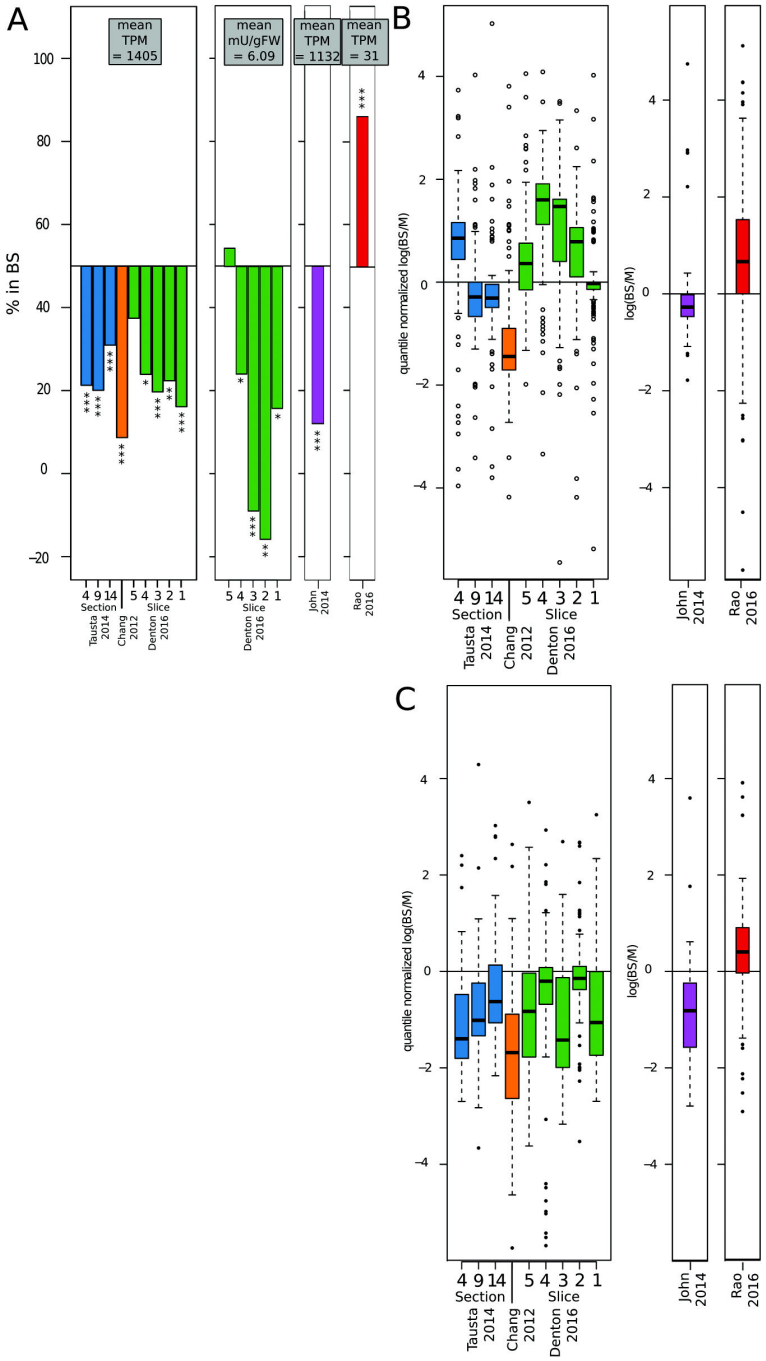


Fig. 6. Biological insights drawn from interstudy comparison. (A) Tissue enrichment of AspAT of transcripts in maize (left), enzyme activity in *Z. mays* (mid-left), transcripts in *S. viridis* (mid-right) and transcripts in *P. virgatum* (right). (B, C) Tissue enrichment of transcripts in the MapMan functional category for 60S ribosomal protein (B) and photosystem II (C) in *Z. mays* (left), *S. italica* (middle), and *P. virgatum* (right). In (A) asterisks denote significance of FDR for transcripts and *P*-values for enzyme activity (**P*<0.05, ***P*<0.01, ****P*<0.001).

method over BS *vs* M differences in the clustering of RNAseq data, it is perhaps unsurprising that the even more labile metabolites continue to pose challenges. Similarly, the low absolute enrichment of this method and the Leegood (1985) method decreases the signal to noise ratio, particularly making identification of low log fold changes between cell types difficult (as seen in the RNAseq). This is likely exacerbated by the division of some metabolites, such as aspartate and

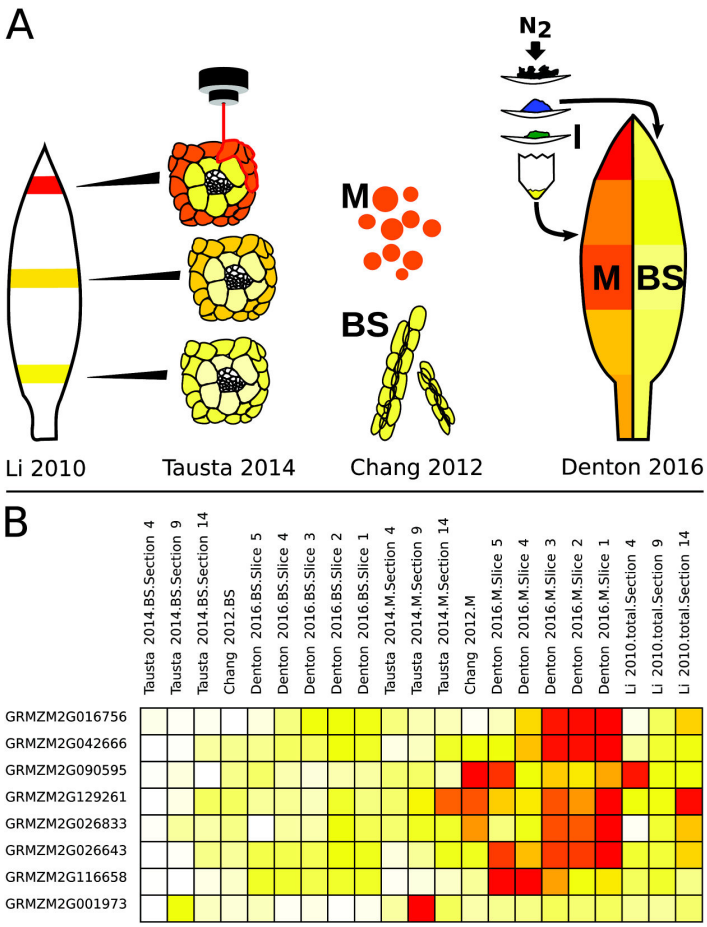


Fig. 7. Web visualization resource. (A) Comparative BS and M separation targeted graphical heatmap view of example gene (GRMZM2G129261). (B) Example gene set visualization of highest confidence M transcription factors.

malate, into active and inactive pools. These inactive pools can be substantial, accounting for about 60% and 80% of the total aspartate and malate, respectively, in the grass *Chloris gayana* (Hatch, 1979). In contrast, the high density of plasmodesmata between M and BS cells in C_4 plants supports diffusion of C_4-cycle metabolites at the rate of carbon fixation (Laisk and Edwards, 2000); it is thus implausible that any cytoplasmic metabolite could build up enrichment levels comparable to transcripts and enzymes. Therefore a study prioritizing understanding metabolic differences between BS and M cells should err on the side of a few more replicates than the five that is the 'industry standard' for metabolic studies (Sumner *et al.*, 2007). Similarly, sequencing a few more than the typical two to three replicates for RNAseq may help compensate for the lower sensitivity of this method.

For any study not targeting metabolites, the higher purity achieved by any of the other methods over the method here has an obvious allure; however, the biases associated with lower quality RNA must be accounted for. As shown here and reported previously (Romero *et al.*, 2014) RNA does not degrade at consistent rates, but rather some RNA molecules, often including transcription factors (Yang *et al.*, 2003), are much more sensitive to degradation. These degradation-sensitive genes are numerous (12.5% of detectable genes showed significantly lower abundance after laser micro-dissection; Li *et al.*, 2010; Tausta *et al.*, 2014). Further, shared genes with bias in Chang *et al.* (2012) and John *et al.* (2014) indicate degradation sensitivity is conserved across species and can masquerade as conserved tissue specificity. For the above reasons, care must be taken not to intermingle any biological signal sensitive to degradation and the biological signal between samples. For instance, the two callose synthases that Chang *et al.* (2012) discussed as being BS specific (GRMZM2G553532 and GRMZM2G004087) appear to be very sensitive to degradation as they are both among the genes significantly less abundant after laser micro-dissection, and one, GRMZM2G553532, is in the conflict set 1 list with strong 3′ bias. This raises the worrisome question of whether this is a case of differential expression, or differential degradation. Future studies may be able to circumvent such problems by including a third and unseparated sample that can be used to detect genes particularly affected by degradation—much

as we've been using the unseparated section 14 from Li *et al.* (2010) as the context for the Tausta *et al.* (2014) separated section. This method has been employed by a recent study using SuperSage on mechanically separated BS and M protoplast in Sorghum (Döring *et al.*, 2016).

A more ideal solution is of course to avoid mixing biological and technical signals by handling RNA in a fashion that preserves RNA quality or at least results in the same amount of degradation in the M and BS samples. Quality control must be performed carefully as a study using the same separation technique as John *et al.* (2014) for qPCR in sorghum achieved very comparable bioanalyser traces for their mechanical BS purified and their leaf-rolled M samples (Covshoff *et al.*, 2013). Indeed, this method was specifically employed for its speed and lack of stress response compared with M protoplast isolation, but still showed distinctly higher 3′ bias in M than BS (John *et al.*, 2014). Thus if a distinct method is to be used for M and BS purification, equivalent RNA needs to be confirmed for the particular species and particular researcher, and not simply assumed based on literature.

While the micro-dissection studies had the strongest overall 3′ bias, there was equivalent bias in the M and BS samples. This resulted in false negatives and lower library complexity, but had no clear link to false positives. In the microdissection studies, an alternative explanation for the 3′ bias is the synthesis of the first strand cDNA using an Arcturus Ribo Amp HS kit, which has been shown to induce a strong 3′ bias in housekeeping genes (Clément-Ziza *et al.*, 2009). This does not, however, nullify the substantial differences and loss of transcript detection seen between the laser micro-dissected (Li *et al.*, 2010; Tausta *et al.*, 2014) and the unseparated samples (Li *et al.*, 2010). There is ongoing research in improving laser-micro-dissection techniques in plants (Ludwig and Hochholdinger, 2014). We recommend that while techniques remain uncertain, researchers invest the necessary time and money in quality control steps and unseparated controls to assure that the bias that is there is traceable.

Use of a different bioinformatics workflow may make a small difference in the measured abundance of genes with a strong 3′ bias, but a perfect solution is not yet available, particularly as tools are not optimized for this. Small additions to a typical workflow, such as flagging discrepancy in 3′ bias between groups (e.g. Chang *et al.*, (2012)'s samples in Fig. 4), could help avoid erroneous conclusions.

Where one study has weaknesses, interstudy comparison can provide a helpful additional opinion. The completion of a third *Z. mays* M and BS separation RNAseq study with a complementary technique here continued to yield new biological results. Particularly in areas where results may seem dubious, consensus between several studies (with different techniques or information gathered) is required to gain confidence. An example of this is AspAT's consistent M localization, which while previously noted (Chang *et al.*, 2012; Tausta *et al.*, 2014), was not taken seriously without the supporting enzyme activity data. It may have a simple explanation such as a higher substrate to product ratio in the BS requiring less enzyme, or a more complex one such as an aspartate pool in the M simply adding stability to CO_2 fixation should diffusion or decarboxylation of malate become temporarily limiting. Either way, this warrants further investigation. Similarly, the switch from BS to M specificity of ribosomal proteins is much easier to trust when identified in two independent studies. Differentiation of veins and the associated BS cells precedes that of the M, and signals from the BS are necessary for M differentiation in Arapidopsis (Kinsman and Pyke, 1998; Lundquist *et al.*, 2014), and the C_4 dicot *Gynandropsis gynandra* shows the same developmental trajectory (Külahoglu *et al.*, 2014). Therefore, we hypothesize the initial BS enrichment in protein synthesis may reflect faster differentiation and photosynthetic ramp-up in the BS cells. As the photosynthetic rate increases along the developing leaf (Pick *et al.*, 2011), the shifting of the protein synthesis towards the M likely supports the high turnover of photosystem II subunits (Rokka *et al.*, 2005). Considering that photosynthesis-related proteins make up over half of mature leaf protein (Friso *et al.*, 2010; Majeran *et al.*, 2010), the distribution of protein synthesis in mature leaf may reflect the balance between the demand from synthesizing photosystem II (when in the M) and synthesizing Rubisco and the other BS-specific CBBC enzymes.

Altogether, the separation technique of choice depends upon the research question. In many cases the weaknesses of one study are compensated for by the strengths of another, particularly when biases are characterized and taken into consideration. This work provides a visual access tool summarizing this study and Li *et al.* (2010), Chang *et al.* (2012) and Tausta *et al.* (2014), tables of all data looked at here (above and Wang *et al.*, 2013; Sekhon *et al.*, 2013; John *et al.*, 2014; Rao *et al.*, 2016), and highlights biological observations drawn from the sum of many studies.

Data deposition

The following data are available at Dryad Digital Repository http://dx.doi.org/10.5061/dryad.tf6q6.

Datasets S1. Enzyme activity and metabolite abundance.
Datasets S2. Functional enrichments.
Datasets S3. Compiled RNAseq data.
Datasets S4. By gene 3′ bias.
Datasets S5. Gene sets of interest.
Fig. S1. Setup and confirmation of separation method.
Fig. S2. Metabolite enrichment.
Fig. S3. Contextual data for interstudy comparison.
Fig. S4. Coverage of BS (Chang *et al.*, 2012) *vs* M (this study) conflict genes.
Fig. S5. Example read coverage.
Fig. S6. Co-purification of additional tissues.
Fig. S7. AlaAT and 40S ribosome distributions.
Table S1. Bioinformatics parameters.
Table S2. Counting significant differences.
Table S3. Harvest and growth conditions.

Acknowledgements

We acknowledge Katrin Weber for GC/MS assistance; Simon Schliesky for data management support; Björn Usadel and Shin-Han Shiu for helpful discussion; and the Deutsche Forschungsgemeinschaft (IRTG 1525 supporting

JM, CK and AKD; EXC 1028 to MJL and APMW; and CRC 680 to MJL) as well as the European Union 7th Framework Program (EU 3to4 to APMW) for financial support.

Competing interests

The authors declare that they have no competing financial interests.

Author contributions

AKD, AB and APMW designed the study. AKD performed wetlab measurements, computational, statistical and general data analysis and wrote the manuscript. JM developed the visualization tool, performed general data analysis, and assisted in wetlab measurements and writing the manuscript. CK, MJL, AB, and APMW contributed to data analysis, interpretation of the results, and writing the manuscript.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25,** 3389–3402.

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics **31,** 166–169.

Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (last accessed 29 November 2016).

Aubry S, Kelly S, Kümpers BM, Smith-Unna RD, Hibberd JM. 2014. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C_4 photosynthesis. PLoS Genetics **10,** e1004365.

Bellasio C, Griffiths H. 2014. The operation of two decarboxylases, transamination, and partitioning of C_4 metabolic processes between mesophyll and bundle sheath cells allows light capture to be balanced for the maize C_4 pathway. Plant Physiology **164,** 466–480.

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 **57,** 289–300.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics **29,** 1165–1188.

Bennetzen JL, Schmutz J, Wang H, *et al.* 2012. Reference genome sequence of the model plant Setaria. Nature Biotechnology **30,** 555–561.

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

Bräutigam A, Kajala K, Wullenweber J, *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, Schliesky S, Külahoglu C, Osborne CP, Weber AP. 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.

Broglie R, Coruzzi G, Keith B, Chua NH. 1984. Molecular biology of C_4 photosynthesis in *Zea mays*: differential localization of proteins and mRNAs in the two leaf cell types. Plant Molecular Biology **3,** 431–444.

Chang Y-M, Liu W-Y, Shih AC-C, *et al.* 2012. Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. Plant Physiology **160,** 165–177.

Clément-Ziza M, Gentien D, Lyonnet S, Thiery JP, Besmond C, Decraene C. 2009. Evaluation of methods for amplification of picogram amounts of total RNA for whole genome expression profiling. BMC Genomics **10,** 246.

Covshoff S, Furbank RT, Leegood RC, Hibberd JM. 2013. Leaf rolling allows quantification of mRNA abundance in mesophyll cells of sorghum. Journal of Experimental Botany **64,** 807–813.

Covshoff S, Szecowka M, Hughes TE, *et al.* 2016. C_4 photosynthesis in the rice paddy: insights from the noxious weed *Echinochloa glabrescens*. Plant Physiology **170,** 57–73.

Deeken R, Ache P, Kajahn I, Klinkenberg J, Bringmann G, Hedrich R. 2008. Identification of *Arabidopsis thaliana* phloem RNAs provides a search criterion for phloem-based transcripts hidden in complex datasets of microarray experiments. The Plant Journal **55,** 746–759.

Denton AK, Maß J, Külahoglu C, Lercher MJ, Bräutigam A, Weber APM. 2016. Data from: Freeze-quenched maize mesophyll and bundle sheath separation uncovers bias in previous tissue-specific RNA-Seq data. Dryad Digital Repository http://dx.doi.org/10.5061/dryad.tf6q6

DOE-JGI. 2016. Panicum virgatum v1.1. https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvirgatum (last accessed 29 November 2016).

Döring F, Streubel M, Bräutigam A, Gowik U. 2016. Most photorespiratory genes are preferentially expressed in the bundle sheath cells of the C_4 grass *Sorghum bicolor*. Journal of Experimental Botany **67,** 3053–3064.

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

Furbank RT. 2011. Evolution of the C_4 photosynthetic mechanism: are there really three C_4 acid decarboxylation types? Journal of Experimental Botany **62,** 3103–3108.

Goodstein DM, Shu S, Howson R, *et al.* 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research **40,** D1178–D1186.

Gowik U, Bräutigam 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 C_4? The Plant Cell **23,** 2087–2105.

Hatch MD. 1979. Mechanism of C_4 photosynthesis in *Chloris gayana*: pool sizes and kinetics of $^{14}CO_2$ 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, Mau S. 1977. Properties of phosphoenolpyruvate carboxykinase operative in C4 pathway photosynthesis. Functional Plant Biology **4,** 207–216.

Huang P, Brutnell TP. 2016. A synthesis of transcriptomic surveys to dissect the genetic basis of C_4 photosynthesis. Current Opinion in Plant Biology **31,** 91–99.

Javelle M, Vernoud V, Depège-Fargeix N, Arnould C, Oursel D, Domergue F, Sarda X, Rogowsky PM. 2010. Overexpression of the epidermis-specific homeodomain-leucine zipper IV transcription factor Outer Cell Layer1 in maize identifies target genes involved in lipid metabolism and cuticle biosynthesis. Plant Physiology **154,** 273–286.

John CR, Smith-Unna RD, Woodfield H, Covshoff S, Hibberd JM. 2014. Evolutionary convergence of cell-specific gene expression in independent lineages of C_4 grasses. Plant Physiology **165,** 62–75.

Kersey PJ, Allen JE, Armean I, *et al.* 2016. Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Research **44,** D574–D580.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology **14,** R36.

Kinsman EA, Pyke KA. 1998. Bundle sheath cells and cell-specific plastid development in Arabidopsis leaves. Development **125,** 1815–1822.

Kodama Y, Shumway M, Leinonen R. 2012. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Research **40,** D54–D56.

Külahoglu C, Denton AK, Sommer M, *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.

Laisk A, Edwards GE. 2000. A mathematical model of C_4 photosynthesis: The mechanism of concentrating CO_2 in NADP-malic enzyme type species. Photosynthesis Research **66,** 199–224.

Leegood RC. 1985. The intercellular compartmentation of metabolites in leaves of Zea mays L. Planta **164,** 163–171.

Leinonen R, Akhtar R, Birney E, *et al*. 2010. The European nucleotide archive. Nucleic Acids Research **39,** D28–D31.

Li P, Ponnala L, Gandotra N, *et al*. 2010. The developmental dynamics of the maize leaf transcriptome. Nature Genetics **42,** 1060–1067.

Lohse M, Nagel A, Herter T, *et al*. 2014. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant, Cell & Environment **37,** 1250–1258.

Ludwig Y, Hochholdinger F. 2014. Laser microdissection of plant cells. Methods in Molecular Biology **1080,** 249–258.

Lundquist PK, Rosar C, Bräutigam A, Weber AP. 2014. Plastid signals and the bundle sheath: mesophyll development in reticulate mutants. Molecular Plant **7,** 14–29.

Majeran W, Cai Y, Sun Q, van Wijk KJ. 2005. Functional differentiation of bundle sheath and mesophyll maize chloroplasts determined by comparative proteomics. The Plant Cell **17,** 3111–3140.

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.

Meyer S, Melzer M, Truernit E, HuÈmmer C, Besenbeck R, Stadler R, Sauer N. 2000. AtSUC3, a gene encoding a new Arabidopsis sucrose transporter, is expressed in cells adjacent to the vascular tissue and in a carpel cell layer. The Plant Journal **24,** 869–882.

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.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Rao X, Lu N, Li G, Nakashima J, Tang Y, Dixon RA. 2016. Comparative cell-specific transcriptomics reveals differentiation of C_4 photosynthesis pathways in switchgrass and other C_4 lineages. Journal of Experimental Botany **67,** 1649–1662.

Rawsthorne S, Hylton CM, Smith AM, Woolhouse HW. 1988. Photorespiratory metabolism and immunogold localization of photorespiratory enzymes in leaves of C_3 and C_3-C_4 intermediate species of Moricandia. Planta **173,** 298–308.

Robinson MD, McCarthy DJ, Smyth GK. 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26,** 139–140.

Rokka A, Suorsa M, Saleem A, Battchikova N, Aro EM. 2005. Synthesis and assembly of thylakoid protein complexes: multiple assembly steps of photosystem II. The Biochemical Journal **388,** 159–168.

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

Romero IG, Pai AA, Tung J, Gilad Y. 2014. RNA-seq: impact of RNA degradation on transcript quantification. BMC Biology **12,** 42.

Rudolf M, Machettira AB, Groß LE, *et al*. 2013. In vivo function of Tic22, a protein import component of the intermembrane space of chloroplasts. Molecular Plant **6,** 817–829.

Sage RF, Zhu XG. 2011. Exploiting the engine of C_4 photosynthesis. Journal of Experimental Botany **62,** 2989–3000.

Schulze E-D, Ellis R, Schulze W, Trimborn P, Ziegler H. 1996. Diversity, metabolic types and $\delta^{13}C$ carbon isotope ratios in the grass flora of Namibia in relation to growth form, precipitation and habitat conditions. Oecologia **106,** 352–369.

Sekhon RS, Briskine R, Hirsch CN, Myers CL, Springer NM, Buell CR, de Leon N, Kaeppler SM. 2013. Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. PloS One **8,** e61005.

Shen Q, Hu J, Jiang N, Hu X, Luo Z, Zhang H. 2016. contamDE: differential expression analysis of RNA-seq data for contaminated tumor samples. Bioinformatics **32,** 705–712.

Stitt M, Heldt HW. 1985. Control of photosynthetic sucrose synthesis by fructose 2,6-bisphosphate: VI. Regulation of the cytosolic fructose 1,6-bisphosphatase in spinach leaves by an interaction between metabolic intermediates and fructose 2,6-bisphosphate. Plant Physiology **79,** 599–608.

Sumner LW, Amberg A, Barrett D, *et al*. 2007. Proposed minimum reporting standards for chemical analysis. Metabolomics **3,** 211–221.

Tausta SL, Li P, Si Y, Gandotra N, Liu P, Sun Q, Brutnell TP, Nelson T. 2014. Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C_4-related processes. Journal of Experimental Botany **65,** 3543–3555.

Tello-Ruiz MK, Stein J, Wei S, *et al*. 2016. Gramene 2016: comparative plant genomics and pathway resources. Nucleic Acids Research **44,** D1133–D1140.

Vidal J, Chollet R. 1997. Regulatory phosphorylation of C_4 PEP carboxylase. Trends in Plant Science **2,** 230–237.

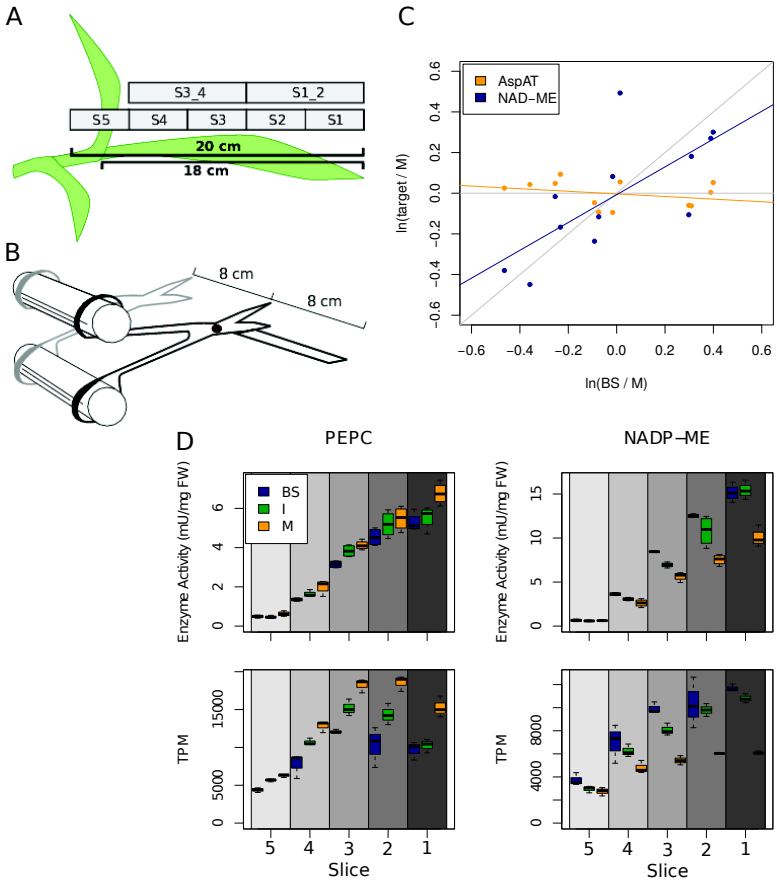
Walker RP, Trevanion SJ, Leegood RC. 1995. Phosphoenolpyruvate carboxykinase from higher plants: purification from cucumber and evidence of rapid proteolytic cleavage in extracts from a range of plant tissues. Planta **196,** 58–63.

Wang Y, Bräutigam A, Weber AP, Zhu XG. 2014. Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis. Journal of Experimental Botany **65,** 3567–3578.

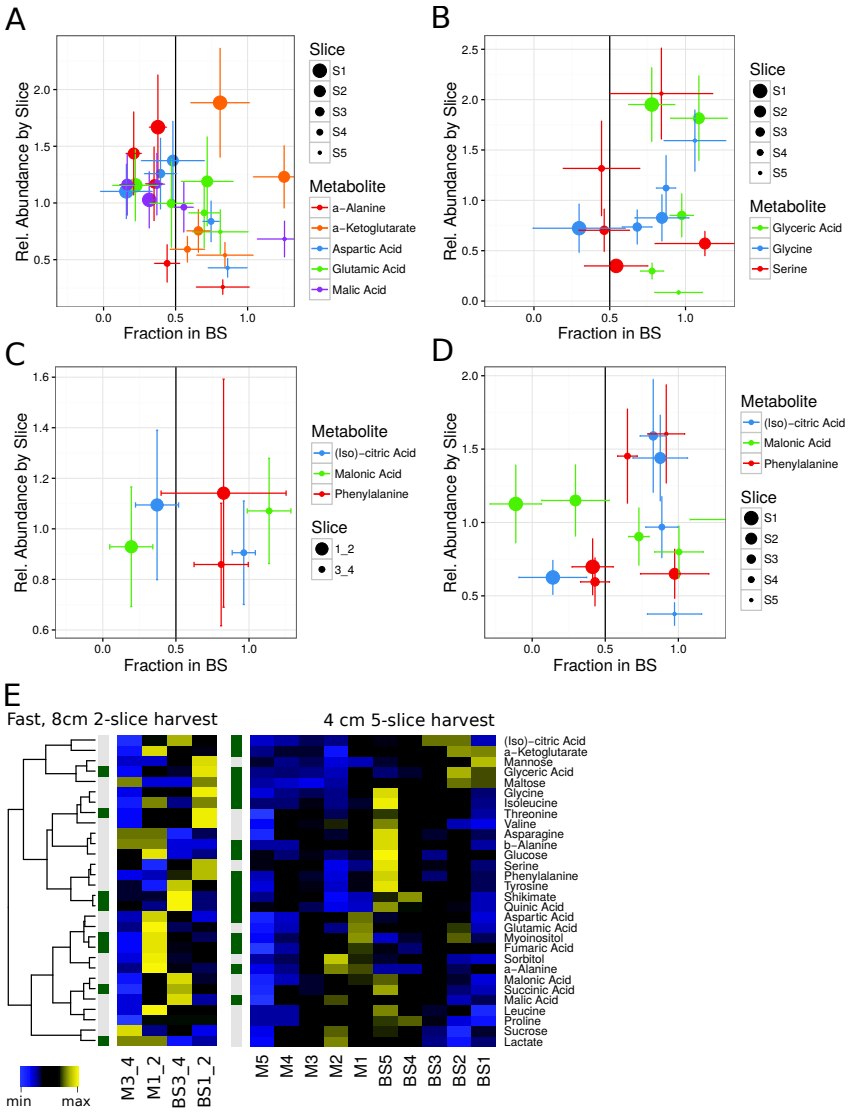
Wang P, Kelly S, Fouracre JP, Langdale JA. 2013. Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C_4 Kranz anatomy. The Plant Journal **75,** 656–670.

Wysoker A, Tibbetts K, Fennell T. 2012. Picard. http://broadinstitute.github.io/picard/ (last accessed 29 November 2016).

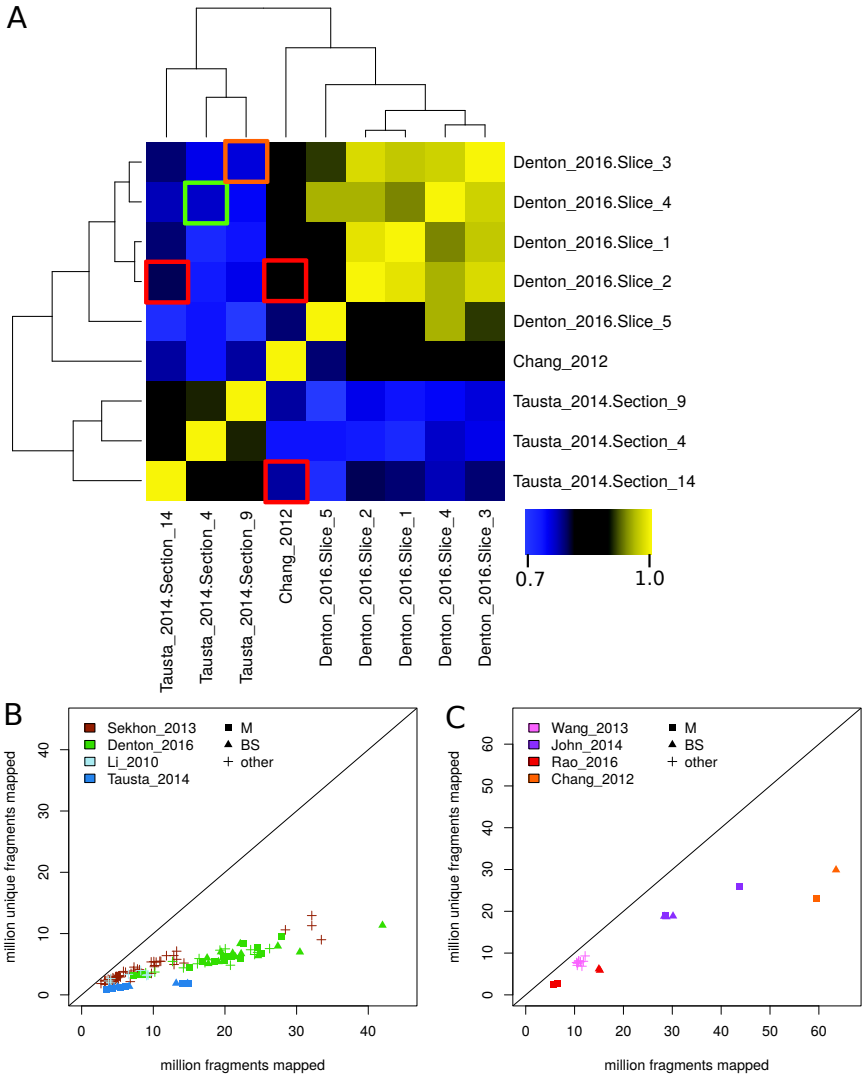
Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE Jr. 2003. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Research **13,** 1863–1872.



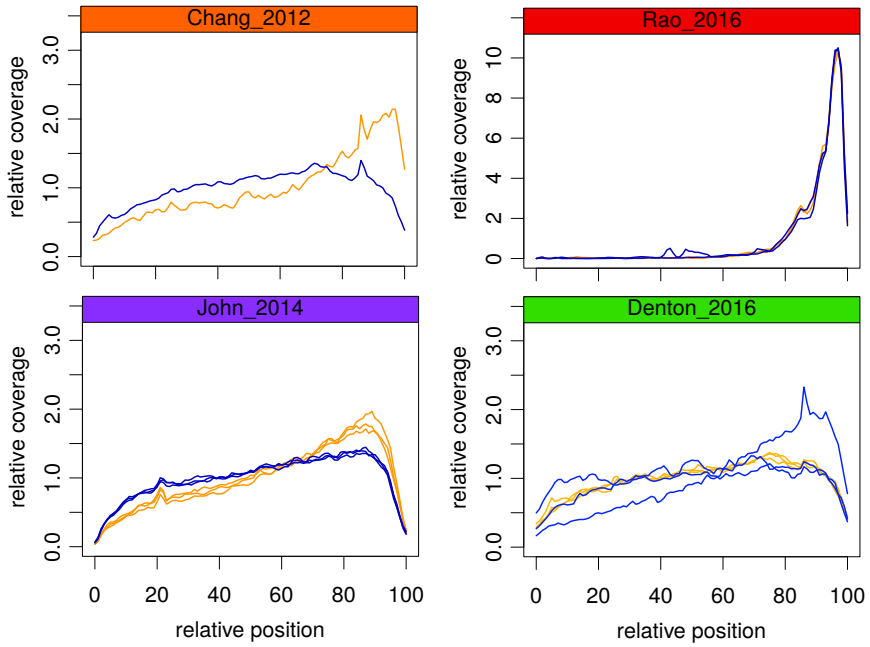
Supplementary Fig. 1: **Setup and Confirmation of Separation Method.** Summary of the leaf slices that were harvested (A) in the fast (less than 1s) two-slice harvest and the full five-slice gradient (10s) harvest. Position was defined by positioning the 2cm mark on the ruler at the leaf 2 ligule. Schematic (B) of paired scissors used for harvesting 8cm slices. Example (C) on how original distribution of target enzymes metabolites (here AspAT and NAD-ME) is estimated based on markers (enzyme activity of NADP-ME for BS, and PEPC for M). Distribution of maker enzyme activities (above) and transcript abundance (below) in raw enrichment fractions. BS-e: bundle sheath enriched fraction, I-e: intermediate enrichment fraction, M-e: mesphyll enriched fraction.



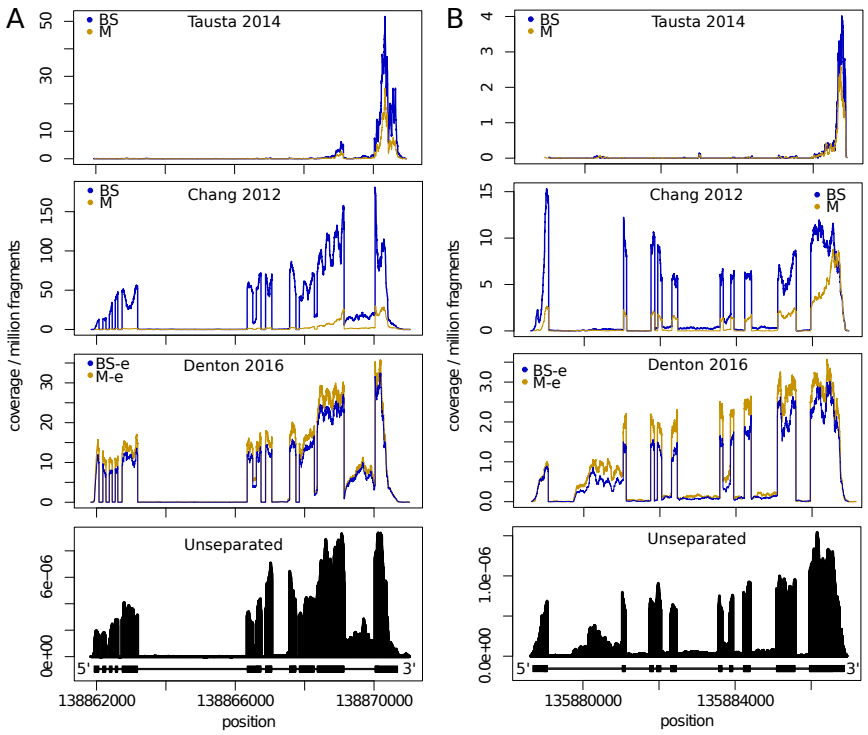
Supplementary Fig. 2: **Metabolite Enrichment** The estimated tissue enrichment and abundance of measureable metabolites associated with the photorespiratory cycle (A) and the C_4 cycle (B) in the slower five-slice harvest (the faster two-slice harvest is shown in maintext Figure 1). The same for three select metabolites (Malonic Acid, (Iso)-Citric Acid, and Phenylalanine) including both the two-slice (C) and five-slice (D) harvests. Error bars indicate standard error. Relative abundance of consistently-detectable metabolites (E), with green-boxes on side denoting metabolites that showed significantly (FDR <0.05) different abundance between at least two adjacent slices.



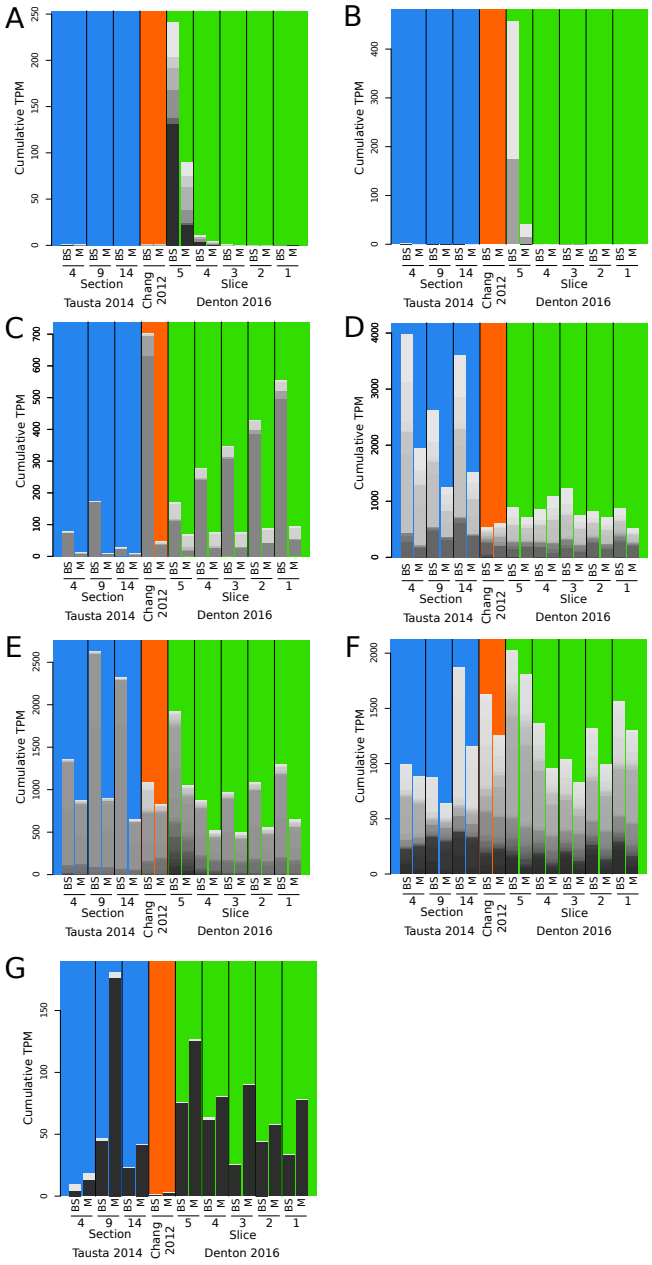
Supplementary Fig. 3: **Context of Inter-Study Comparison.** Spearman correlation between studies (A) was used to identify the most comparable samples for mature (Chang et al. 2012; Section 14 - Tausta et al., 2014; Slice 2 - Denton et al., 2016; red boxes) intermediate (Section 9 - Tausta et al., 2014; Slice 3 - Denton et al., 2016; orange box), and immature (Section 4 - Tausta et al.; Slice 4 - Denton et al., 2016; green box) tissues. Comparing the number of mapping vs uniquely mapping single end reads (B) or read pairs (C) provides an estimate of the total complexity of the RNAseq library.



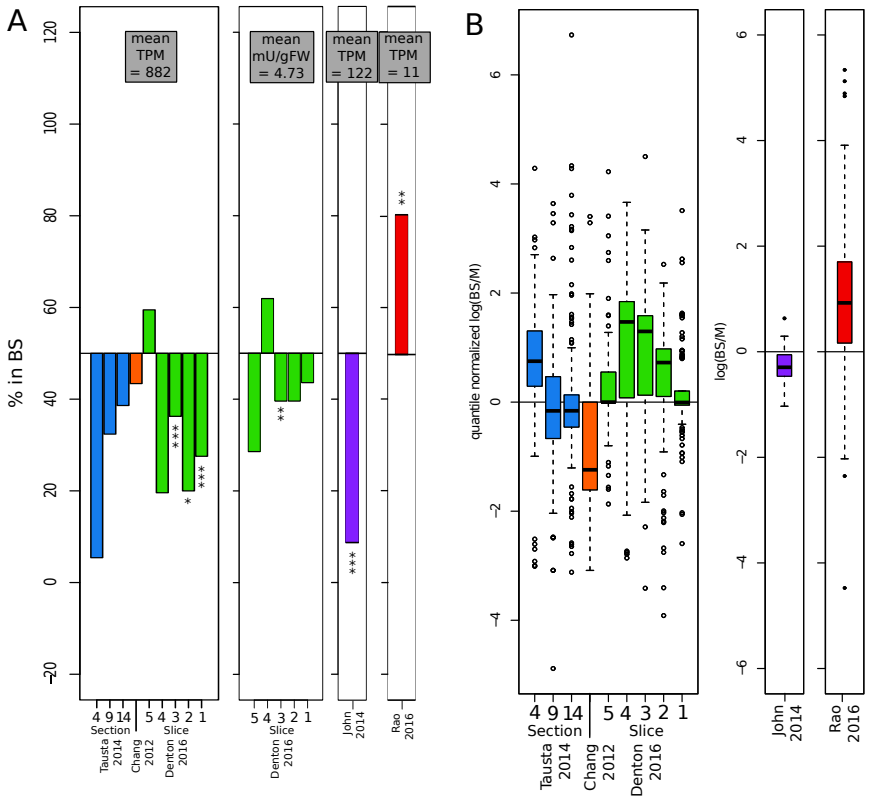
Supplementary Fig. 4: **Transcript coverage of genes with conflicting enrichments.** Relative coverage of *Z. mays* transcripts significantly enriched in the BS in Chang et al. (2012) and the M in Slice_2 (this study), and their orthologs in *P. virgatum* (Rao et al., 2016) and *S. viridis* (John et al., 2014).
Blue denotes BS, Yellow M, only the most comparable mature stage is shown.



Supplementary Fig. 5: **Coverage of Example Loci** Coverage of example genes (A) GRMZM2G095727 and (B) GRMZM2G053925 (which are both in conflict set 1) in the various M and BS *Z. mays* separation studies, and in unseparated samples of (Li et al., 2010; Wang et al., 2013; Sekhon et al., 2013).



Supplementary Fig. 6: **Evaluating Co-purification of additioan tissues.** Fully normalized abundance of genes that included the word "LAC17" (A), "XCP" (B), "SUT2" (C; AtSUT3 homologs), "xylem" (E), or "vascular" (F) in their MapMan description. Further fully normalized abundance of putatively phloem transported transcripts (D) from table 1 of (Deeken et al., 2008), and genes (G) that were identified by Javelle et al., 2010 as being expressed in the epidermis but not in the mesophyll (GRMZM2G345700; higher abundance) and (GRMZM2G387360).



Supplementary Fig. 7: **Biological insights drawn from inter-study comparison.** Tissue enrichment of AlaAT (A) of transcripts in *Z. mays* (left), enzyme activity in *Z. mays* (mid-left), transcripts in *S. viridis* (mid-right) and transcripts in *P. virgatum* (right). Tissue enrichment of transcripts in the MapMan functional category for 40S ribosomal protein (B) in *Z. mays* (left), *S. italica* (mid), and *P. virgatum* (right). In panel (A) stars denote significance of FDR for transcripts and P-values for enzyme activity (* <0.05, ** <0.01, *** <0.001).

2.3 Manuscript 3: Expression divergence following gene duplication contributes to the evolution of the complex trait C_4 photosynthesis.

Overview

Title: Expression divergence following gene duplication contributes to the evolution of the complex trait C_4 photosynthesis.

Authors: Alisandra K. Denton*, Janina Maß*, Canan Külahoglu, Martin Lercher, Andrea Bräutigam and Andreas P.M. Weber

* These authors contributed equally to this manuscript

Co-first authorship

Contributions

- Assistance in wet lab work (see 2.2)
- Phylogenetic analyses (pipeline automation), incl.
  - homology detection
  - multiple sequence alignment
  - outlier filtering
  - phylogenetic tree reconstruction
  - selective pressure (dN/dS) determination
- Discussion and interpretation of data
- Editing of manuscript
- Additional bioinformatic support

Expression divergence following gene duplication enables the evolution of the complex trait C_4 photosynthesis

Alisandra K. Denton[1†], Janina Maß[2†], Canan Külahoglu[1], Martin J. Lercher[2], Andrea Bräutigam[1], and Andreas P.M. Weber[1]

[1]Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS), and iGRAD-*plant* Program, Heinrich-Heine-University, 40225 Düsseldorf, Germany.
[2]Institute for Computer Science, Cluster of Excellence on Plant Sciences (CEPLAS), and iGRAD-*plant* Program, Heinrich-Heine University, 40225 Düsseldorf, Germany.
[†]These authors contributed equally to this manuscript

Abstract

Whole genome duplications are hypothesized to promote the evolution of morphological diversity and complex traits. However, to date support for the hypothesis is limited to the timing of radiation shortly after whole genome duplications and to examples from individual gene families. Here we test on a genome-wide scale how expression divergence following gene duplication facilitated the evolution of a complex trait, C_4 photosynthesis. Genes encoding the core C_4 cycle and belonging to functional categories related to C_4 anatomy and energy balance show expression divergence between the highly specialized mesophyll and bundle sheath tissues. Similarly, genes from larger gene families tended to be more significantly differentially expressed between mesophyll and bundle sheath. This held in duplicates originating tens of millions of years before the evolution of C_4 photosynthesis, providing strong evidence that whole genome duplications both facilitate the evolution of C_4 photosynthesis and can contribute to complex trait evolution.

Whole genome duplications (WGD) are proposed to have facilitated the evolution of morphological diversity in lineages such as vertebrates [1-3] and flowering plants (Angiosperms) [4,5]. In the Angiosperm lineage, a WGD event took place prior to the radiation and diversification of monocots and dicots [4]. Similarly, the early vertebrate lineage underwent two rounds of WGD after its divergence from the amphioxus lineage and before the radiation of the main vertebrate lineages [6, 7]. However, timing of ancient WGDs should be interpreted with caution, as there are many known WGDs that are not apparently associated with major morphological specialization or radiation events [8,9]. There is anecdotal evidence for the contribution of ancient WGDs to lineage specific features, such as the expression of neural regulatory paralogs in the neural crest, a tissue type only found in vertebrates [3]. There is also a limited amount of experimental evidence linking gene duplication to complex trait evolution on a broader scale; for instance, genes with stress-responsive expression are enriched in tandem duplicates in *Arabidopsis thaliana* [10].

Quantitative analysis of expression divergence enables rigorous statistical testing of the functional importance and consequences of gene duplication on a genome-wide scale. Analyses of the transcriptomes of fungi, animals, and plants have elucidated basic trends in duplicate expression divergence [11]. These trends include greater expression divergence in duplicates from small scale duplications than WGDs [12,13] and a correlation of expression divergence with synonymous (*dS*) and, in young duplicates, non-synonymous (*dN*) substitutions rates [14-16]. To test the contribution of gene and genome duplications to the evolution of novel traits, we examine how expression divergence has contributed to the evolution of the complex trait C_4 photosynthesis on a genome wide scale.

The C_4 trait is an evolutionary add-on to the ancestral C_3 photosynthetic type that helps plants thrive in hot and arid environments [17]. The core of C_4 photosynthesis is a biochemical cycle that pumps CO_2 from the outer mesophyll (M) tissue to the interior bundle sheath (BS) tissue. However, the integrated trait involves extensive changes compared to the ancestral C_3 state, including specialized anatomy and partitioning of metabolism between M and BS tissues. Comparisons between closely related C_3 and C_4 species indicate that hundreds or perhaps thousands of genes undergo expression changes in leaf tissue during the evolution of C_4 photosynthesis [18-21]. Meta-comparison confirms many of these expression changes are shared between independent C_4 origins [22]. Despite this complexity, C_4 photosynthesis has evolved in more than 66 different Angiosperm lineages [17]. This degree of convergent evolution is possible, in part, because all of the enzymes in the biochemical cycle are already present in a C_3 plant, but with a different function [23]. Gene duplication is thought to precondition the evolution of C_4 photosynthesis by allowing one copy to maintain ancestral gene function, while the second copy can be recruited to C_4 photosynthesis. However, the role of gene duplication in C_4 evolution has recently been questioned [24,25] and largely investigated only for the core genes of the C_4 cycle [26].

Here, we investigate whether and how expression divergence following gene duplication facilitates the evolution of the C_4 trait in *Zea mays. Z. mays* is a highly duplicated C_4 species that underwent two readily traceable WGDs: the pan-grass duplication (~70 mya) and the *Z. mays* specific tetraploidy event (5-12 mya) [27]. The evolution of C_4 photosynthesis in the *Z. mays* lineage occurred between these two WGDs ~20mya [28]. Integrating transcriptomics [29-35] and phylogenetics, we characterize how the

expression divergence of known C_4 genes and functions contribute to the specialized biochemistry, energy balance, and anatomy of the trait. We find that paralogs from larger gene families show more divergent expression, and more BS and M tissue specificity genome wide, even when duplications occurred long before the evolution of C_4 photosynthesis.

Results

Measurement and compilation of expression data in grasses to cover important C_4 tissues.

To evaluate how gene duplication and subsequent expression divergence may facilitate C_4 evolution, we first needed a transcriptomic dataset capturing the development of the key M and BS tissues.

To this end, we harvested contiguous slices along a developmental gradient in *Z. mays* leaves (Supplementary Fig. 1; Fig. 1b) [36]; enriched BS and M tissues [37]; and measured levels of metabolites, transcripts, and enzyme activity (Supplementary Dataset 1, 2). The mechanical BS and M separation method employed here provides high quality RNA and metabolites as tissues are kept frozen from harvest until extraction; however, it results in only partial tissue enrichment. To estimate the original distributions of metabolites, enzymes and transcripts in M and BS tissues from the partial enrichment data, we "deconvoluted" the data based on markers (see methods; Supplementary Fig. 2). The deconvolution included a test for whether a target transcript, enzyme activity, or metabolite was significantly closer in distribution to either the M or BS marker. The deconvoluted data for mature tissues was consistent with previous studies with M and BS specific transcriptomes [38-40] at the level of individual genes (Supplementary Fig. 3, Supplementary Table 1) and functional categories (Supplementary Fig. 4,5), indicating that the separation method was effective.

To investigate the functional differences between M and BS tissues throughout development, we performed functional enrichment analysis for clusters and for differentially expressed genes. Of the eight *k*-means clusters, six showed simple patterns, and were high in the M, BS, or both tissue types in either the leaf tip or base (3 = "M-tip", 5 = "BS-tip", 6 = "even-tip", 4 = "M-base", 2 = "BS-base", and 7 = "even-base"); while the remaining two clusters were less distinct (1 = "mixed-BS" and 8 = "mixed-M"; Supplementary Fig. 6). The functional enrichments in mature tissue have been well described in previous separation studies (Supplementary Fig. 4, 5, supplementary note; [41, 44, 45]). In developing tissue, the "BS-base" cluster was enriched (Fisher's exact test, FDR < 0.05) in categories including cell, cell wall, and lignin biosynthesis. The "M-base" cluster was enriched in categories including lipid biosynthesis and tetrapyrrole biosynthesis (Supplementary Dataset 3).

To provide evolutionary perspective, published transcriptome data was collected from *Zea mays* and additional C_4 (*Sorghum bicolor*, *Setaria italica*, *Setaria viridis*) and C_3 (*Brachypodium distachion*, *Oryza sativa*) grass species [29-35]. This public data included photosynthetic and non-photosynthetic tissues; and, where possible, also included developing leaf (*Z. mays*, *S. italica*) and resolution of BS and M tissues (*Z. mays*, *S. viridis*, *O. sativa*). In each grass, the total collected data was sufficient to show a clear pattern for photosynthetic genes (Supplementary Fig. 7, 8).

Known C_4 genes show high expression divergence, photosynthetic expression patterns, and tissue specificity.

To investigate how gene duplication and expression divergence may contribute to the evolution of C_4 photosynthesis, we first examined how much and what sort of expression divergence occurs between the known core-C_4 cycle genes and their paralogs. Therefore, we qualitatively and quantitatively compared the expression divergence of core-C_4 genes (those in Fig. 1a) to the genome-wide background.

The core-C_4 genes shared distinct expression characteristics with each other that were not shared with their nearest homologs. First, the C_4 enzymes and transporters were expressed very highly (>300 FPKM; Fig. 2a). Second, all C_4 paralogs had expression patterns peaking in mature leaf tissue (Fig. 1a) that were photosynthetic-like (Fig. 2c), that is, highly correlated with photosynthesis genes (defined from MapMan category [41]; Supplementary Fig. 7). Third, many core-C_4 genes are known to be tissue specific to orchestrate the pumping of CO_2 from M to BS [42], which was consistent with our data (Fig. 1a, Fig. 2b Supplementary Fig. 9, 10, 11, 12, supplementary note). Similarly, core-C_4 genes required in both tissue types were fairly evenly expressed (triose phosphate transporter; TPT) or showed fairly even enzyme activity (Alanine Aminotransferase; AlaAT) between tissues, with the exception of Aspartate Aminotransferase (Fig. 1). Notably, these characteristics: high, photosynthetic-like, and M or BS specific expression, were held by just the C_4 paralog, and were not shared with other members of their gene families regardless of phylogenetic proximity (Fig. 2a-d). Indeed, four of the C_4 genes had a paralog with the opposite tissue specificity (based on clustering).

Quantitatively, C_4 genes and their paralogs were exceptionally divergent. The C_4 genes were significantly more divergent in expression pattern (divergence measured by Pearson's correlation coefficient, r_p, $p=0.016$) and level (divergence measured by natural log of the ratio of max expression, $p=1.01*10^{-5}$) than the background of other paralogs in the BS & M gradient (Fig. 2e,f). These correlations held when datasets with other photosynthetic, but not heterotrophic tissues were used to calculate divergence (Supplementary Table 3). We used multiple regression to check whether the high divergence between C_4 genes and their paralogs could be explained by known factors: C_4 or not, collinear or not, *dS*, *dN*, # *Z. mays* paralogs. The C_4 genes were sufficiently divergent from their paralogs that the "C_4 or not" factor significantly improved (pattern $p=6.88*10^{-6}$; level $p=0.0014$) a multiple regression model already including the afore-mentioned factors (Supplementary Fig. 13; Supplementary Table 4, 5).

To establish when adoption of core-C_4 expression patterns likely occurred, we examined expression patterns in a phylogenetic context. For each C_4 gene tree, we selected the (non-C_4) homologs in each targeted grass species that were the most closely related to the *Z. mays* C_4 gene and compared their expression to the remaining homologs (Fig. 2d). No significant differences were found between the nearest and remaining homologs in expression level, correlation to photosynthetic pattern, or tissue specificity (Fig. 2a-c; Supplementary Table 2). Thus the high divergence occurred very specifically between C_4 and non-C_4 paralogs. Indeed, even between C_4 genes and their young, syntenic paralogs there are large changes in expression pattern and level (Supplementary Fig. 14, 15, 16). These included paralogs derived from the maize tetraploidy event about 10 my after the evolution of C_4 photosynthesis (Supplementary Fig. 14, 15),

and in these cases the divergence most likely represents loss of the high, photosynthetic-like C_4 pattern in one of the paralogs.

As the protein sequences of many core C_4 genes are known to have evolved under positive selection [26,43], and the core C_4 genes were highly divergent in expression; we asked if there is a general relationship between selective pressure and expression pattern divergence. While no significant relation between pairwise dN/dS and expression divergence was observed (pattern p=0.38; level p=0.73), there was a significant positive correlation between dN and expression divergence in duplicates originating in the *Z. mays* tetraploidy, which are all of the same age (r_p^2=0.036, p=6.57*10^{-13}; Supplementary Fig. 17). Positive selection (dN/dS significantly > 1) can be more readily and reliably identified when more sequence information is included [44]; therefore, we compared dN/dS to expression divergence in a test set of 64 whole gene families. While there was a negative correlation (Spearman's $R\ (r_s)$ = - 0.08) between the p-value dN/dS > 1 at a branch and the mean of pairs for expression divergence between this branch and its sister branch, this was not significant (p=0.11). In summary, of the three measures for sequence level positive selection, only the dN of duplicated genes was significantly positively correlated to expression divergence. However, both of the other two measures—pairwise dN/dS and significance of dN/dS at tree branch—trended in the same direction. Thus, the relationship between sequence level positive selection and expression divergence remains undefined; however, the strong co-occurrence of positive selection and expression divergence seen in core C_4 genes appears to be an exceptional case.

Divergent expression between paralogs relates to specialization in C_4 anatomy and energy balance.

An integrated C_4 trait requires modifications to metabolism and anatomy that go far beyond the establishment of the core C_4 cycle, which is reflected in the high number of genes consistently differentially expressed across independent C_4 origins [22]. If WGDs, and not just gene duplications, are important for the evolution of C_4 photosynthesis, we expect gene duplication to contribute to the greater complexity of the C_4 trait.

We asked whether any C_4-related gene functions (MapMan categories) showed a tendency towards particular patterns of expression divergence. We used a graph theory approach to categorize the patterns of expression divergence, with the *k*-means expression clusters as nodes, and the number of paralog pairs found between clusters as edge weights (Fig. 3). For example, a pair of paralogs expressed in clusters 1 and 2 were assigned to the edge connecting clusters 1 and 2 (1_2), while a pair of paralogs that were both expressed in cluster 1, were assigned to the edge connecting cluster 1 to itself (not plotted), that is the “loop”, 1_1. To reduce noise, we excluded paralogs in different clusters that were more similar in expression to each other than to their cluster centers. Then we tested all edges for functional enrichments (MapMan categories). While most significant enrichments (FDR < 0.05) were found in loops, and these were unsurprisingly very similar to the enrichment of all genes in the cluster (Supplementary Dataset 3, 4); there were 76 significant enrichments in non-loop edges.

Paralogs with functions related to balancing energy in photosynthetic tissues were enriched between “M-tip” and “BS-tip” clusters. Specifically the photosynthesis category was enriched in the edge 3 5

("M-tip" to "BS-tip") and a closer look showed edge 3_5 contained subunits of photosystem I, and ATP-consuming enzymes from the Calvin-Benson-Bassham (CBB) and photorespiratory cycle (Supplementary Table 6). Maize has a complex energy balance between cell types, with photosystem II restricted to M cells, and several cycles shuttling reducing equivalents into the BS. The use of two decarboxylation enzymes is proposed to add stability to the energy balance between subtypes in fluctuating light conditions [45, 46], and sub- or neo-functionalization of the ATP-consuming enzymes between mature BS and M tissue could add further robustness or fine-regulation to energy balance. Alternatively, two of these enzymes have their highest expression in the M, and the secondary BS specific paralog could provide an overflow mechanism if and when diffusion were to become limiting to the photorespiratory or CBB cycles.

Paralogs with functions related to C_4 anatomy were enriched between various immature clusters. The modifications in vascular patterning required for C_4 photosynthesis are thought to require changes in auxin levels and perception [47, 48], and both auxin response transcription factors and their downstream targets were enriched in the edge 2_7 ("BS-base" to "even-base"; Supplementary Dataset 4). The same edge is further enriched in cell wall categories, which could support the specialized anatomy observed in the BS cell wall [49]. Genes classified under miscellaneous gluco- galacto- and mannosidases were enriched in the edge 4_7 ("M-base" to "even-base"; Supplementary Dataset 4). These genes included 1,3 beta-galactosidases and various cellulases, which are often associated with loosening or modification of the cell wall [50, 51]. Thus, paralogs with functions related to C_4 anatomy, show specialized developmental expression patterns consistent with the anatomical specialization of BS and M tissues in C_4 plants.

Evidence for a preconditioning effect of gene duplication on changes in photosynthetic expression pattern and tissue-specific expression patterns.

C_4 photosynthesis involves extensive anatomical and metabolic changes to leaf tissue, in particular specialization of functions between M and BS tissues. We hypothesized gene duplication may contribute to expression-level specialization in general photosynthetic, M and BS tissues.

The evolution of photosynthesis involves the recruitment of genes to expression in photosynthetic tissues. This is found for both the core- C_4 genes [18] (Fig. 2c) and for regulatory genes thought to shape C_4-anatomy [29, 52, 53]. If the increasing expression divergence observed in larger gene families (Supplementary Fig. 13; Supplementary Table 5) facilitates the evolution of C_4 photosynthesis, we further expect expression divergence to include recruitment of genes to a photosynthetic pattern. To test this, we first classified the expression pattern of every gene in every species as photosynthetic or not. This classification was based on the bi-modal distribution of r_p between each gene's expression pattern and the expression pattern of the photosynthesis genes (MapMan category; Supplementary Fig. 18). This was used to test whether duplication level (# paralogs) related to how frequently photosynthetic-like gene expression patterns were unique to one species (as a parsimony-based proxy for gain), present all but one species (proxy for loss) or shared between all species. We found that higher levels of gene duplication were associated with species specificity in both presence and absence of photosynthetic pattern compared to conserved photosynthetic pattern across all species (Fig. 4ab; Supplementary Fig. 19). This indicates that the general correlation

between gene duplication level and expression divergence (Supplementary Fig. 13; Supplementary Table 5) includes both recruitment to, and loss of, photosynthetic pattern.

To test whether gene duplication promotes or preconditions the tissue specificity that is characteristic of C_4 photosynthesis, we tested for a correlation between gene family size and tissue specificity. The average p-value for tissue specificity along the developmental leaf gradient was negatively correlated (r_s=-0.071; p<0.001) with gene family size (Fig. 4c). To determine if this was specific to C_4 photosynthesis, we compared the p-value for tissue specificity in rice microarray data [35] to gene family size and found the opposite pattern (Fig. 4d; r_s=0.029; p<0.05). Thus, in the *Z. mays* data, but not in the rice data, larger gene families show higher tissue specificity.

We asked if the afore-mentioned correlations between gene family size and C_4-related expression changes arose in duplicates from before and after C_4-evolution. We identified 'ancient' orthologous groups, which had not further expanded after the time of the pan-grass genome duplication (minimum pairwise *dS* > 1), and 'young' orthogroups, which have expanded entirely since the time of the *Z. mays* tetraploidy (maximum pairwise *dS* < 0.3; Supplementary Fig. 20). The association between gene family size and increased change in photosynthetic expression pattern mostly held in both ancient (p < 0.05 for 7 of 10 comparisons) and young (p < 0.05 for 6 of 10 comparisons) gene families (Supplementary Fig. 21, 22). The increase in tissue-specificity with gene family size was found in ancient gene families; however, the opposite correlation was found in young gene families (Supplementary Fig. 23). This may relate to silencing of younger duplicates, as genes with low expression level are generally harder to detect and reliably measure tissue specificity. In summary, ancient gene duplications prior to C4 trait evolution are associated with increased changes in gene expression patterns relevant to the evolution of C_4, in particular increased tissue specificity, but not younger duplications.

Discussion

Ancient whole genome duplications are thought to have promoted the evolution of the morphological diversity observed in vertebrates and angiosperms today. However, few studies link gene duplication to evolutionary traits on a genome wide scale. Here, we have tested how gene duplication (including the WGDs at 70mya and 5-12mya) [27] and the following expression divergence have contributed to the evolution and integration of the complex trait C_4 photosynthesis that emerged ~20mya in the ancestors of *Z. mays* [28].

We find high expression divergence in the core-C_4 genes, in particular recruitment of the C_4 paralog to a high amplitude, photosynthetic-like, tissue-specific expression pattern. This expression divergence was significantly higher than expected, especially when accounting for other factors related to expression divergence (dN, dS, collinearity and *Z. mays* gene family size). Similarly, there is a striking co-occurrence of positive selection on the amino acid sequence [26,43] and expression divergence in core-C_4 genes, but only a weak association on a wider scale.

Expression divergence following gene duplication is associated with the specialization required for C_4 photosynthesis. Duplicates with functional annotations related to C_4 photosynthesis show specific

patterns of expression divergence. Namely, ATP-consuming photosynthetic enzymes diverge between mature M and BS tissue, where they have a likely role in energy balance; while various anatomy-related categories diverge between BS, M and even base clusters. More generally, gene duplication was associated with gain and loss of photosynthetic-like expression pattern, which could promote the changes in expression of hundreds to thousands of genes observed between mature leaves of closely related C_3 and C_4 species [18–22]. Supporting the importance of specialization in leaf tissue, the divergence of core- C_4 genes appears greater in photosynthetic tissues (Supplementary Table 3) and *Z. mays* duplicates have previously been found to show greater expression divergence in foliar leaves than husks [54].

Further specialization between duplicates is found in BS and M tissue. During C_4 evolution BS tissue takes on a major photosynthetic role in addition to its ancestral function as a “smart pipe” regulating access to the vasculature [55, 56]. Both M and BS undergo metabolic specialization as many functions are divided between them [34, 38–40, 57, 58]. The positive correlation between gene-family size and BS or M specificity in *Z. mays* (Fig. 4c) indicates duplicated genes may be more able to evolve tissue specificity. This correlation held in gene families that have not expanded since the pan-grass duplication roughly 50 million years before C_4 evolution [27, 28]. This raises the question of how a large reservoir of duplicates was maintained for ~50 million years before being incorporated in the C_4-related patterns. However, the phenomenon is not unusual: a large portion of duplicates from the more recent tetraploidy (5-12 mya) have yet to be lost in the ongoing process of diploidization [59]. Thus, ancient genome duplications tens of millions of years prior to trait evolution can facilitate the evolution of expression patterns important for C_4 photosynthesis.

An important question is how duplication mechanistically facilitates changes in expression. Small-scale duplications show more divergent expression, which has been attributed to duplication without the regulatory sequence. While for WGDs, it could be due to more rapid accumulation of single nucleotide polymorphisms associated with the reduced purifying selection seen in duplicates [60,61]. Alternatively, in a species like *Z. mays*, which is still undergoing diploidization and massive genome arrangement after the tetraploidy event [62,63], expression divergence may result from the rearrangement or loss of neighboring genes and *cis*-regulatory regions. Notably, the core-C_4 duplicates with high divergence despite their youth and colinearity (PEPCK, PPDK, and PPDK_RP) all show rearrangement of genes in the immediate upstream region (Supplementary Fig. 24) [64]. We hypothesize that such rearrangements may have contributed to the high expression divergence of C_4 genes. Finally, where transcriptional regulators show expression divergence (e.g. Auxin Response Factors diverging between “BS-base” and “even-base” clusters; Supplementary Dataset 4), their down-stream targets will be affected. Duplicated transcription factors may increase the robustness of the regulatory network [65], show subfunctionalization in their targets [66], or become antagonistic [67]. An antagonistic paralog would be one way to achieve the observed tissue specificity of some C_3 *cis*-regulatory regions when heterologously expressed in C_4 species [24].

Overall, this study connects genome-wide changes in expression to the evolution of a complex trait, showing how duplication facilitates C_4 evolution. It builds a bridge between the numerous single-gene family studies and large-scale correlation based studies to improve our understanding of evolutionary processes. To

further understand the contribution of WGDs to complex trait evolution, it will be important to perform additional large-scale, yet function-oriented studies. In particular, examining species with relatively recent WGDs paired with relatively recently evolved traits of interest would increase specificity and potentially allow for a more mechanistic understanding of divergence after WGD.

Accession Numbers

The reads related to this article have been deposited in the Sequence Read Archives under the accession number SRP052802.

Acknowledgements

We acknowledge Katrin L. Weber for assistance with GC/MS; Simon Schliesky for data management support; Shin-Han Shiu and the Michigan State University High Performance Computing Cluster team for a good computing experience; and the German Research Foundation for financial support (IRTG 1525 supporting J.M., C.K. and A.K.D.; EXC 1028 to M.J.L. and A.P.M.W.; and CRC 680 to M.J.L.).

Competing Interests

The authors declare that they have no competing financial interests.

Correspondence

Correspondence and requests for materials should be addressed to A.P.M. Weber (email: Andreas.Weber@uni-duesseldorf.de).

Author Contributions

A.K.D., A.B. and A.P.M.W. designed the study. A.K.D. performed wetlab measurements, all computational and statistical data analysis not related to phylogenetics, general data analysis and wrote the manuscript. J.M. performed all computational and statistical data analysis related to phylogenetics, general data analysis, and assisted in wetlab measurements and writing the manuscript. C.K., M.J.L., A.B., and A.P.M.W. contributed to data analysis, interpretation of the results, and writing the manuscript.

References

1. Holland, P. W., Garcia-Fernández, J., Williams, N. A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Development* **1994**, 125–133 (1994).
2. Holland, L. Z. Evolution of new characters after whole genome duplications: insights from amphioxus. *Seminars in cell & developmental biology* **24**, 101–9 (2013).
3. Van Otterloo, E., Cornell, R. A., Medeiros, D. M. & Garnett, A. T. Gene regulatory evolution and the origin of macroevolutionary novelties: insights from the neural crest. *Genesis* (New York, N.Y. : 2000) **51**, 457–70 (2013).
4. Jiao, Y. *et al*. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100

(2011).

5. De Bodt, S., Maere, S. & Van de Peer, Y. Genome duplication and the origin of angiosperms. *Trends in ecology & evolution* **20**, 591–7 (2005).

6. Sidow, A. Gen (om) e duplications in the evolution of early vertebrates. *Current opinion in genetics & development* **6**, 715–722 (1996).

7. Putnam, N. H. et al. The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064–1071 (2008).

8. Wood, T. E. *et al*. The frequency of polyploid speciation in vascular plants. *Proceedings of the national Academy of sciences* **106**, 13875–13879 (2009).

9. Cannon, S. B. *et al*. Polyploidy did not predate the evolution of nodulation in all legumes. *PLoS One* **5**, e11630 (2010).

10. Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K. & Shiu, S.-H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant physiology* **148**, 993–1003 (2008).

11. Maere, S. & de Peer, Y. V. Duplicate retention after small and large scale duplications. In Dittmar, K. & Liberles, D. (eds.) *Evolution after gene duplication*, chap. 3, 31–56 (Wiley-Blackwell, Hoboken, New Jersey, 2010).

12. Casneuf, T., De Bodt, S., Raes, J., Maere, S. & Van de Peer, Y. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome biology* **7**, R13 (2006).

13. Kim, J., Shiu, S.-H., Thoma, S., Li, W.-H. & Patterson, S. E. Patterns of expansion and expression divergence in the plant polygalacturonase gene family. *Genome biology* **7**, R87 (2006).

14. Chung, W.-Y., Albert, R., Albert, I., Nekrutenko, A. & Makova, K. D. Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. *BMC bioinformatics* **7**, 46 (2006).

15. Gu, Z., Nicolae, D., Lu, H. H.-S. & Li, W.-H. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics* **18**, 609–613 (2002).

16. Makova, K. & Li, W. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome research* **13**, 1638–1645 (2003).

17. Sage, R. F., Sage, T. L. & Kocacinar, F. Photorespiration and the evolution of C_4 photosynthesis. *Annual Review of Plant Biology* **63**, 19–47 (2012).

18. Külahoglu, C. *et al*. Comparative transcriptome atlases reveal altered gene expression modules between two cleomaceae C_3 and C_4 plant species. *The Plant Cell* **26**, 3243–3260 (2014).

19. Bräutigam, A. *et al*. An mRNA blueprint for C_4 photosynthesis derived from comparative transcriptomics of closely related C_3 and C_4 species. *Plant Physiology* **155**, 142–56 (2011).

20. Gowik, U., Bräutigam, A., Weber, K. L., Weber, A. P. M. & Westhoff, P. Evolution of C_4

photosynthesis in the genus Flaveria: how many and which genes does it take to make C_4 *The Plant Cell* **23**, 2087–105 (2011).

21. Mallmann, J. *et al*. The role of photorespiration during the evolution of C_4 photosynthesis in the genus flaveria. *Elife* **3**, e02478 (2014).

22. Bräutigam, A., Schliesky, S., Külahoglu, C., Osborne, C. P. & Weber, A. P. M. Towards an integrative model of C_4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C_4 species. *Journal of Experimental Botany* **65**, 3579–93 (2014).

23. Aubry, S., Brown, N. J. & Hibberd, J. M. The role of proteins in C(3) plants prior to their recruitment into the C(4) pathway. *Journal of experimental botany* **62**, 3049–59 (2011).

24. Williams, B. P., Aubry, S. & Hibberd, J. M. Molecular evolution of genes recruited into C_4 photosynthesis. *Trends in Plant Science* **17**, 213–20 (2012).

25. van den Bergh, E. *et al*. Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae. *Current Plant Biology* **1**, 2–9 (2014).

26. Wang, X. *et al*. Comparative genomic analysis of C_4 photosynthetic pathway evolution in grasses. *Genome biology* **10**, R68 (2009).

27. Schnable, J. C., Freeling, M. & Lyons, E. Genome-wide analysis of syntenic gene deletion in the grasses. *Genome biology and evolution* **4**, 265–77 (2012).

28. Christin, P.-A., Salamin, N., Kellogg, E. a., Vicentini, A. & Besnard, G. Integrating phylogeny into studies of C_4 variation in the grasses. *Plant physiology* **149**, 82–7 (2009).

29. Wang, P., Kelly, S., Fouracre, J. P. & Langdale, J. a. Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C_4 Kranz anatomy. *The Plant Journal : For Cell and Molecular Biology* **75**, 656–70 (2013).

30. Sekhon, R. S. *et al*. Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. *PloS one* **8**, e61005 (2013).

31. Davidson, R. M. *et al*. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *The Plant journal: for cell and molecular biology* **71**, 492–502 (2012).

32. Bennetzen, J. L. *et al*. Reference genome sequence of the model plant Setaria. *Nature biotechnology* **30**, 555–61 (2012).

33. Zhang, G. *et al*. Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nature biotechnology* **30**, 549–54 (2012).

34. John, C. R., Smith-Unna, R. D., Woodfield, H., Covshoff, S. & Hibberd, J. M. Evolutionary convergence of cell-specific gene expression in independent lineages of C_4 grasses. *Plant physiology* **165**, 62–75 (2014).

35. Jiao, Y. *et al*. A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nature genetics* **41**, 258–63 (2009).

36. Pick, T. R. *et al*. Systems analysis of a maize leaf developmental gradient redefines the current C_4 model and provides candidates for regulation. *The Plant Cell* **23**, 4208–20 (2011).

37. Stitt, M. & Heldt, H. W. Control of photosynthetic sucrose synthesis by fructose-2, 6-

bisphosphate: Intercellular metabolite distribution and properties of the cytosolic fructose-bisphosphatase in leaves of Z*ea mays* l. *Planta* **164**,179–188 (1985).

38. Tausta, S. L. *et al*. Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C_4-related processes. *Journal of Experimental Botany* **65**, 3543–55 (2014).

39. Li, P. *et al*. The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* **42**, 1060–1067 (2010).

40. Chang, Y.-M. *et al*. Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. *Plant Physiology* **160**, 165– 177 (2012).

41. Lohse, M. *et al*. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, cell & environment* **37**, 1250–1258 (2014).

42. Sheen, J. C_4 Gene Expression. *Annual Review of Plant Physiology and Plant Molecular Biology* **50**, 187–217 (1999).

43. Christin, P.-A. *et al*. Oligocene CO_2 decline promoted C_4 photosynthesis in grasses. *Current Biology : CB* **18**, 37–43 (2008).

44. Anisimova, M., Bielawski, J. P. & Yang, Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular biology and evolution* **18**, 1585–1592 (2001).

45. Bellasio, C. & Griffiths, H. The operation of two decarboxylases, transamination, and partitioning of C_4 metabolic processes between mesophyll and bundle sheath cells allows light capture to be balanced for the maize C_4 pathway. *Plant Physiology* **164**, 466–80 (2014).

46. Wang, Y., Bräutigam, A., Weber, A. P. M. & Zhu, X.-G. Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis. *Journal of Experimental Botany* **65**, 3567–78 (2014).

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

48. Denton, A. K., Simon, R. & Weber, A. P. C_4 Photosynthesis: From Evolutionary Analyses to Strategies for Synthetic Reconstruction of the Trait. Current Opinion in Plant Biology 16, 315–321 (2013).

49. Eastman, P. A. K., Dengler, N. G. & Peterson, C. A. Suberized bundle sheaths in grasses (Poaceae) of different photosynthetic types I. anatomy, ultrastructure and histochemistry. *Protoplasma* **142**, 92–111 (1988).

50. Goulao, L. F. & Oliveira, C. M. Cell wall modifications during fruit ripening: when a fruit is not the fruit. *Trends in Food Science & Technology* **19**, 4–25 (2008).

51. Levy, I., Shani, Z. & Shoseyov, O. Modification of polysaccharides and plant cell wall by endo-1, 4-β-glucanase and cellulose-binding domains. *Biomolecular engineering* **19**, 17– 30 (2002).

52. Slewinski, T. L., Anderson, A. a., Zhang, C. & Turgeon, R. Scarecrow plays a role in estab-

lishing Kranz anatomy in maize leaves. *Plant & cell physiology* **53**, 2030–7 (2012).

53. Slewinski, T. L. Using evolution as a guide to engineer kranz-type C_4 photosynthesis. *Frontiers in plant science* **4**, 212 (2013).

54. Hughes, T. E., Langdale, J. A. & Kelly, S. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. Genome research 24, 1348–1355 (2014).

55. Shatil-Cohen, A. & Moshelion, M. Smart pipes: the bundle sheath role as xylem-mesophyll barrier. *Plant signaling & behavior* **7**, 1088–1091 (2012).

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

57. Friso, G., Majeran, W., Huang, M., Sun, Q. & van Wijk, K. J. Reconstruction of metabolic pathways, protein expression, and homeostasis machineries across maize bundle sheath and mesophyll chloroplasts: large-scale quantitative proteomics using the first maize genome assembly. *Plant physiology* **152**, 1219–50 (2010).

58. Majeran, W. *et al*. Structural and metabolic transitions of C_4 leaf development and differentiation defined by microscopy and quantitative proteomics in maize. *The Plant cell* **22**, 3509–42 (2010).

59. Hirsch, C. N. *et al*. Insights into the maize pan-genome and pan-transcriptome. The Plant Cell Online 26, 121–135 (2014).

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

61. Hellsten, U. *et al*. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. *BMC Biology* **5**, 31 (2007).

62. Lai, J. *et al*. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature genetics* **42**, 1027–30 (2010).

63. Springer, N. M. *et al*. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS genetics* **5**, e1000734 (2009).

64. Proost, S. *et al.* Plaza 3.0: an access point for plant comparative genomics. *Nucleic acids research* **43**, D974–D981 (2015).

65. Choi, S. H. *et al*. Gene duplication of type-B ARR transcription factors systematically extends transcriptional regulatory structures in Arabidopsis. Scientific reports 4 (2014).

66. Pougach, K. *et al*. Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. Nature communications 5 (2014).

67. Floyd, S. K. *et al*. Origin of a novel regulatory module by duplication and degeneration of an ancient plant transcription factor. Molecular phylogenetics and evolution 81, 159–173 (2014).

Figure legends

Fig. 1: The core C_4 cycle: abundance and distribution of transcripts and enzyme activities (a). The bars from left to right are immature to mature M, followed by immature to mature BS, as summarized in (b). Bars show transcript abundance in FPKM with colors denoting different paralogs. Red lines represent relative enzyme activity. Inside: schematic summary of the core C_4 cycle with enzymes as pentagons, transporters as circles, and regulatory proteins as stars. Chloroplastic enzymes are in green compartments and the (putative) mitochondrial reactions in the purple compartments. Abrieviations: Asp = aspartate, Mal = malate, OAA = oxaloacetate, Pyr = pyruvate, PEP = phospho*enol*pyruvate, PPDK = pyruvate phosphate dikinase, PPDK RP = PPDK regulatory protein, PEPC = PEP carboxylase, PEPCK = PEP carboxykinase, PPCK = PEPC kinase, CA = carbonic anhydrase, PPT = phosphate/PEP tanslocator, ASPAT = aspartate amino transferase, ALAAT2 = alanine amino transferase 2, TPT = triose phosphate translocator, NAD(P) = nicotinamide adenine dinucleotide (phosphate), MDH = malate dehydrogenase, ME = malic enzyme.

Fig. 2: Expression characteristics and divergence of the core C_4 genes. The absolute expression level (excluding regulatory genes)(a), the BS or M tissue specificity (for C_4 genes that are tissue specific in Z. mays only)(b), and the similarity (r_p) in expression to the PS (photosynthesis) MapMan category (c) between C_4 genes (green), their phylogenetically closest homologs in each species (blue), and the remaining homologs (red). Example classification of homologs on a perfect, no-loss gene tree (d). Where there was only one non-C_4 homolog in any species (grey), homologs could not be classified as closest nor remaining and were excluded. Quantification of the divergence in expression pattern (e) and level (f) between the C_4 genes and their paralogs vs between all other paralogs.

Fig. 3: Paralogs in the edges between expression clusters and their functional enrichments. Example's of significantly enriched functions (fdr <0.05) are shown with an arrow to edge between clusters in which they are enriched (all significant enrichments included in Supplementary Dataset 4). Clusters plotted as z-scores with M base to tip followed by BS base to tip from left to right. The width of the lines connecting clusters is relative to the number of pairs in the edge connecting the respective clusters, while the color indicates whether the edges are larger (blue) or smaller (yellow) than expected based on the total number of pairs in non-loop edges of the connected clusters.

Fig. 4: The relationship between paralog number and expression characteristics related to C_4 evolution. The relation between photosynthetic pattern evolution and group size (# paralogs in orthogroup in respective species)(a-b). Cases where all 5 species show a photosynthetic-like expression pattern (see Supplementary Fig. 18) are considered conserved, while cases where 4 of 5 or 1 of 5 species show a photosynthetic-like expression pattern are considered gain or loss,

respectively. The odd species out is *Z. mays* in (a) or *O. sativa* in (b). The significance of tissue specificity (average p-value) vs the group size in (c) *Z. mays* and (d) *O. sativa*.

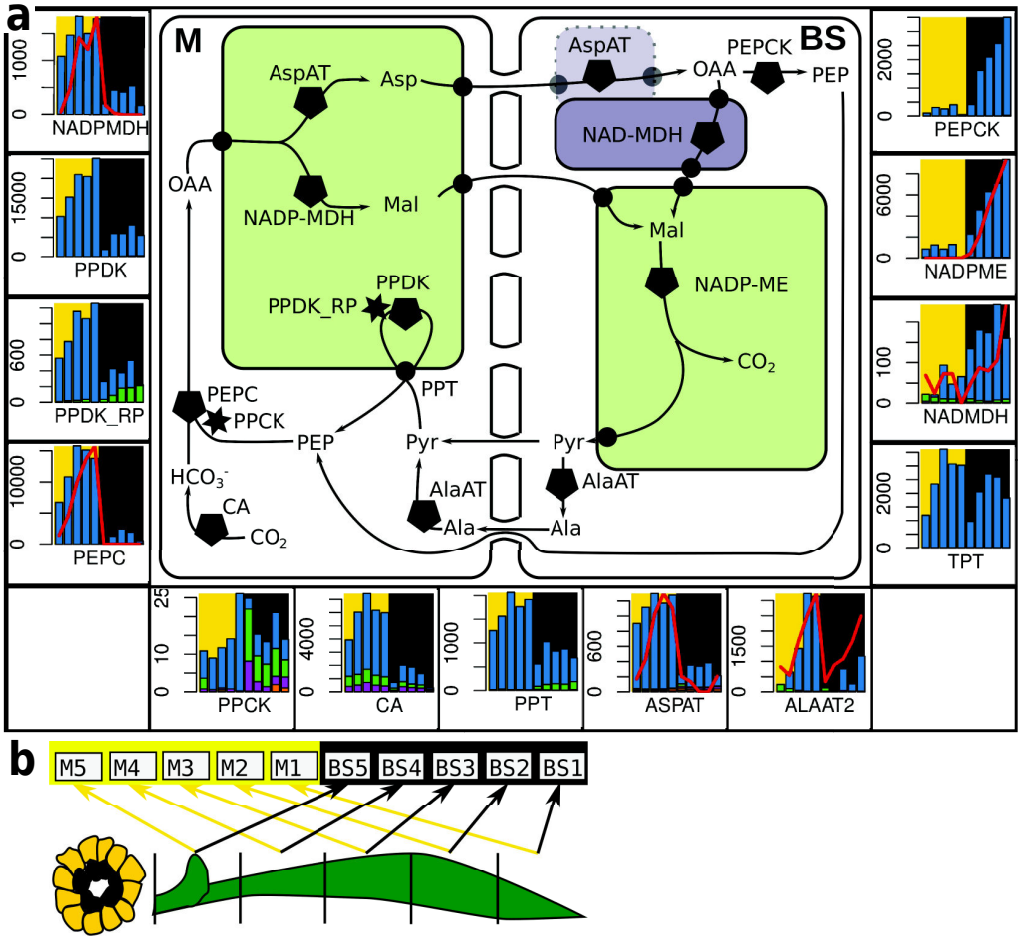


Fig. 1: The core C_4 cycle: abundance and distribution of transcripts and enzyme activities (a). The bars from left to right are immature to mature M, followed by immature to mature BS, as summarized in (b). Bars show transcript abundance in FPKM with colors denoting different paralogs. Red lines represent relative enzyme activity. Inside: schematic summary of the core C_4 cycle with enzymes as pentagons, transporters as circles, and regulatory proteins as stars. Chloroplastic enzymes are in green compartments and the (putative) mitochondrial reactions in the purple compartments. Abrieviations: Asp = aspartate, Mal = malate, OAA = oxaloacetate, Pyr = pyruvate, PEP = phospho*enol*pyruvate, PPDK = pyruvate phosphate dikinase, PPDK_RP = PPDK regulatory protein, PEPC = PEP carboxylase, PEPCK = PEP carboxykinase, PPCK = PEPC kinase, CA = carbonic anhydrase, PPT = phosphate/PEP tanslocator, ASPAT = aspartate amino transferase, ALAAT2 = alanine amino transferase 2, TPT = triose phosphate translocator, NAD(P) = nicotinamide adenine dinucleotide (phosphate), MDH = malate dehydrogenase, ME = malic enzyme.

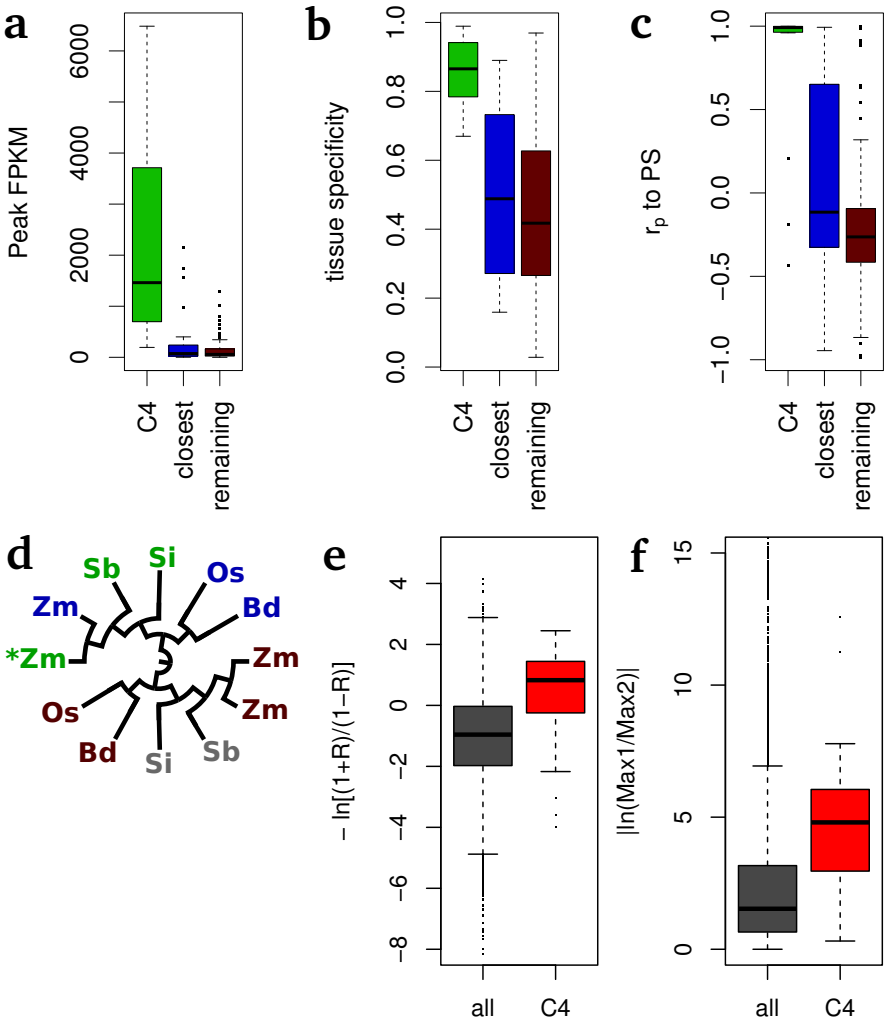


Fig. 2: Expression characteristics and divergence of the core C_4 genes. The absolute expression level (excluding regulatory genes)(a), the BS or M tissue specificity (for C_4 genes that are tissue specific in *Z. mays* only)(b), and the similarity (r_p) in expression to the PS (photosynthesis) MapMan category (c) between C_4 genes (green), their phylogenetically closest homologs in each species (blue), and the remaining homologs (red). Example classification of homologs on a perfect, no-loss gene tree (d). Where there was only one non-C_4 homolog in any species (grey), homologs could not be classified as closest nor remaining and were excluded. Quantification of the divergence in expression pattern (e) and level (f) between the C_4 genes and their paralogs vs between all other paralogs.

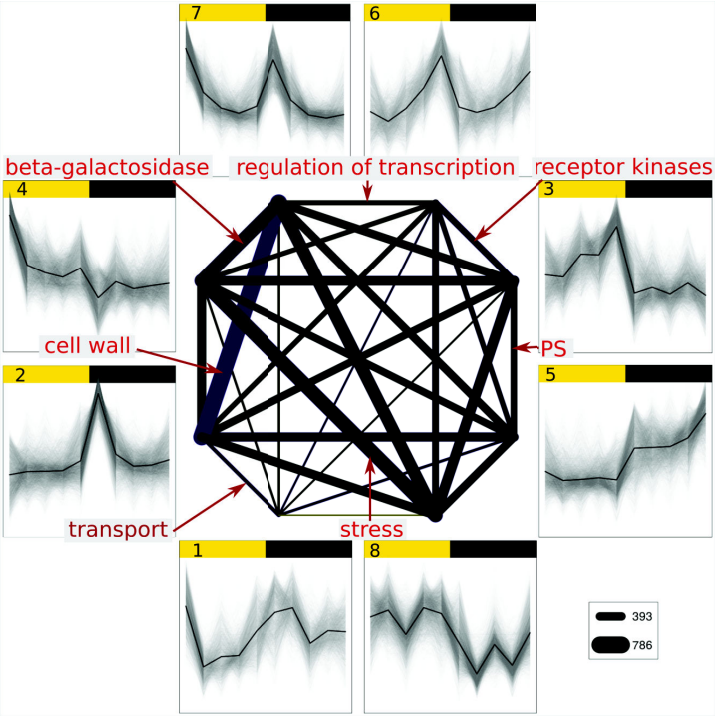


Fig. 3: Paralogs in the edges between expression clusters and their functional enrichments. Example's of significantly enriched functions (fdr <0.05) are shown with an arrow to edge between clusters in which they are enriched (all significant enrichments included in Supplementary Dataset 4). Clusters plotted as z-scores with M base to tip followed by BS base to tip from left to right. The width of the lines connecting clusters is relative to the number of pairs in the edge connecting the respective clusters, while the color indicates whether the edges are larger (blue) or smaller (yellow) than expected based on the total number of pairs in non-loop edges of the connected clusters.

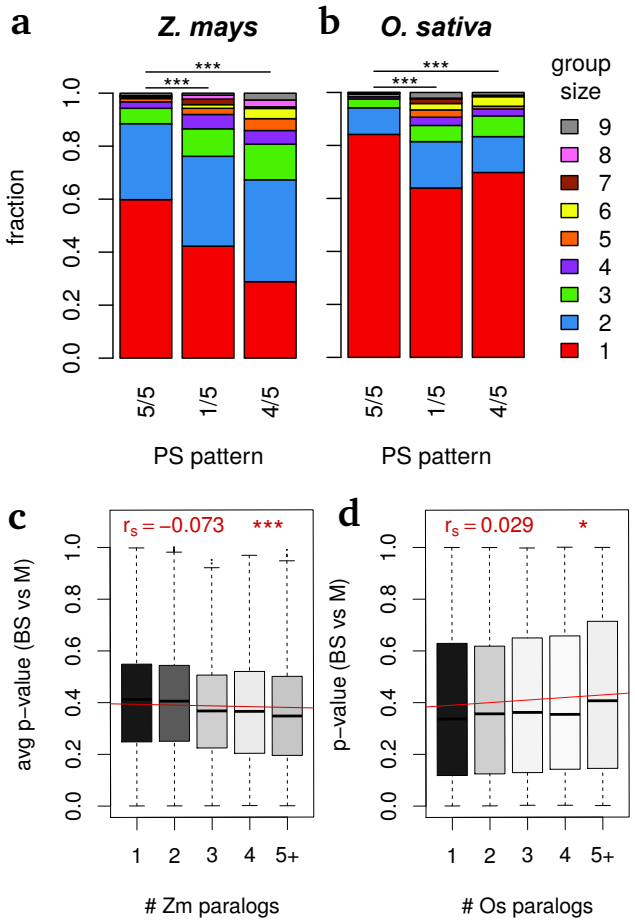


Fig. 4: The relationship between paralog number and expression characteristics related to C_4 evolution. The relation between photosynthetic pattern evolution and group size (# paralogs in orthogroup in respective species)(a-b). Cases where all 5 species show a photosynthetic-like expression pattern (see Supplementary Fig. 18) are considered conserved, while cases where 4 of 5 or 1 of 5 species show a photosynthetic-like expression pattern are considered gain or loss, respectively. The odd species out is *Z. mays* in (a) or *O. sativa* in (b). The significance of tissue specificity (average p-value) vs the group size in (c) *Z. mays* and (d) *O. sativa*.

Methods

Statistical notes. Unless otherwise noted, all statistical analysis was performed in the R statistical environment. Whenever a test was performed more than 20 times, the false discovery rate [68] was calculated from the resulting p-value.

Obtaining and processing plant Genome Data. Genome and gene-model data was downloaded for 12 grasses with available genomes and for banana as an outgroup from Phytozome 10.0 (*Z. mays*, *S. bicolor*, *S. italica*, *O. sativa*, *B. distachyon*, *Panicum halli*, *Panicum virgatum*; [69]) or Gramene V40 (*Oryza brachyantha*, *Oryza glaberrima*, *Triticum aestivum*, *Triticum urartu*, *Hordeum vulgare*, *Musa acuminata*; [70]). In cases with multiple gene models, the longest protein sequence was used for further analysis.

Defining homology. Three methods were used to define homologous genes as appropriate for the context. First, BLAST [71] was used to define pairs of homolgous genes by reciprocal best hits as well as the 'best' ortholog for a *Z. mays* gene by one-directional best blast hit. Second, OrthoMCL [72] was used to more inclusively define whole orthogroups/gene families. Third, we used paralogs which were previously found to have originated from the pan-grass WGD, the *Z. mays* specific tetraploidy, or from tandem duplications [27].

Mapping between species and genome annotations. Combining data for this study required confident mapping of gene identity between different genome releases. As not all genes with the same identifier show any homology, we used a combination of BLAST and provided mappings (i.e. matching IDs, ftp://ftp.gramene.org/pub/gramene/maizesequence.org/release-5a/working-set/4a_discontinued_ids.txt) to obtain confident mappings. Mappings were given a score of 0 for a provided mapping and a reciprocal best BLAST hit, 2 for only a reciprocal best blast hit, 3 for a provided mapping and best BLAST hit from *Z. mays* 6a to the other genome, and 5 for only a best BLAST hit from 6a to the other genome. Ties were broken randomly. The same scoring was used for interspecies mappings, but without provided mappings. Finally, before using the annotated duplicate origins [27] we filtered pairs that didn't pass a final quality check to see if the mapped WGD derived duplicates showed collinearity using McscanX [73] and if the mapped tandem duplicates occurred within 40 genes of each other.

Phylogenetic analysis. Multiple sequence alignment for orthologous groups was performed with prank [74], and in the case of pairwise *Z. mays* sequences with MAFFT [75]. The ungapped alignment area of the resulting multiple sequence alignment was maximized by filtering poorly aligned and gap-causing sequences with seqSieve (https://pypi.python.org/pypi/seqSieve/0.9.1). Resulting protein alignments were translated to codons with pal2nal [76]. Phylogentic trees were constructed with RaxML [77]. Plots were produced using the ete2 python package [78]. For display only, we manually corrected the PPDK tree so that the paralogs originating from the *Z. mays* tetraploidy were sister to each other. Pairwise estimates for the synonymous and non-synonymous substitution rate (dS and dN) were calculated using codeml from the PAML package [79]. In a test set (described at end of methods) the signature of positive selection (dN/dS >1) was tested using

the branch site model, and significance calculated with a likelihood ratio test [80]. This test was performed at all *Z. mays* genes, their parental branches, and the parental branches there of.

Plant Growth conditions and harvest. *Z. mays* B73 were grown in the summer of 2012 in the same green house and conditions as previously described [36]. The 3rd leaf was harvested when it measured 18 cm from the 2nd ligule to the leaf tip. Two different harvesting methods were performed. In the first, a leaf gradient consisting of 5 sequential developmental slices (4 cm each) were harvested simultaneously using the "leaf guillotine" [36]. This method required 10s to extract the 3rd leaf and properly align it, which does not allow for reliable estimates of the metabolite distributions for high-turnover photosynthetic metabolites. Therefore, a second harvesting method was performed, in which the plants were positioned above two liquid nitrogen containers and two 8 cm slices were cut with connected scissors (Supplementary Fig. 1). With this method there was a delay of less than 1s between slicing and quenching. The full, five slice gradient was used for RNA sequencing, and the faster two slice gradient was used for metabolite extraction.

Tissue enrichment. Mesohpyll and bundle sheath tissues were mechanically enriched by serial filtration on liquid nitrogen using a method modified from [37]. Ground material was filtered through 250, 80, and $41\mu M$ meshes on liquid nitrogen. Three fractions were selected for further analysis because they showed the most enrichment of bundle sheath tissue (did not pass through $80\mu M$ mesh), most enrichment in mesophyll tissue (passed through $41\mu M$ mesh) or intermediate, but consistent proportions of tissues (did not pass through $41\mu M$ mesh).

Extraction and abundance measurements metabolites/enzymes. Enzymes were extracted and desalted as described in [22] from the three enrichment fractions, and the enzyme activity was measured through chlorometric assays as described in [81, 82]. Metabolites were extracted and quantified via gas chromatography/electron-impact time-of-flight mass spectrometry as described in [83]. To consistently exclude data where the peak was hard to distinguish from the background, low-signal metabolites were excluded. Further individual replicates with a raw % abundance in BS of more than 3 standard deviations from the mean were excluded. The integrated peaks were divided by the area of the ribitol (internal standard) peak and the fresh weight, and to further reduce noise and compensate for FW/DW differences between the cell types by the mean abundance for the replicate. Therefore, normalized differences between metabolites represent not absolute distribution, but distribution relative to the other metabolites, particularly sucrose and the other highly-abundant metabolites.

Sequencing and estimating transcriptional abundances. RNA was extracted with QIAGEN RNeasy Plant kits, according to the manufactures instructions except for the addition of an extra wash step in 80% EtOH. Libraries were prepped from RNA with a RNA integrity number >8 and sequenced with the Illumina HiSeq 2000 platform. All additional reads were downloaded from the Sequence Read Archives [84]. Illumina adaptors were trimmed using cutadapt [85] and trimmed for quality using FASTX (Hannon Lab). Trimmed reads were mapped to the 6a release of the *Z. mays* B73 genome (or the respective species' genome, as available from Phytozome 10.0, so *S. viridis* reads were mapped to *S. italica* genome) with Tophat2 [86] and transcripts abundance

calculated with Cufflinks [87]. However, one study [39] used for minor comparisons was mapped only to the 5a genome. For the one microarray study included [35] data was downloaded from Gene Expression Ombibus [88], and the expression and significance calculated with GEO2R, which uses the Limma R package [89]. Non-default parameters used for all bioinformatics programs are provided (Supplementary Table 7).

Estimation of initial tissue specificity by "deconvolution". The abundance of metabolites, enzymes and transcripts was compared to abundance of BS and M markers to estimate the original tissue specificity by a method modified from [37]. First, to allow for comparison of data with different absolute expression levels, all data was converted into fraction of total transcript in developmental slice. Second, marker transcript (or marker enzyme) levels were used as proxies for the amount of M and BS tissue in each enrichment fraction. The natural log of the BS marker/M marker was plotted against the natural log of a target unknown/M marker across all samples, and the slope of a regression line between these two log ratios estimated the fraction of target gene transcripts that are localized to the BS Supplementary Fig. 2. To determine if target unknowns were more related to either of the tissue markers, we tested whether the slope of this line was significantly different from 0.5 (corresponding to a null enrichment of 50% M, 50% BS). This was automated with a linear regression in R and calculated for every non-marker enzyme, metabolite, and every gene that had a minimum FPKM >0. Tissue specificity was estimated independently in each developmental slice. We assumed the average abundance between the raw values of all enrichment fractions was equal to the average abundance between M and BS. Therefore, to estimate the "pure" abundance values the estimated fraction in BS and M (1 - fraction BS) were multiplied by 2 x the average FPKM value for the developmental slice. For enzyme and metabolite data, the enzyme activity of PEPC and NADPME were used as markers for M and BS respectively. For RNA sequencing data, Lipoxygenase 2 (GRMZM2G015419) and the sum of Ribulose-phosphate 3-epimerase (GRMZM2G026807) and Phenylalanine ammonia-lyase 1 (GRMZM2G074604) were used as M and BS markers, because these markers showed similar enrichment to-, but more steady enrichment than- PEPC and NADPME throughout development.

K-means clustering. K-means clustering was performed to get an overview of the data and allow qualitative categorization of divergence between paralogs. K-means clustering was performed on all genes where the initial tissue specificity could be estimated in every developmental slice (minimum raw FPKM >0). To choose the number of clusters, the sum of standard error (SSE) of clusters with the original data was compared to the SSE of clusters with scrambled data [90]. We proceeded with 8 clusters as this provided a fairly low SSE for the original data and a large difference in SSE between original and scrambled data Supplementary Fig. 25. Clustering was repeated 10,000 times and the solution with the lowest SSE was selected. Each cluster was tested for functional enrichment in all distinct MapMan [41] categories with a Fisher's Exact test.

Defining divergence. We employ two quantitative methods and one qualitative method to estimate divergence. First, we use transformed Pearson correlation between expression patterns as an interval scaled variable for the amount of divergence in expression pattern. The transformation is performed to provide an unbounded and more normally distributed value. The transformation

of Pearson's r (r_p) is equal to $ln(\frac{1+r_p}{1-r_p})$. Second, to measure divergence in expression level, we recorded the absolute value of the natural log for the ratio between peak expression of the paralogs ($|ln(\frac{peakFPKM1}{peakFPKM2})|$). Finally, to evaluate divergence in a qualitative fashion, we developed a clustering based method to track particular patterns of divergence. Using graph theory, we considered the k-means clusters nodes, and pairs of paralogs formed edges either between them or, when both paralogs occurred in the same cluster, loops. To avoid assigning pairs of paralogs to a divergent (non-loop) edge, if they had a conserved expression pattern that was intermediate between the clusters, we excluded "boundary" pairs from further analyses. Boundary pairs were defined as pairs in a non-loop edge where the r_p between the expression pattern of the two paralogs was higher than the the r_p of either pair to its cluster center.

Regression analysis. Multiple linear regression was used to compare the expression divergence (for both pattern and level) of two paralogs to their other characteristics (dN, dS, dN/dS, number of *Z. mays* paralogs in orthogroup, whether either paralog was a C_4 gene or not). The calculated p-value represents the chance of seeing the observed improvement in model fit of adding the factor in question to a model already containing all the other factors if the null hypothesis (no relation) is true. When comparing values that did not approach a normal distribution (e.g. p-values, FPKM, % abundance in BS or M) we performed Spearman rank correlation.

Controlling pairwise counting bias. Some analyses could be sensitive to a bias resulting from counting the pairwise combinations of different sized orthogroups. For instance, there are three pairwise combinations of the group "a", "b", "c" ("a-b", "a-c", and "b-c") and every group member is counted twice; however, add 'd' to the group and there are six pairwise combinations ("a-b", "a-c", "a-d", "b-c", "b-d", "c-d") and every group member is counted three times. To control for this, without introducing other bias by sub setting the data (e.g. taking reciprocal best blast hits selects for young paralogs from small gene families), we scrambled the data to get an empirical p-value accounting for this bias. Specifically we scrambled expression information, but held gene family information constant and counted the number of instances where the result was as, or more, extreme than the original to obtain an empirical p-value.

We expected this bias to be most problematic for two analyses: the correlation between number of *Z. mays* genes in orthogroup and the expression divergence and the functional enrichment in edges between clusters. To test the significance of the correlation between the number of *Z. mays* genes in the orthogroup and expression divergence we scrambled the expression patterns of genes and re-calculated rp between the afore mentioned values 3200 times. To test for an enrichment (or depletion) in edges between clusters and MapMan functional categories, we scrambled the cluster assignment of 'divergent' and 'conserved' pairs 63999 times, and counted the number of cases where each functional category was more or less enriched than the original in each cluster pair using the Python Language [91].

Calculating divergence on phylogenetic tree We used a mean of pairs method to calculate the divergence for nodes of a phylogenetic trees. Pairs consisted of any genes originating from the same species and occurring on different daughter branches of the node. The mean divergence

across all pairs was taken as the divergence at the node. The test set where this was calculated consisted of 64 orthogroups of 60 genes or less with at least one divergent pair of *Z. mays* genes and one conserved pair of *Z. mays* genes. The orthogroups were sorted by the expression of lowest paralog contributing to the conserved or divergent pair, and the 64 most highly expressed were chosen.

Methods References

68. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 289–300 (1995).
69. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40,** D1178–D1186 (2012).
70. Monaco, M. K. *et al.* Gramene 2013: comparative plant genomics resources. *Nucleic acids research* **42,** D1193–D1199 (2014).
71. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25,** 3389–3402 (1997).
72. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13,** 2178–2189 (2003).
73. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research* **40,** e49–e49 (2012).
74. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 10557–10562 (2005).
75. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics* **9,** 286–298 (2008).
76. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research* **34,** W609–W612 (2006).
77. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
78. Huerta-Cepas, J., Dopazo, J. & Gabaldón, T. ETE: a python Environment for Tree Exploration. *BMC bioinformatics* **11,** 24 (2010).
79. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24,** 1586–91 (Aug. 2007).
80. Yang, Z. & Dos Reis, M. Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution* **28,** 1217–1228 (2011).

81. Walker, R. P., Trevanion, S. J. & Leegood, R. C. Phosphoenolpyruvate carboxykinase from higher plants: purification from cucumber and evidence of rapid proteolytic cleavage in extracts from a range of plant tissues. *Planta* **196,** 58–63 (1995).

82. Hatch, M. & Mau, S. Properties of phosphoenolpyruvate carboxykinase operative in C_4 pathway photosynthesis. *Functional Plant Biology* **4,** 207–216 (1977).

83. Rudolf, M. *et al.* In vivo function of Tic22, a protein import component of the intermembrane space of chloroplasts. *Molecular plant,* sss114 (2012).

84. Kodama, Y., Shumway, M. & Leinonen, R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic acids research* **40,** D54–D56 (2012).

85. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17,** pp–10 (2011).

86. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14,** R36 (2013).

87. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7,** 562–578 (2012).

88. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research* **41,** D991–D995 (2013).

89. Smyth, G. K. in *Bioinformatics and computational biology solutions using R and Bioconductor* 397–420 (Springer, 2005).

90. Peeples, M. *R Script for K-Means Cluster Analysis* 2011. <http://www.mattpeeples.net/kmeans.html>.

91. Van Rossum, G. & Drake Jr, F. L. *Python reference manual* (1995).

Supplementary Information

Supplemental Note

C_4 cycle. The key feature of C_4 photosynthesis is the biochemical pump which concentrates CO_2 at the site of Rubisco and suppress the costly process of photorespiration. This can result in a 50% increase in photosynthetic efficiency [92]. To achieve this, C_4 plants use the non-oxygen sensitive enzyme, Phospho*enol*pyruvate Carboxylase (PEPC), to fix CO_2 onto Phospho*enol*pyruvate (PEP) in the M. The resulting 4-carbon acid must diffuse to the BS, and be decarboxylated, releasing CO_2. In *Z. mays*, the primary decarboxylating enzyme is NADP Malic Enzyme (NADPME); however, around 15% of the carbon appears to flow through the secondary decarboxylating enzyme PEP Carboxykinase (PEPCK) [36, 93, 94]. The resulting 3-carbon acid diffuses back to the M and is regenerated to PEP, as necessary, completing the cycle. In addition to the carbon shuttle, a small part of the Calvin Benson Bassham cycle is localized to the M and the rest to the BS, resulting in a triphosphate based redox shuttle transporting reducing equivalents to the BS. Both the C_4 cycle and redox shuttle require upregulation of metabolite transporters to support the high flux of metabolites in and out of subcellular compartments. However, only two transporters have been fully characterized in *Z. mays*. Here-after, when we refer to the core-C_4 cycle, we are referring to the enzymes of the primary and secondary C_4 cycle, the known transporters Phosphoenolpyruvate/Phosphate Translocator (PPT) and Triose Phosphate Transporter (TPT), and the two established regulatory proteins PEPC Kinase (PPCK) and Pyruvate Phosphate Dikinase - Regulatory Protein (PPDK-RP).

The elements of the key C_4 cycle are well distributed in our data. Transcripts, and where available enzyme activity, for the enzymes responsible for regenerating PEP (Pyruvate Phosphate Dikinase, PPDK), converting (Carbonic Anhydrase; CA) and fixing the CO_2 (PEPC), and converting the resulting oxaloacetate (OAA) to the transfer acid Mal (NADP Malate Dehydrogenase; NADPMDH) are higher in the M as expected ($p < 0.05$, enzymes; fdr < 0.05, transcripts in Slice 3 - 1; except NADPMDH in Slice 2 where fdr = 0.058; Supplementary Fig. 10, 11. The decarboxylation enzymes are both higher in the BS (fdr < 0.05, transcripts in Slice 4 - 1; Supplementary Fig. 10, 11), and the enzyme which can convert OAA that was transported as aspartate to malate (NAD Malate Dehydrogenase; NADMDH) showed a preference for the BS ($p < 0.05$, enzyme in Slice 1; fdr < 0.05, transcripts in Slice 3 - 2);Supplementary Fig. 10). Several activities in the cycle are expected to be balanced between tissue types, including the TPT transporter, and the Aspartate- and Alanine-Amino Transferase (AspAT and AlaAT). TPT expression is quite even between tissues Supplementary Fig. 12, while AlaAT is enriched in the M at the level of transcripts but not enzyme activity Supplementary Fig. 10. In contrast, for AspAT both enzyme activity and transcripts are strongly enriched in the M. However, we find a M specific paralog with high expression level, and a BS specific paralog with a low expression level, which is very consistent with the previous studies [39, 40], and even with *S. italica* [34].

Metabolites. The metabolic data is hard to interpret as separation was not sufficient to produce significant results after multiple hypothesis correction. However, as there is very little data available for the separation of metabolites between BS and M cells, we want to describe the data anyways to provide information that may help in the design or analysis of future studies.

This study shows the care that will be required to confidently measure the values of photosynthetically active metabolites. The major advantage of the employed technique, is the immediate shock freezing, and frozen processing of tissue, which allows very little time for changes in leaf metabolome. Unfortunately, the employed thecnique allows for only modest enrichment of tissues, and in contrast to enzymes and transcripts there are no known internal metabolite controls that are close-to-perfectly tissue specific, and as small molecular weight metabolites can readily diffuse across the plasmodesmata, there are unlikely to be any fully tissue specific and cytoplasmic metabolites. Therefore, enzyme activity was used for normalization.

Although nothing was significant, we will try to briefly summarize the trends in the data. Metabolites in the core-C_4 cycle all behaved similarly in our data, with a tendency towards BS enrichment in slice 3_4 and a tendency towards M enrichment in slice 1_2 (Supplementary Fig. 26). The mature tendency towards M matches expectation for aspartate and malate, which need to diffuse from the M to the BS. Two previous studies [37, 95] also estimated that concentrations of malate were higher in the M than the BS (Supplementary Fig. 27). Glutamate and α-ketoglutarate are not expected to show a net flux between tissues, and the tendency towards M in mature tissue is therefore unexpected; however, the estimated % M is surprisingly consistent with that reported by [95]. In contrast α-alanine showed a tendency opposite to that of the expected concentration gradient, and incosistent with the previously reported even distribution [95] (Supplementary Fig. 27). Notably, there were also major differences between the fast 2-slice harvest and slower 5-slice harvest (e.g. α-ketoglutarate; Supplementary Fig. 27). The lack of statistically significant enrichments, differences between the developmental stages, differences in slow harvest vs fast harvest, and inconsistency with previous data (Supplementary Fig. 27), point to, if nothing else, the lability of metabolites. The same lability that makes metabolites hard to measure between experiments and sensitive for instance to shading or cooling, means the plant must be able to tolerate a non-continuous distribution of metabolites between tissues.

All the measured photorespiratory metabolites had a tendency towards BS enrichment, as is expected with the BS specific localization of the photorespiratory cycle (Supplementary Fig. 28). Other categories of sugars (Supplementary Fig. 29), amino- (Supplementary Fig. 30) and other organic acids (Supplementary Fig. 31) showed a variety of distributions, with frequent change both in level and tendency towards tissue specificity between the two slices. Indeed the metabolites appeared to show more frequent changes in tissue preference than the enzymes or transcripts. While this may simply reflect the generally high error and low-significance, it may also, in part, reflect how dynamic the metabolome is.

Transcription factors of interest There is strong interest in engineering the C_4 trait into C_3 crop species to increase photosynthetic efficiency and ultimately growth and yield. However, the complexity of the C_4 trait goes well beyond the capabilities of even the most successful current engineering methods. However, the highly convergent nature of C_4-evolution provides hope that extreme changes may be facilitated by comparatively simple changes in regulatory architecture. Therefore, we used the compiled expression data to highlight some top-candiate transcription factors of interest to understanding C_4 photosynthesis and its evolution.

Individual studies targeting transcription factors of interest to the C_4 trait in *Z. mays*, have provided candidate lists from X-Y members [29, 36, 38–40]. While this remains a very ambitious number for individual characterization, taking the intersection of various studies is an extremely strict measure, that results in 0 remaining candidates [38]. Therefore, we take a more permissive and inclusive approach to find transcription factors that are of interest in understanding C_4 photosynthesis supported by four or more of the following six criteria relating to C_4 photosynthesis, its evolution, and kranz anatomy. The criteria were: 1) significantly associated with either the M or BS marker in all 5 slices in this study; 2) consistent direction of enrichment across all samples and studies (all BS >M or all M >BS; 3) the FPKM in *Z. mays* leaf ("V5_Tip_s-2_Leaf", [30] was at least twice that of both *B. distachyon* and *O. sativa* leaves [31]; 4) The peak expression in floral primordia was at least 1.5 times that of husk primodia [29]; 5) expressed at least 20 FPKM in floral primordia; And 6) show a correlation to the PS expression pattern (r_p) higher that 0.4 in *Z. mays*, but not in *B. distachyon* or *O. sativa*. In total, 19 transcription factors met these criteria (Supplementary Dataset 5).

Among the identified transcription factors are ones with particularly interesting orthologs in *A. thaliana*. Three DOF transcription factors were selected (GRMZM2G114998, AC233935.1_FG005, and GRMZM2G179069), all of which had higher FPKM in maize and the other C_4 species than either C_3 species, a photosynthetic-like expression pattern in *Z. mays* but not in either C_3 species, and were more highly expressed in floral than husk primordia. Further, in concordance with the enrichment of the whole DOF family among BS specific genes, all three selected DOF genes were higher in the BS of every comparison, and significantly higher in the BS in every slice of our leaf gradient. The *A. thaliana* ortholog of GRMZM2G114998, AT4G24060 or DOF4.6, is expressed at the sites of early vein formation [96], making DOF4.6 an interesting candidate in understanding the narrower vein spacing in C_4 species. The other two DOF family transcription factoers, GRMZM2G179069 and AC233935.1_FG005, share their closest *A. thaliana* homolog, AT3G55370 or OBP3, which is a mediator of phytochrome signaling [97]. Phytochrome signaling is a major regulator of photomorphogenesis or how a plant develops in response to light [98]. Another mediator of phytochome signaling, the COP9 signalosome, has been putatively linked to the differences in leaf development seen between C_3 and C_4 sister species [18].

Two auxin response regulators were identified, both of which were higher in *Z. mays* than either C_4 species, and higher in floral than husk primordia. Further ARF3 was expressed highly in the floral primordia, and consistently higher in M than BS; while AXR2 had a photosynthetic pat-

tern in *Z. mays* and *S. bicolor* that was not shared with the C_3 species, and was consistently higher in BS than M. In *A. thaliana*, ARF3 (AT2G33860) helps mediate the specification of abaxial and adaxial fate [99, 100]. In a study in grasses, the C_4 leaves showed more asymmetry, and modified M/BS ratios between abaxial and adaxial regions, while the C_3 leaves did not [101]. AXR2 (AT3G23050) is involved in the interplay between ABA and auxin response [102]. Auxin is a major hormone for specifying vascular cell fate [103], and modifications in auxin signaling, through modifications in synthesis, transport, perception and timing, are thought to be related to the specialized vein pattering in C_4 species [47]. Finally, in relation to the enhanced secondary cell walls in BS, MYB52 (GRMZM2G455869) is an exceptionally interesting candidate. MYB52 showed over twice the FPKM in the C_4 species compared to the C_3 species, showed a photosynthetic-like expression pattern specifically in the C_4 species, was expressed more highly in the BS in every comparison, and significantly so across the leaf gradient. A. thaliana over expressing MYB52 (AT1G17950) show hypersensitivity to ABA and increased drought tolerance [104]. MYB52 was further identified in a "post-genomic" screen for secondary cell wall related proteins, and it's mutant showed hyper-lignification [105]. In summary, the transcription factors discussed here and the rest from (Supplementary Dataset 5) are highly interesting candidates, which warrant further investigation to see if their promising expression patterns and annotations might help drive any of the features of BS or M tissue specificity in C_4 species.

Advancements in understanding the differences between BS and M cells. To determine if this separation method was consistent with previous studies at a functional level, we tested sets of genes significantly co-regulated with M and BS markers and our k-means clusters for enrichments in MapMan functional categories. To facilitate the comparison of the various M and BS separation studies, we re-ran enrichment testing for all provided [38, 39] or described [40] gene sets that were considered differentially regulated between BS and M cells. For comparability, each set was compared to a background of the 6a genome release. Enrichments in genes specific to the BS were quite consistent between studies (Supplementary Fig. 5), with a handful of categories shared between all samples and studies. Many of these categories are well understood (e.g. the Calvin Benson Bassham cycle) or have hypothesized benefits (e.g. S- assimilation, the DOF transcription factor family; [39, 40, 57, 106]. However, one previously un-examined category, misc.myrosinases-lectin-jacalin, was consistently enriched in the BS. An *A. thaliana* homolog (AT4G19840) of the *Z. mays* myrosinases-lectin-jacalins is a phloem sap protein with a putative role in defense [107, 108], indicating that this category may relate to conserved BS functions and not C_4 photosynthesis. In addition to functions that were consistent across all tissues, many sub categories of protein synthesis were enriched in BS specific genes specifically in three comparable younger tissues (slice 4, slice 3, and section -1 from [38]).

Enrichments in M specific genes showed greater variability between studies (Supplementary Fig. 4). While no categories were enriched in every sample, there was still a strong bias for particular categories. For instance, 20 categories were enriched in seven or more of the ten samples. These included several subcategories of the photosynthesis light reactions, particularly photosystem II; lipid metabolism and lipid transfer proteins; isoprenoid/carotenoid synthesis; and light signaling.

Interestingly, transport was consistently enriched in both M and BS specific genes, indicating it is a category generally undergoing specialization between tissues.

The above analyses compared genes differentially expressed in each developmental slice individually, and to integrate gradient and tissue specificity patterns we performed a functional enrichment analysis on k-means clusters. As clustering was performed only on genes expressed sufficiently to be "deconvoluted" (min FPKM >0), but compared, as above, to the unfiltered 6a genome; some major categories such as RNA, protein, and signaling were enriched in most to all clusters, and not assigned.unknown was frequently depleted. Therefore, we focus on the smaller categories and more specific enrichments.

Clusters 2 and 5 consisted of genes with high expression in the BS base and tip, respectively, and showed a distinct set of enrichments. In cluster 2 many developmental and structural categories were enriched; including cell and cell organization; cell wall proteins and precursor synthesis; lignin synthesis; and categories likely related to the cell wall such as β- 1,3 glucan hydrolases. Additionally several regulation related categories were enriched, such as hormone metabolism with jasmonate and auxin response, and a few transcription factor families. In contrast, in cluster 5, with expression high in the BS tip, the enrichments were dominated by major energy and metabolism categories. In relation to energy production, cluster 5 was enriched in the photosynthetic categories of Calvin Benson Bassham cycle, and photorespiration, as well as mitochondrial electron transport and the TCA cycle. Related to metabolism, cluster 5 was enriched in major and minor carbohydrate metabolism, sulfur metabolism, nucleotide metabolism, secondary nitrogen metabolism, polyamine metabolism, and the oxidative pentose phosphate pathway. Finally, cluster 5 was enriched in a set of regulatory categories distinct from that of cluster 2, including ethylene metabolism and response, and six transcription factor families, of which, only basic Helix-Loop-Helix was shared with cluster 2. In summary, while genes and categories significantly up-regulated in the BS were fairly constant across the leaf (Supplementary Fig. 5) [38], strong differences could be seen in functions of clusters peaking in the BS base or tip, with the base more specialized in development, cell wall and lignification, while the tip was more specialized in photosynthesis and metabolism. Both BS base and tip were enriched in regulatory genes, but largely distinct subcategories of hormone metabolism/response, and transcription factor families.

Similarly, in the M we observed distinct enrichments in the M base cluster (4) and the M tip cluster (3). In the M base cluster 4, enrichments included lipid metabolism and some development, protein and signaling categories, as well as tetrapyrolle synthesis. While in the M tip cluster 3, there were strong enrichments in photosynthesis including both photosystem I and photosystem II, and a concomitant enrichment in light stress. In addition, cluster 3 was enriched in isoprenoid and flavenoid biosynthesis, and the often down-stream-of-photoreceptors family, CONSTANS.

Clusters expressed highly in both tip tissues (6) or both base tissues (7) showed enrichments distinct from the individual tissue types. Most striking in cluster 6, were not the few enrichments,

such as heat stress, that were specific to this cluster; but the lack of an enrichment in the PS categories that was so characteristic of the tissue specific tip clusters 3 and 5. The even base cluster 7, shared cell wall enrichments with the BS base cluster 2, showed distinct auxin related enrichments (auxin response factor (ARF) and Aux/IAA family instead of the auxin.induced-regulated-responsive-activated in cluster 2), and was the only cluster enriched in brasinosteroid metabolism.

The "mixed" categories 1 and 8 appear to contain biological information despite their weird appearance. The deconvolution method is such that it can induce a small pattern in a fairly evenly expressed gene. Double checking the raw data for these clusters, we see that cluster 1 can be described as expressed evenly high in the base, and otherwise slightly higher in the BS than M, while cluster 8 can be described as expressed highest in tip and base, and shows mild M enrichment in some slices (Supplementary Fig. 32). Cluster 1 shared enrichments with other more basal clusters, like cell wall and protein synthesis, as well as histones. Cluster 8 was enriched in cytoskeleton, lipid degradation, minor CHO metabolism, protein degradation and targeting, various regulation of transcription and signaling pathways, and various stress categories.

Bundle fraction contains not only BS but also tracheary elements. Consistent with expectations for the enrichment method, the transcriptome reflects co-enrichment of the vascular tissue with the BS tissue. Ethylene response is enriched in the BS in every slice, which has been implicated in triggering cambial cell division and xylem growth in populus and Zinna cell cultures [109, 110]. We observed many positive regulators of tracheary elements with peak expression in the basal BS slice (BS5). Both vascular cells and BS cells have highly developed and lignified secondary cell walls, which would be difficult to tease apart from each other in the enrichments in cell wall and lignin biosynthesis in basal BS up-regulated genes. However, LAC17 is necessary for lignification of the protoxylem elements in *A. thaliana* [111], and three of it's homologs in *Z. mays* are expressed (21-309 FPKM), and BS specific (fdr <0.05) in the basal slice. In the vasculature, programmed cell death is induced after secondary cell wall deposition [112]. Among genes associated with programmed cell death we find XYLEM CYSTEIN PROTEASE (XCP) 1 (GRMZM2G066326) and 2 (GRMZM2G367701) highly (664 and 398 FPKM) and specifically (fdr <0.01) expressed in BS5.

Supplemental Datasets

Supplementary Dataset 1: Spreadsheet with transcriptional, annotation, and mapping information for *Z. mays* genes

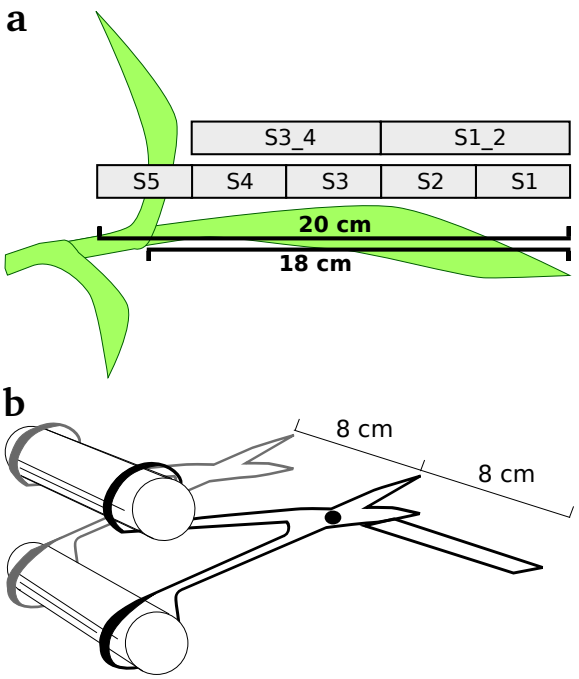
Supplementary Dataset 2: Spreadsheet with metabolic and enzyme activity data

Supplementary Dataset 3: Spreadsheet with significant enrichments for tissue specific genes in each slice and for k-means clusters

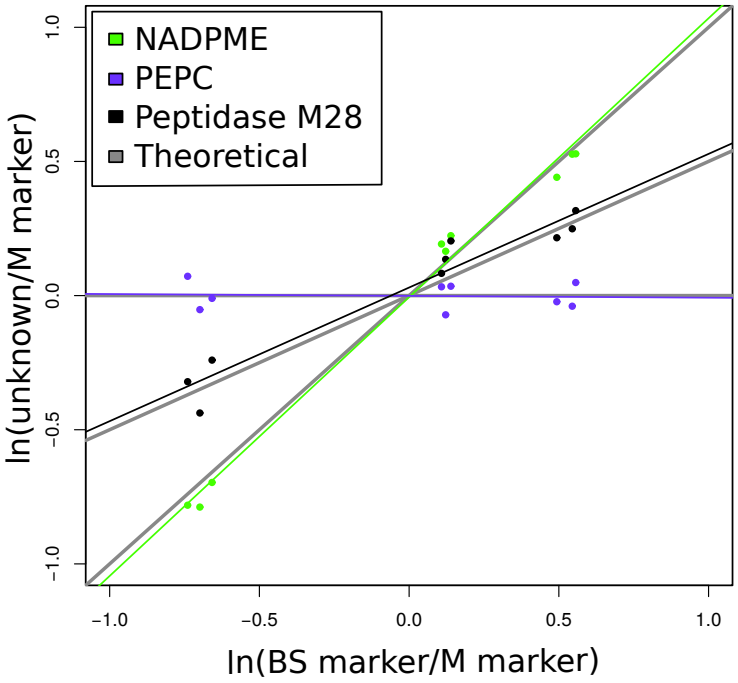
Supplementary Dataset 4: Spreadsheet with significant enrichments for edges between clusters

Supplementary Dataset 5: Spreadsheet with transcription factors meeting the criteria of interest

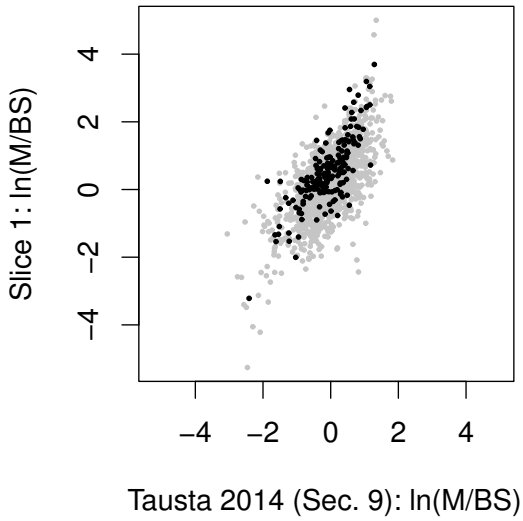
Supplemental figures



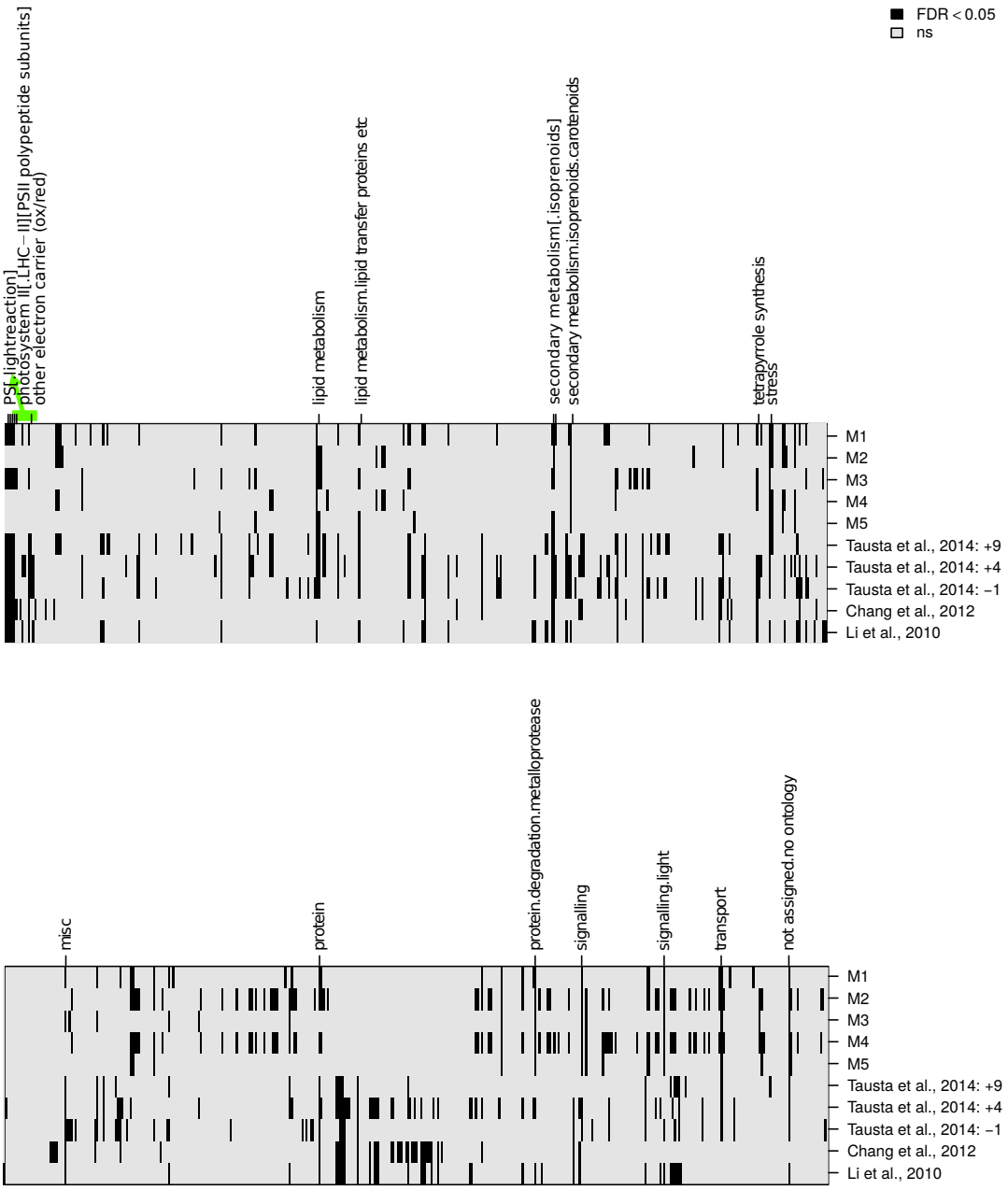
Supplementary Fig. 1: Visual summary of tissues (a) and harvest method (b). The five 4 cm slices (a) were harvested for transcriptome analysis using the leaf "guilotine" [36], while the two 8 cm (a) slices were harvested for metabolite analysis using two pairs of attached scissors (b).



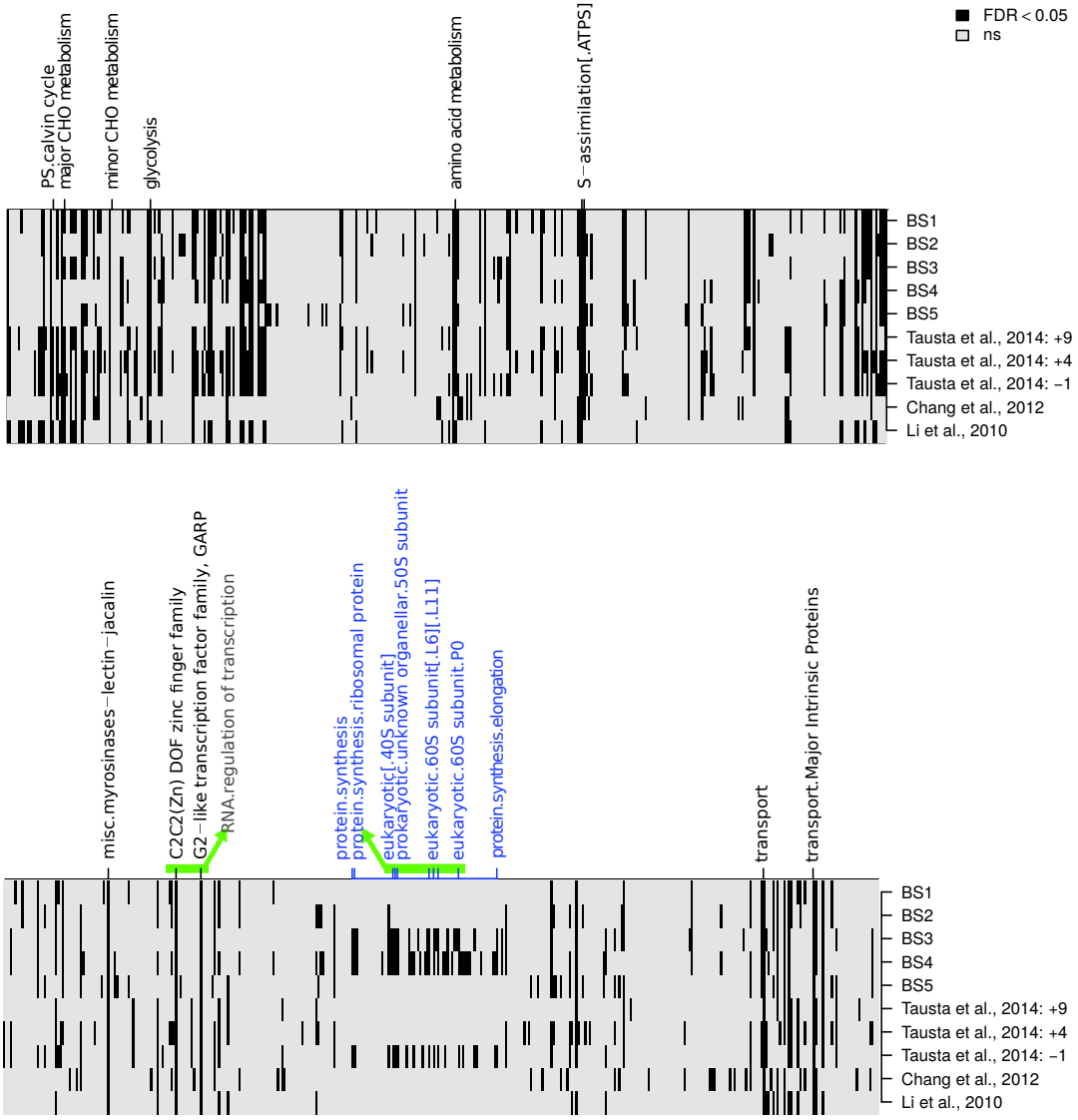
Supplementary Fig. 2: Example comparision between target genes and markers used to "deconvolute" data, that is, estimate the original distribution of target abundance between BS and M cells [37]. PEPC (GRMZM2G083841) as an example of a M specific target, NADPME (GRMZM2G085019) as an example of a BS specific target, and a peptidase M28 (GRMZM2G159171) as an example of a non-enriched target. The slope of the linear regression line yields the estimated fraction of abundance in BS.



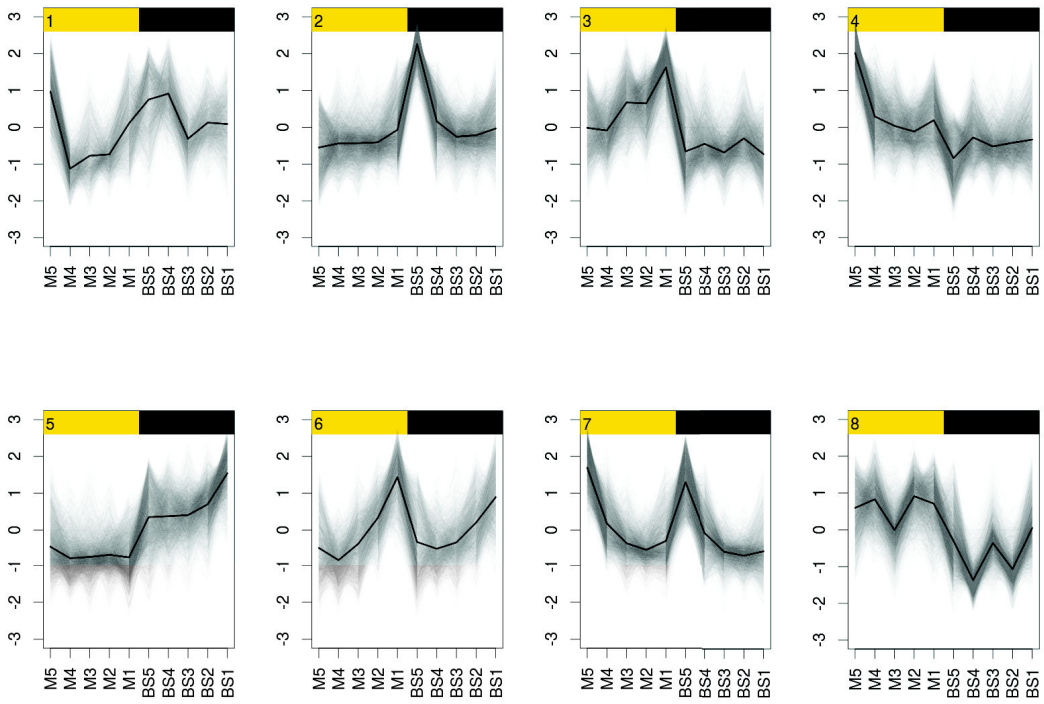
Supplementary Fig. 3: Comparison between transcript BS/M ratios in mature leaf in "deconvoluted" data (Slice 1) and a previous study using laser micro disection (Section +9) [38].



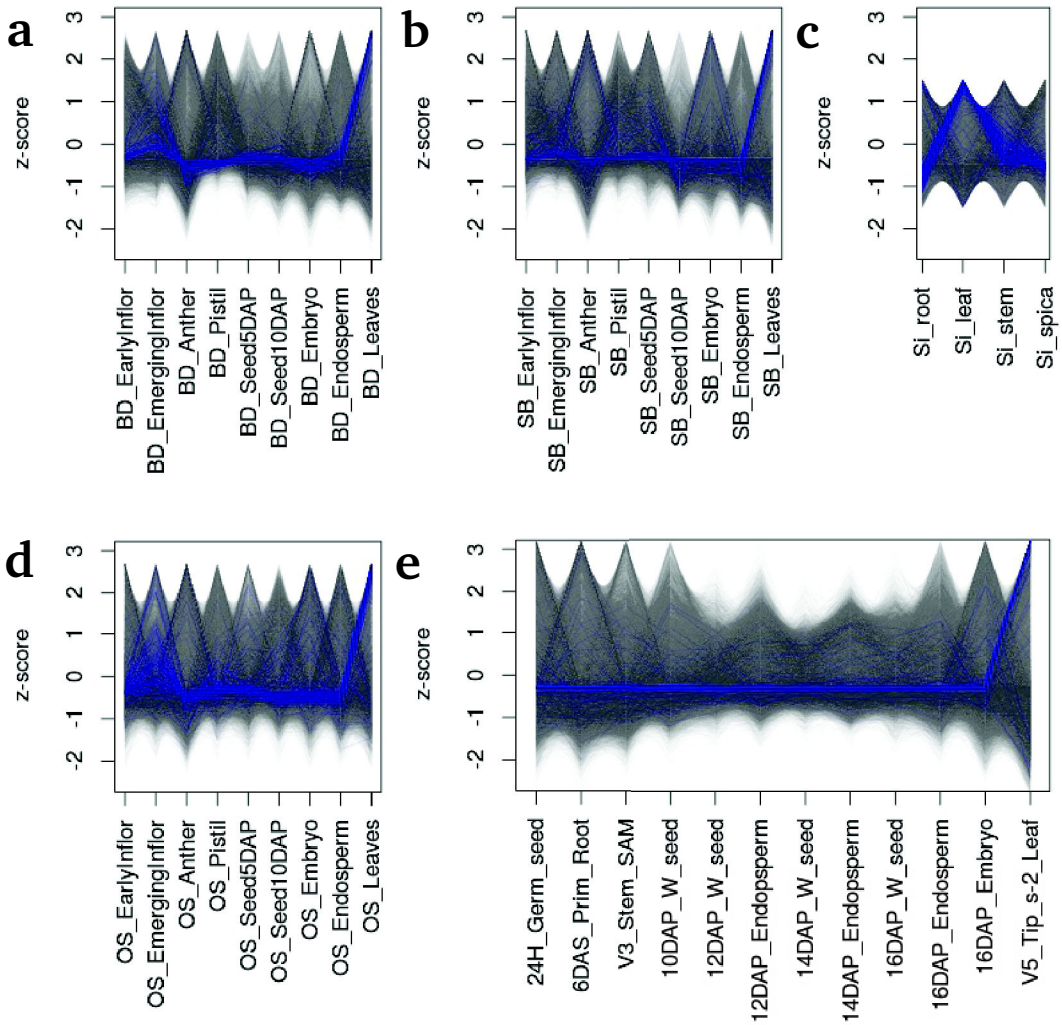
Supplementary Fig. 4: Comparison of tissue enrichments between studies. For comparability all enrichments were calculated with the significantly tissue-specific genes (as defined in each study) in the foreground and the remainder of the unfiltered 6a genome in the background for a Fisher's exact test. The most-consistent enrichments (those enriched in at least 7 samples) are labeled. For the sake of compact display, green bars indicate highlighted categories are subcategories of the more basal category indicated with an arrow, and square bracets indicate enrichment of both basal category and [sub category] 36



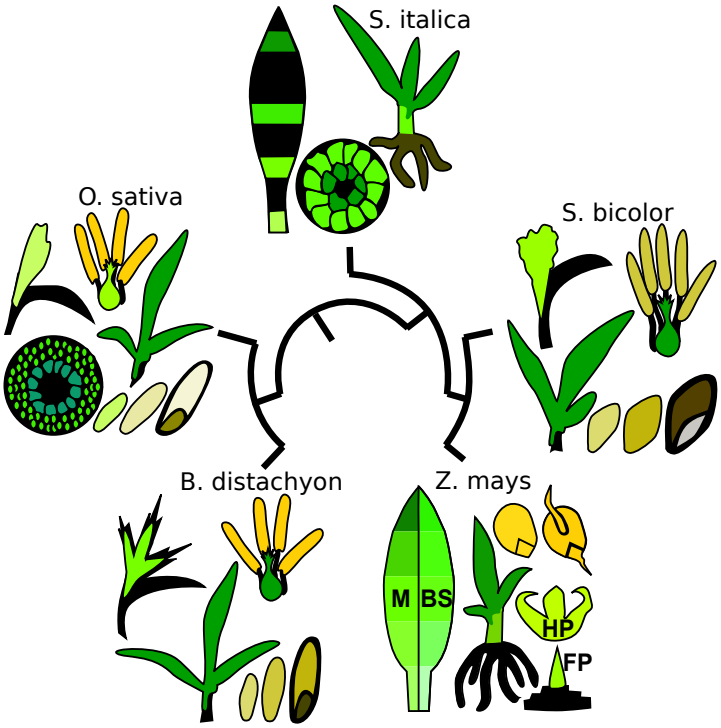
Supplementary Fig. 5: Comparison of tissue enrichments between studies. For comparability all enrichments were calculated with the significantly tissue-specific genes (as defined in each study) in the foreground and the remainder of the unfiltered 6a genome in the background for a Fisher's exact test. The most-consistent enrichments are labeled. Categories enriched in every sample are labeled in black, and those only enriched in the most similar immature tissues (Slice 4 and 3 here, and section -1 from [38]) are labeled in blue. For the sake of compact display, green bars indicate highlighted categories are subcategories of the more basal category indicated with an arrow, and square bracets indicate enrichment of both basal category and [sub category].



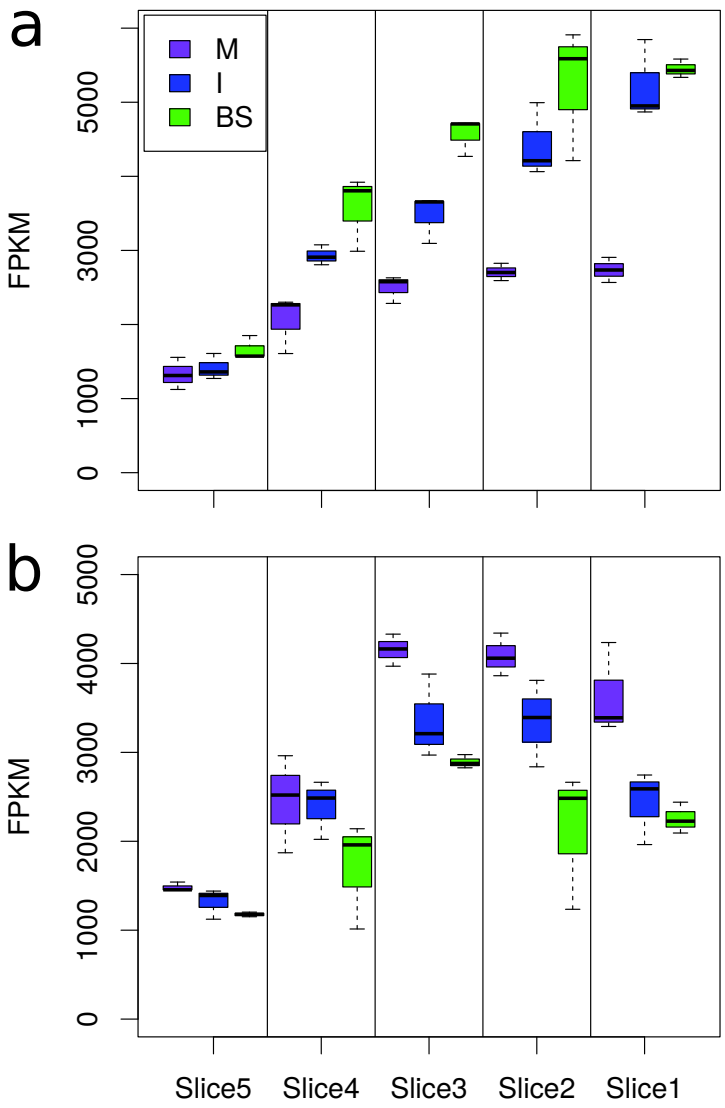
Supplementary Fig. 6: K-means clustering of BS & M gradient, individual genes in grey and centers in black.



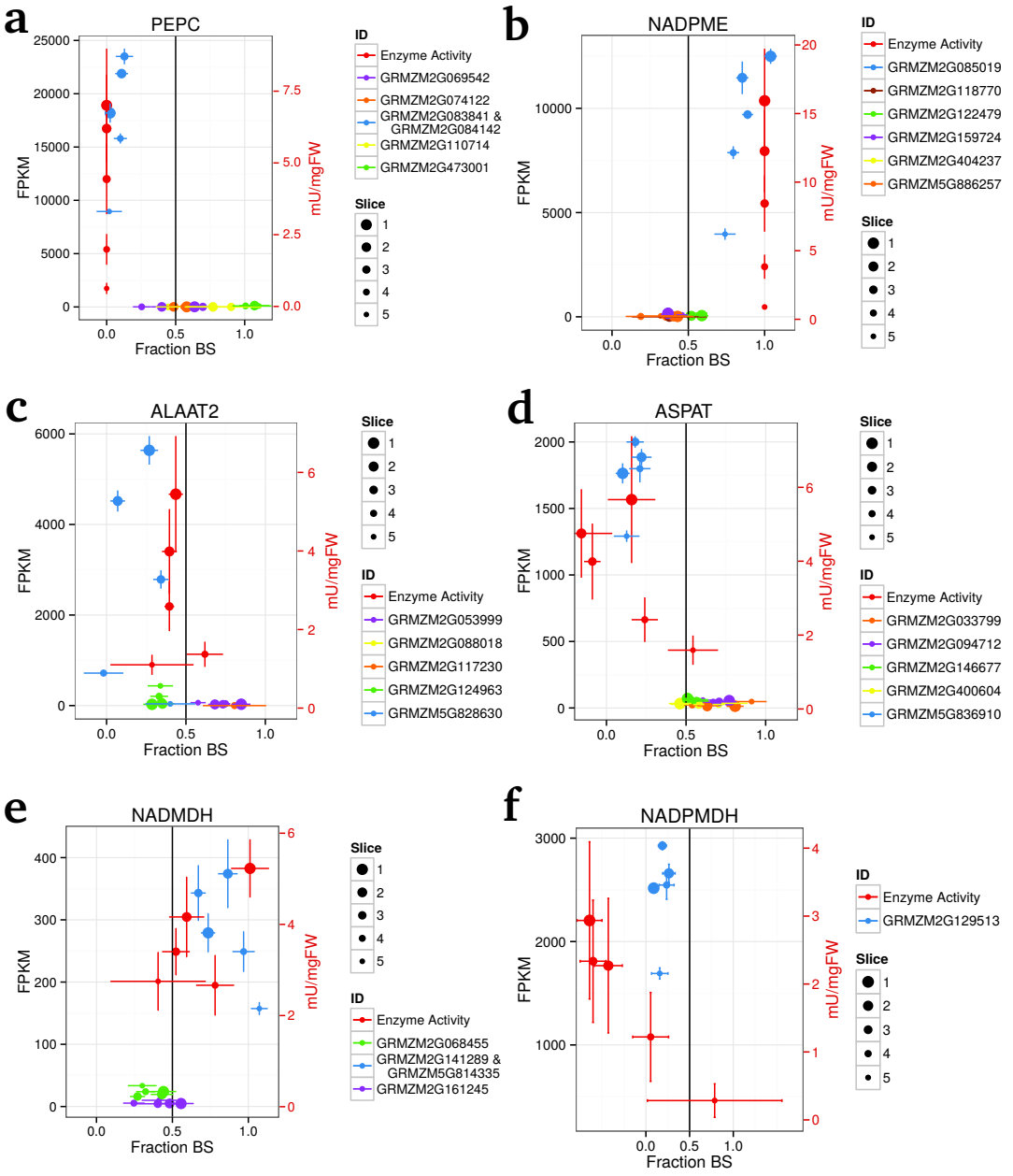
Supplementary Fig. 7: Photosynthetic expression pattern (blue) in tissues used to define it in (a) *B. distachyon*, (b) *S. bicolor*, (c) *S. italica*, (d) *O. sativa*, and (e) *Z. mays*.



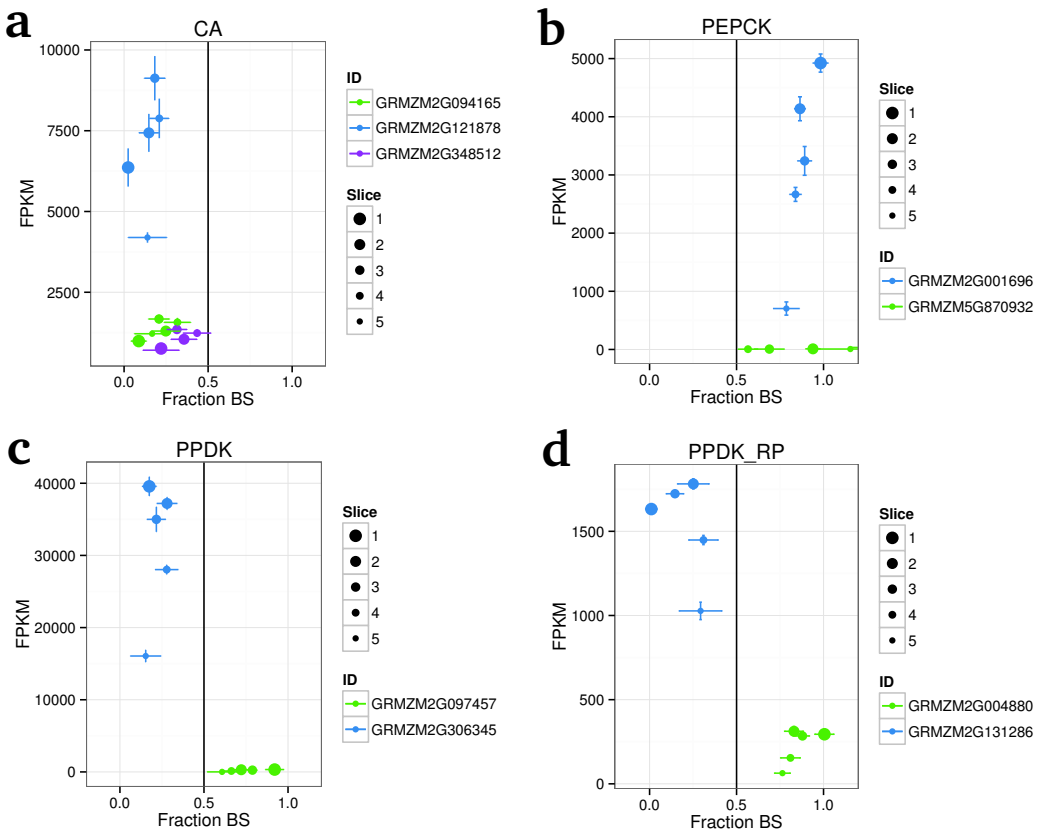
Supplementary Fig. 8: Cartoon summary of the types of tissues coverd by the expression data in each species.



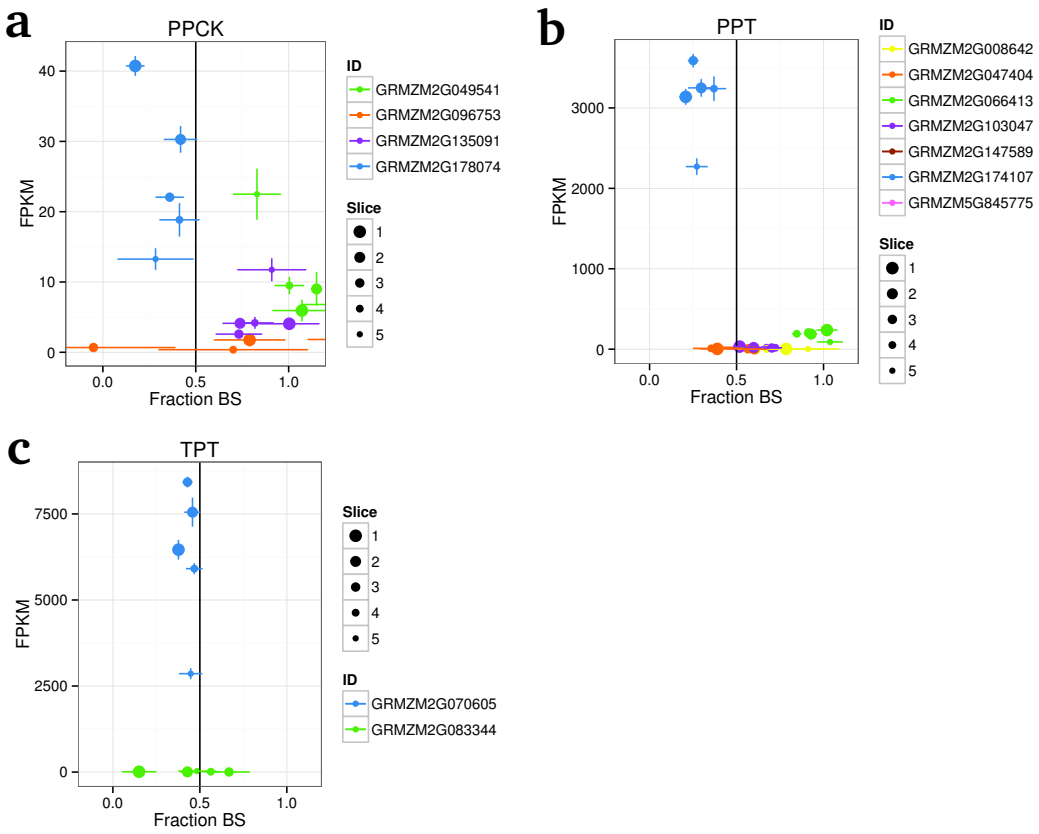
Supplementary Fig. 9: Distribution of transcript abundance of genes with known tissue specificity in raw data. In the BS (a), NADPME (GRMZM2G085019), and in the M (b) PEPC (GRMZM2G083841).



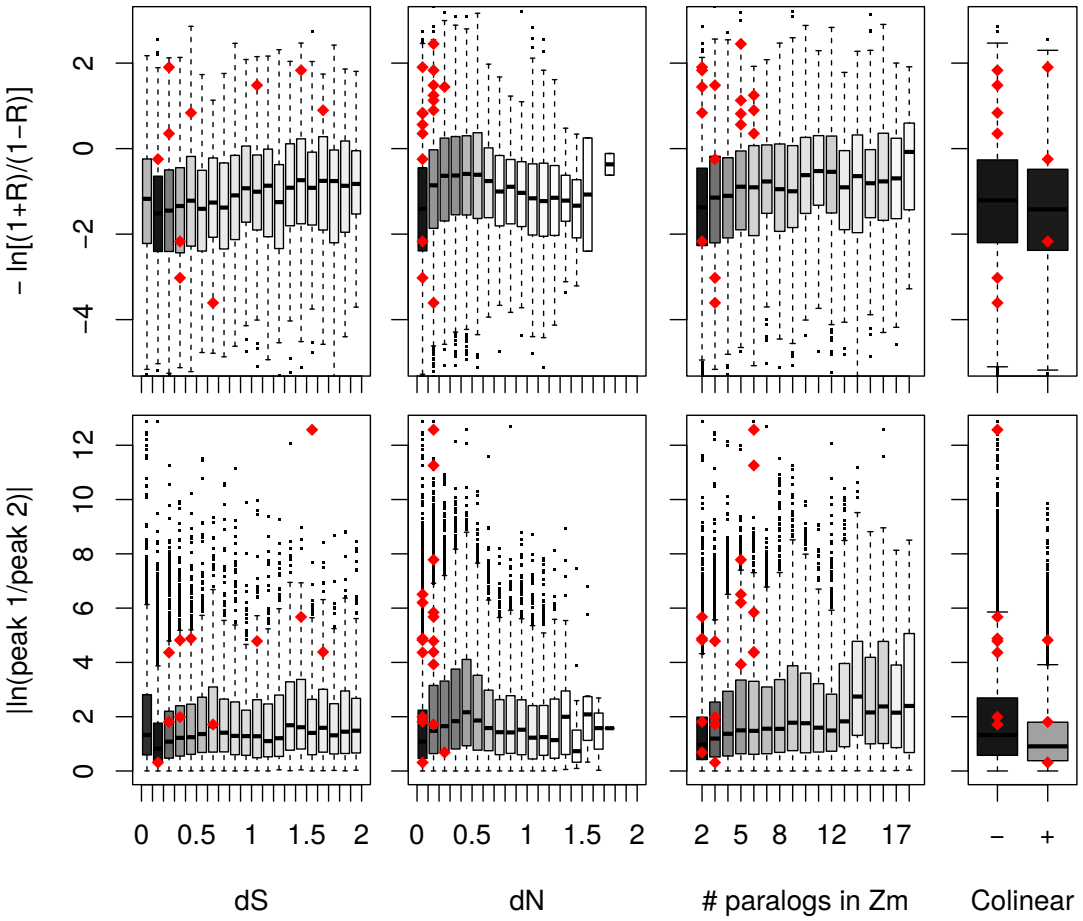
Supplementary Fig. 10: Abundance and specificity of enzyme activity (red) and trancripts (non-red colors represent different paralogs) in core C_4 gene families. Error bars show standard error. In PEPC (a) and NADPME (b) the enzyme activities were used as the markers, and therefore are defined at 0 and 1 fraction in BS, respectively. The remaining families are AlaAT2 (c), AspAT (d), NADMDH (e), NADPME (f). Two identifiers per color indicate the sum of the genes annotated on positive and negative strand at same loci is used.



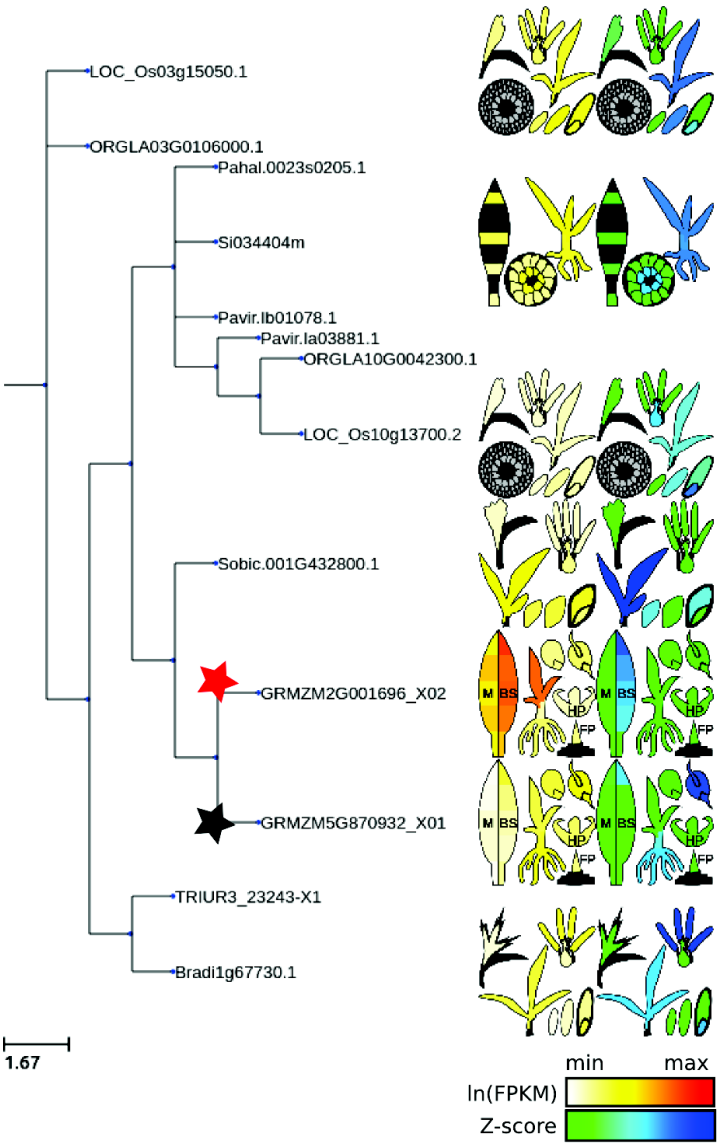
Supplementary Fig. 11: Abundance and specificity of trancripts (colors represent different paralogs) in core C_4 gene families. Error bars show standard error. Families are CA (a), PEPCK (b), PPDK (c) and PPDK-RP (d).



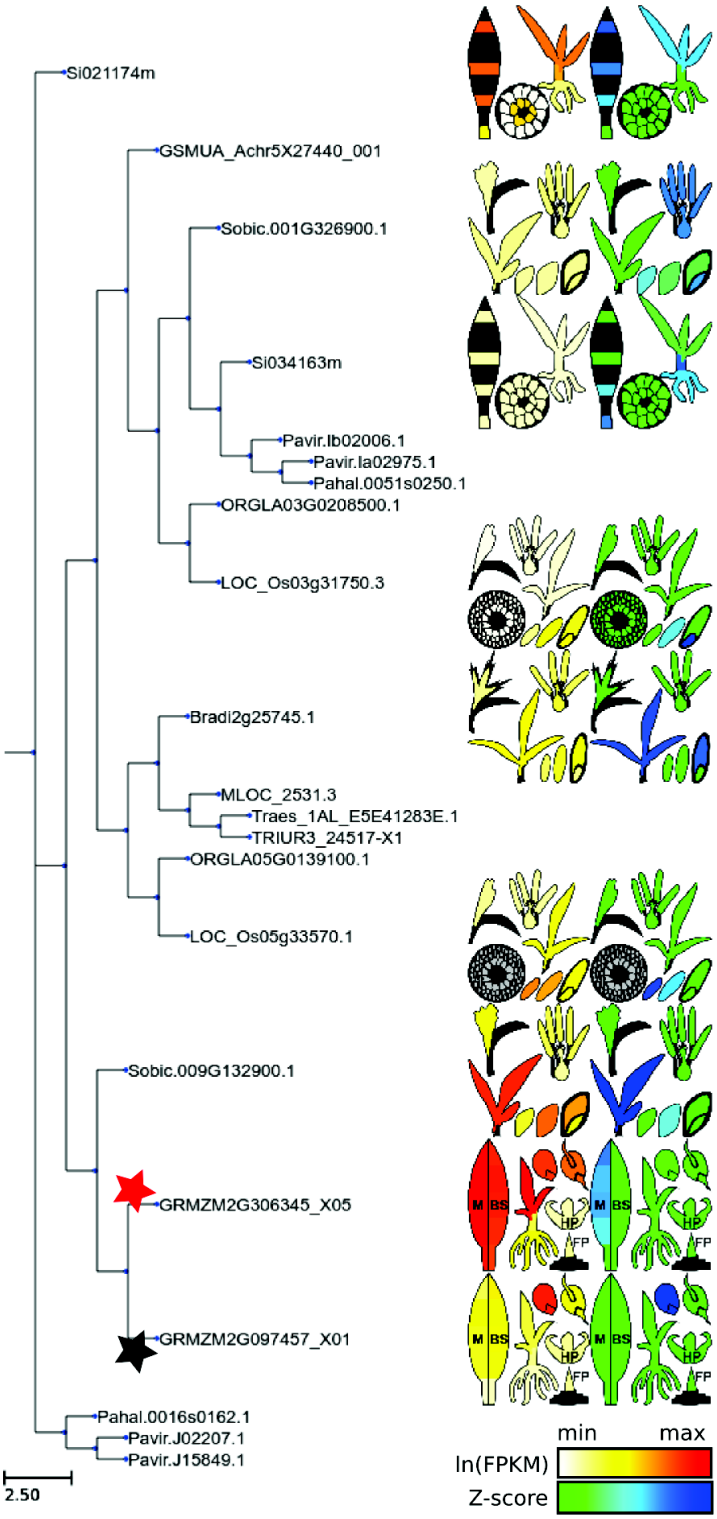
Supplementary Fig. 12: Abundance and specificity of trancripts (colors represent different paralogs) in core C_4 gene families. Error bars show standard error. Families are PPCK (a), PPT (b), and TPT (c).



Supplementary Fig. 13: Comparison of sequence and gene family factors to divergence in pattern (transformed pearson correlation; above) and level divergence ($|ln(peakFPKM1/peakFPKM2)|$; below). Shading indicates number of pairs in bin, relative to the largest bin. The red diamonds indicate pairs including core C_4 genes. To control age prior to plotting colinearity, pairs were filterd to those with dS <1.6, after which the mean dS of colinear (0.397) and non-colinear (0.402) pairs did not significantly differ (t-test $p = 0.44$).

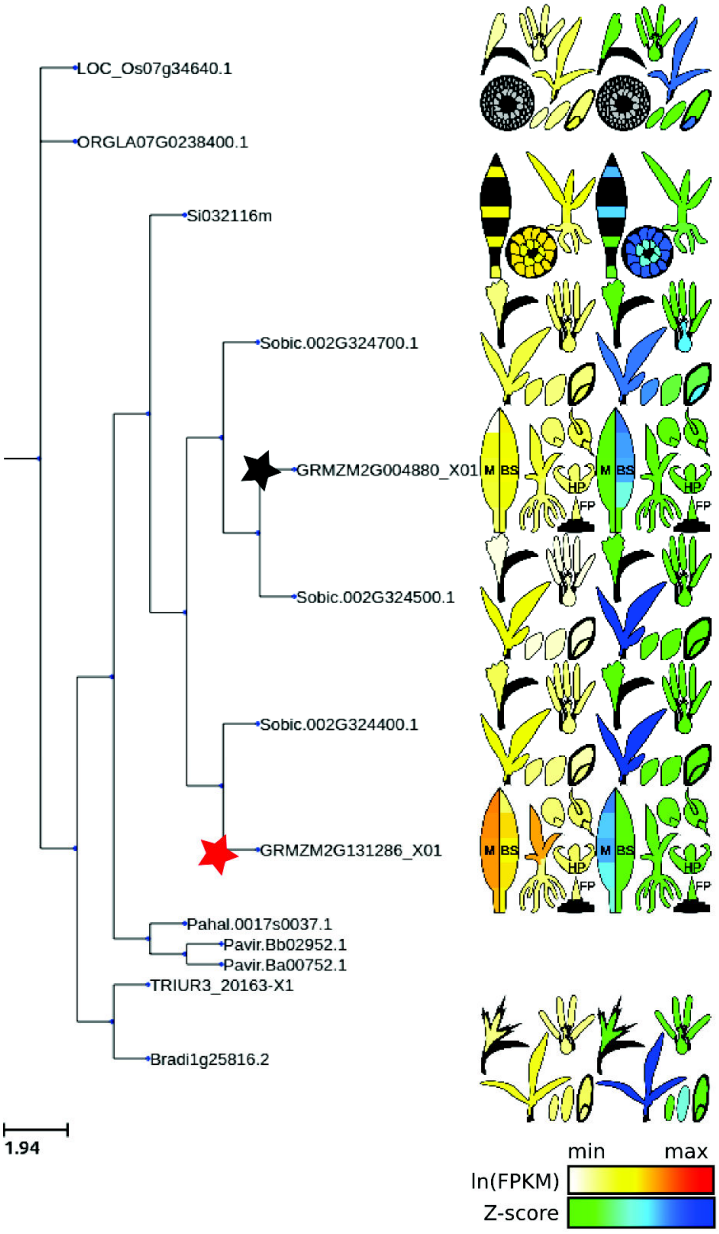


Supplementary Fig. 14: Expression pattern of PEPCK gene family on phylogeny. The C_4 gene and its young, syntenic paralog are marked with red and black stars, respectively.

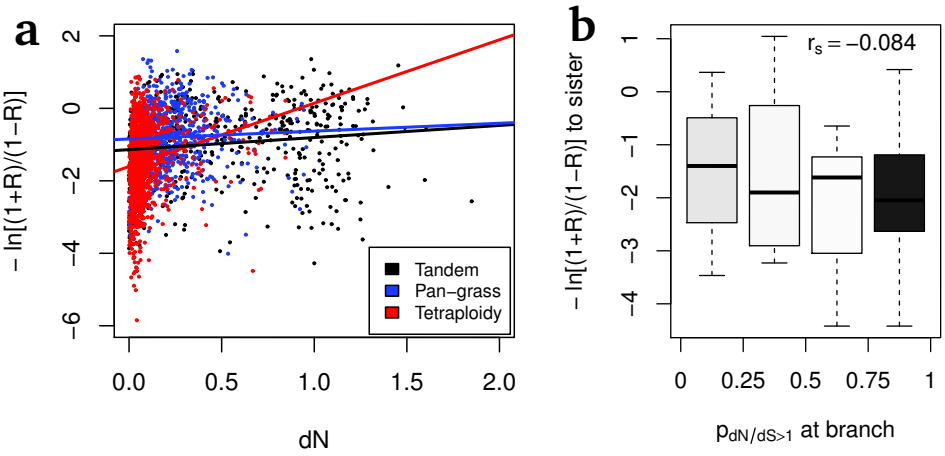


47

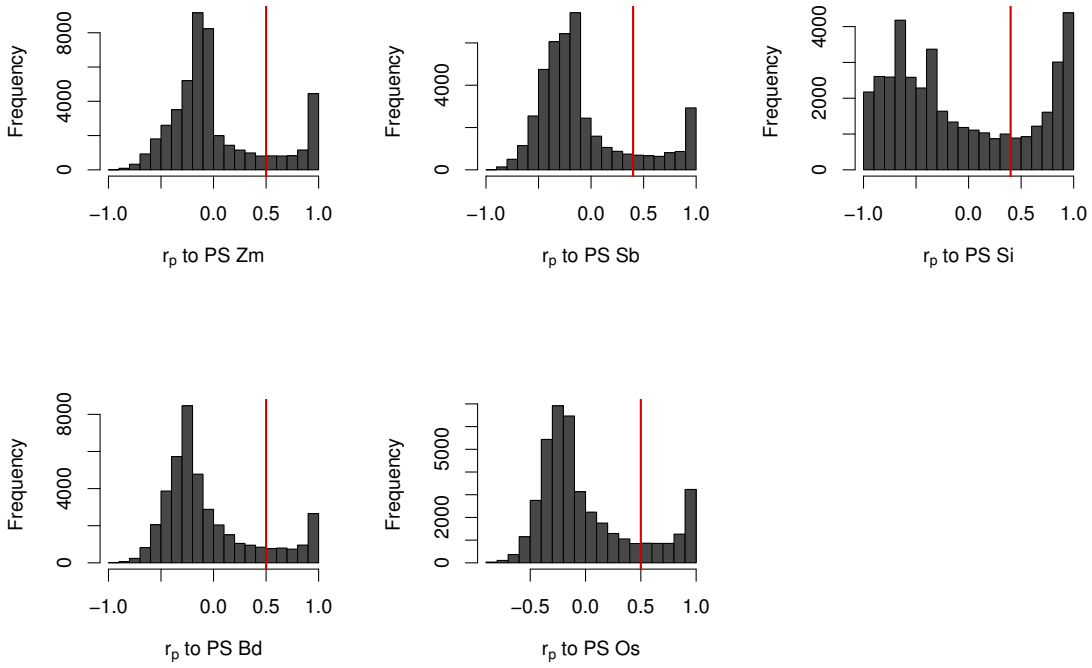
Supplementary Fig. 15: Expression pattern of PPDK gene family on phylogeny. The C_4 gene and its young, syntenic paralog are marked with red and black stars, respectively.



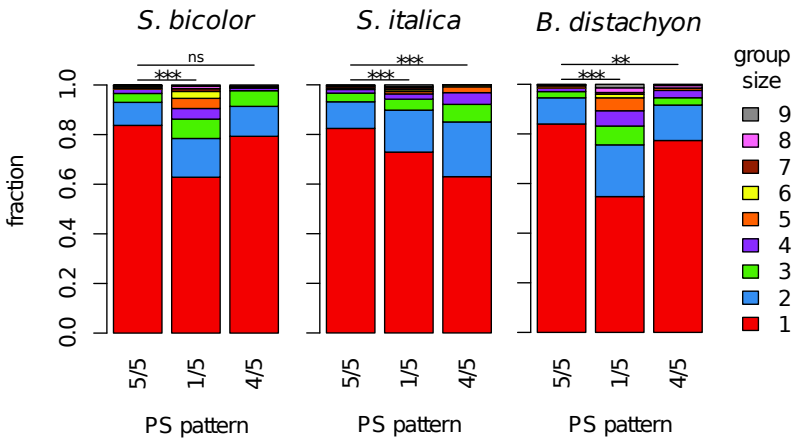
Supplementary Fig. 16: Expression pattern of PPDK-RP gene family on phylogeny. The C_4 gene and its young, syntenic paralog are marked with red and black stars, respectively.



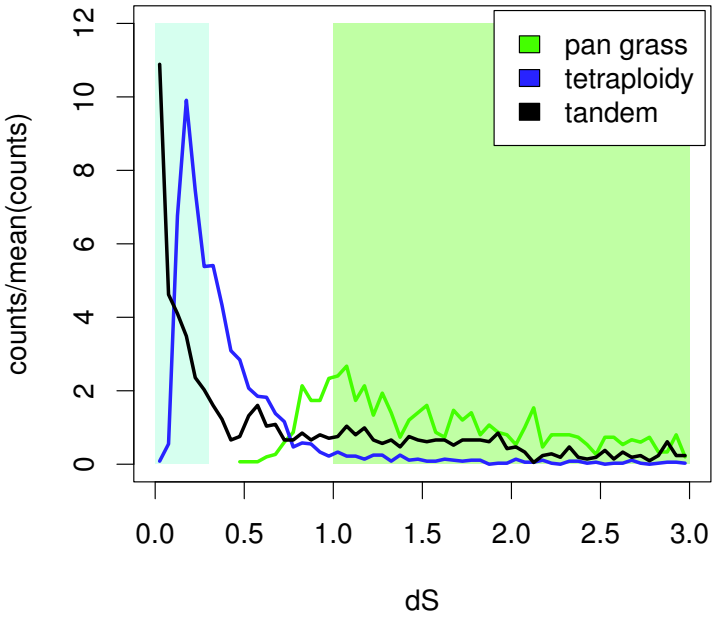
Supplementary Fig. 17: The relationship between expression pattern divergence and possible indicators of selection pressure. In (a) divergence is compared to the dN of genes with different annotated origins [27] including those where all duplicates are of the same age (the *Z. mays* tetraploidy and pan-grass WGDs). In (b) divergence at a branch on a phylogenetic tree is compared to the significance of positive selection at the same branch of the tree.



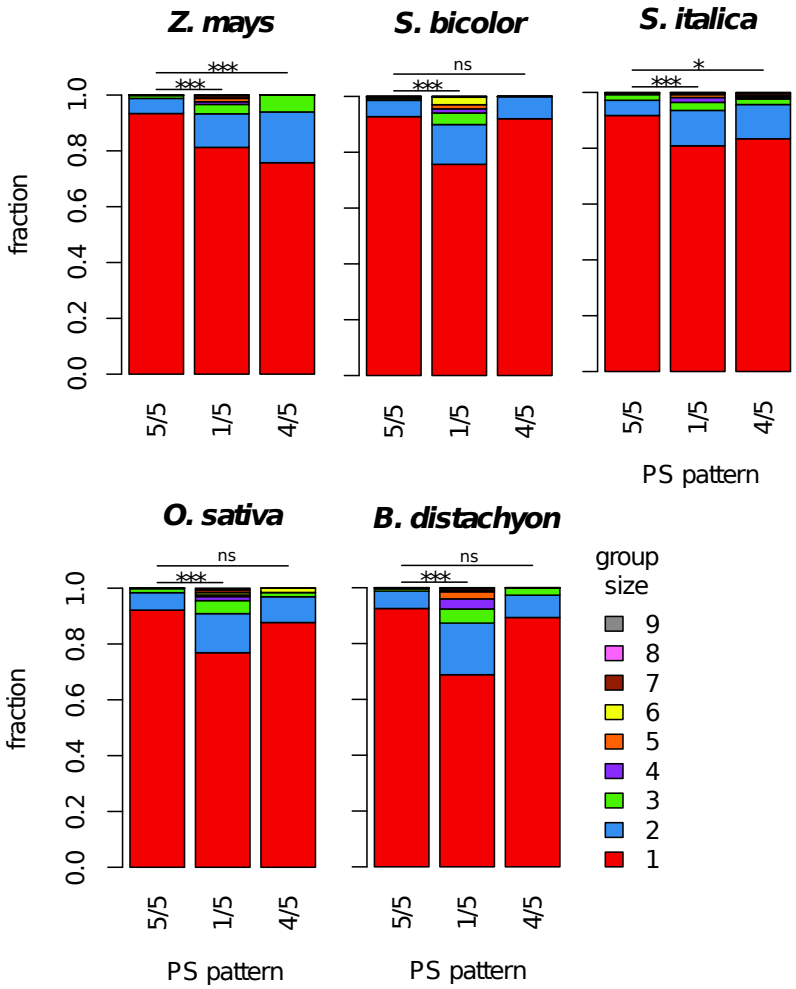
Supplementary Fig. 18: The categorization of expression patterns into photosynthesis-like or not. For the tissues in (Supplementary Fig. 7; Supplementary Table 8) the r_p between the expression pattern of each gene and the mean z-score of the genes in PS (photosynthesis) MapMan [41] category. Threshholds (red) were set to divide the resulting bi-modal distributions.



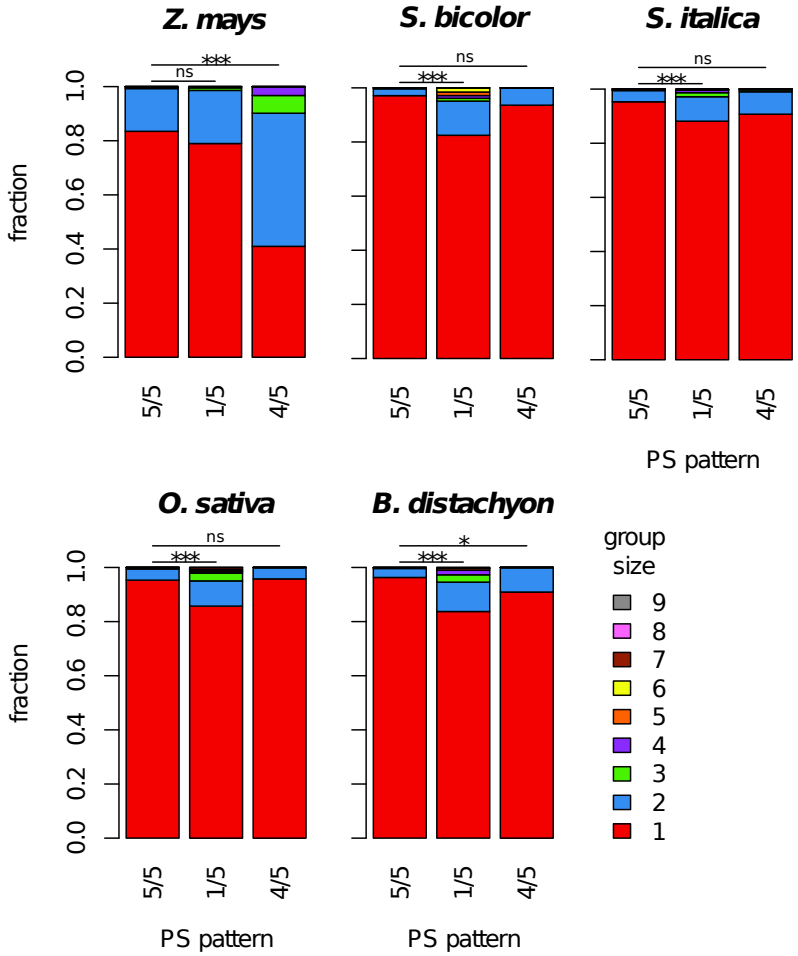
Supplementary Fig. 19: The relationship between group size (# paralogs in orthogroup in respective species) and photosynthetic pattern evolution. Cases where all 5 species show a photosynthetic-like expression pattern (see Supplementary Fig. 18) are considerd conserved, while cases where 4 of 5 or 1 of 5 species show a photosynthetic-like expression pattern are considered gain or loss, respectively. The odd species out is *S. bicolor* in (a), *S. italica* in (b) and *B. distachtyon* in (c).



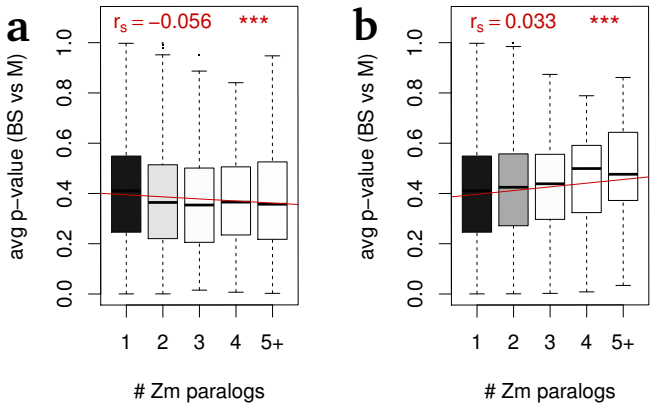
Supplementary Fig. 20: The dS (proxy for age) of duplicates from different WGDs or tandem duplications. Lines depict scaled histogram of duplicates originating from the pangrass duplication ∼70mya (green), from the maize tetraploidy event ∼5-12mya (blue), and from ongoing tandem duplications (black). The green and blue boxes represent the age ranges of duplicates in orthogroups defined as "old" (dS >1) and "young" (dS <0.3), respectively.



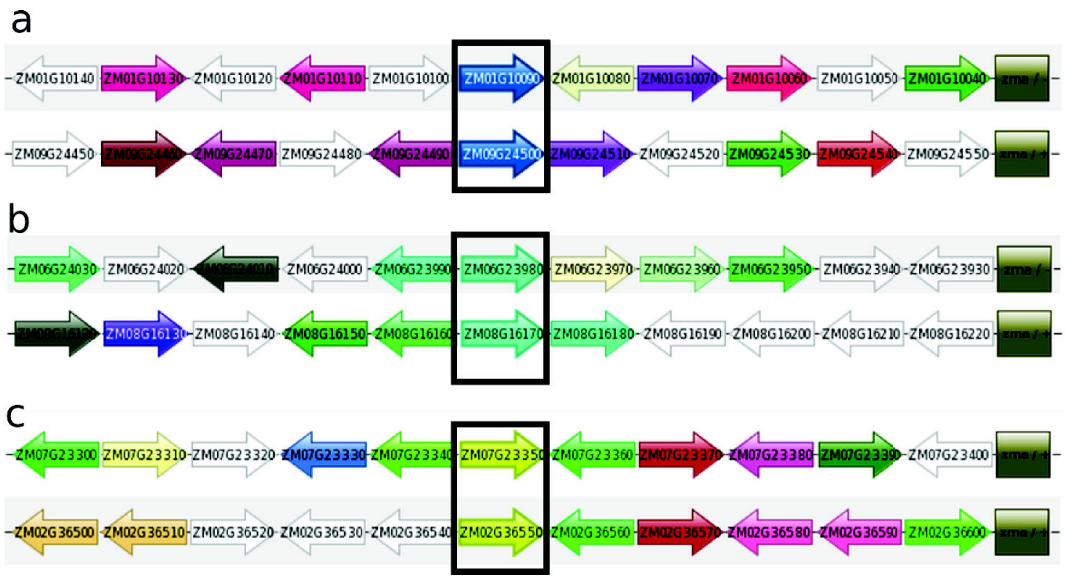
Supplementary Fig. 21: The relationship between group size (# paralogs in orthogroup in respective species) and photosynthetic pattern evolution in "ancient orthogroups" (min dS >1). cases where all 5 species show a photosynthetic-like expression pattern (see Supplementary Fig. 18) are considerd conserved, while cases where 4 of 5 or 1 of 5 species show a photosynthetic-like expression pattern are considered gain or loss, respectively. The odd species out is *Z. mays* in (a), *s. bicolor* in (b), *S. italica* in (c), *o. sativa* in (d), and *B. distachtyon* in (e).



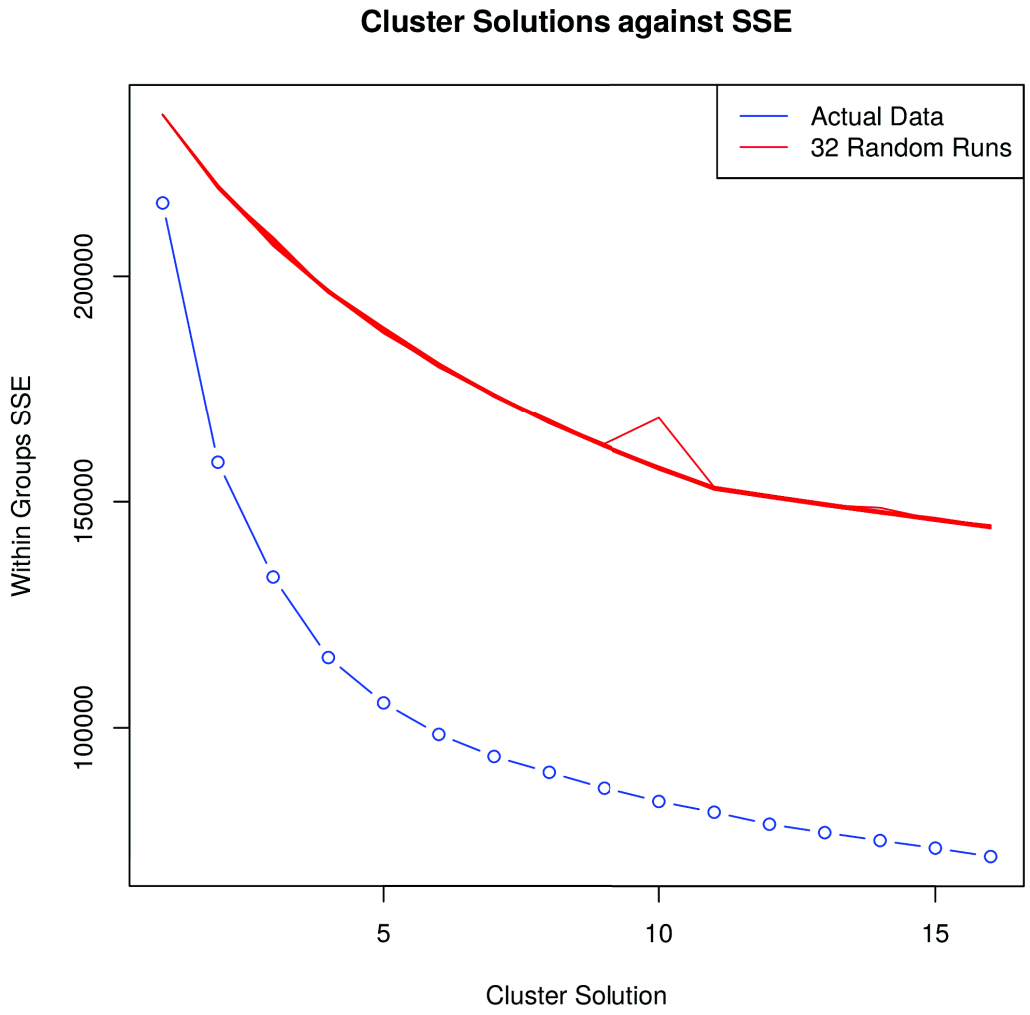
Supplementary Fig. 22: The relationship between group size (# paralogs in orthogroup in respective species) and photosynthetic pattern evolution in "young orthogroups" (max dS <0.3). Cases where all 5 species show a photosynthetic-like expression pattern (see Supplementary Fig. 18) are considerd conserved, while cases where 4 of 5 or 1 of 5 species show a photosynthetic-like expression pattern are considered gain or loss, respectively. The odd species out is *Z. mays* in (a), *S. bicolor* in (b), *S. italica* in (c), *O. sativa* in (d), and *B. distachtyon* in (e).



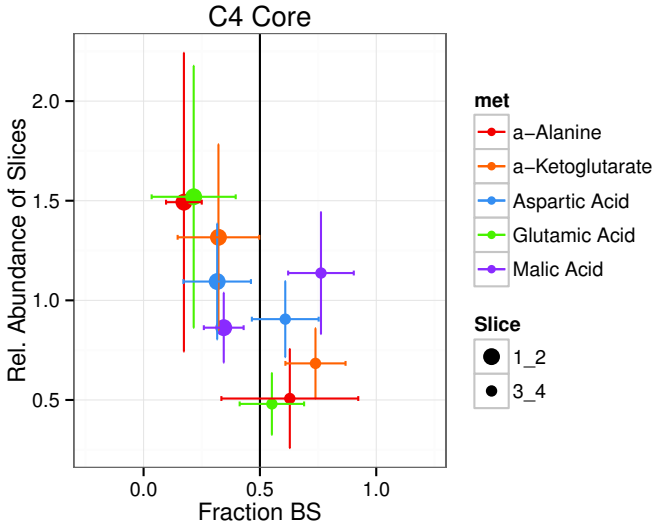
Supplementary Fig. 23: The relationship between group size (# paralogs in orthogroup in respective species) and significance of tissue specificity (average p-value) in (a) "ancient orthogroups" (min dS >1) and (b) "young orthogroups" (max dS <0.3)



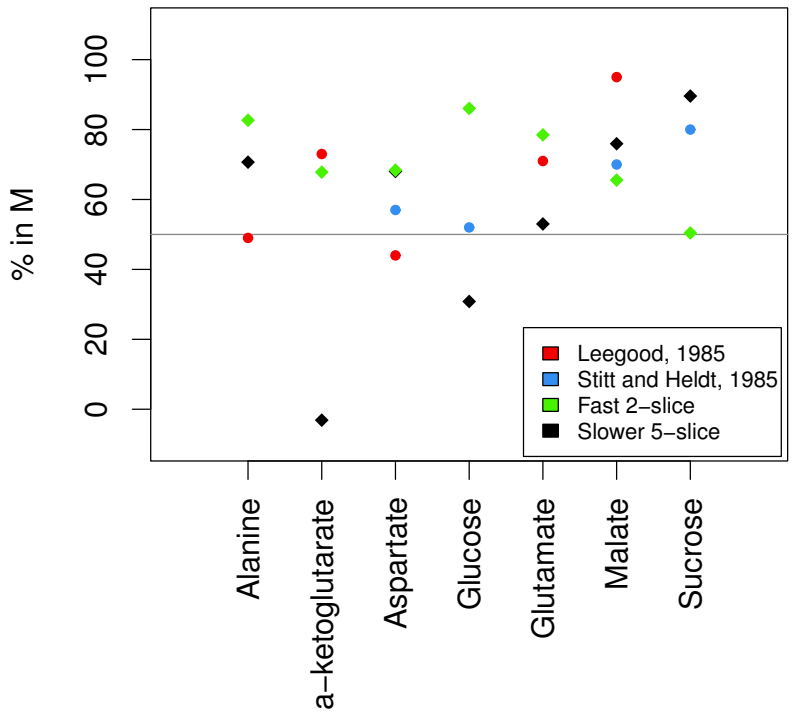
Supplementary Fig. 24: Local gene organization of syntenic, young duplicates with high divergence (in boxes). Namely, (a) PEPCK, (b) PPDK, and (c) PPDK-RP. Data and visualization from Plaza 3.0 [64]. Different colors or shades denote different homologous groups.



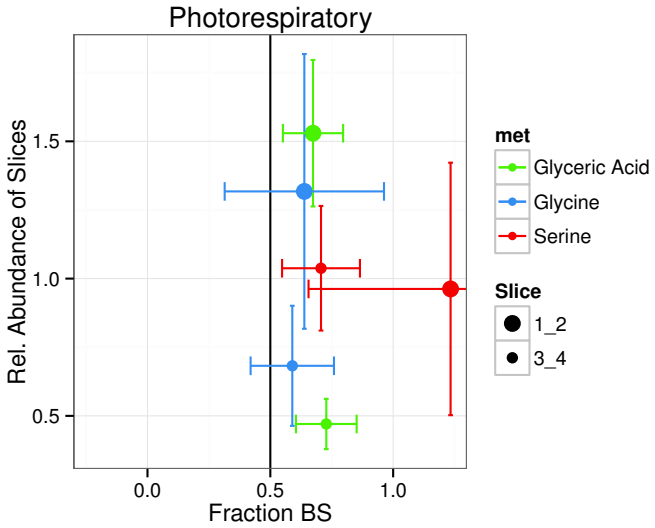
Supplementary Fig. 25: Sum of standard error compared between original and random data with a different number of cluster centers.



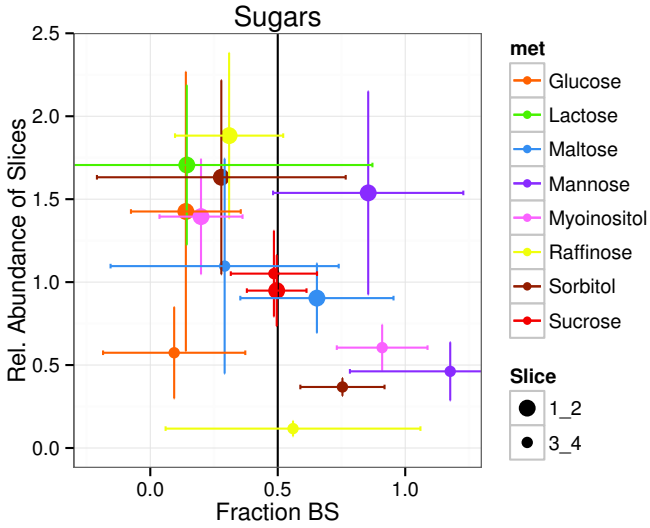
Supplementary Fig. 26: Distribution of measured core C4 metabolites between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute).



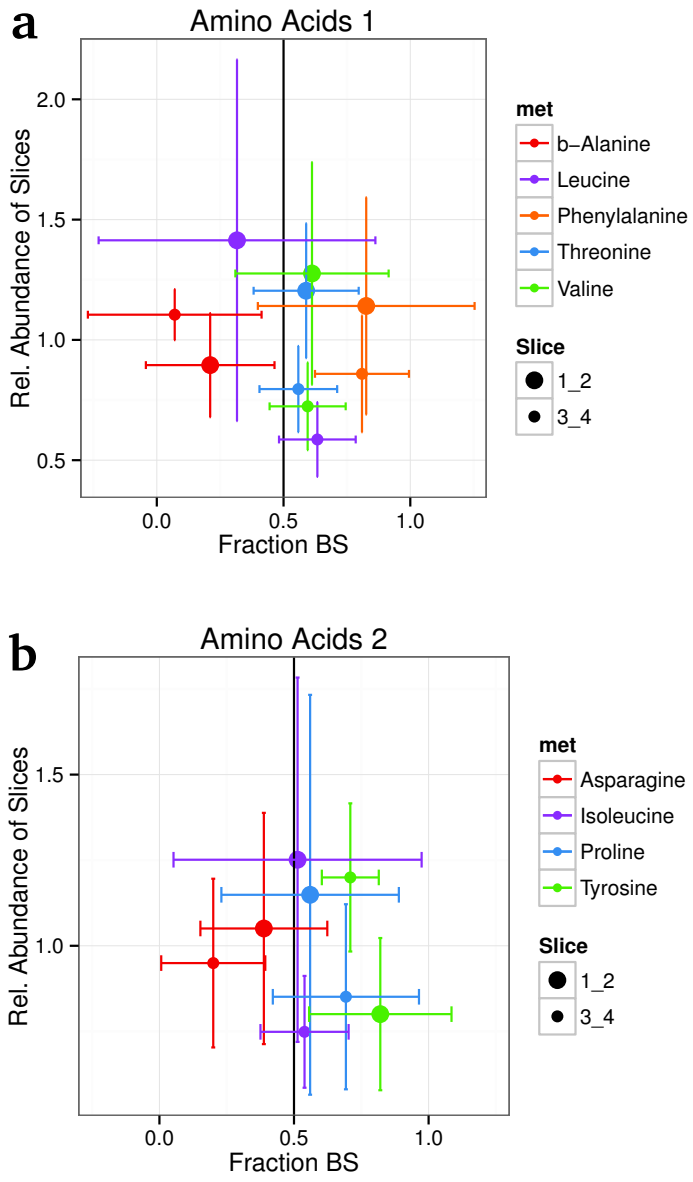
Supplementary Fig. 27: Comparison between the cell specificity of metabolites that have been measured in previous studies [37, 95]; the fast, 2-slice metabolite havest; and the slower, 5-slice gradient harvest.



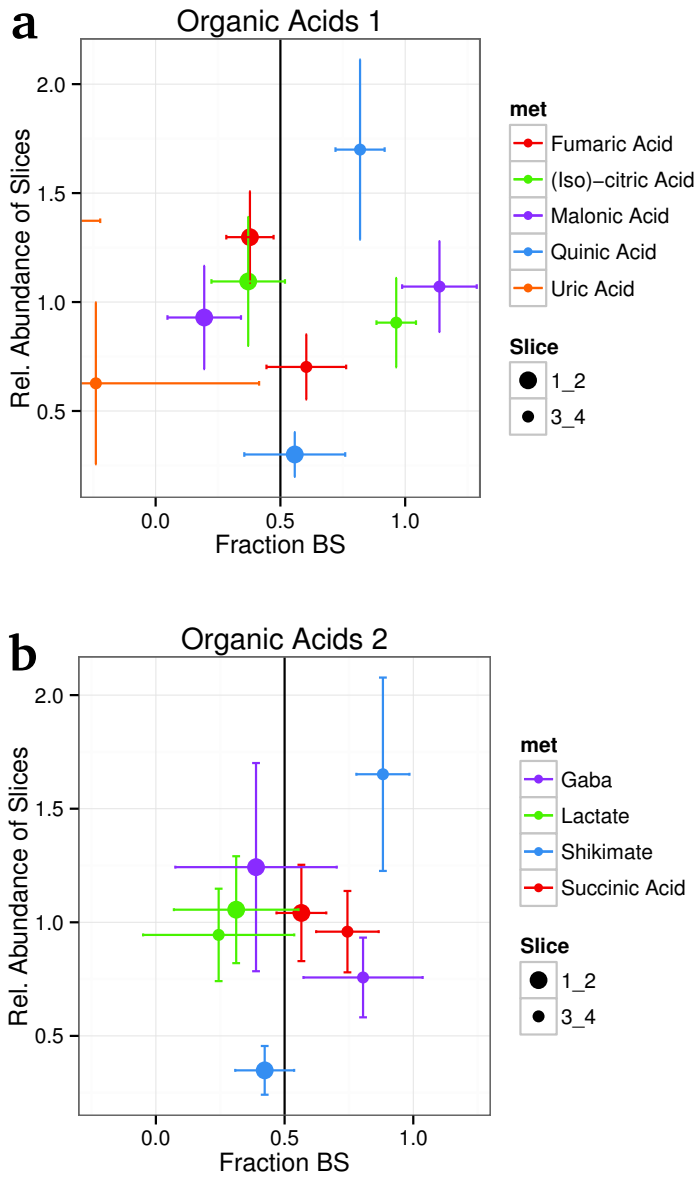
Supplementary Fig. 28: Distribution of measured photorespiratory metabolites between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute).



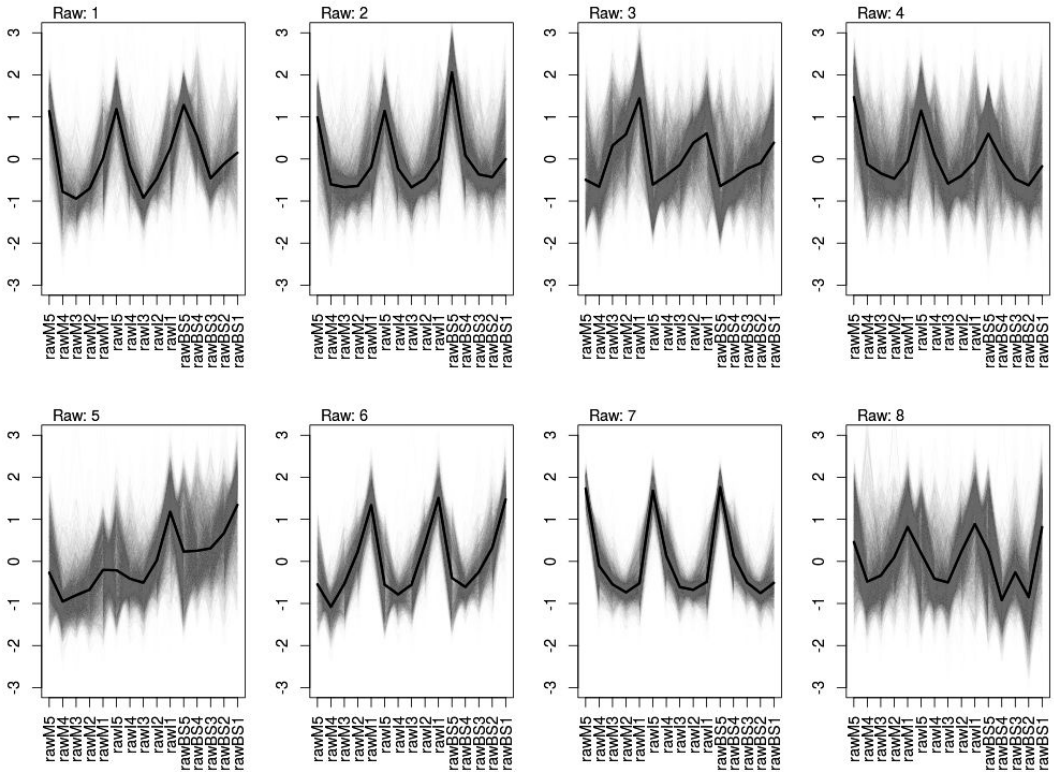
Supplementary Fig. 29: Distribution of measured sugars between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute).



Supplementary Fig. 30: Distribution of measured non-C4 core nor photorespiratory amino acids between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute). Metabolites split arbitrarily into (a) and (b) for plotting clarity.



Supplementary Fig. 31: Distribution of measured non-C4 core nor photorespiratory organic acids between slice 3_4 and 1_2 and between M and BS. Values are relative between slices for each metabolite (mean = 1), and normalized by sum peak area (so distributions are relative to other metabolites, and not absolute). Metabolites split arbitrarily into (a) and (b) for plotting clarity.



Supplementary Fig. 32: K-means clustering of BS & M gradient, plotted with the z-score of the raw (not deconvoluted) expression data. Individual genes in grey and centers in black.

Supplemental Tables

Supplementary Table 1: Comparison of BS/M values between studies by linear regression. The study using enzymatic and mechanical separation [40] reported the purest tissues, while the studies using laser micro dissection [38, 39] are more able to separate BS from vascular bundle. In mature tissue Slice 1 and Section + 9 [38] were compared; while in immature tissue Slice 4, 5 and Section -1 [38] were compared.

			Min > 10 FPKM			Min > 100 FPKM		
	x.study	y.study	slope	p.value	r^2	slope	p.value	r^2
Mature	[40]	[39]	0.32	1.21×10^{-91}	0.35	0.44	1.32×10^{-27}	0.64
	[40]	[38]	0.33	3.45×10^{-161}	0.36	0.39	8.71×10^{-31}	0.51
	[40]	S1	0.40	4.91×10^{-252}	0.27	0.71	9.22×10^{-54}	0.63
	[39]	[38]	0.96	~ 0	0.94	0.99	8.72×10^{-186}	0.97
	[39]	S1	0.94	2.22×10^{-115}	0.43	1.13	2.54×10^{-25}	0.64
	[38]	S1	0.94	9.65×10^{-220}	0.42	1.16	4.79×10^{-40}	0.59
Immature	[38]	S5	0.61	7.21×10^{-102}	0.24	0.68	1.99×10^{-30}	0.42
	[38]	S4	0.70	1.39×10^{-157}	0.37	0.68	2.92×10^{-30}	0.45

Supplementary Table 2: P-values from wilcox-rank test for differences between the C_4 genes, their closest homologs, and their remaining homologs.

	C_4 vs closest	C_4 vs remaining	remaining vs closest	notes on data
r_p to photosynthesis	1.21×10^{-5}	1.84×10^{-7}	0.15	see Table S.Tissues
peak FPKM	4.77×10^{-6}	7.84×10^{-9}	0.37	no PPDK_RP, PPCK
tissue specificity	9.32×10^{-3}	3.22×10^{-4}	0.61	$C_4 > 0.7$ only

Supplementary Table 3: P-values for a t-test of the divergence between the C4 core genes and all their paralogs vs all other paralog pairs in genome.

	Pattern divergence $ln(\frac{1+r_p}{1-r_p})$	**Level divergence** $\lvert ln(\frac{peakFPKM1}{peakFPKM2})\rvert$
M & BS gradient	0.016	1.01×10^{-5}
primordial leaf/husk gradient [29]	0.638	0.085
Atlas including leaves [30]	0.048	0.130
All non-leaf/husk tissues from above	0.683	0.417
All leaf/husk tissues from above	4.72×10^{-4}	1.40×10^{-6}

Supplementary Table 4: Multiple regression of expression level divergence vs sequence and gene family features.

	features	**estimate**	**standard error**	**p-value**	**r-squared**
deconvoluted BS & M gradient	dS	0.029	0.0099	3.28×10^{-03}	0.0045
	dN	-0.017	0.0102	9.76×10^{-02}	
	Colinearity	-0.050	0.0097	2.90×10^{-07}	
	# paralogs *Z. mays*	0.007	0.0094	4.48×10^{-01}	
	dN/dS	-0.003	0.0087	7.37×10^{-01}	
	C4 or not	0.027	0.0087	1.67×10^{-03}	
non-photosynthetic atlas tissues[30]	dS	0.053	0.0054	1.02×10^{-22}	0.00461
	dN	-0.012	0.0056	3.81×10^{-02}	
	Colinearity	-0.051	0.0055	7.35×10^{-21}	
	# paralogs *Z. mays*	-0.026	0.0055	1.77×10^{-06}	
	dN/dS	0.007	0.0052	2.11×10^{-01}	
	C4 or not	0.003	0.0052	5.41×10^{-01}	
all (developing) photosynthetic tissues (deconvoluted, primordial[29] and atlas[30]	dS	0.040	0.0099	4.89×10^{-05}	0.0061
	dN	-0.020	0.0102	5.33×10^{-02}	
	Colinearity	-0.051	0.0097	1.76×10^{-07}	
	# paralogs *Z. mays*	0.020	0.0094	3.78×10^{-02}	
	dN/dS	0.001	0.0087	8.80×10^{-01}	
	C4 or not	0.025	0.0087	3.46×10^{-03}	

Supplementary Table 5: Multiple regression of expression pattern divergence vs sequence and gene family features.

	features	estimate	standard error	p-value	r-squared
deconvoluted BS & M gradient	dS	0.060	0.0098	5.77×10^{-10}	0.034
	dN	0.029	0.0100	4.30×10^{-03}	
	Colinearity	-0.096	0.0095	8.56×10^{-24}	
	# paralogs Z. mays	0.068	0.0093	2.71×10^{-13}	
	dN/dS	-0.007	0.0086	3.88×10^{-01}	
	C4 or not	0.039	0.0086	6.52×10^{-06}	
non-photosynthetic atlas tissues[30]	dS	0.029	0.0054	9.58×10^{-08}	0.025
	dN	0.027	0.0055	9.52×10^{-07}	
	Colinearity	-0.065	0.0054	2.29×10^{-33}	
	# paralogs Z. mays	0.108	0.0054	3.76×10^{-87}	
	dN/dS	-0.009	0.0052	9.69×10^{-02}	
	C4 or not	0.005	0.0051	3.51×10^{-01}	
all (developing) photosynthetic tissues (deconvoluted, primordial[29], and atlas[30])	dS	0.108	0.0095	6.97×10^{-30}	0.085
	dN	-0.029	0.0097	2.69×10^{-03}	
	Colinearity	-0.148	0.0093	1.78×10^{-56}	
	# paralogs Z. mays	0.160	0.0090	2.54×10^{-69}	
	dN/dS	-0.009	0.0084	2.68×10^{-01}	
	C4 or not	0.041	0.0084	1.18×10^{-06}	

Supplementary Table 6: The paralog pairs of the MapMan PS category, which occurred non-ambiguously in the edge of Clusters 3 “M-tip” and 5 “BS-tip”, and whether they consume ATP

Paralogs	MapMan bincode	MapMan subcategory of PS	Uses ATP	Clusters
		calvin cycle		
GRMZM2G089136, GRMZM2G382914	1.3.3	phosphoglycerate kinase	Y	3, 5
GRMZM2G026024, GRMZM2G463280	1.3.12	PRK	Y	5, 3
GRMZM2G162529, GRMZM2G463280	1.3.12	PRK	Y	5, 3
		photorespiration		
GRMZM2G018786, GRMZM2G054663	1.2.7	glycerate kinase	Y	3, 5
GRMZM2G076239, GRMZM2G129246	1.2.2	glycolate oxydase	N	3, 5
		lightreaction		
GRMZM2G010555, GRMZM2G102349	1.1.40	cyclic electron flow-chlororespiration	N	5, 3
GRMZM5G885392, GRMZM5G896082	1.1.40	cyclic electron flow-chlororespiration	N	3, 5
GRMZM2G048313, GRMZM2G122337	1.1.5.2	other electron carrier (ox/red).ferredoxin	N	5, 3
GRMZM2G329047, GRMZM2G377855	1.1.2.2	photosystem I.PSI polypeptide subunits	N	5, 3

Supplementary Table 7: Non-default parameters used for bioinformatics programs.

tophat2	For studies with reads shorter than 50 bases **--segment-length=N** (N = read length/2) was set so that reads were mapped in at least 2 segments **--b2-very-sensitive** and **--read-realign-edit-dist=0** were set to increase sensitivity **-G** <**file.gtf**> was used to guide mappings to annotated transcriptome
cufflinks2	**-u** was set to improve distribution of reads mapping to more than one position **-G** <**file.gtf**> was used to guide assembly to annotated transcriptome
cutadapt	**-e0.1** was used to set the maximum fraction of errors for a match **-O5** was used to require an adaptor match to be at least 5 bases long
fastq-quality-trimmer	**-Q33** indicates the quality encoding **-l25** was used to discard trimmed reads shorter than 25 bases **-t28** was set for the quality score threshhold
blastall	**-p blastn** was used for BLAST searches in nucleotide space between *Z. mays* genome releases, while **-p blastp** was used for BLAST searches in protein space between species **-m8** was set for a tabular output **-FF** was set to turn off quality filtering, and thereby allow avoid excluding perfect matches between different *Z. mays* genome releases **-e1e-1** was set to skip any matches of a quality where 0.1 or more would be expected by chance based on database size
McscanX	**-w1**, **-k300**, **-m50**, and **-g-0.5** were set to err on the sensitive side while detecting colinearity
Prank	**+F** was set as recommended for sequences with many insertions or deletions
Mafft	**--auto** was used
RaxML	**-m PROTGAMMAIJTT** was set to employ the JTT amino acid substitution matrix with optimized substitution rates, and a gamma model of rate heterogenity including invariant sites. **-k** was used to print branch lengths **-NautoMR** was used to stop bootstrapping after convergence **-b 123** is used to set a seed for random numbers while bootstrapping **-p 12345** was used to set a seed for random numbers in parsimony inference
codeml	**runmode = -2**, **model = 0**, and **Nssites = 0 F3x4 model** were used to estimate pairwise dN and dS **runmode = 0**, **seqtype = 1**, **CodonFreq = 2**, **model = 2**, **Nssites = 2**, **fix_kappa = 0**, and **kapa = 2** were used for both null and alternative branch site models The negative log likelihood of the null model with parameters **fix_omega = 1** and **omega = 1** in the null model was compared to the alternative model with parameters **fix_omega = 0** and **omega = 1.5** to determine significance of dN/dS signature >1 for the branch site models

Supplementary Table 8: Tissues used for differnet analyses

Z. mays	r_p PS	Leaf & husk	mature tissue specificity
6DAS_Prim_Root [30]	Y		
24H_Germ_seed [30]	Y		
16DAP_Embryo [30]	Y		
V3_Stem_SAM [30]	Y		
12DAP_W_seed [30]	Y		
10DAP_W_seed [30]	Y		
16DAP_W_seed [30]	Y		
14DAP_W_seed [30]	Y		
14DAP_Endopsperm [30]	Y		
12DAP_Endopsperm [30]	Y		
16DAP_Endosperm [30]	Y		
V9_13th_Leaf [30]		Y	
V9_11th_Leaf [30]		Y	
V9_Immature_Leaves [30]		Y	
R2_13th Leaf [30]		Y	
VT_13th Leaf [30]		Y	
V9_8th_Leaf [30]		Y	
V5_Tip_s-2_Leaf [30]	Y	Y	
All primordia samples [29]		Y	
M5		Y	
M4		Y	
M3		Y	
M2		Y	Y
M1		Y	Y
BS5		Y	
BS4		Y	
BS3		Y	
BS2		Y	Y
BS1		Y	Y
S. italica	r_p PS		
M [34]			
BS [34]			
leaf ligule 4 + 1 [32]			
leaf ligule 3 - 1 [32]			
leaf ligule 3 + 2 [32]			
leaf tip - 1 [32]			
root [33]	Y		
stem [33]	Y		
leaf [33]	Y		
spica [33]	Y		

Supplemental References

92. Amthor, J. S. From sunlight to phytomass: on the potential efficiency of converting solar radiation to phyto-energy. *New Phytologist* **188,** 939–959 (2010).

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

94. Hatch, M. D. The C_4 -pathway of photosynthesis. Evidence for an intermediate pool of carbon dioxide and the identity of the donor C_4-dicarboxylic acid. *The Biochemical journal* **125,** 425–32 (Nov. 1971).

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

96. Gardiner, J., Sherr, I. & Scarpella, E. Expression of DOF genes identifies early stages of vascular development in Arabidopsis leaves. *International Journal of Developmental Biology* **54,** 1389 (2010).

97. Ward, J. M., Cufr, C. A., Denzel, M. A. & Neff, M. M. The Dof transcription factor OBP3 modulates phytochrome and cryptochrome signaling in Arabidopsis. *The Plant Cell Online* **17,** 475–485 (2005).

98. Nemhauser, J. & Chory, J. Photomorphogenesis. *The Arabidopsis book/American Society of Plant Biologists* **1** (2002).

99. Kelley, D. R., Arreola, A., Gallagher, T. L. & Gasser, C. S. ETTIN (ARF3) physically interacts with KANADI proteins to form a functional complex essential for integument development and polarity determination in Arabidopsis. *Development* **139,** 1105–1109 (2012).

100. Iwasaki, M. *et al.* Dual regulation of ETTIN (ARF3) gene expression by AS1-AS2, which maintains the DNA methylation level, is involved in stabilization of leaf adaxial-abaxial partitioning in Arabidopsis. *Development* **140,** 1958–1969 (2013).

101. Soares-Cordeiro, A. S. *et al.* Variations in the dorso-ventral organization of leaf structure and Kranz anatomy coordinate the control of photosynthesis and associated signalling at the whole leaf level in monocotyledonous species. *Plant, cell & environment* **32,** 1833–1844 (2009).

102. Belin, C., Megies, C., Hauserová, E. & Lopez-Molina, L. Abscisic acid represses growth of the Arabidopsis embryonic axis after germination by enhancing auxin signaling. *The Plant Cell Online* **21,** 2253–2268 (2009).

103. Teale, W. D., Paponov, I. A. & Palme, K. Auxin in action: signalling, transport and the control of plant growth and development. *Nature Reviews Molecular Cell Biology* **7,** 847–859 (2006).

104. Park, M. Y., Kang, J.-y. & Kim, S. Y. Overexpression of AtMYB52 confers ABA hypersensitivity and drought tolerance. *Molecules and cells* **31,** 447–454 (2011).

105. Cassan-Wang, H. *et al.* Identification of novel transcription factors regulating secondary cell wall formation in Arabidopsis. *Frontiers in plant science* **4** (2013).

106. Weckopp, S. C. & Kopriva, S. Are changes in sulfate assimilation pathway needed for evolution of C_4 photosynthesis? *Frontiers in Plant Science* **5,** 773 (2015).

107. Lee, J. R., Boltz, K. A. & Lee, S. Y. Molecular chaperone function of Arabidopsis thaliana phloem protein 2-A1, encodes a protein similar to phloem lectin. *Biochemical and biophysical research communications* **443,** 18–21 (2014).

108. Zhang, C. *et al.* Harpin-induced expression and transgenic overexpression of the phloem protein gene AtPP2-A1 in Arabidopsis repress phloem feeding of the green peach aphid Myzus persicae. *BMC plant biology* **11,** 11 (2011).

109. Love, J. *et al.* Ethylene is an endogenous stimulator of cell division in the cambial meristem of Populus. *Proceedings of the National Academy of Sciences* **106,** 5984–5989 (2009).

110. Pesquet, E. & Tuominen, H. Ethylene stimulates tracheary element differentiation in Zinnia elegans cell cultures. *New Phytologist* **190,** 138–149 (2011).

111. Schuetz, M. *et al.* Laccases direct lignification in the discrete secondary cell wall domains of protoxylem. *Plant physiology* **166,** 798–807 (2014).

112. Groover, A. & Jones, A. M. Tracheary element differentiation uses a novel mechanism coordinating programmed cell death and secondary cell wall synthesis. *Plant Physiology* **119,** 375–384 (1999).

Chapter 3

Co-Author Manuscripts

3.1 Manuscript 4: An mRNA Blueprint for C_4 Photosynthesis Derived from Comparative Transcriptomics of Closely Related C_3 and C_4 Species

Overview

Title: An mRNA Blueprint for C_4 Photosynthesis Derived from Comparative Transcriptomics of Closely Related C_3 and C_4 Species
Authors: Andrea Bräutigam, Kaisa Kajala, Julia Wullenweber, Manuel Sommer, David Gagneul, Katrin L. Weber, Kevin M. Carr, Udo Gowik, Janina Maß, Martin J. Lercher, Peter Westhoff, Julian M. Hibberd, and Andreas P.M. Weber
Published in Plant Physiology, January 2011

Co-authorship

Contributions

- Bioinformatics support
- Evaluation of methods for sequence analysis

An mRNA Blueprint for C_4 Photosynthesis Derived from Comparative Transcriptomics of Closely Related C_3 and C_4 Species[1][W][OA]

Andrea Bräutigam[2], Kaisa Kajala[2], Julia Wullenweber, Manuel Sommer, David Gagneul, Katrin L. Weber, Kevin M. Carr, Udo Gowik, Janina Maß, Martin J. Lercher, Peter Westhoff, Julian M. Hibberd[2], and Andreas P.M. Weber[2]*

Institute of Plant Biochemistry (A.B., J.W., M.S., D.G., K.L.W., A.P.M.W.), Institute of Plant Molecular and Developmental Biology (U.G., P.W.), and Institute of Informatics (J.M., M.J.L.), Heinrich-Heine University, 40225 Duesseldorf, Germany; Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom (K.K., J.M.H.); and Bioinformatics Core, Research Technology Support Facility, Michigan State University, East Lansing, Michigan 48824 (K.M.C.)

C_4 photosynthesis involves alterations to the biochemistry, cell biology, and development of leaves. Together, these modifications increase the efficiency of photosynthesis, and despite the apparent complexity of the pathway, it has evolved at least 45 times independently within the angiosperms. To provide insight into the extent to which gene expression is altered between C_3 and C_4 leaves, and to identify candidates associated with the C_4 pathway, we used massively parallel mRNA sequencing of closely related C_3 (*Cleome spinosa*) and C_4 (*Cleome gynandra*) species. Gene annotation was facilitated by the phylogenetic proximity of *Cleome* and Arabidopsis (*Arabidopsis thaliana*). Up to 603 transcripts differ in abundance between these C_3 and C_4 leaves. These include 17 transcription factors, putative transport proteins, as well as genes that in Arabidopsis are implicated in chloroplast movement and expansion, plasmodesmatal connectivity, and cell wall modification. These are all characteristics known to alter in a C_4 leaf but that previously had remained undefined at the molecular level. We also document large shifts in overall transcription profiles for selected functional classes. Our approach defines the extent to which transcript abundance in these C_3 and C_4 leaves differs, provides a blueprint for the NAD-malic enzyme C_4 pathway operating in a dicotyledon, and furthermore identifies potential regulators. We anticipate that comparative transcriptomics of closely related species will provide deep insight into the evolution of other complex traits.

C_4 photosynthesis is a complex biological trait that enables plants to either accumulate biomass at a much faster rate or live in adverse environments compared with "ordinary" plants (Hatch, 1987; Osborne and Freckleton, 2009). These C_4 plants have added a CO_2 concentration mechanism on top of their regular photosynthetic carbon fixation that makes them not only more efficient at assimilating inorganic carbon; they frequently also have higher water and nitrogen use efficiencies (Black, 1973; Oaks, 1994; Osborne and Freckleton, 2009). Beyond the basic biochemistry, our understanding of C_4 photosynthesis is limited.

The principle of C_4 photosynthesis is deceivingly simple: instead of using Rubisco as the primary carbon-fixing enzyme, C_4 plants use phospho*enol*pyruvate carboxylase (PEPC). Unlike Rubisco, PEPC is more specific for inorganic carbon (Hatch, 1987). Since the C_4 cycle is an add-on rather than a replacement for Rubisco and the Calvin-Benson cycle, the prefixed CO_2 is transported in a bound form, a C_4 acid (hence the name), to the site of Rubisco. The C_4 cycle generates high concentrations of CO_2 around Rubisco (Hatch, 1987), and this increases the rate of photosynthesis because competition between CO_2 and oxygen at the active site of Rubisco is reduced (Jordan and Ogren, 1984). In most C_4 plants, concentrating CO_2 around Rubisco involves the reactions of photosynthesis being partitioned between bundle sheath (BS) and mesophyll (M) cells as well as changes to cell biology and leaf development (Hatch, 1987; Sage, 2004), although in some lineages, C_4 photosynthesis operates within individual cells (Reiskind et al., 1989; Keeley, 1998; Voznesenskaya et al., 2001, 2002, 2003).

In all known C_4 plants, CO_2 enters M cells and is converted into bicarbonate by carbonic anhydrase.

[1] This work was supported by the German Research Council (grant nos. WE 2231/4–1 to A.P.M.W., SFB TR1 to P.W. and A.P.M.W., and IRTG 1525/1 to P.W. and A.P.M.W.) and the Leverhulme Trust and Isaac Newton Trust (to J.M.H.).

[2] These authors contributed equally to the article.

* Corresponding author; e-mail andreas.weber@uni-duesseldorf.de.

The authors 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.plantphysiol.org) are: Julian M. Hibberd (julian.hibberd@plantsci.cam.ac.uk) and Andreas P.M. Weber (andreas.weber@uni-duesseldorf.de).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.110.159442

PEPC then combines HCO_3^- with PEP to generate the C_4 oxaloacetic acid, which is rapidly converted into either Asp or malate. These C_4 acids then diffuse to the site of Rubisco through abundant plasmodesmata, where C_4 acid decarboxylases release CO_2 (Hatch, 1987). Three distinct C_4 acid decarboxylases, known as NADP-dependent malic enzyme (NADP-ME), NAD-dependent malic enzyme (NAD-ME), and PEP carboxykinase, have been coopted into the C_4 pathway, and this has been used to define three biochemical subtypes of C_4 photosynthesis. The three-carbon compound released after decarboxylation diffuses back to the M cells and is converted to PEP catalyzed by pyruvate,orthophosphate dikinase (PPDK; Hatch and Slack, 1968). Because the enzymes involved in the C_4 cycle are found in the cytosol, chloroplasts, and mitochondria, a significant amount of transport across organellar membranes is required for the C_4 cycle to operate. However, few genes encoding transporters that allow the increased intracellular flux of metabolites required for C_4 photosynthesis have been identified (Bräutigam et al., 2008a; Majeran and van Wijk, 2009). In addition, we have a very limited understanding of the mechanisms controlling the altered cell biology and morphology associated with C_4 leaves. The C_4 cycle likely affects not only the relatively small number of enzymes and transport proteins needed to perform the core reactions but, given the consequences on the ecological performance of the plants, also a range of other processes.

The gaps in our understanding of the mechanisms underlying C_4 photosynthesis limit insight into a metabolic pathway that has evolved repeatedly at least 45 times in plants (Sage, 2004) and so is of interest in terms of understanding a remarkable example of convergent evolution. In addition, because C_4 plants are among the most productive on the planet and the pathway is associated with increased water and nitrogen use efficiencies (Brown, 1999), it has been suggested that characteristics of C_4 photosynthesis should be placed into C_3 crops (Matsuoka et al., 2001; Mitchell and Sheehy, 2006; Hibberd et al., 2008). A more complete understanding of genes involved in C_4 photosynthesis is fundamental to attempts at placing components of the C_4 pathway into C_3 crops to increase yield.

Recently, a new set of tools has become available to analyze species without sequenced genomes on a genomic scale: next generation sequencing (NGS) technology (summarized in Metzker, 2010). With NGS, the transcriptome of a tissue can be sequenced and quantified at the same time (RNA-Seq; Wang et al., 2009). The 454 FLX genome sequencer provides a quarter million sequence reads of 230 bases in each run from a cDNA template generated from mRNA (http://www.454.com/; Metzker, 2010). The resulting reads can be mapped onto a closely related reference to quantify the number of reads matching a gene locus, thus providing a measure of transcript abundance (Flicek and Birney, 2009; Bräutigam and Gowik, 2010). We chose to compare the C_4 plant *Cleome gynandra* with the C_3 plant *Cleome spinosa*, since they are members of the same genus and are closely related to Arabidopsis (*Arabidopsis thaliana*; Brown et al., 2005; Marshall et al., 2007). Given the close phylogenetic relationship, we can take advantage of the well-annotated Arabidopsis genome (Swarbreck et al., 2008) and its known genome history (Bowers et al., 2003; Haberer et al., 2004; Thomas et al., 2006) to identify and quantify the biological functions regulated at the level of transcript abundance in the C_4 species compared with the C_3 species. Although the experiment will also capture variation in the abundance of transcripts associated with differences between the species that do not relate to C_4 photosynthesis, the close proximity of the *Cleome* species should reduce this effect. We chose to use mature fully differentiated leaves for the analysis, since we wanted to minimize the influence of species-specific effects during leaf differentiation but rather focus on transcript profiles when C_4 photosynthesis is fully operational. Once this profile is defined, analysis of developmental stages may reveal how the profile is achieved during differentiation.

By comparing the transcriptomes of closely related C_3 and C_4 species, we will test (1) whether cross-species transcriptomic comparisons are feasible, (2) the degree to which the core C_4 cycle enzymes and transport proteins are regulated at the level of transcript abundance, and (3) whether the changes in metabolism associated with C_4 photosynthesis are associated with additional unexpected shifts in transcript profiles in leaves of C_4 compared with C_3 plants, and (4) define candidates for additional functions critical to C_4 photosynthesis based on unbiased observation of the data. By analyzing the complete transcriptome, we define the maximal extent to which the C_4 pathway alters leaf transcript profiles.

RESULTS

Physiological Analysis of C_3 and C_4 Leaves Confirms C_4 Metabolism in *C. gynandra*

To confirm that the *C. spinosa* and *C. gynandra* leaves we used for transcriptomic analysis were using C_3 and C_4 photosynthesis, respectively, we analyzed the steady-state levels of metabolites associated with the C_4 cycle. For example, large quantities of Asp, Ala, and pyruvate are produced in M and BS cells of NAD-ME C_4 leaves, and they were 19, 3.9, and 3.6 times more abundant, respectively, in *C. gynandra* compared with *C. spinosa* (Supplemental Table S1). In contrast, and in agreement with the lower demand for the photorespiration in C_4 leaves, glycerate and glycolate, intermediates of the photorespiratory cycle, were 4.5 and 1.9 times more abundant in *C. spinosa* (Supplemental Table S1). We also determined the extractable activities of PEPC, aspartate aminotransferase (AspAT), NAD-dependent malate dehydrogenase (NAD-MDH), NAD-ME, and alanine aminotransferase (AlaAT). Except for NAD-MDH, significantly higher activities of the enzymes required for the C_4 cycle were measured in

C. gynandra leaf extracts (Supplemental Fig. S1). The metabolite profiling of leaf extracts using gas chromatography-electron impact-time of flight (GC-EI-TOF) and the enzyme activity assays showed that the plants we used for digital gene expression analysis had clear differences in their metabolite profiles and enzyme activities, and these were consistent with functional C_3 and C_4 photosynthesis operating in leaves of *C. spinosa* and *C. gynandra*, respectively.

The Leaf Transcriptomes for Closely Related C_3 and C_4 Species Are Qualitatively Similar

To obtain sequence tags for digital gene expression (DGE) analysis from *C. spinosa* (C_3) and *C. gynandra* (C_4), RNA was isolated from mature leaves of each species and prepared for 454 sequencing. One sequencing run on a Genome Sequencer FLX (GS FLX; Roche) sequencing system was conducted on leaf cDNA isolated from either *C. gynandra* or *C. spinosa*. From *C. spinosa*, we obtained 70,564,592 nucleotides, and from *C. gynandra*, 91,851,136 nucleotides of raw sequence were obtained; after quality control, these corresponded to 65,525,139 and 85,681,233 nucleotides, respectively (Table I). The mean read length of the cleaned sequence reads was 232 nucleotides for *C. gynandra* and 230 nucleotides for *C. spinosa* (Table I).

To exclude program-specific mapping artifacts and to test whether the *C. gynandra* and *C. spinosa* libraries behave robustly during mapping, two different programs, BLAST and BLAT (BLAST-Like Alignment Tool), were used to align the reads to Arabidopsis as the reference genome. To define the most suitable mapping parameters, an array of parameters for mappings in both the DNA and protein space were tested (Table II). Neither the *C. gynandra* nor the *C. spinosa* library mapped well to Arabidopsis cDNAs in the DNA space using BLAT or BLAST, although the differences are more dramatic for BLAT (Table II). In the protein space, however, the proportion of mapped reads increased dramatically. When 75% amino acid sequence identity was required, three-quarters of the reads could be mapped with BLAT, resulting in 1.48 and 1.57 average mappings per read, respectively. Even with the most lenient mapping parameters, the proportion of mapped reads did not exceed 83% with BLAT and 78.8% with BLAST (Table II). In all mapping attempts, the *C. gynandra* and *C. spinosa* read libraries yielded qualitatively similar mapping results, irrespective of mapping program or parameters.

To obtain a stringent yet inclusive mapping, the mapping conducted in protein space at 75% or greater identity with BLAT was chosen, and this mapping file was parsed by in-house scripts to keep only the read match with the highest number of matching bases. For a more lenient mapping, a BLAST mapping at a cutoff of $1e^{-5}$ was chosen and parsed to keep only the best BLAST hit for each read. For each Arabidopsis Genome Initiative (AGI) code, the number of matching reads was counted and the hit count was then transformed to reads per million (RPM) to normalize for the number of reads available for each species. After parsing, the sequenced libraries matched between 50.5% and 55.3% of the genes in the Arabidopsis reference (Supplemental Table S2).

To assess whether the data sets for the two different species and the two different mappings were qualitatively similar, we tested the coverage of the functional classes. Overall, about 50% of all genes were represented in both species with the BLAT (Fig. 1A) and the BLAST mapping (Fig. 1B). Although the majority of gene classes were represented by more than 50% of genes in each class for both mappings, the classes function unknown, putative lipid transfer protein, storage protein, and defense were underrepresented compared with all genes (Fig. 1). Genes present in the organellar genomes were not well represented (Supplemental Table S3). Genes classified into primary metabolism including photosynthesis, central carbon, nitrogen metabolism, amino acid, and nucleotide metabolism as well as many cellular processes were well-represented categories, and about four-fifths of genes predicted to be involved in the C_4 pathway were detected in both species. Overall, the pattern of detection in the different gene classes was similar for both species and independent of the program used for the mapping (Fig. 1).

Table I. *Massively parallel signature sequencing allows large-scale assembly of transcripts in both C. spinosa and C. gynandra after comparison with the TAIR 8 Arabidopsis database*

One GS FLX sequencing run allowed significant generation of sequence for both species, and the vast majority of these could be used to assemble contigs and then matched to Arabidopsis genes.

Data	*C. spinosa*	*C. gynandra*
Raw reads	313,807	402,674
Raw nucleotides	70,564,592	91,851,136
Raw mean length	225	228
Clean reads	284,318	368,333
Clean nucleotides	65,525,139	85,681,233
Clean mean length	230	232
Contigs	17,655	18,992
Total length (nucleotides)	7,746,894	9,062,043
Total reads	245,324	319,732
Percent assembled	86.3	86.8

Transcripts of Known C_4 Genes Are More Abundant with One Exception

Detailed analysis of known C_4 genes showed that all but one gene necessary for the core C_4 cycle of NAD-ME-type plants were massively up-regulated in *C. gynandra* compared with *C. spinosa*. Transcripts encoding PEPC were up-regulated 78-fold, those encoding AspAT were up-regulated 343-fold, the transcripts for the two isoforms of NAD-ME were up-regulated 27- and 21-fold, respectively, and AlaAT were up-regulated 29-fold (Table III). The results for the BLAT

Table II. *Mapping the sequence reads with different BLAT and BLAST parameters to empirically determine suitable mapping conditions*

The percentage of AGI codes with at least one mapped read and the average mappings per read were determined prior to parsing the tables to retain only the best match. Suitable mapping conditions are printed in bold; for BLAT, the cutoff value is the minimal number of matching bases; for BLAST, it is the minimal accepted e-value.

Mapping Program	Library	Search Space	Cutoff Value	Percentage Reads with at Least One Hit in the Reference	Percentage AGI Codes with at Least One Mapped Read	Average Mappings per Read
BLAT	*C. gynandra*	DNA	60	40.9	42.0	1.19
			75	40.7	41.7	1.19
			85	30.2	35.8	1.15
			90	7.7	19.5	1.09
		Protein	25	82.6	70.4	2.35
			50	82.6	70.4	2.35
			75	**75.4**	**62.6**	**1.48**
			80	56.4	52.2	1.27
	C. spinosa	DNA	60	40.8	38.9	1.29
			75	40.6	38.5	1.28
			85	29.7	32.3	1.21
			90	8.5	17.1	1.15
		Protein	25	83.0	67.7	2.49
			50	83.0	67.7	2.46
			75	**76.0**	**58.9**	**1.57**
			80	57.9	48.4	1.32
BLAST	*C. gynandra*	DNA	1e-05:	68.9	56.5	30.7
			1e-10:	58.8	49.1	27.7
			1e-30:	29.6	30.5	18.9
			1e-50:	9.9	15.9	11.5
		Protein	**1e-05:**	**78.0**	**76.9**	**106.6**
			1e-10:	67.8	71.0	64.6
			1e-30:	29.0	39.5	22.9
			1e-50:	0.1	0.3	7.5
	C. spinosa	DNA	1e-05:	69.7	53.0	28.2
			1e-10:	59.6	46.3	25.1
			1e-30:	29.4	28.3	16.2
			1e-50:	9.8	14.4	9.8
		Protein	**1e-05:**	**78.8**	**75.3**	**93.7**
			1e-10:	68.3	68.7	56.4
			1e-30:	29.3	36.0	21.2
			1e-50:	0.1	0.3	4.6

and the BLAST mappings were similar with one exception. In the BLAST mapping, the reads mapping to PEPC were split onto two genes in the Arabidopsis reference genome, whereas they mapped to only one gene in the BLAT mapping (Table III). Transcripts encoding mitochondrial malate dehydrogenases were increased only 1.3-fold (Supplemental Table S3). Not only were genes associated with the C_4 pathway upregulated compared with C_3, but they also had high absolute read counts between 1,800 and 4,806 RPM.

The Leaf Transcriptomes for Closely Related C_3 and C_4 Species Are Quantitatively Different

Before undertaking detailed analysis of differences in transcript abundance between *C. gynandra* and *C. spinosa*, we used quantitative (q)PCR to confirm estimates of transcript abundance identified by RNA-Seq. We chose genes whose transcript abundance differed over 4 orders of magnitude and used qPCR to assess their abundance. qPCR was performed on both the cDNA used for RNA-Seq and cDNA generated from RNA isolated from leaves in a separate experiment. This approach provided strong support for the differences in abundance of transcripts between the two species that we determined from RNA-Seq (Fig. 2). Overall, this showed that the ratios of transcript abundance obtained by RNA-Seq-based DGE are suitable for calling differentially expressed genes between two related species.

Of the 13,662 transcripts for which we captured quantitative data (Supplemental Table S3), we identified 583 (BLAT) or 603 (BLAST) transcripts whose abundance differed significantly ($P \leq 0.01$) between *C. spinosa* and *C. gynandra*, with 256/258 (1.2%/1.2%) transcripts being more abundant in *C. gynandra* and

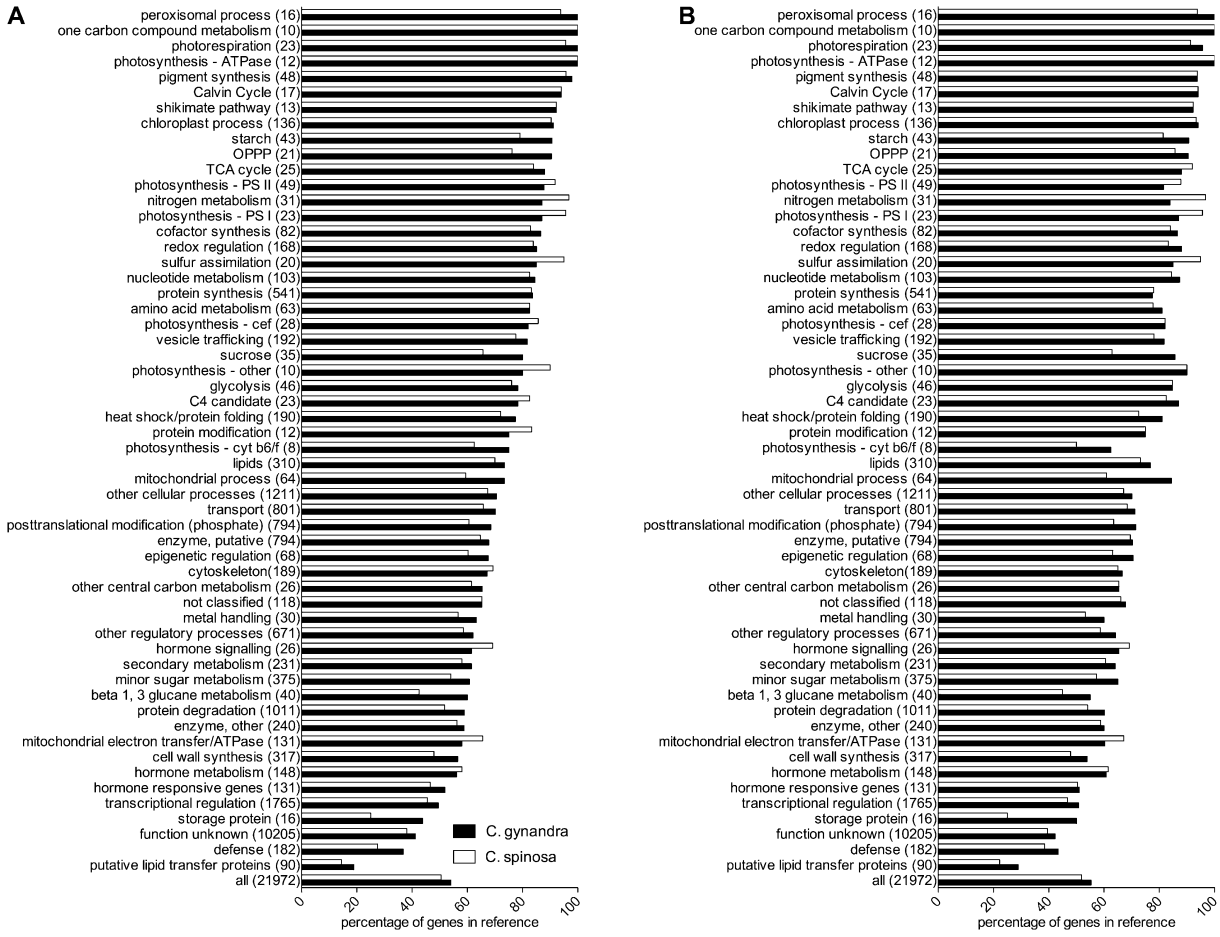


Figure 1. The qualitative patterns of transcript abundance between *C. gynandra* and *C. spinosa* are very similar, with the same classes underrepresented and overrepresented in both libraries. A, Analysis based on BLAT mapping. B, Analysis based on BLAST mapping. Black bars refer to the C_4 plant *C. gynandra*, and white bars refer to the C_3 plant *C. spinosa*.

327/345 (1.5%/1.6%) transcripts being more abundant in *C. spinosa* (Fig. 3, "all"). We tested whether significantly changed transcripts are enriched in functional categories and whether they were more highly expressed in the C_4 or the C_3 species. While the qualitative classification of detected genes showed a very similar pattern between *C. spinosa* and *C. gynandra* (Fig. 1), the quantitative analysis revealed massive differences in representation between gene classes in the C_3 and the C_4 species (Fig. 3). The transcript profile generated by the BLAT mapping (Fig. 3A) is similar to the one generated by the BLAST mapping (Fig. 3B), although not all genes called as significantly regulated were identical (Supplemental Table S3). The classes containing the highest percentage of changed genes are the photosynthetic classes as well as the C_4 cycle, Calvin-Benson cycle, and photorespiration (Fig. 3). The latter two have lower steady-state mRNA levels in C_4 leaf tissue (Fig. 3, bottom), while the photosynthetic classes of PSI, cyclic electron flow, and cytochrome b_6/f complex as well as the C_4 cycle have higher levels in C_4 leaf tissue (Fig. 3, top). A number of classes involved in primary metabolism also have lower steady-state transcript levels in C_4 tissues: one-carbon compound metabolism, other central carbon metabolism, shikimate pathway, and amino acid metabolism. Protein synthesis also has lower steady-state transcript levels, which are limited to cytosolic and plastidic protein synthesis genes (Supplemental Fig. S3). Among the classes with higher steady-state transcript levels are starch metabolism, cofactor synthesis, putative lipid transfer proteins, nitrogen metabolism, and β-1,3 glucan metabolism. The quantitative pattern (Fig. 3) is similar to the qualitative pattern (Fig. 1) with regard to the influence of the mapping program; the BLAT and BLAST mappings look remarkably similar with the exception of shikimate metabolism.

Table III. *Transcript abundance of C_4 cycle genes that have significantly higher transcript abundance in C_4 leaf tissue*
Asterisks denote changes significant only in BLAST mapping.

Enzyme	Locus	BLAT Mapping			BLAST Mapping		
		C. gynandra RPM	*C. spinosa* RPM	Fold Change	*C. gynandra* RPM	*C. spinosa* RPM	Fold Change
AspAT	AT2G30970	4,806	14	343.3	4,601	18	257.9
PPDK	AT4G15530	3,262	14	233.0	3,216	13	240.3
PEPC	AT2G42600	9,702	124	78.2	8,321	169	49.1
AlaAT	AT1G17290	7,610	267	28.5	7,242	259	28.0
NAD-ME1	AT4G00570	1,357	51	26.6	1,326	49	27.0
NAD-ME2	AT2G13560	1,800	87	20.7	1,723	85	20.3
PEPC kinase	AT1G08650	230	37	6.2	226	36	6.3
NADP-ME*	AT1G79750	227	60	3.8	216	45	4.8
PEPC*	AT1G53310	94	248	0.4	950	192	5.0
PPDK regulatory protein*	AT4g21210	148	32	4.6	198	27	7.3

Transcripts with Similar Patterns of Abundance Compared with Bona Fide C_4 Genes and Rubisco

The list of 13,662 transcripts detected in either *C. spinosa* or *C. gynandra* tissues and the list of 603 transcripts that are differentially regulated between both species (Supplemental Table S3, BLAST mapping) prompted us to determine which transcripts showed changes in abundance similar to the core C_4 genes or Rubisco subunit-encoding genes. Such transcripts display both a large fold change between the C_3 and the C_4 plants and large absolute read numbers. For example, among the transcripts encoding putative transport proteins, three plastidic transport proteins, the PEP phosphate translocator PPT, a putative bile acid:sodium symporter, and a putative proton:sodium antiporter, two mitochondrial dicarboxylate carriers, and one plasma membrane intrinsic protein were massively up-regulated in C_4 *C. gynandra* (Table IV). No transcripts encoding transport proteins were found to be down-regulated to a comparable degree. Among metabolic genes, two cytosolic carbonic anhydrases, one of which (CA4; Table IV) is likely tethered to the plasma membrane, an adenylate kinase, and a pyrophosphatase were up-regulated at levels comparable to those of C_4 cycle genes. Many proteins of unknown function showed differential expression, the most striking case being a putative lipid transfer protein, also annotated as an extensin-like protein. Based on annotation and differential expression pattern, several transcripts predicted to encode known C_4 functions that have not yet been assigned to genes, such as CHLOROPLAST UNUSUAL POSITIONING1 (CHUP1) and actin for chloroplast positioning or callose-degrading enzymes for regulating plasmodesmatal opening, were identified (Table IV).

Regulatory Genes That Are Significantly Changed

The transcript profiles of these C_3 and C_4 species identify a number of regulatory proteins that are candidates for maintaining C_4 status. Among transcripts encoding proteins with regulatory functions, 43 were significantly up-regulated in either *C. gynandra* or *C. spinosa* (Fig. 3). These include bona fide transcription factors, protein phosphatases and kinases, and the regulatory proteins of the pyruvate dehydrogenase complex (up-regulated in C_4), of PPDK (up-regulated in C_4), and of Rubisco (down-regulated in C_4). Only 17 transcription factors are significantly changed; seven of those have higher steady-state mRNA levels compared with the C_3 leaf tissue, while 10 have lower steady-state mRNA levels (Table V).

In addition to the detailed quantitative and qualitative analysis of read mappings to generate ESTs for both species, contigs were assembled from cleaned reads for each species as described previously (Weber et al., 2007; Bräutigam et al., 2008b) and then annotated by BLASTX versus The Arabidopsis Information Resource (TAIR) 9 protein models. A total of 18,992 and 17,655 contigs representing total sequence lengths of 9,062,043 and 7,746,894 nucleotides were obtained for *C. gynandra* and *C. spinosa*, respectively (Table I).

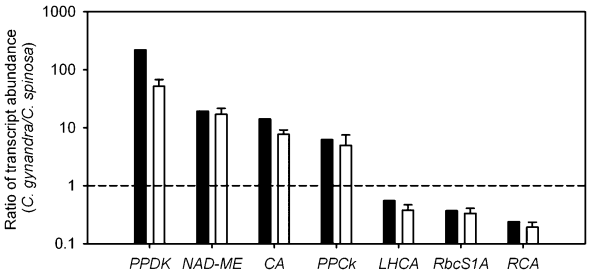


Figure 2. Massively parallel sequencing of mRNAs (RNA-Seq) and qPCR generate similar profiles of transcript abundance in *C. gynandra* and *C. spinosa*. Ratios of transcript abundance in *C. gynandra* and *C. spinosa* were calculated, and transcripts selected for this analysis spanned 4 orders of magnitude. CA, Carbonic anhydrase; PPCk, PEPC kinase; LHCA, light-harvesting complex subunit A; RbcS1a, ribulose bisphosphate carboxylase oxygenase 1a; RCA, Rubisco activase. Black bars represent data from RNA-Seq, and white bars represent data from qPCR. The horizontal dashed line represents a ratio of 1 and indicates no difference in transcript abundance between the two species.

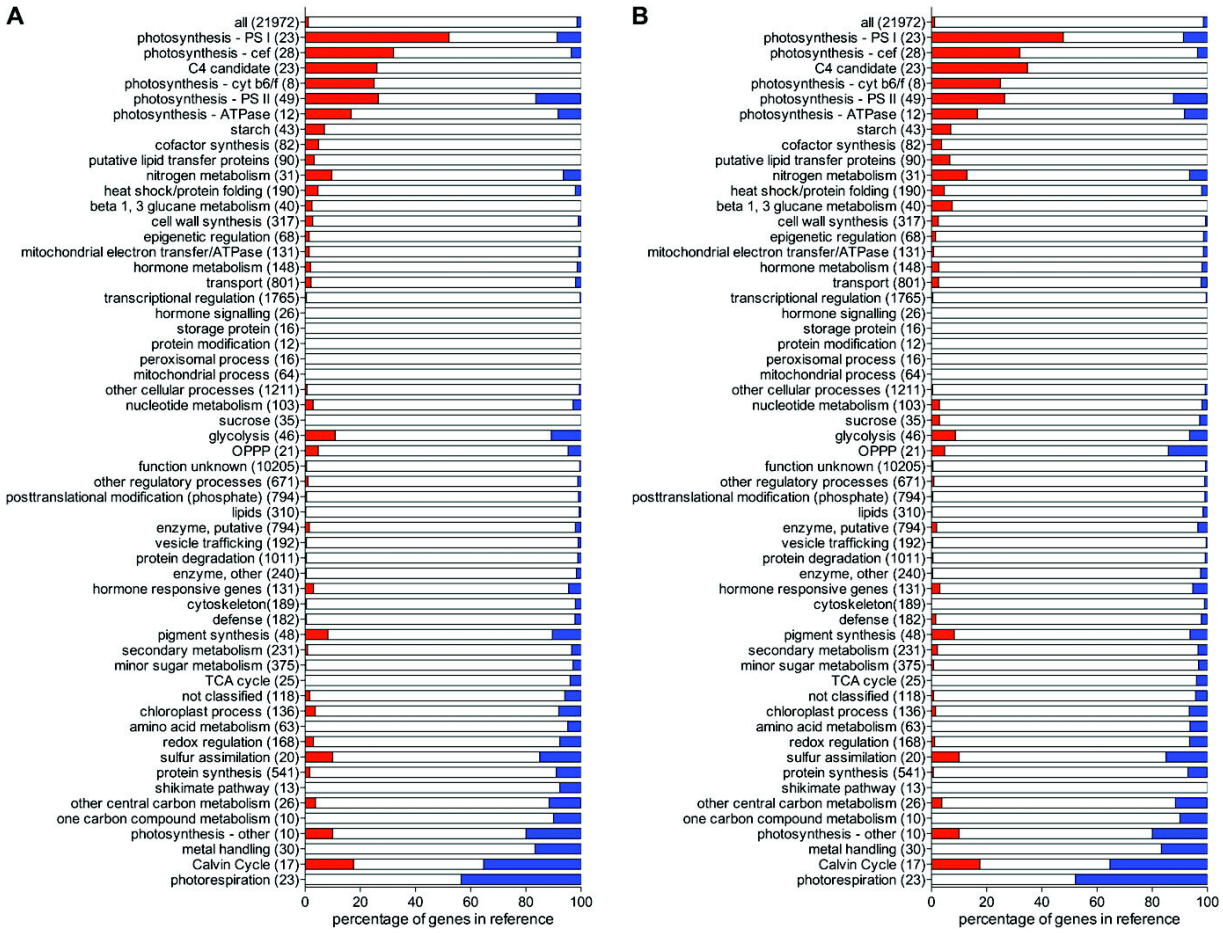


Figure 3. The quantitative patterns of transcript accumulation in *C. gynandra* and *C. spinosa* are distinct. A, Analysis based on BLAT mapping. B, Analysis based on BLAST mapping. Shown are the percentages of genes with significantly higher abundance of transcripts in C_4 (red bars), unchanged (white bars, including genes not detected), and significantly lower abundance of transcripts in C_4 (blue bars) based on the total number of genes in each annotation class (in parentheses on the *y* axis).

DISCUSSION

Transcriptomic Comparisons of Different Species with NGS Technology Are Feasible

Read mapping by alignment is a well-established tool to quantify transcript abundance and thus determine mRNA steady-state levels (Wall et al., 2009; Metzker, 2010). The concept of mapping to a cross-species reference has also been established theoretically (Palmieri and Schlotterer, 2009), although the potential has not been experimentally explored to date (Bräutigam and Gowik, 2010).

To explore cross-species mapping, the transcriptome sequencing was carried out using 454 FLX, a long-read technology, since theoretical work had established that at least BLAT is capable of mapping reads that contain alterations in comparison with the reference if the reads are at least 100 bases long (Palmieri and Schlotterer, 2009). We also established a reference database, which removes the genome history of Arabidopsis as far as it is known (Bowers et al., 2003; Haberer et al., 2004; Thomas et al., 2006). Tandem duplicated genes and segmentally duplicated genes (remnants of the last whole genome duplications) were removed to prevent genome history from interfering with comparative quantitative mapping (Bräutigam and Gowik, 2010).

Both BLAT and BLAST mappings indicate that using a minimal reference does not diminish read mappings (Supplemental Table S4) while avoiding mapping problems based on genome history (Bräutigam and Gowik, 2010). The mappings in protein space allowed more successful read mappings, because protein sequences diverge more slowly than nucleotide sequences. Although the proportion of reads mapped varied with changing mapping parameters (Table II;

Table IV. *Transcript abundance of selected genes with an expression similar to that of C_4 cycle genes and Rubisco*
All changes are significant at $P \leq 0.01$. n/a, Not available.

Function	Locus	Annotation (TAIR 9)	*C. gynandra* RPM	*C. spinosa* RPM	Ratio
Transport proteins					
	AT2G26900	Bile acid:sodium symporter family protein	4,774	55	86.8
	AT2G22500	Mitochondrial dicarboxylate carrier	324	0	n/a
	AT4G24570	Mitochondrial dicarboxylate carrier	148	0	n/a
	AT2G45960	Plasma membrane intrinsic protein subfamily protein	2,686	133	20.2
	AT5G33320	Phospho*enol*pyruvate/phosphate translocator	1,955	97	20.2
	AT1G49810	Member of Na^+/H^+ antiporter family	1,321	83	15.9
Metabolism					
	AT3G52720	α-Carbonic anhydrase 1	227	152	1.5
	AT1G23730	β-Carbonic anhydrase 4	497	87	5.7
	AT5G35170	Adenylate kinase family protein	1,994	235	8.5
	AT5G09650	Inorganic pyrophosphatase	2,664	833	3.2
Proteins of unknown function					
	AT1G12090	Extensin-like protein (ELP)	6,278	147	42.7
Callose-degrading enzymes					
	AT3G57240	Member of glycosyl hydrolase family 17, likely β-1,3 glucanase	436	0	n/a
	AT1G32860	Member of glycosyl hydrolase family 17, likely β-1,3 glucanase	50	0	n/a
	AT5G42100	Plasmodesmal-associated β-1,3-glucanase	173	32	5.4
Cell biology					
	AT3G25690	CHUP1	22	170	0.13
	AT3G12110	ACTIN	122	727	0.2

Supplemental Table S4), the *C. spinosa* and *C. gynandra* libraries yielded similar results, indicating that, evolutionarily, both species are approximately equally distant from Arabidopsis, with mapping incurring similar penalties depending on parameters.

Since no read alignment program has emerged as the consensus program for NGS data analysis, two different programs were used for mapping and the output was compared in all cases. The output proved robust against changing the mapping program both qualitatively and quantitatively. When we mapped the quarter million reads obtained from each species of *Cleome* to a minimized TAIR 9 release of the Arabidopsis genome, they corresponded to approximately 11,000 loci. As the minimized TAIR 9 data set contains 21,972 gene loci, the reads we collected in *C. gynandra* and *C. spinosa* represent approximately 50% of the transcriptome. In Arabidopsis seedlings, approximately 60% of the loci represented in the TAIR 8 release were detectable (Weber et al., 2007); hence, we have likely captured a large proportion of the transcripts associated with leaves of *C. spinosa* and *C. gynandra*.

The qualitative representation of gene classes detected reflects that leaf tissues were analyzed. While photosynthetic genes as well as primary metabolism are well represented in all data sets, genes implicated in cell walls, secondary metabolism, and defense responses are underrepresented (Fig. 1). These classes contain genes that are likely specific to certain tissues, developmental stages, or environmental challenges. For example, cell wall genes may be better represented if our sampling had included expanding leaf or stem material (Schmid et al., 2005), and stress-response genes may be better represented if plants were sampled after exposure to extreme conditions (Kilian et al., 2007).

Likewise, certain pathways of secondary metabolism are likely restricted to defined tissues or developmental stages, making it unlikely that we would pick up many of these genes when profiling leaf libraries. Based on the gene detection pattern, the two plant species did not encounter different biotic or abiotic stresses or were not in different stages of growth, as very similar genes were detected in both species (Figs. 1 and 3).

Finally, only a very small proportion of transcripts showed significant differences in abundance between the two different species (Supplemental Tables S2 and S3), and these changes were enriched in a limited number of functional classes (Fig. 3). We conclude that cross-species mapping in protein space is a feasible strategy to compare different species as long as an equidistant reference is available.

Transcripts Derived from Core C_4 Cycle Genes Are More Abundant in the C_4 Species

C_4 photosynthesis has evolved convergently in many different lineages of plants (Sage, 2004), and in many cases the alterations to expression of specific genes has been related to transcriptional regulation (summarized in Sheen, 1999). Our genome-scale analysis allowed us to compare the steady-state transcript levels for all candidate C_4 genes at the same time. For all of the enzymes where a change in total extractable activity could be shown (Supplemental Fig. S1), a higher mRNA level of at least one isoform as judged from the read count was also present (Table III). The only enzyme showing no changes in transcript level is the mitochondrial NAD-MDH. Possibly, the activity of the mitochondrial NAD-MDH is high enough already

Table V. *Transcription factors that are significantly changed between the leaf tissue samples*
Asterisks denote changes significant only in BLAST mapping. n/a, Not available.

Locus	Transcription Factor Type	BLAT Mapping			BLAST Mapping			Segmentally Duplicated?
		C. gynandra RPM	*C. spinosa* RPM	Ratio	*C. gynandra* RPM	*C. spinosa* RPM	Ratio	
AT1G25560	AP2-EREBP	176	9	19.6	219	9	24.3	Yes
AT5G07580	AP2-EREBP	223	51	4.4	292	36	8.1	Yes
AT1G53910	AP2-EREBP*	32	138	0.2	84	268	0.3	Yes
AT5G10570	bHLH	0	83	n/a	0	112	n/a	Yes
AT3G21330	bHLH*	0	74	n/a	0	107	n/a	
AT3G62420	bZIP	11	138	0.1	10	138	0.1	
AT2G20570	G2-like	220	0	n/a	292	0	n/a	
AT1G72030	GNAT	11	179	0.1	10	330	0.0	
AT2G22430	HB	515	106	4.9	505	116	4.4	Yes
AT1G10200	LIM	22	230	0.1	21	205	0.1	
AT4G30410	Not specified*	0	32	n/a	0	76	n/a	
AT1G32700	PLATZ	176	9	19.6	115	4	28.8	
AT5G02810	Pseudo ARR-B	0	106	n/a	10	112	0.1	
AT2G36990	Sigma70-like	130	0	n/a	143	0	n/a	
AT1G48500	Tify	11	147	0.1	10	174	0.1	Yes
AT1G17380	Tify*	18	110	0.2	24	161	0.1	Yes
AT3G02790	Zinc finger	374	87	4.3	407	112	3.6	Yes

in C_3 plants to support a C_4-type metabolic flux. The only transport protein known to date that is involved in the C_4 cycle, the PEP phosphate translocator (Fischer et al., 1997; Bräutigam et al., 2008a), is also up-regulated 20-fold, indicating that this transport protein is regulated at the level of mRNA abundance. Based on similarities in transcript abundance to known C_4 genes, our comparative RNA-Seq also identified likely additional components needed for C_4 photosynthesis. When PPDK was characterized, it was proposed that adenylate kinase as well as inorganic pyrophosphatase need to be abundant in C_4 chloroplasts (Hatch and Slack, 1968). RNA-Seq confirmed this prediction and showed that the up-regulation also occurs at the level of transcript abundance. Taken together, we found that almost all transcripts encoding the proteins required for the core C_4 cycle have higher steady-state mRNA levels, and we propose that, at least in *C. gynandra*, the activity of C_4 cycle enzymes and transport proteins is controlled at least partially at the level of transcript abundance.

Alterations to the Abundance of Transcripts Associated with Other Metabolic Processes

Changes in the abundance of transcripts that are not associated with the core C_4 cycle are also detectable in leaves of *C. gynandra* and *C. spinosa*. The high-flux C_4 cycle poses additional demands in terms of ATP and reduction equivalents on the light reaction (Hatch, 1987). Specifically, the recycling of the initial CO_2 acceptor PEP requires additional ATP molecules (Hatch, 1987). In C_4 leaf tissue, one-third to one-half of the genes in the photosynthetic gene classes that contribute to ATP production by cyclic electron flow are up-regulated compared with C_3 leaf tissue: PSI, the cytochrome b_6/f complex, and the genes mediating cyclic electron flow themselves (Fig. 3). It remains an open question whether these higher steady-state levels are caused by higher ATP demand or whether C_4 photosynthesis requires up-regulation of these genes to meet the ATP demand prior to establishing C_4 photosynthesis.

On the other hand, the classes of Calvin-Benson cycle genes and photorespiratory genes are those with the highest number of genes with significantly lower steady-state mRNA levels. It is a well-established fact that most C_4 plants have less Rubisco protein compared with C_3 plants (Ku et al., 1979) and that flux through the photorespiratory pathway is reduced compared with C_3 species (Chollet and Ogren, 1975; Leegood, 2002). Transcripts encoding the large and small subunits of Rubisco were reduced from 22,968 and 15,442 RPM to 6,984 and 4,900 RPM in *C. spinosa* and *C. gynandra*, respectively. Overall, the trend for Calvin-Benson cycle genes was for them to be down-regulated in *C. gynandra* compared with *C. spinosa* (Fig. 3). Likewise, a large number of genes encoding photorespiratory proteins, proteins involved in one-carbon compound metabolism, and the genes involved in ammonia reassimilation, Gln synthetase, and Glu synthase have lower steady-state transcriptional levels (Fig. 3; Supplemental Table S3). The reduced flow through the photorespiratory pathway obviously decreases the demand on the expression system to maintain high steady-state levels of mRNA for many Calvin-Benson cycle and photorespiratory genes. The photosynthetic genes, the Calvin-Benson cycle and photorespiratory genes (in C_3), and the C_4 cycle genes (in C_4) are those with the highest read counts of the genes with known function (Supplemental Table S3). Although it is currently not possible to quantify

absolute transcript levels, since the genome of neither *Cleome* species has been sequenced, the high read counts obtained for the genes of central carbon metabolism and photosynthesis indicate that the steady-state levels of transcripts are high. Since the most altered gene classes are also those that contain the genes with the highest absolute read counts, it is not clear whether C_4 photosynthesis lowers or raises the demand on protein synthesis and accessory pathways such as amino acid synthesis. However, both the protein synthesis and the amino acid metabolism classes contain more genes that have lower steady-state levels in C_4 leaf tissue (Fig. 3). Within the protein synthesis gene class, many transcripts encoding structural components of plastidic and cytosolic ribosomes were reduced (Supplemental Fig. S3). This was not the case for components of mitochondrial ribosomes (Supplemental Fig. S3), indicating that there is not a general effect on translation but that the effect is likely specific to ribosomes involved in translation for the Calvin-Benson cycle and photorespiration. The protein-to-fresh weight ratio is also lower in C_4 leaf tissue compared with C_3 leaf tissue (Supplemental Fig. S2). We propose that plastidic ribosomes are relieved of the high translation load associated with the large subunit of Rubisco and that the cytosolic ribosomes need to translate fewer transcripts associated with central carbon metabolism as well as the small subunit of Rubisco. The reduced production of proteins in the leaves of C_4 plants is considered important in increasing nitrogen use efficiency, because the rate of photosynthesis per unit of nitrogen in the leaf is increased (Oaks, 1994). Our data indicate that there is also likely a significant saving in the nitrogen provision in the leaf, because fewer ribosomes as well as fewer proteins for central carbon metabolism are required.

The data set contains two additional gene classes, β-1,3 glucan metabolism and putative lipid transfer proteins, that showed differences in transcript abundance between *C. gynandra* and *C. spinosa* that could be explained within the current framework of knowledge of C_4 photosynthesis. The C_4 pathway requires efficient exchange of metabolites between M and BS cells via large numbers of plasmodesmata connecting both cell types, while the BS cell wall of many C_4 plants is suberized to reduce diffusion of CO_2 away from Rubisco (Hatch, 1987). Transcripts encoding three distinct glucan 1,3-β-glucosidases (Table IV) involved in governing plasmodesmatal conductivity by regulating the turnover of the β-1,3-glucan callose (Levy et al., 2007) were up-regulated in leaves of *C. gynandra* compared with *C. spinosa*. Therefore, it is possible that these genes are involved in increasing the open probability of plasmodesmata (Roberts and Oparka, 2003), which allows the efficient flux of organic acids between M and BS cells required during C_4 photosynthesis (Evert et al., 1977; Botha, 1992; Roberts and Oparka, 2003). A transcript annotated as a putative lipid transfer protein is among those that are most highly up-regulated in *C. gynandra* compared with *C. spinosa*. Lipid transfer proteins are required for the export of lipids to the cell wall during cutin biosynthesis (DeBono et al., 2009). Interestingly, in Arabidopsis, some lipid transfer proteins are exclusively and abundantly expressed in the root endodermis, where suberin biosynthesis is required to establish the Casparian strip.

There are additional changes in the transcript profile that are less easily explained. Among the gene classes containing more genes with significantly higher transcript levels in C_4 leaf tissue are starch metabolism, cofactor synthesis and nitrogen metabolism, and heat shock/protein folding (in order of decreasing number of significantly different genes). On the other hand, it is difficult to conceive why genes involved in metal handling are frequently lower in transcript level in C_4 leaf tissues (Fig. 3). These changes may be connected to currently unknown phenomena relating to the C_4 pathway or may be part of differences not relating to C_4 photosynthesis between the two species. Overall, the global analysis of transcription on the level of functional classes reveals unexpected shifts in transcript profiles that can be explained based on the current knowledge about the C_4 pathway, while a range of smaller changes remain enigmatic.

Finally, our global transcriptional analysis of C_4 and C_3 leaf tissues not only allows testing hypotheses about the C_4 pathway on a global scale but also allows genes with expression patterns similar to those of known C_4 genes to be identified. The phylogenetic proximity of the Cleomaceae to Arabidopsis allows the identification of the orthologs in Arabidopsis, which will facilitate translational research into the model species (Brown et al., 2005).

Candidates for Additional C_4-Related Genes

The identification of transport proteins involved in the C_4 cycle lags behind that of enzymes, considering that the C_4 cycle requires the intracellular transport of pyruvate, PEP, Asp, and Ala across different organellar membranes (Bräutigam and Weber, 2011). A wide range of C_4 plants take up pyruvate into chloroplasts from the M in cotransport with sodium (Aoki et al., 1994; Aoki and Kanai, 1997), which might explain the requirement for sodium as a micronutrient in many C_4 species (Brownell and Crossland, 1972). Since the rate of pyruvate transport into C_4 M cell chloroplasts occurs at or exceeds the apparent rate of CO_2 assimilation, sodium-coupled pyruvate import implies a large influx of sodium into these chloroplasts, but the transporter has not yet been identified at the molecular level (Aoki and Kanai, 1997). Our finding that a putative plastidic proton:sodium symporter (NHD1) is 16-fold up-regulated in *C. gynandra* prompts us to hypothesize that it functions in exporting sodium from the chloroplast in order to maintain the sodium gradient required for import of pyruvate. In addition, we found strong up-regulation of a putative bile acid: sodium cotransporter in *C. gynandra*. Interestingly, up-

Table VI. *Comparison of alterations in transcript abundance in C_4 and C_3 leaves with those induced by cold, sugar feeding, attack by pests or pathogens, diurnal changes to light, or circadian rhythms*

Cause	Estimated Change in Transcriptome	Change	Reference
		%	
Cold treatment	514 (24,000) ATH1	2.1	Vogel et al. (2005)
C_4 leaves and C_3 leaves	583/603 (13,443/13,662)	2.7/2.8	This study
Glc feeding	978 (22,500) ATH1	4.4	Price et al. (2004)
Pseudomonas syringae	2,034 (23,750) ATH1	8.6	De Vos et al. (2005)
Myzus persicae	2,181(23,750) ATH1	9.1	De Vos et al. (2005)
Diurnal regulation	1,115 (11,521) cDNA array	11	Schaffer et al. (2001)
Circadian regulation	2,282 (18,890) Galbraith	12	Dodd et al. (2007)

regulation of the putative bile acid:sodium cotransporter or of NHD1 was not observed in maize (*Zea mays*; Bräutigam et al., 2008a), which belongs to a group of C_4 plants that show proton-dependent, not sodium-dependent, transport of pyruvate into M cell chloroplasts (Aoki et al., 1994; Aoki and Kanai, 1997). PEP generated from pyruvate in M cell chloroplasts is exported from these chloroplasts by PPT, thereby providing the substrate for the cytosolic PEPC reaction. Accordingly, transcripts encoding PPT are 20-fold up-regulated in *C. gynandra*, likely reflecting the increased requirement for transport of PEP (Table III). In contrast to what has been observed for the NADP-ME-type C_4 plant maize by quantitative proteomic analysis (Bräutigam et al., 2008a), we did not detect increased transcript abundance of the putative M chloroplast oxaloacetate/malate exchanger DiT1 (Taniguchi et al., 2002, 2004; Renne et al., 2003; Supplemental Table S3). This is consistent with the fact that oxaloacetic acid/malate shuttling across the M cell chloroplast envelope membrane is not required for NAD-ME-type C_4 photosynthesis (Weber and von Caemmerer, 2010; Bräutigam and Weber, 2011). The mitochondrial dicarboxylate carriers are prime suspects for the C_4 acid importer into the mitochondria, where decarboxylation takes place (Table IV). The initial uptake of inorganic carbon and its conversion to bicarbonate may be facilitated by the concerted action of a membrane intrinsic protein channeling the gas and a carbonic anhydrase that is predicted to be membrane bound (Table IV).

Chloroplasts in the BS of *C. gynandra* are larger than those in the BS of C_3 species and, as in many other C_4 plants, are positioned in a strictly centripetal pattern (Marshall et al., 2007; Voznesenskaya et al., 2007). Transcripts derived from the *GIANT CHLOROPLAST1* (*GC1*) gene were more abundant in *C. gynandra* than in *C. spinosa* (Table IV). Although overexpression of *GC1* in Arabidopsis is reported not to effect chloroplast division (Maple et al., 2004), it is possible that it does so in *C. gynandra*. In addition, we also detected reduced accumulation of transcripts derived from the *CHUP1* and *ACTIN11* genes. In Arabidopsis, the outer chloroplast envelope membrane protein CHUP1 contains an actin-binding motif and is required for preventing chloroplast aggregation (Oikawa et al., 2003). Differential positioning of chloroplasts in BS and M cells of the C_4 plants finger millet (*Eleusine coracana*) and maize requires the actomyosin system (Kobayashi et al., 2009). Since AtCHUP1 is involved in positioning chloroplasts at the periclinal plasma membrane during the weak-light acclimation response via a coiled-coil domain and interaction with the cytoskeleton (Oikawa et al., 2003), it is possible that the centripetal positioning of chloroplasts in BS cells is linked to lower expression of the *CgCHUP1* and *ACTIN11* genes.

Controlling and Maintaining a C_4 State in Leaf Tissue

Our estimate that around 603 transcripts accumulate differentially in leaves of C_3 and C_4 species provides insight into the extent to which gene expression profiles change in C_4 leaves. For example, the fact that 258 transcripts were more abundant in the leaves of C_4 compared with C_3 species indicates that about 2.8% of the leaf transcriptome differentially accumulates in C_4 leaves (Supplemental Tables S2 and S6). To compare the complexity of the C_4 pathway with other multigenic traits, we assessed the number of transcripts that are known to be regulated by sugars, cold, diurnal and

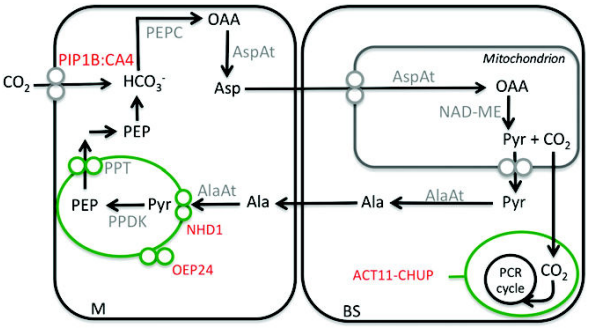


Figure 4. Schematic of components associated with the C_4 cycle in the NAD-ME subtype based on interpretation of RNA-Seq. Proteins that have been described previously are in gray, and novel proteins are marked in red. Metabolites are in black. PIP1B:CA4, PIP1B plasma membrane aquaporin:membrane-tethered carbonic anhydrase; OAA, oxaloacetic acid; ACT11-CHUP11, ACTIN11-CHUP1 complex; Pyr, pyruvate; OEP24, chloroplast outer envelope protein 24.

circadian rhythms, as well as attack by pests and pathogens (Table VI). Interestingly, the alterations in transcript abundance of leaves of *C. gynandra* compared with those of *C. spinosa* were greater than those observed in response to cold treatment and lower than those induced by Glc feeding, those occurring during pathogen attack, and the response to both diurnal and circadian rhythms. As significant progress has been made in understanding sugar signaling (Rolland et al., 2006), pathogen attack (Wise et al., 2007), and the control of gene expression in response to the diurnal cycle and circadian rhythms (Imaizumi et al., 2007), it should be possible to identify the regulators responsible for these alterations in transcript abundance in a C_4 leaf compared with a C_3 leaf. The changes in transcript abundance that we document in a C_4 leaf compared with a C_3 leaf likely overrepresent the changes in transcript abundance actually associated with C_4 photosynthesis on a whole leaf basis, as some differences in gene expression are likely due to the phylogenetic distance between *C. gynandra* and *C. spinosa*. A more confident estimate of the extent to which the leaf transcriptome is altered in association with C_4 photosynthesis will be generated when additional congeneric pairs of C_3 and C_4 species are subjected to deep transcriptome analysis and shared transcripts are identified. Between M and BS cells, the alterations in gene expression may be greater than those that we have defined for whole leaves. For example, up to 18% of genes are estimated to be differentially expressed between M and BS cells of maize (Sawers et al., 2007). However, it is not clear how different the transcript profiles of M and BS cells are in a dicot C_3 leaf, and until this is defined, it is not possible to infer the extent to which transcript abundance alters in these cell types in association with C_4 photosynthesis.

As we sampled from mature leaves to capture the differences between C_3 and C_4 leaves at the point of fully differentiated pathways, we likely also captured regulatory genes needed to maintain C_4 architecture and metabolism in mature leaves. Of the 17 transcription factors significantly altered (Table V), *GOLDEN2-LIKE1* (*GLK1*) has previously been implicated in regulating genes important in C_4 photosynthesis. In maize, GOLDEN2 controls functional differentiation of chloroplasts in BS cells (Langdale and Kidner, 1994), and GLK1 has been implicated in the expression of photosynthesis genes in M cells (Rossini et al., 2001). The fact that *GLK1* transcripts are significantly more abundant in leaves of *C. gynandra* would not necessarily be predicted, as previous work indicates that it becomes specialized in BS cells of C_4 leaves but not that its abundance is altered significantly. This implies that the increase in abundance of *GLK1* transcripts may not simply be due to its involvement in C_4 photosynthesis. When overexpression of *GLK1* was induced in Arabidopsis, the abundance of 114 transcripts was altered (Waters et al., 2009). We assessed the extent to which the genes that are controlled by *GLK1* change in abundance in leaves of *C. gynandra* compared with *C. spinosa* and found that only 19 genes were shared between the two data sets. This may be due to a number of factors that could include the following: that there are differences in the targets of GLK1 in Arabidopsis and *C. gynandra*; that a number of other transcriptional regulators are more important than GLK1 in maintaining patterns of photosynthesis gene expression in *C. gynandra*; and that a rapid induction of *GLK1* gene expression has more impact than increasing the steady-state level of *GLK1*. This analysis is also subject to the caveat that in neither case was the amount of GLK1 protein measured.

In all of our analyses, differences in transcript abundance between the leaves of *C. gynandra* and *C. spinosa* may reflect the operation of the C_4 and C_3 photosynthetic pathways; alternatively, they may be due to differences in metabolism and cell biology associated with the phylogenetic distance between the two species. However, in many cases, it is striking that our analysis has identified differences in the abundance of transcripts derived from genes that have been documented to be involved in processes known to alter in a C_4 leaf. Taken together, the analysis allows us to significantly extend the number of C_4-related genes controlled at the level of transcript abundance and to extend the current model for C_4-related processes in NAD-ME C_4 plants (Fig. 4). Analysis of additional pairs of C_3 and C_4 species will likely facilitate the identification of genes specifically involved in the C_4 pathway and exclude genes that are modified for other reasons.

MATERIALS AND METHODS

Plant Material and 454 Sequencing

Cleome spinosa and *Cleome gynandra* plants for transcript profiling by RNA-Seq were grown in standard potting mix in a glasshouse in August and September 2007. To obtain sequence tags for DGE analysis from *C. spinosa* and *C. gynandra*, total RNAs were isolated from fully expanded leaves sampled from 56-d-old plants of each species. mRNA was reverse transcribed to cDNA after two consecutive rounds of oligo(dT) purification and prepared for 454 sequencing as described previously (Weber et al., 2007).

Mapping and Quantification of the Sequence Reads

Evolution did not stop in the lineage to the reference genome of Arabidopsis (*Arabidopsis thaliana*) after the Cleomaceae branch diverged. Hence, there may be genes that were tandem duplicated or retained after the whole genome duplication event of the Brassicaceae that are absent in either of the Cleomaceae species (Bräutigam and Gowik, 2010). To avoid mapping problems such as splitting of reads or mapping errors due to differential retention of genes in either Cleomaceae or Arabidopsis, we created a minimal genome for mapping. The remnants of the last whole genome duplication in the lineage of the Brassicaceae (Bowers et al., 2003; Thomas et al., 2006) and the tandem duplicated genes (Haberer et al., 2004) were reduced to one representative for each based on the TAIR 9 coding sequence set. In each case, the gene with the lowest AGI code was retained for mapping. For each gene, the Supplemental Data store whether there are duplicates and which duplicates match the gene (Supplemental Tables S3 and S5). We recommend recovery of the associated duplicated genes followed by a detailed analysis with phylogenetic trees to define the true ortholog when translating the results of Cleomaceae analyses to Arabidopsis research.

The 454 sequence reads were mapped onto coding sequences of the minimalized TAIR 9 genome by BLAT (Kent, 2002) and BLAST (Altschul et al.,

1997) with varying parameters, and the output was parsed with in-house PERL scripts to retain only the best matching AGI codes for each sequence read and the best BLAST hit, respectively. Differentially expressed transcripts were identified using the Poisson statistics developed by Audic and Claverie (1997) followed by a Bonferroni correction to account for the accumulation of α-type errors when conducting multiple pair-wise comparisons (Audic and Claverie, 1997).

Plant Material and qPCR Analysis

Both species were grown in a growth chamber in long-day conditions (16 h of light/8 h of dark) under 350 μmol photons m^{-2} s^{-1}, at 22°C, and 65% relative humidity prior to samples being taken for qPCR. qPCR was conducted on the same samples used for RNA-Seq and also on mature leaves collected at noon grown in the growth cabinet. For qPCR, RNA was isolated using TriPure reagent (Roche Applied Science). RNA was treated with DNase I (Promega) and purified with the RNeasy Mini Kit (Qiagen). First-strand cDNA was then synthesized with SuperScriptII reverse transcriptase (Invitrogen) using 4 μg of RNA and oligo(dT) primers (Roche Applied Science). Quantitative reverse transcription-PCR was carried out with 96-well plates using a DNA Engine thermal cycler, Chromo4 real-time detector (Bio-Rad), SYBR Green JumpStart Taq Ready Mix (Sigma), and 15-fold dilution of the cDNA as a template. Initial denaturation was carried out at 94°C for 2 min, followed by 40 cycles of 94°C for 20 s, 60°C for 30 s, 72°C for 30 s, and 75°C for 5 s. Primers were designed to have melting temperatures of 60°C ± 0.5°C and to produce amplicons of 91 to 189 bp. The specificity of the primers and lack of primer dimers in the PCR were verified using agarose gel electrophoresis and melting curve analysis. For each product, the threshold cycle CT, where the amplification reaction enters the exponential phase, was determined for three technical replicates and four independent biological replicates per species. The comparative $2^{-\Delta\Delta CT}$ method was used to quantify relative abundance of transcripts (Livak and Schmittgen, 2001). *ACTIN7* was chosen as a reference because the 454 sequencing data showed equal, intermediate levels of *ACTIN7* transcripts in both species. For the qPCR, SE values were calculated from $2^{-\Delta\Delta CT}$ values of each combination of biological replicates.

Polar Metabolite, Chlorophyll, Protein, and Enzyme Activity Analyses

For metabolite analysis, mature leaves from 56-d-old plants were collected in the middle of the light period and immediately frozen in liquid nitrogen. Three independent biological replicates were used. The tissues were ground in a mortar, and a 50-mg fresh weight aliquot was extracted using the procedure described by Lee and Fiehn (2008). Ribitol was used as an internal standard for data normalization. For GC-EI-TOF analysis, samples were processed and analyzed according to Lee and Fiehn (2008). Enzyme activities, chlorophyll, and protein content were determined according to Hausler et al. (2001).

The *Cleome* read data have been submitted to the National Center for Biotechnology Information Short Read Archive: *C. spinosa* = SRS002743.1 and *C. gynandra* = SRS002744.2.

Supplemental Data

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

Supplemental Figure S1. Quantitation of marker enzyme activities in leaf extracts of *C. spinosa* and *C. gynandra*.

Supplemental Figure S2. Protein-to-fresh weight and protein-to-chlorophyll ratios in leaves of *C. gynandra* and *C. spinosa*.

Supplemental Figure S3. Changes in transcript abundance for ribosomal proteins.

Supplemental Table S1. Relative abundance of predominant metabolites detected by GC-EI-TOF in *C. gynandra* and in *C. spinosa*.

Supplemental Table S2. Number of gene loci and number of differentially expressed genes detected with BLAT and BLAST.

Supplemental Table S3. Quantitative information for all reads mapped onto the reference genome from Arabidopsis.

Supplemental Table S4. Comparison of mapping parameters.

Supplemental Table S5. Segmental and tandem duplicates in the Arabidopsis genome.

ACKNOWLEDGMENTS

We thank Tom Hardcastle for help with R.

Received May 18, 2010; accepted June 9, 2010; published June 11, 2010.

LITERATURE CITED

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Aoki N, Kanai R (1997) Reappraisal of the role of sodium in the light-dependent active transport of pyruvate into mesophyll chloroplasts of C_4 plants. Plant Cell Physiol **38:** 1217–1225

Aoki N, Ohnishi JI, Kanai R (1994) Proton/pyruvate cotransport into mesophyll chloroplasts of C-4 plants. Plant Cell Physiol **35:** 801–806

Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res **7:** 986–995

Black CC (1973) Photosynthetic carbon fixation in relation to net CO_2 uptake. Annu Rev Plant Physiol Plant Mol Biol **24:** 253–286

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

Bowers JE, Chapman BA, Rong JK, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433–438

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

Bräutigam A, Hofmann-Benning S, Weber APM (2008a) Comparative proteomics of chloroplast envelopes from C_3 and C_4 plants reveals specific adaptations of the plastid envelope to C_4 photosynthesis and candidate proteins required for maintaining C_4 metabolite fluxes. Plant Physiol **148:** 568–579

Bräutigam A, Shrestha RP, Whitten D, Wilkerson CG, Carr KM, Froehlich JE, Weber APM (2008b) Comparison of the use of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. J Biotechnol **136:** 44–53

Bräutigam A, Weber APM (2011) Transport processes: connecting the reactions of C_4 photosynthesis. *In* AS Raghavendra, RF Sage, eds, C_4 Photosynthesis and Related CO_2 Concentrating Mechanisms. Advances in Photosynthesis and Respiration, Vol 32. Springer, Berlin, pp 199–219

Brown NJ, Parsley K, Hibberd JM (2005) The future of C_4 research: maize, Flaveria or Cleome? Trends Plant Sci **10:** 215–221

Brown RH (1999) Agronomic implications of C_4 photosynthesis. *In* RF Sage, RK Monson, eds, C_4 Plant Biology. Academic Press, San Diego, pp 473–507

Brownell PF, Crossland CJ (1972) The requirement for sodium as a micronutrient by species having the C_4 dicarboxylic photosynthetic pathway. Plant Physiol **49:** 794–797

Chollet R, Ogren WL (1975) Regulation of photorespiration in C_3 and C_4 species. Bot Rev **41:** 137–179

DeBono A, Yeats TH, Rose JKC, Bird D, Jetter R, Kunst L, Samuelsa L (2009) *Arabidopsis* LTPG is a glycosylphosphatidylinositol-anchored lipid transfer protein required for export of lipids to the plant surface. Plant Cell **21:** 1230–1238

De Vos M, Van Oosten VR, Van Poecke RM, Van Pelt JA, Pozo MJ, Mueller MJ, Buchala AJ, Metraux JP, Van Loon LC, Dicke M, et al (2005) Signal signature and transcriptome changes of Arabidopsis during pathogen and insect attack. Mol Plant Microbe Interact **18:** 923–937

Dodd AN, Gardner MJ, Hotta CT, Hubbard KE, Dalchau N, Love J, Assie JM, Robertson FC, Jakobsen MK, Goncalves J, et al (2007) The Arabidopsis circadian clock incorporates a cADPR-based feedback loop. Science **318:** 1789–1792

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

Fischer K, Kammerer B, Gutensohn M, Arbinger B, Weber A, Hausler RE, Flugge UI (1997) A new class of plastidic phosphate translocators: a putative link between primary and secondary metabolism by the phosphoenolpyruvate/phosphate antiporter. Plant Cell **9:** 453–462

Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods **6:** S6–S12

Haberer G, Hindemitt T, Meyers BC, Mayer KFX (2004) Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. Plant Physiol **136:** 3009–3022

Hatch MD (1987) C_4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. Biochim Biophys Acta **895:** 81–106

Hatch MD, Slack CR (1968) A new enzyme for interconversion of pyruvate and phosphopyruvate and its role in C_4 dicarboxylic acid pathway of photosynthesis. Biochem J **106:** 141–147

Hausler RE, Rademacher T, Li J, Lipka V, Fischer KL, Schubert S, Kreuzaler F, Hirsch HJ (2001) Single and double overexpression of C_4-cycle genes had differential effects on the pattern of endogenous enzymes, attenuation of photorespiration and on contents of UV protectants in transgenic potato and tobacco plants. J Exp Bot **52:** 1785–1803

Hibberd JM, Sheehy JE, Langdale JA (2008) Using C_4 photosynthesis to increase the yield of rice: rationale and feasibility. Curr Opin Plant Biol **11:** 228–231

Imaizumi T, Kay SA, Schroeder JI (2007) Circadian rhythms: daily watch on metabolism. Science **318:** 1730–1731

Jordan DB, Ogren WL (1984) The CO_2/O_2 specificity of ribulose 1,5-bisphosphate carboxylase oxygenase: dependence on ribulosebisphosphate concentration, pH and temperature. Planta **161:** 308–313

Keeley JE (1998) C_4 photosynthetic modifications in the evolutionary transition from land to water in aquatic grasses. Oecologia **116:** 85–97

Kent WJ (2002) BLAT: the BLAST-Like Alignment Tool. Genome Res **12:** 656–664

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J **50:** 347–363

Kobayashi H, Yamada M, Taniguchi M, Kawasaki M, Sugiyama T, Miyake H (2009) Differential positioning of C_4 mesophyll and bundle sheath chloroplasts: recovery of chloroplast positioning requires the actomyosin system. Plant Cell Physiol **50:** 129–140

Ku MSB, Schmitt MR, Edwards GE (1979) Quantitative determination of ribulose bisphosphate carboxylase oxygenase protein in leaves of several C_3 and C_4 plants. J Exp Bot **114:** 89–98

Langdale JA, Kidner CA (1994) Bundle-sheath defective, a mutation that disrupts cellular differentiation in maize leaves. Development **120:** 673–681

Lee DY, Fiehn O (2008) High quality metabolomic data for *Chlamydomonas reinhardtii*. Plant Methods **4:** 7

Leegood RC (2002) C_4 photosynthesis: principles of CO_2 concentration and prospects for its introduction into C_3 plants. J Exp Bot **53:** 581–590

Levy A, Erlanger M, Rosenthal M, Epel BL (2007) A plasmodesmata-associated beta-1,3-glucanase in Arabidopsis. Plant J **49:** 669–682

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. Methods **25:** 402–408

Majeran W, van Wijk KJ (2009) Cell-type-specific differentiation of chloroplasts in C_4 plants. Trends Plant Sci **14:** 100–109

Maple J, Fujiwara MT, Kitahata N, Lawson T, Baker NR, Yoshida S, Moller SG (2004) GIANT CHLOROPLAST 1 is essential for correct plastid division in Arabidopsis. Curr Biol **14:** 776–781

Marshall DM, Muhaidat R, Brown NJ, Liu Z, Stanley S, Griffiths H, Sage RF, Hibberd JM (2007) Cleome, a genus closely related to Arabidopsis, contains species spanning a developmental progression from C_3 to C_4 photosynthesis. Plant J **51:** 886–896

Matsuoka M, Furbank RT, Fukuyama H, Miyao M (2001) Molecular engineering of C_4 photosynthesis. Annu Rev Plant Physiol Plant Mol Biol **52:** 297–314

Metzker ML (2010) Applications of next-generation sequencing technologies: the next generation. Nat Rev Genet **11:** 31–46

Mitchell PL, Sheehy JE (2006) Supercharging rice photosynthesis to increase yield. New Phytol **171:** 688–693

Oaks A (1994) Efficiency of nitrogen utilization in C_3 and C_4 cereals. Plant Physiol **106:** 407–414

Oikawa K, Kasahara M, Kiyosue T, Kagawa T, Suetsugu N, Takahashi F, Kanegae T, Niwa Y, Kadota A, Wada M (2003) CHLOROPLAST UNUSUAL POSITIONING1 is essential for proper chloroplast positioning. Plant Cell **15:** 2805–2815

Osborne CP, Freckleton RP (2009) Ecological selection pressures for C_4 photosynthesis in the grasses. Proc R Soc Lond B Biol Sci **276:** 1753–1760

Palmieri N, Schlotterer C (2009) Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. PLoS ONE **4:** 10

Price J, Laxmi A, St Martin SK, Jang JC (2004) Global transcription profiling reveals multiple sugar signal transduction mechanisms in *Arabidopsis*. Plant Cell **16:** 2128–2150

Reiskind JB, Berg RH, Salvucci ME, Bowes G (1989) Immunogold localization of primary carboxylases in leaves of aquatic and a C_3-C_4 intermediate species. Plant Sci **61:** 43–52

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. Plant J **35:** 316–331

Roberts AG, Oparka KJ (2003) Plasmodesmata and the control of symplastic transport. Plant Cell Environ **26:** 103–124

Rolland F, Baena-Gonzalez E, Sheen J (2006) Sugar sensing and signaling in plants: conserved and novel mechanisms. Annu Rev Plant Biol **57:** 675–709

Rossini L, Cribb L, Martin DJ, Langdale JA (2001) The maize *Golden2* gene defines a novel class of transcriptional regulators in plants. Plant Cell **13:** 1231–1244

Sage RF (2004) The evolution of C_4 photosynthesis. New Phytol **161:** 341–370

Sawers RJH, Liu P, Anufrikova K, Hwang JTG, Brutnell TP (2007) A multi-treatment experimental system to examine photosynthetic differentiation in the maize leaf. BMC Genomics **8:** 12

Schaffer R, Landgraf J, Accerbi M, Simon V, Larson M, Wisman E (2001) Microarray analysis of diurnal and circadian-regulated genes in *Arabidopsis*. Plant Cell **13:** 113–123

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet **37:** 501–506

Sheen J (1999) C_4 gene expression. Annu Rev Plant Physiol Plant Mol Biol **50:** 187–217

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36:** D1009–D1014

Taniguchi M, Taniguchi Y, Kawasaki M, Takeda S, Kato T, Sato S, Tahata S, Miyake H, Sugiyama T (2002) Identifying and characterizing plastidic 2-oxoglutarate/malate and dicarboxylate transporters in *Arabidopsis thaliana*. Plant Cell Physiol **43:** 706–717

Taniguchi Y, Nagasaki J, Kawasaki M, Miyake H, Sugiyama T, Taniguchi M (2004) Differentiation of dicarboxylate transporters in mesophyll and bundle sheath chloroplasts of maize. Plant Cell Physiol **45:** 187–200

Thomas BC, Pedersen B, 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

Vogel JT, Zarka DG, Van Buskirk HA, Fowler SG, Thomashow MF (2005) Roles of the CBF2 and ZAT12 transcription factors in configuring the low temperature transcriptome of Arabidopsis. Plant J **41:** 195–211

Voznesenskaya EV, Edwards GE, Kiirats O, Artyusheva EG, Franceschi VR (2003) Development of biochemical specialization and organelle partitioning in the single-cell C_4 system in leaves of *Borszczowia aralocaspica* (Chenopodiaceae). Am J Bot **90:** 1669–1680

Voznesenskaya EV, Franceschi VR, Kiirats O, Artyusheva EG, Freitag H, Edwards GE (2002) Proof of C_4 photosynthesis without Kranz anatomy in *Bienertia cycloptera* (Chenopodiaceae). Plant J **31:** 649–662

Voznesenskaya EV, Franceschi VR, Kiirats O, Freitag H, Edwards GE (2001) Kranz anatomy is not essential for terrestrial C_4 plant photosynthesis. Nature **414:** 543–546

Voznesenskaya EV, Koteyeva NK, Chuong SDX, Ivanova AN, Barroca J, Craven LA, Edwards GE (2007) Physiological, anatomical and biochemical characterisation of photosynthetic types in genus Cleome (Cleomaceae). Funct Plant Biol **34:** 247–267

Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang HY, Landherr L, Tomsho LP, Hu Y, Carlson JE, et al (2009) Comparison

of next generation sequencing technologies for transcriptome characterization. BMC Genomics **10:** 347

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10:** 57–63

Waters MT, Wang P, Korkaric M, Capper RG, Saunders NJ, Langdale JA (2009) GLK transcription factors coordinate expression of the photosynthetic apparatus in *Arabidopsis*. Plant Cell **21:** 1109–1128

Weber APM, von Caemmerer S (2010) Plastid transport and metabolism of C_3 and C_4 plants: comparative analysis and possible biotechnological exploitation. Curr Opin Plant Biol **13:** 256–264

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. Plant Physiol **144:** 32–42

Wise RP, Moscou MJ, Bogdanove AJ, Whitham SA (2007) Transcript profiling in host-pathogen interactions. Annu Rev Phytopathol **45:** 329–369

3.2 Manuscript 5: Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C_3 and C_4 Plant Species

Overview

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

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

Published in The Plant Cell, August 2014

* These authors contributed equally to this work.

Co-authorship

Contributions

- Assistance with data analysis
- Hierarchical clustering
- Assistance in *de novo* assembly evaluation

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

RESEARCH ARTICLES

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

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

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

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

INTRODUCTION

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

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

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

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

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

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

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

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

RESULTS

Selection of Tissues Featured in the Comparative Atlases

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

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

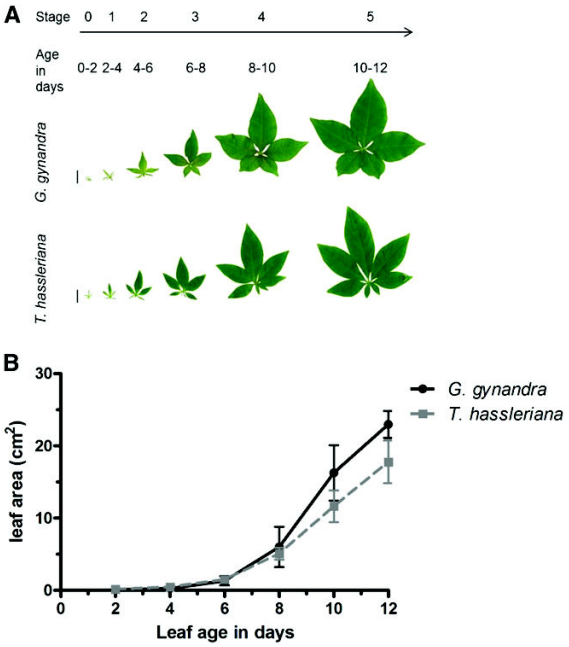


Figure 1. Overview of Leaf Shape and Expansion Rate in *G. gynandra* and *T. hassleriana*.

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

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

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

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

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

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

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

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

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

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

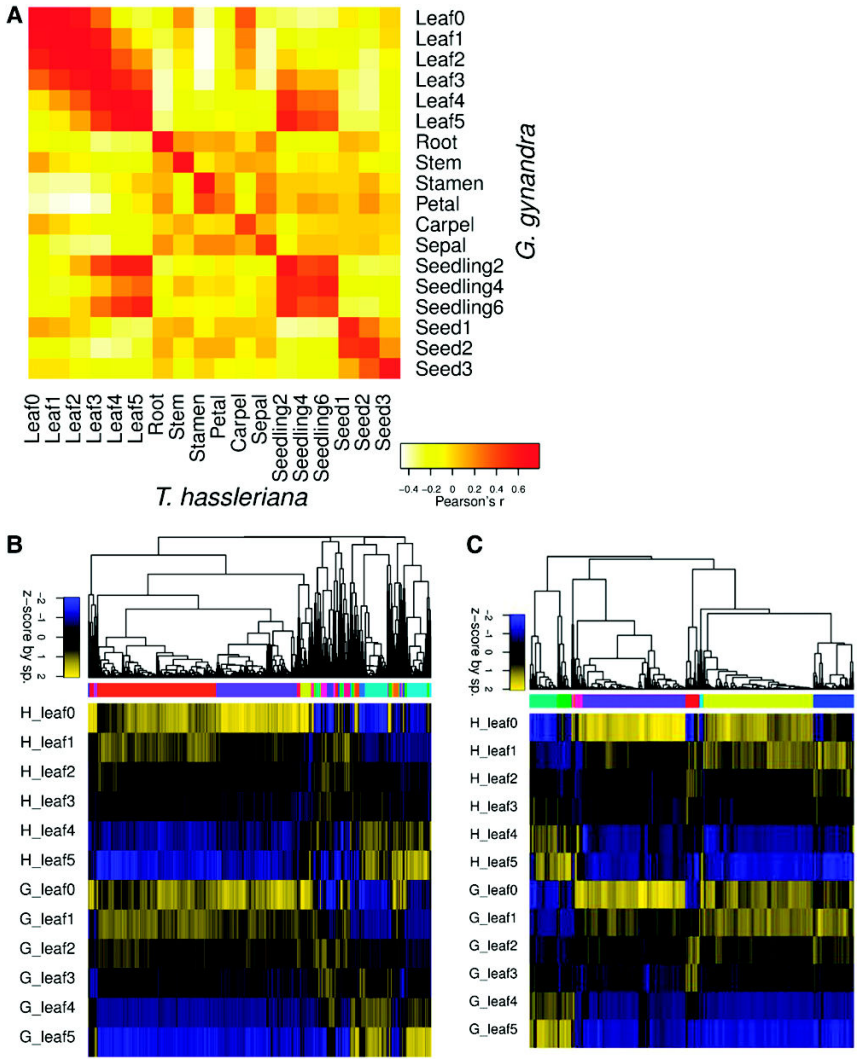


Figure 2. Comparative Tissue Dynamics and Gene Expression Pattern between *G. gynandra* and *T. hassleriana*.

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

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

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

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

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

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

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

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

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

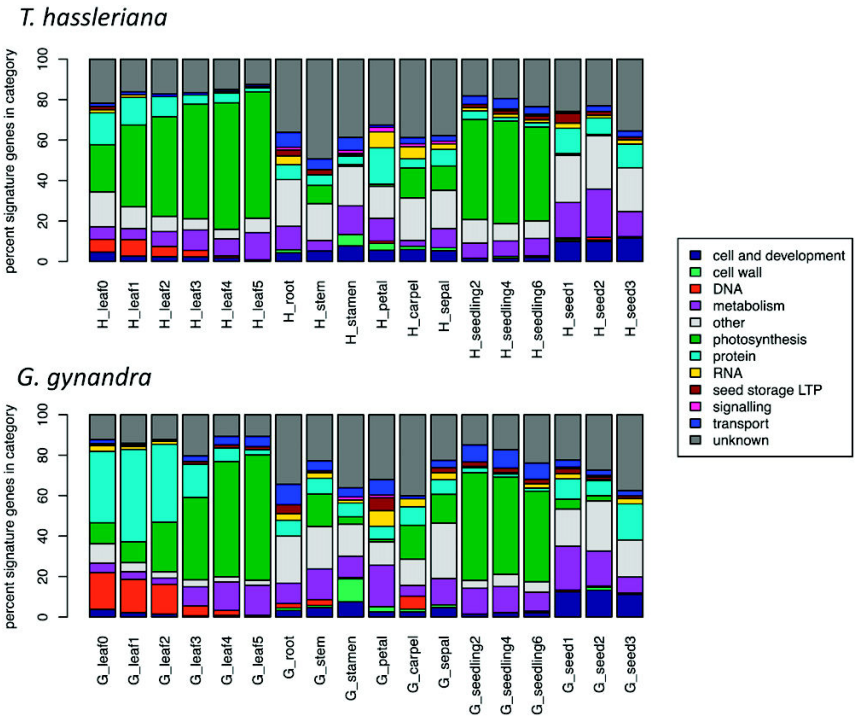


Figure 3. Distribution of Signature Genes in Each Tissue in *G. gynandra* and *T. hassleriana*.

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

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

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

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

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

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

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

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

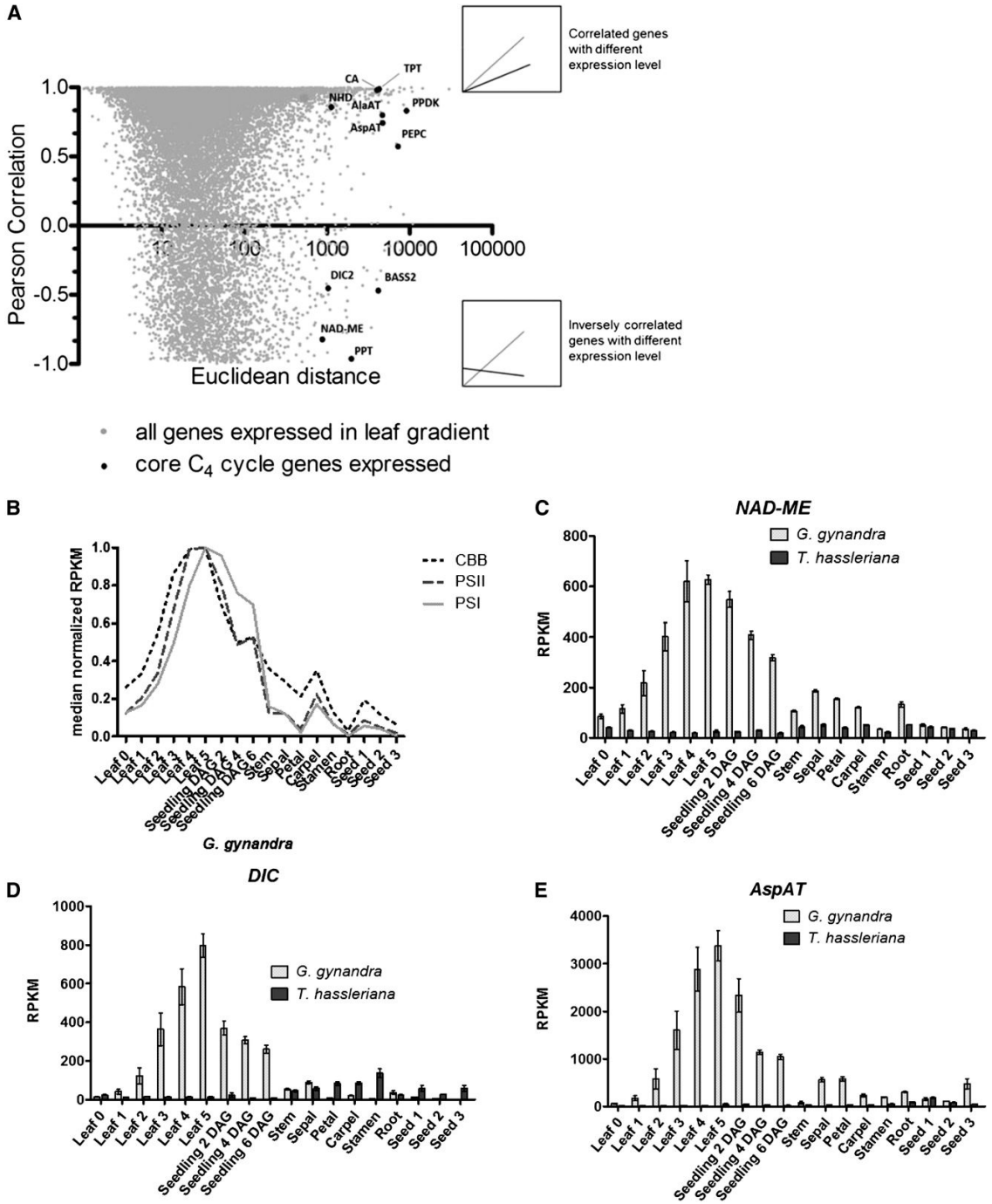


Figure 4. Comparison of Gene Expression Dynamics within the Leaf Gradient of Both Species.

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

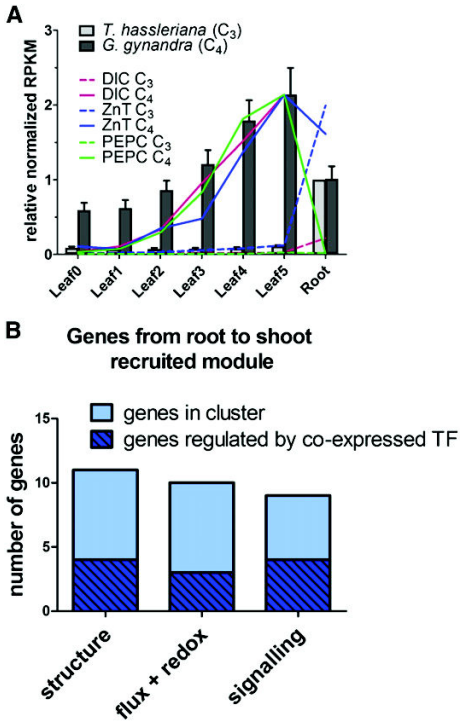


Figure 5. Recruitment of Genes from the Root to Leaf Expression Domain in the C_4 Plant *G. gynandra*.

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

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

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

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

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

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

Figure 4. (continued).

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

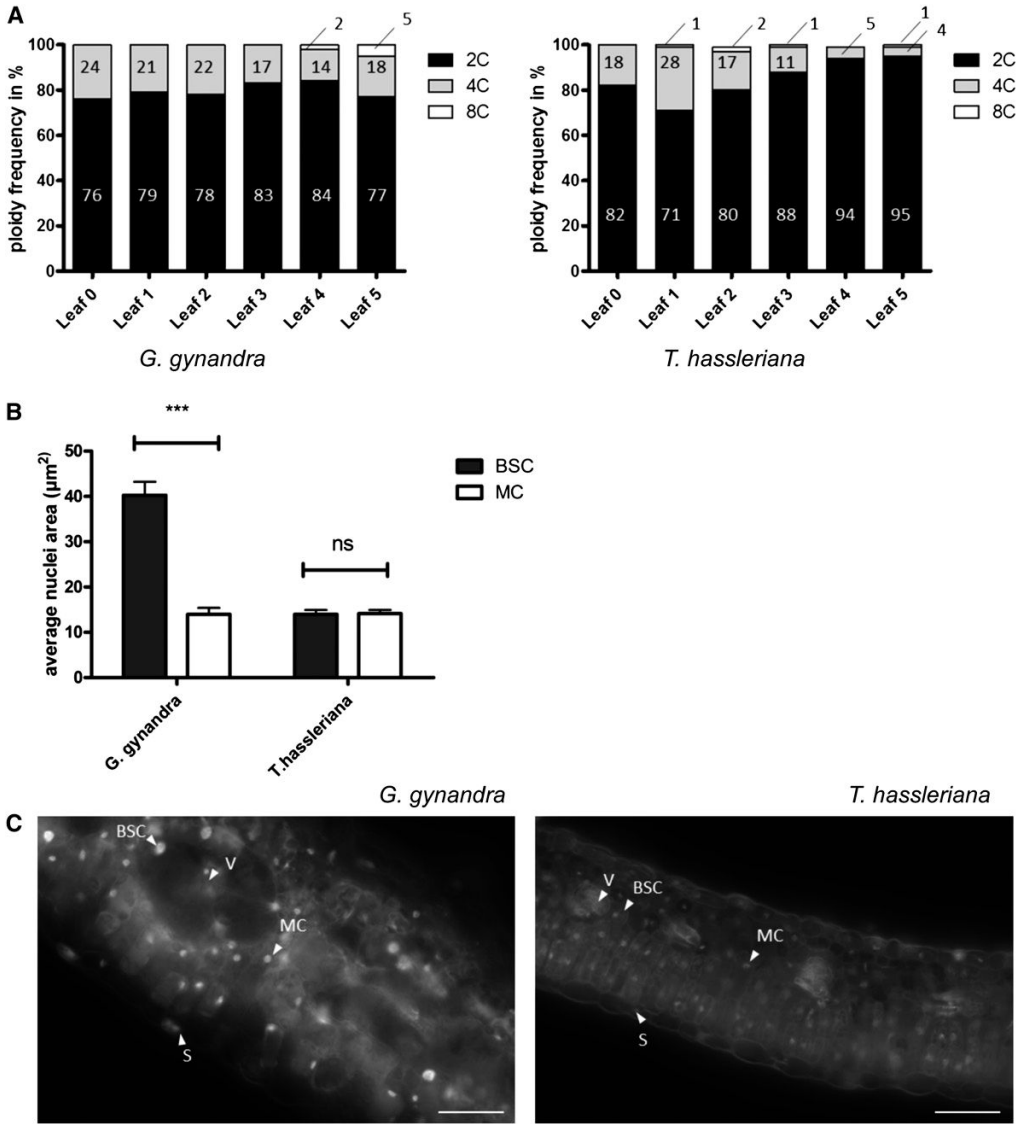


Figure 6. Distribution of Ploidy Levels during Leaf Development and Nuclei Area of BSC and MC between *G. gynandra* and *T. hassleriana*.

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

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

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

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

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

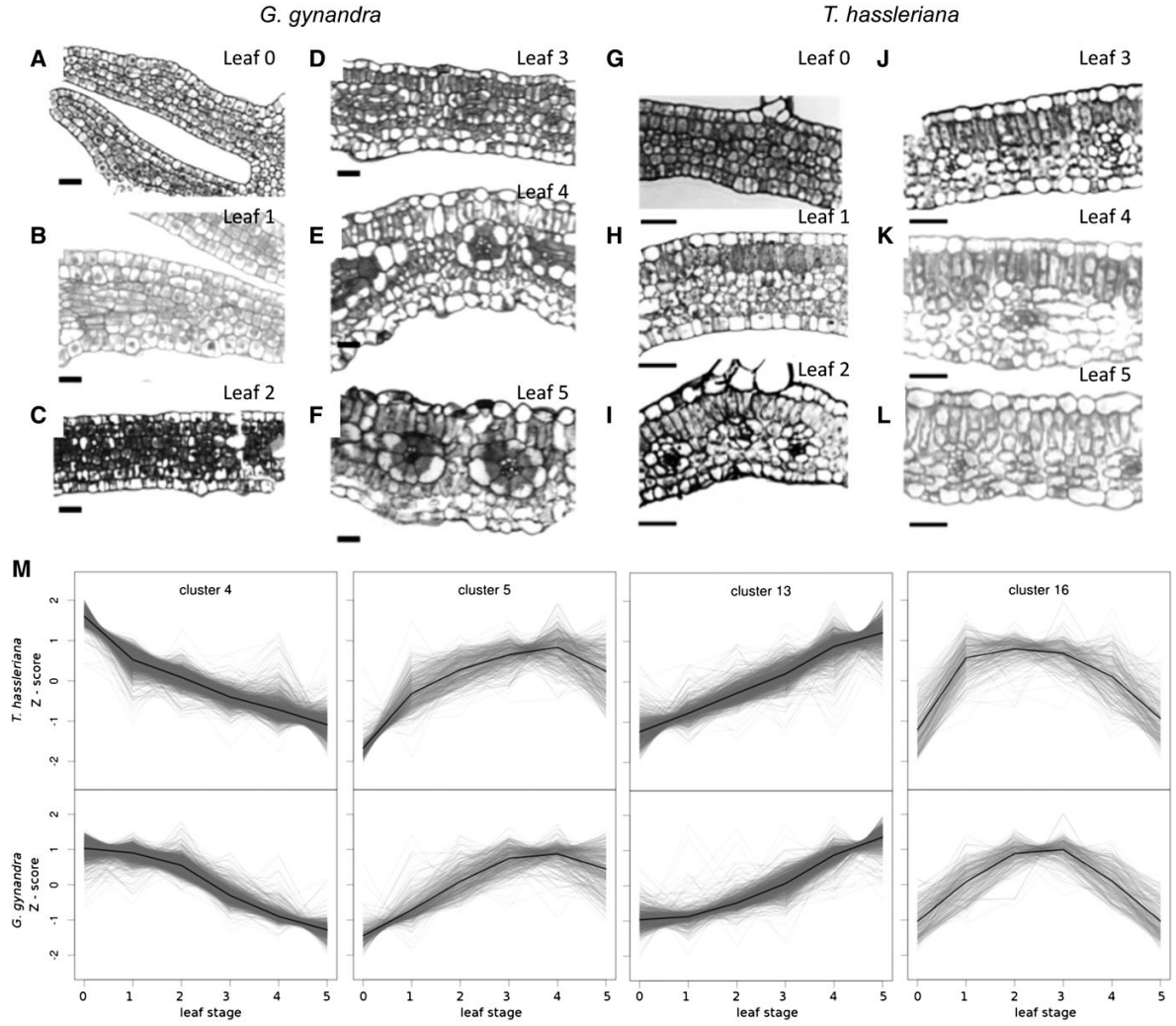


Figure 7. Analysis of Shifted Gene Expression Pattern and Leaf Anatomy during Leaf Ontogeny.

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

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

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

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

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

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

DISCUSSION

Comparative Transcriptome Atlases Provide a Powerful Tool for Understanding C_4 Photosynthesis

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

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

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

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

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

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

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

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

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

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

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

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

Late Differentiation of Mesophyll Tissue Allows Denser Venation

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

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

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

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

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

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

Conclusions

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

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

METHODS

Plant Material and Growth Conditions

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

Leaf Expansion Rate

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

Leaf Cross Sections for Anatomical Studies

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

Flow Cytometry

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

Measurement of Nuclei from Mature Leaves

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

RNA Extraction, Library Construction, and Sequencing

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

Gene Expression Profiling

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

Data Analysis

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

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

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

Accession Numbers

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

Supplemental Data

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Supplemental Table 2. Cross-Species Mapping Results.

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

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

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

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

Supplemental Methods.

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

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

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

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

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

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

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

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

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

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

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

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

ACKNOWLEDGMENTS

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

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

AUTHOR CONTRIBUTIONS

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

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

Bräutigam, A., et al. (2011). An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. Plant Physiol. **155:** 142–156.

Bräutigam, A., Schliesky, S., Külahoglu, C., Osborne, C.P., and Weber, A.P.M. (2014). Towards an integrative model of C_4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C_4 species. J. Exp. Bot. **65:** 3579–3593.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Wang, P., Kelly, S., Fouracre, J.P., and Langdale, J.A. (2013). Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz anatomy. Plant J. **75:** 656–670.

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

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

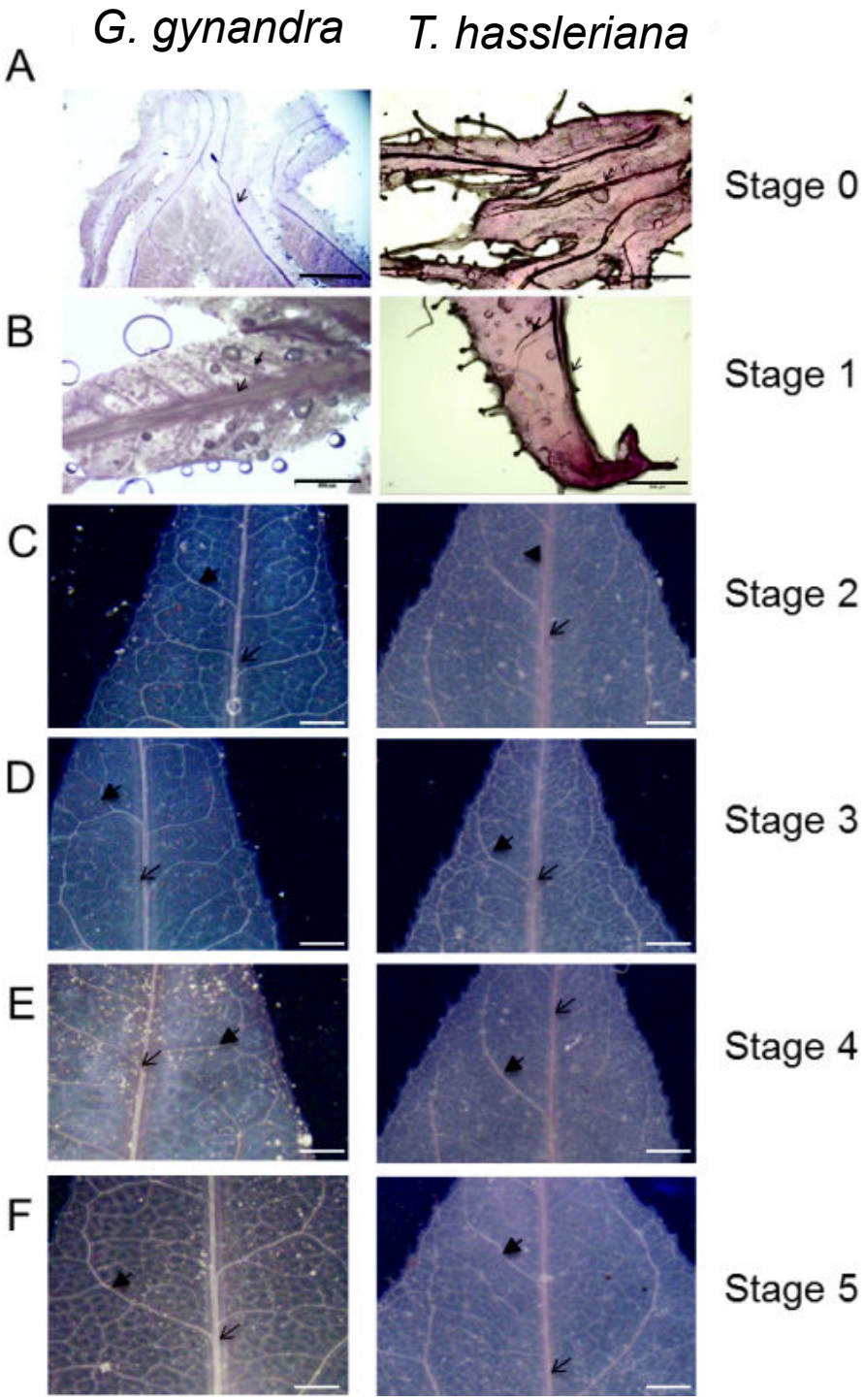
Xu, W., Purugganan, M.M., Polisensky, D.H., Antosiewicz, D.M., Fry, S.C., and Braam, J. (1995). Arabidopsis TCH4, regulated by hormones and the environment, encodes a xyloglucan endotransglycosylase. Plant Cell **7:** 1555–1567.

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

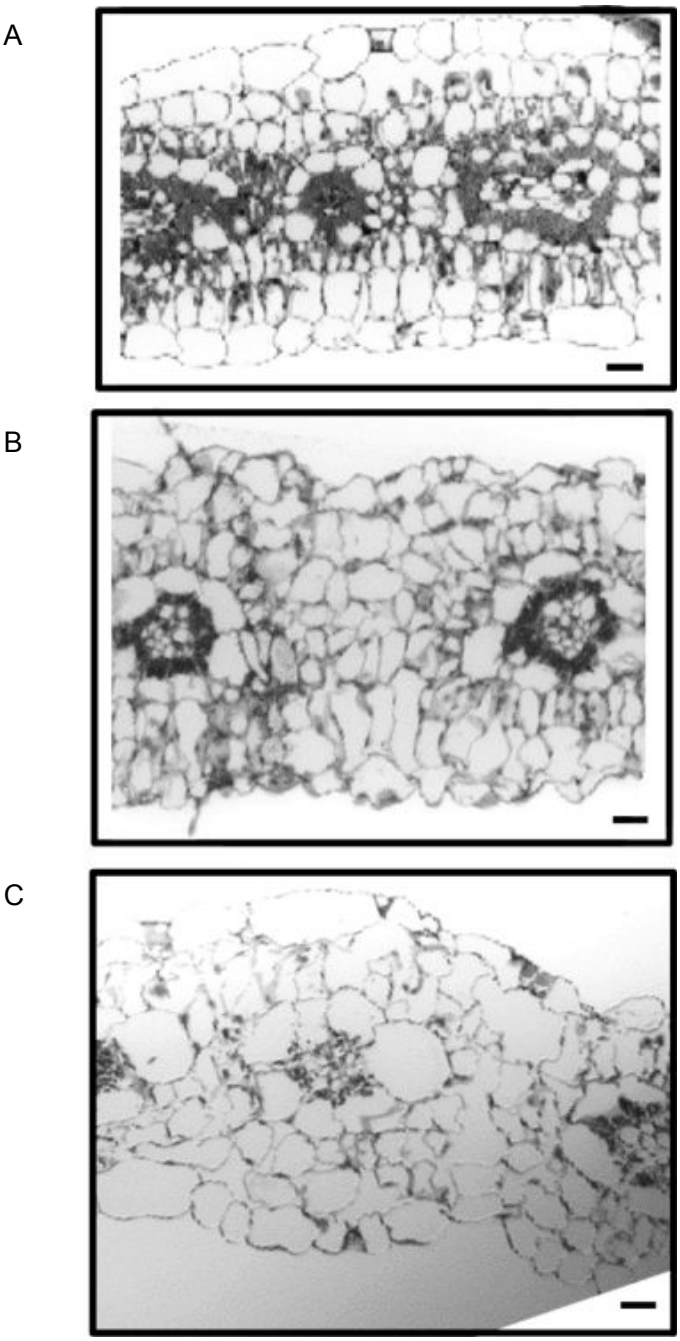
Yekutieli, D., and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J. Stat. Plan. Inference **82:** 171–196.

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

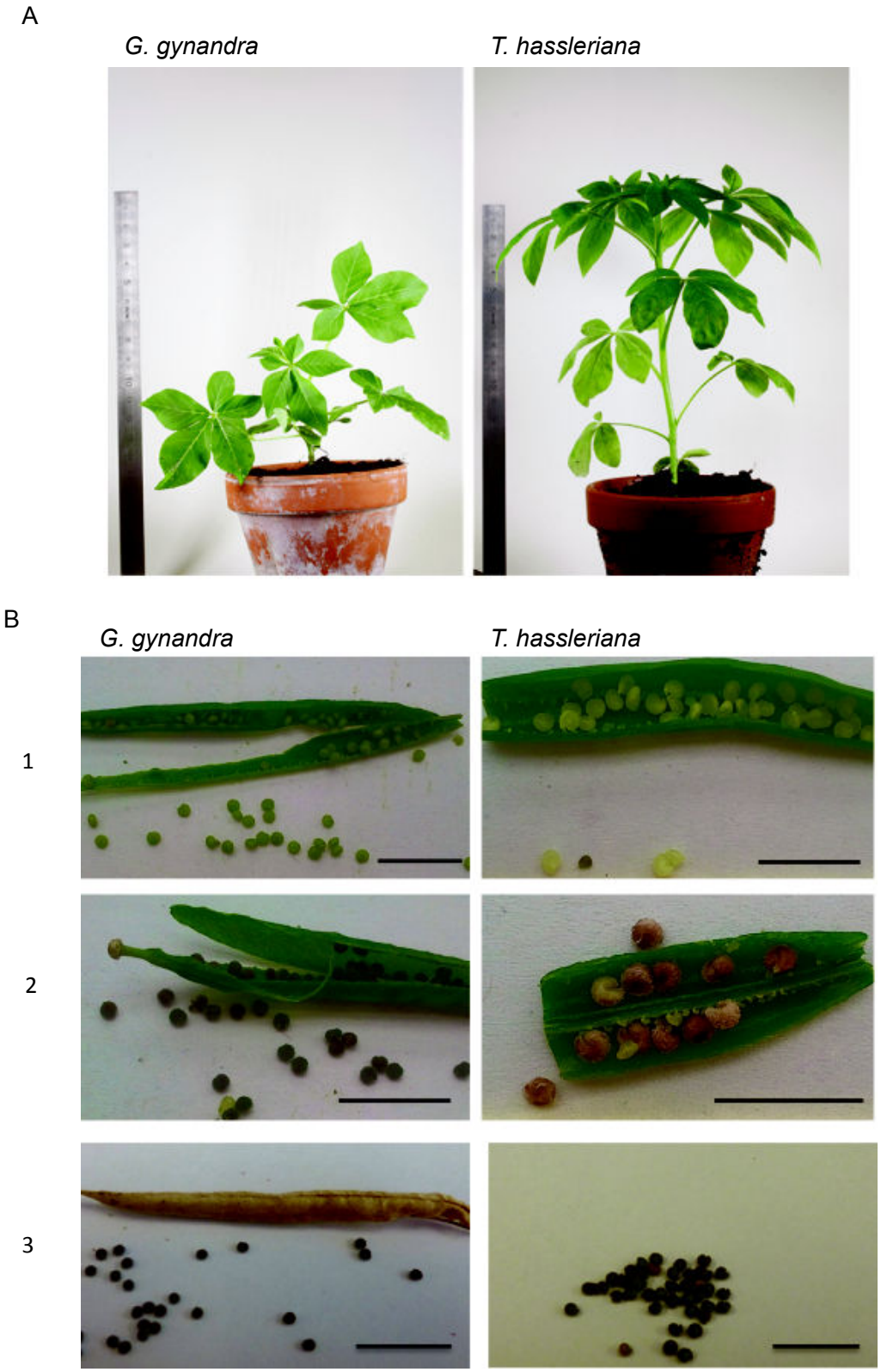
Zhu, X.G., Long, S.P., and Ort, D.R. (2008). What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? Curr. Opin. Biotechnol. **19:** 153–159.



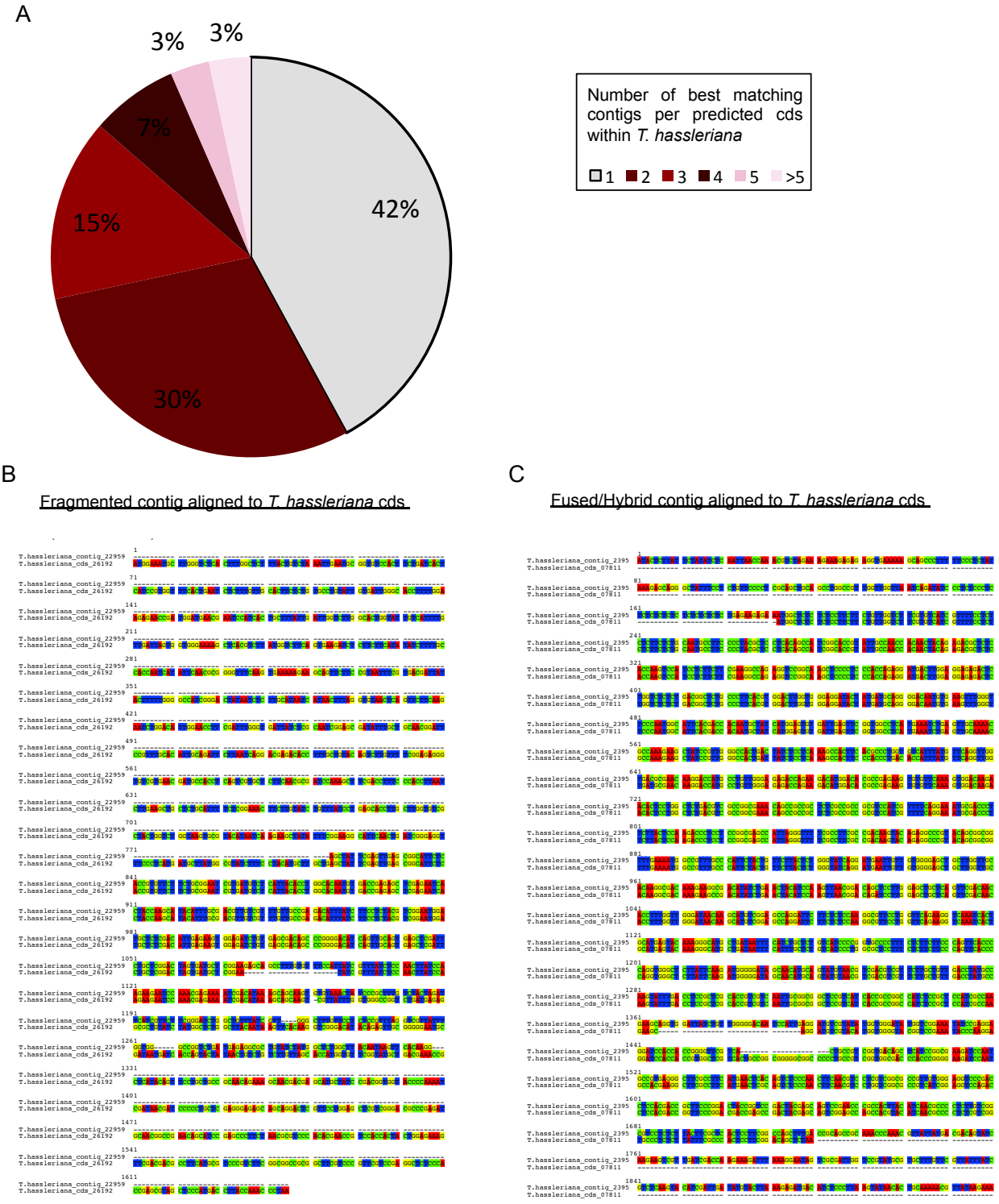
Supplemental Figure 1. Venation patterning during leaf development of *G. gynandra* and *T. hassleriana*.
(A-B) Cleared safranine stained leaves of stage 0 and 1 (n=3; scale bar 0.5 mm)
(C-F) Cleared leaves of stage 2, 3, 4 and 5 respectively (n=3; scale bar 1 mm)
Open arrows indicate the midvein (1°)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



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

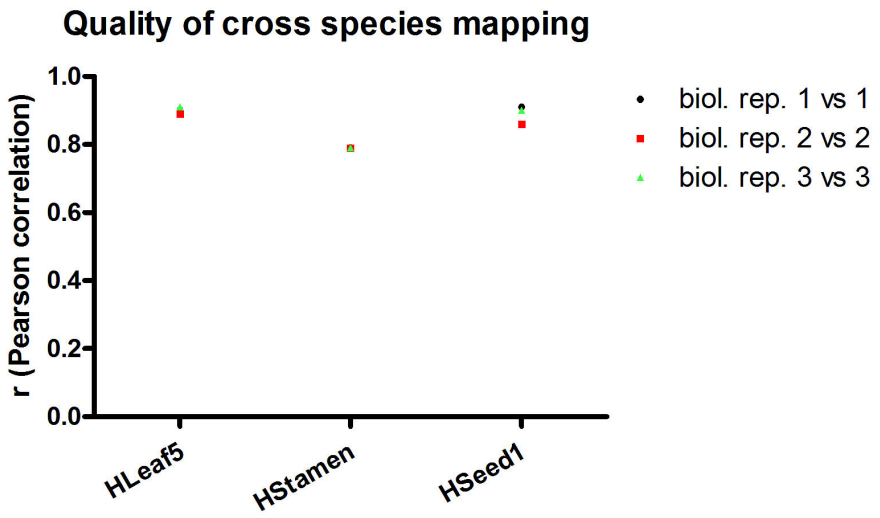


Supplemental Figure 3. Images of tissues harvested for RNA-seq in *G. gynandra* and *T. hassleriana*. (A) Photographic image of *G. gynandra* and *T. hassleriana* 8-week old plants, from which leaf gradient, stem and root system were harvested **(B) Seed coat development from harvested developmental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

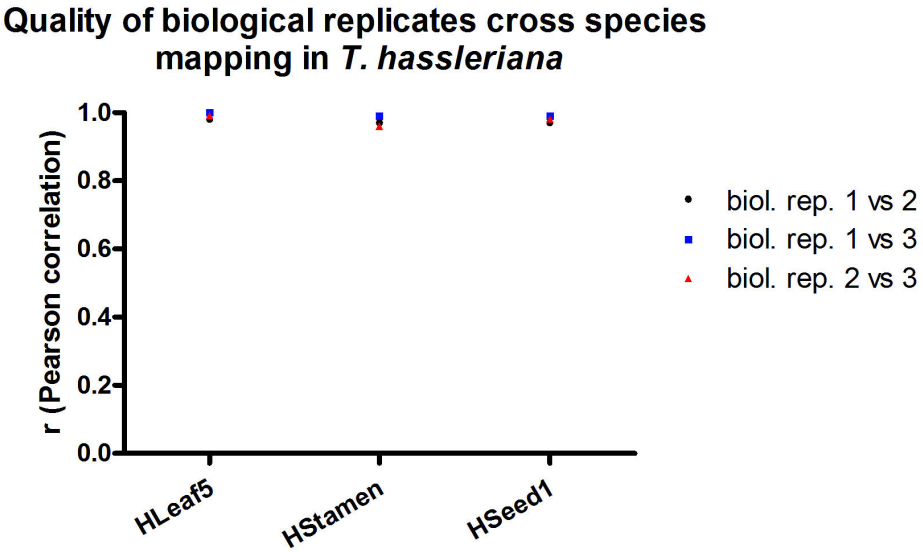


Supplemental Figure 4. Quality assessment of Velvet/OASES assembled *T. hassleriana* contigs against predicted corresponding cds from *T. hassleriana* genome.
(A) Percentage of contig number per predicted cds (Cheng et al., 2013) showing redundancy in assembled contigs.
(B) ClustalW alignment of fragmented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

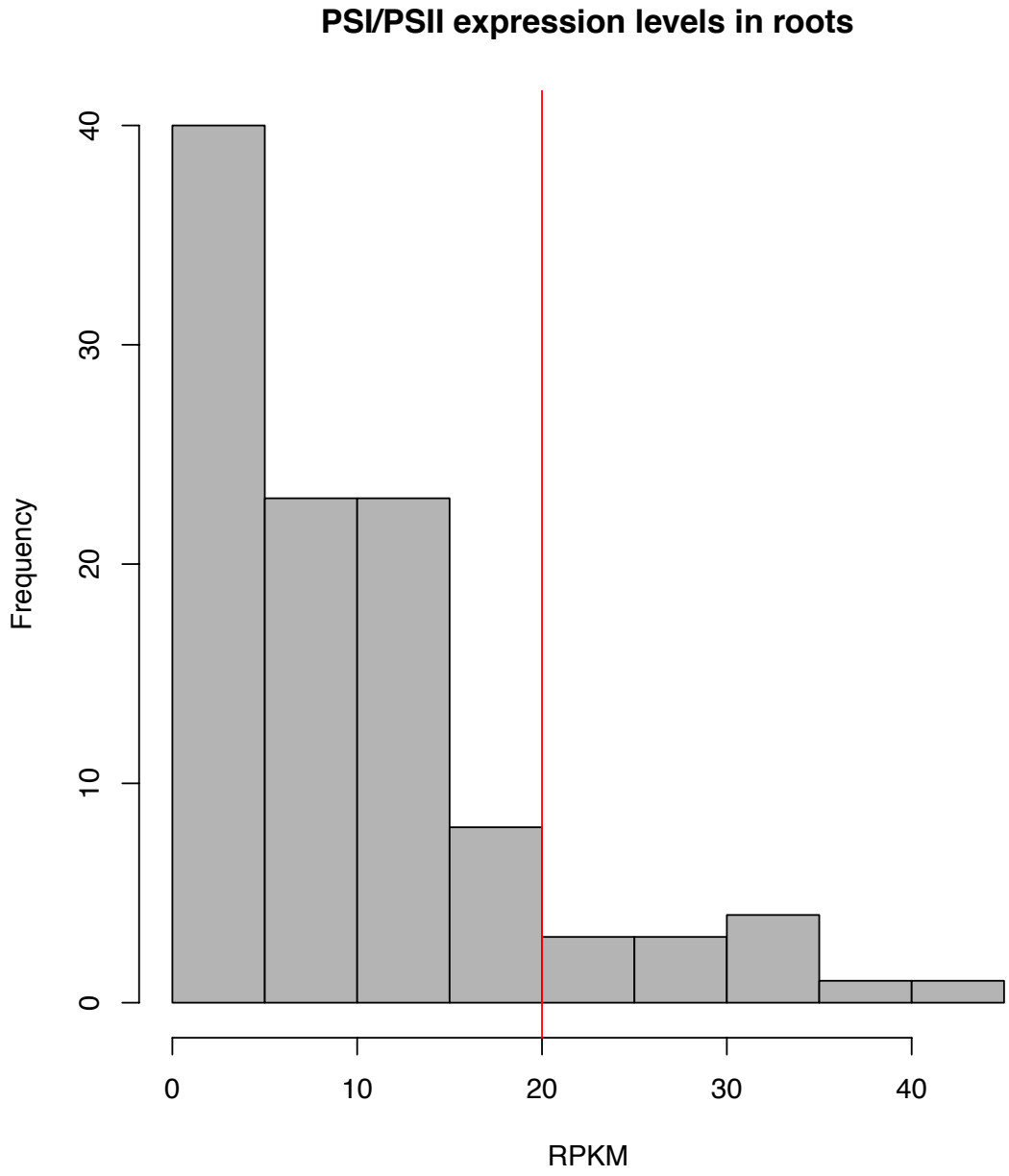
A



B

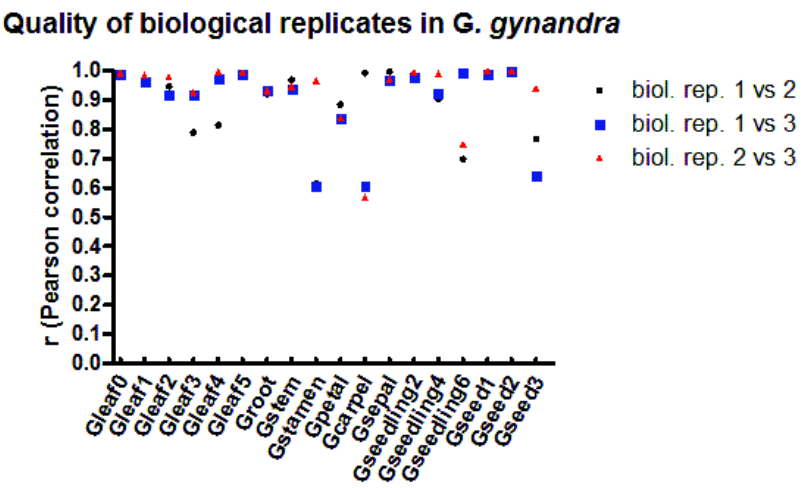


Supplemental Figure 5. Quality assessment of the biological replicates of *T. hassleriana* libraries mapped to *A. thaliana* and mapping similarity of *T. hassleriana* libraries mapped to *A. thaliana* and to its own cds.
(A) Pair-wise Pearson's correlation (*r*) was calculated for all three pairs of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

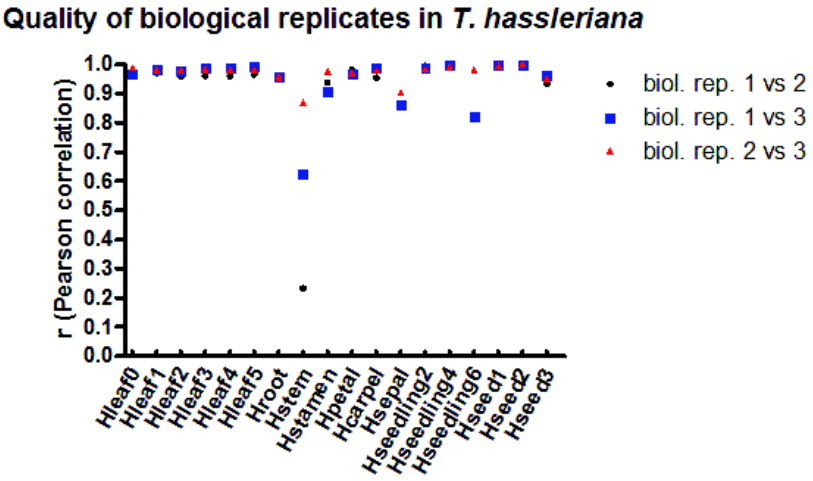


Supplemental Figure 6. Determination of base line gene expression via a histogram of photosystem (PS) I and II transcript abundances reads per mappable million (RPKM) in the *G. gynandra* root.
Y- axis shows frequency and Y- axis depicts RPKM level of PSI and PSII transcript abundance. Red line indicates where threshhold of base line expression was set.

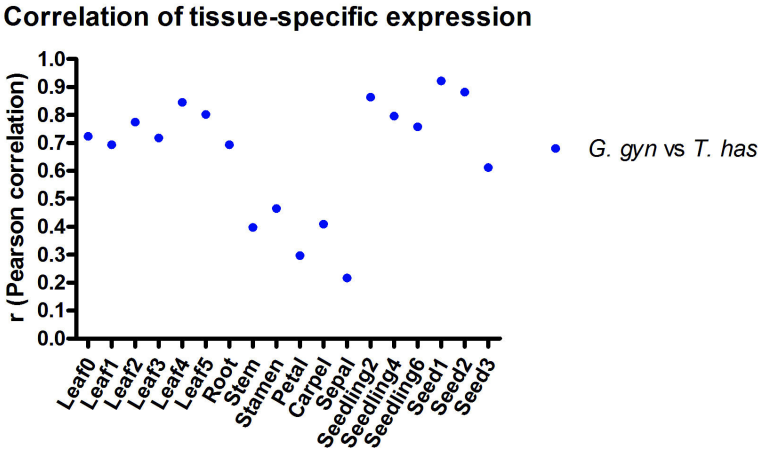
A



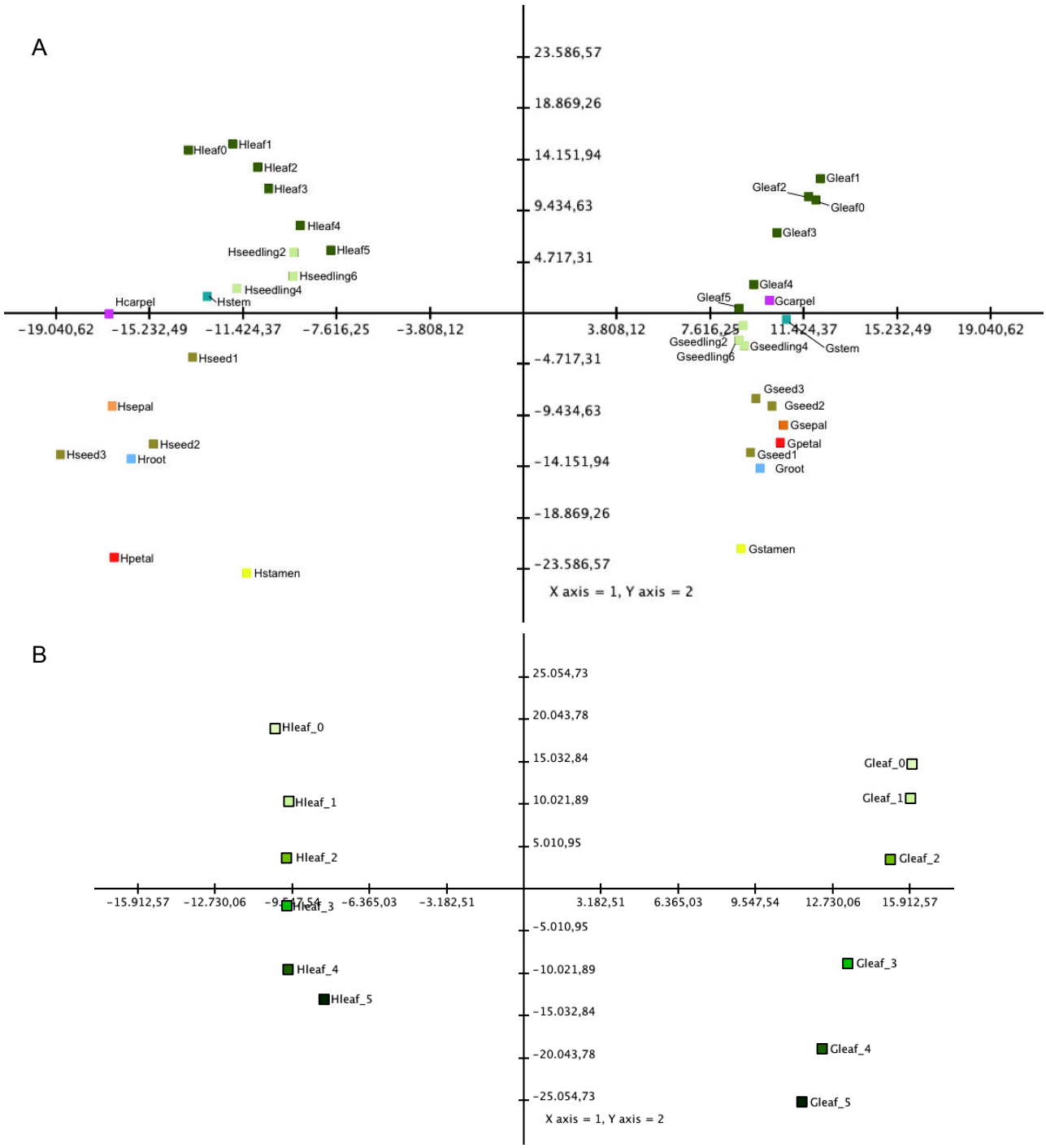
B



C

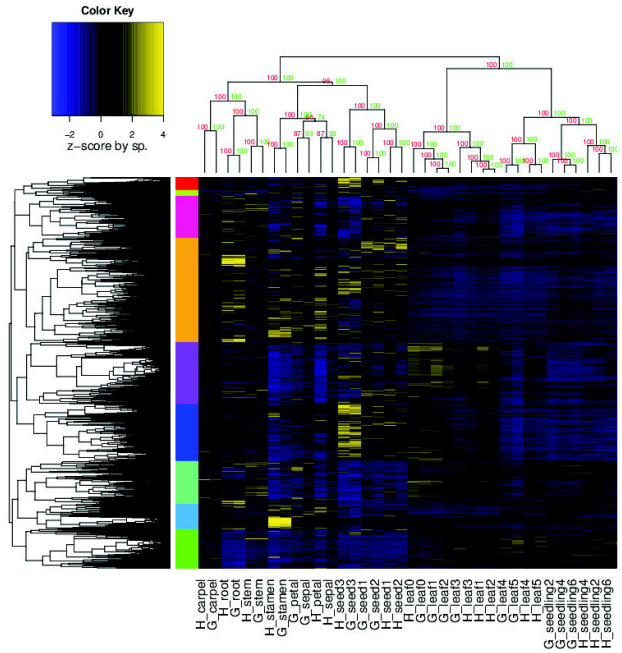


Supplemental Figure 7. Quality assessment of the biological replicates within each species and tissue similarity between *G. gynandra* and *T. hassleriana*. (A) Pair-wise Pearson's correlation (*r*) was calculated for all three pairs of biological replicates for each tissue (n=3) in *G. gynandra*. **(B)**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

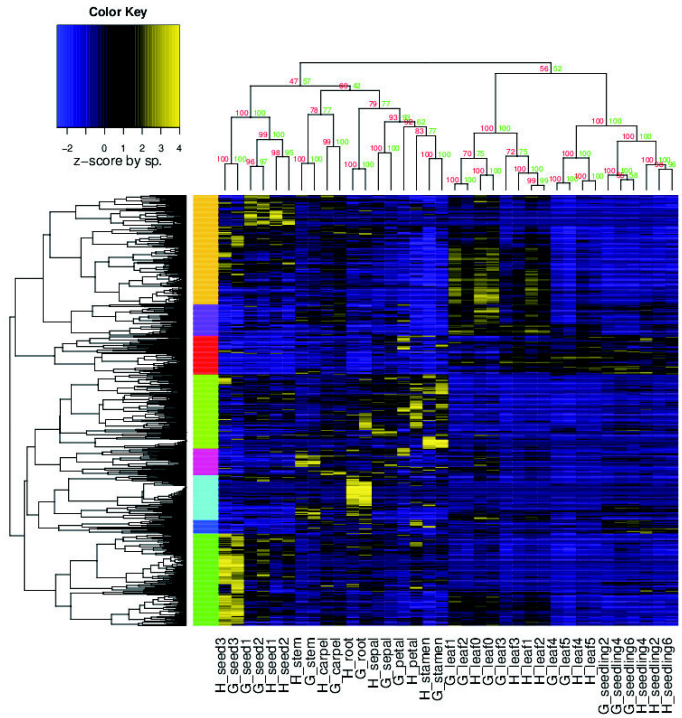


Supplemental Figure 8. Principle component analysis between *G. gynandra* and *T. hassleriana.*
(A) Plot shows all averaged tissues from *G. gynandra* (G) and *T. hassleriana* (H) sequenced (n=3). The first component describes 15% of all data variablility seperating both species. The second component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

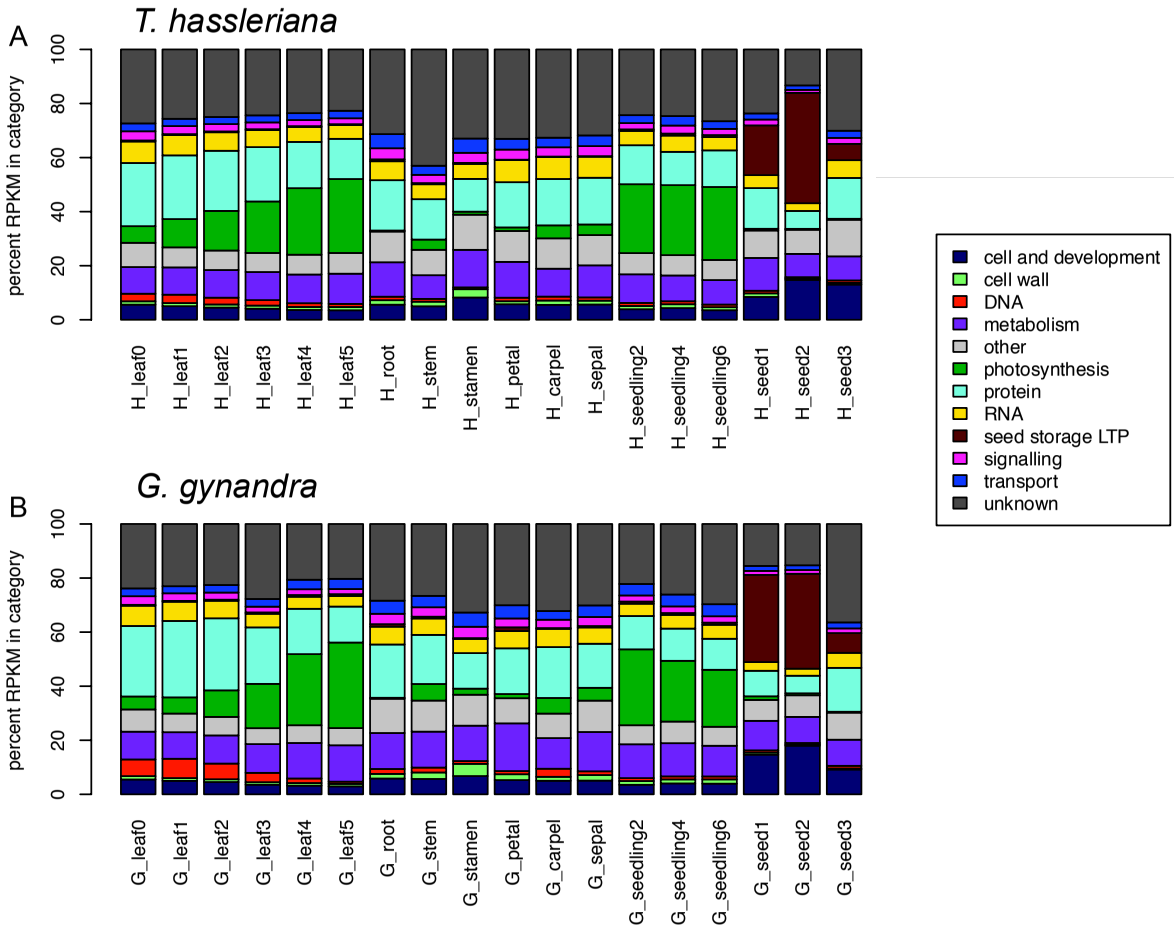
A



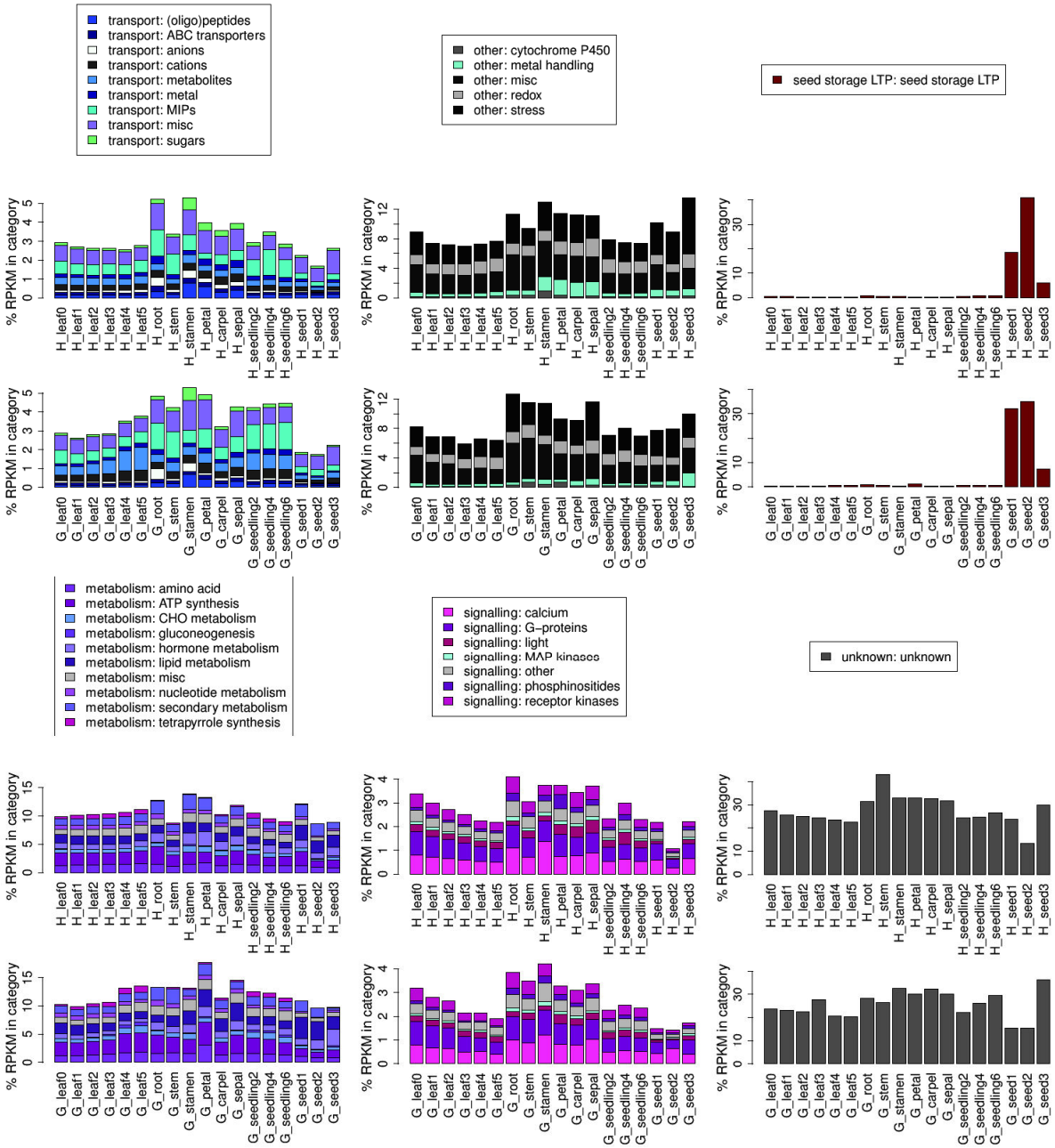
B



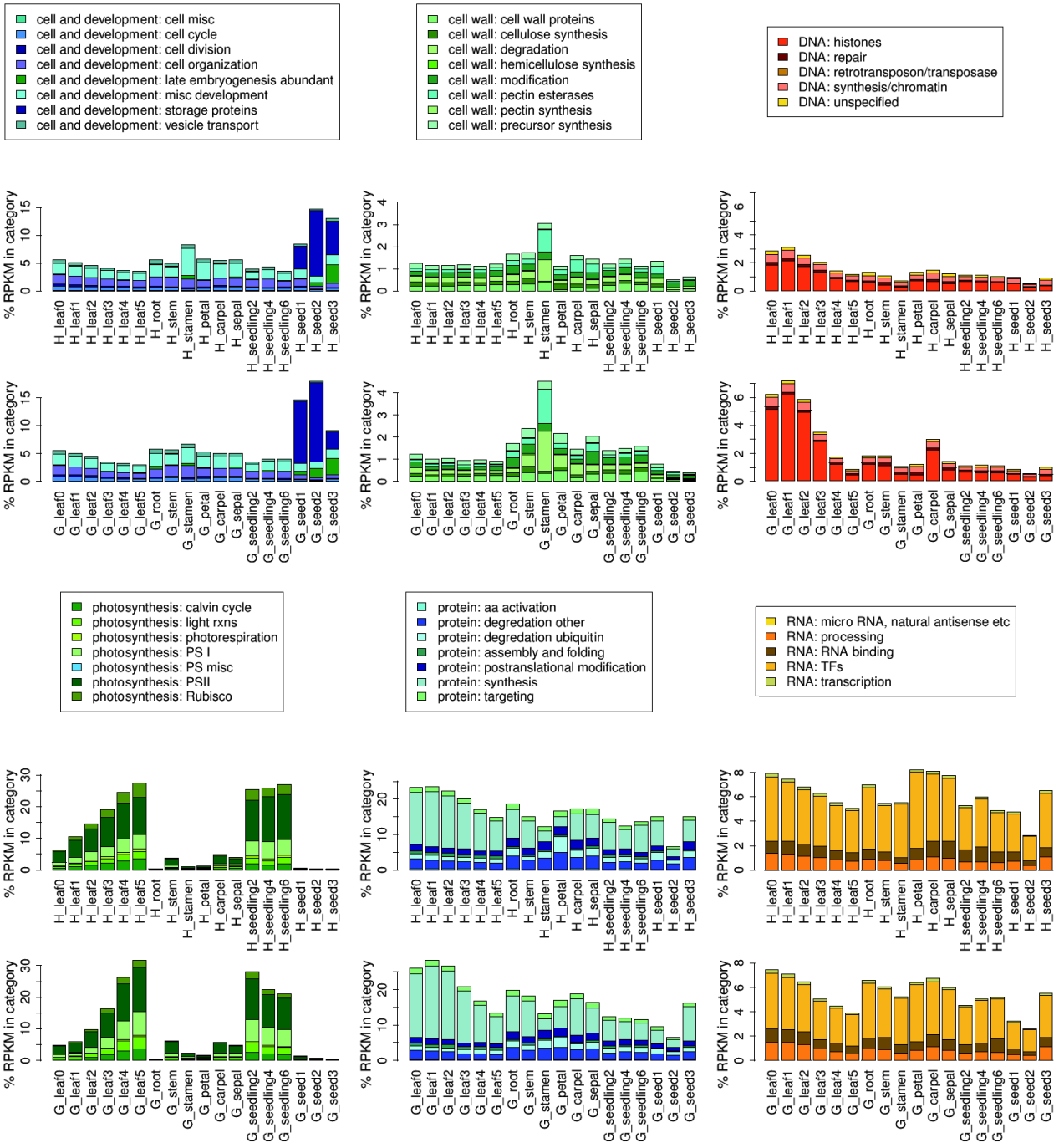
Supplemental Figure 9. Hierarchical cluster analysis with bootstrapped samples of *G. gynandra* and *T. hassleriana*. Numbers above the nodes show the approximately unbiased p-value (red) and the bootstrap probability (green). Blue is lowest expression and yellow highest expression. Left-hand vertical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



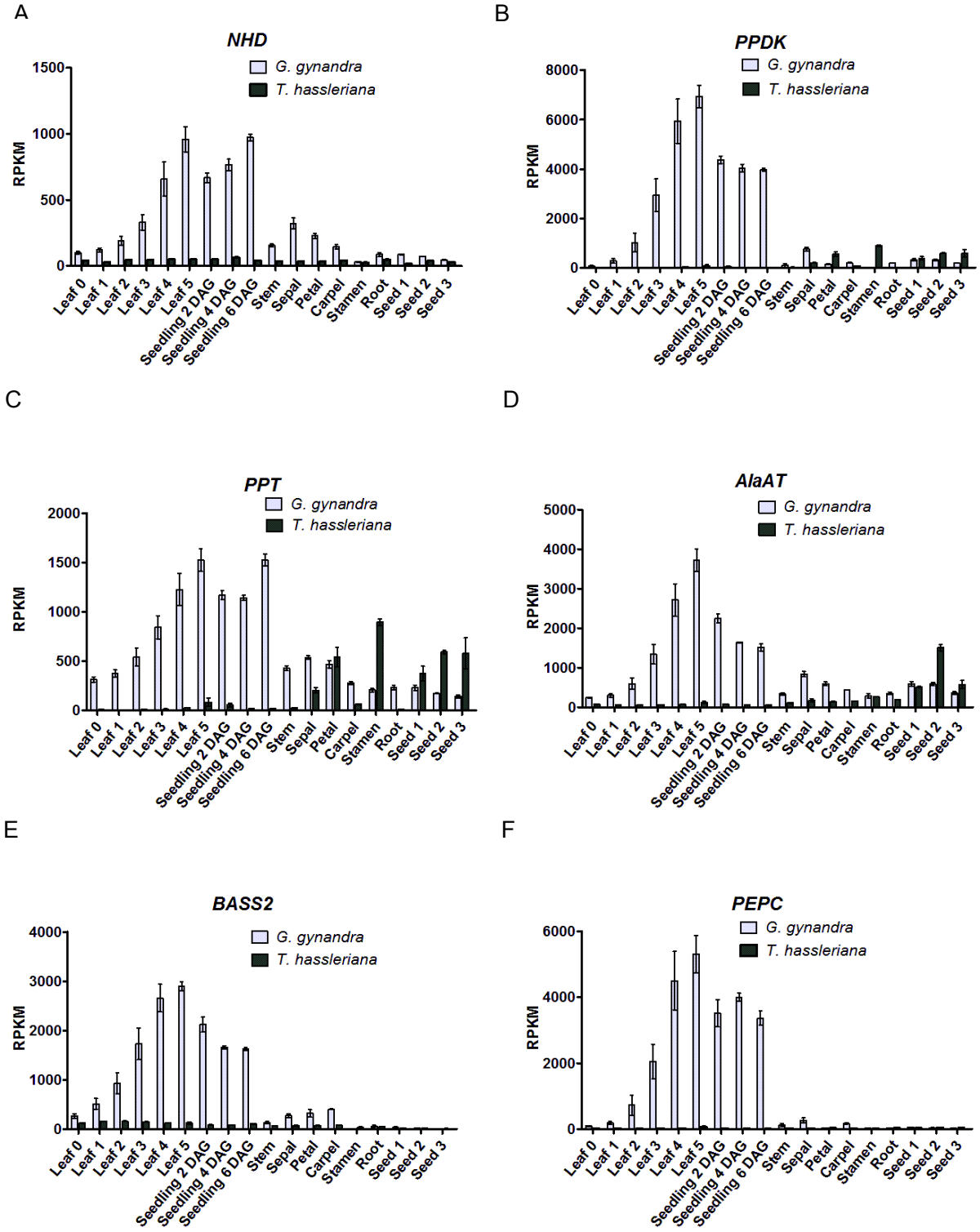
Supplemental Figure 10. Transcriptional investment of each tissue compared in both species. Cumulative average RPKMs in percent for basal Mapman categories for each tissue in *G. gynandra* and *T. hassleriana*.



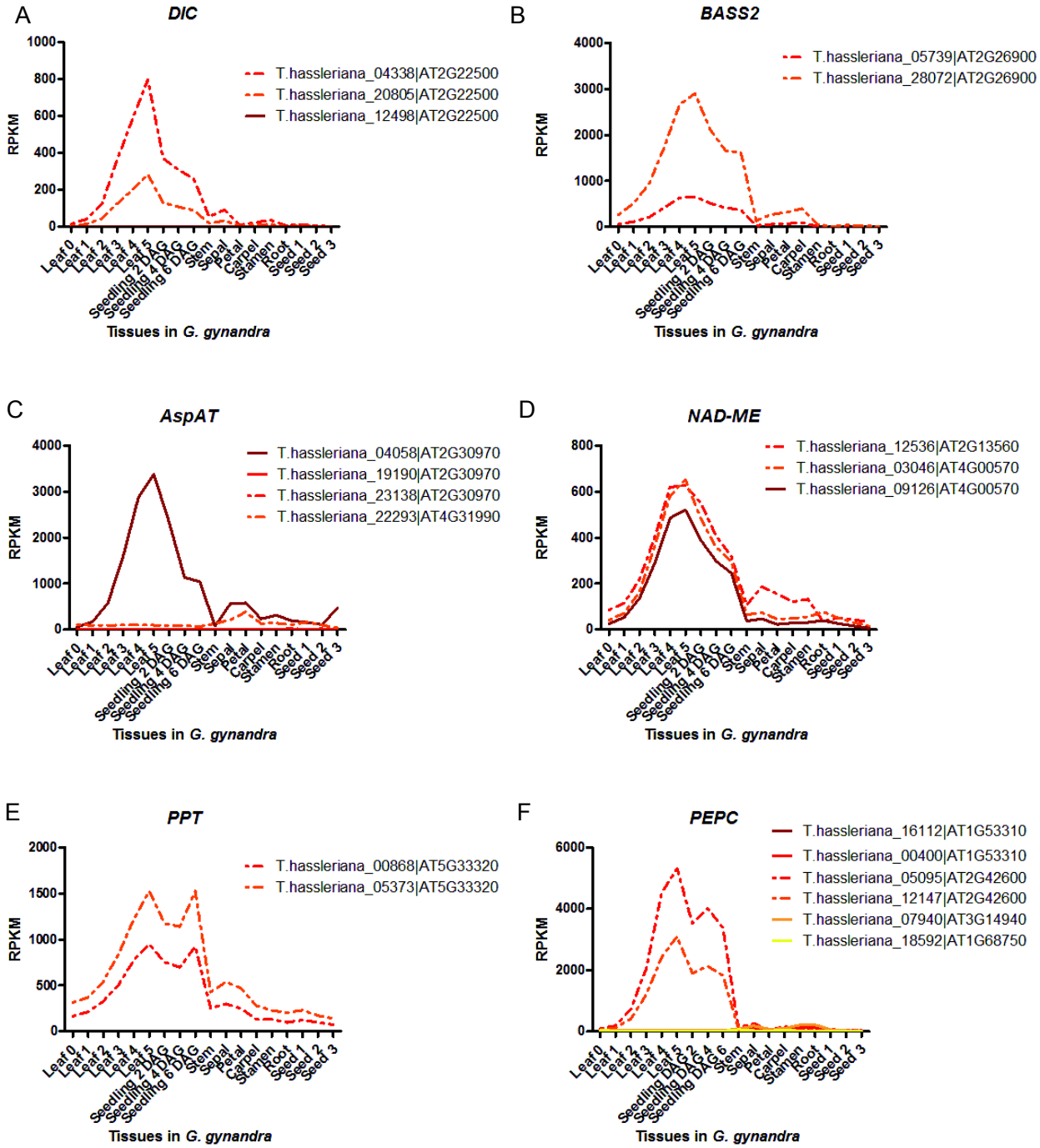
Supplemental Figure 11.1. Transcriptional investment at secondary Mapman category of each tissue compared in both species (Part 1). Distribution of the Mapman categories in each tissue in *G. gynandra* and *T. hassleriana*. Plot shows percent of average RPKMs of the 12 customized secondary Mapman bins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



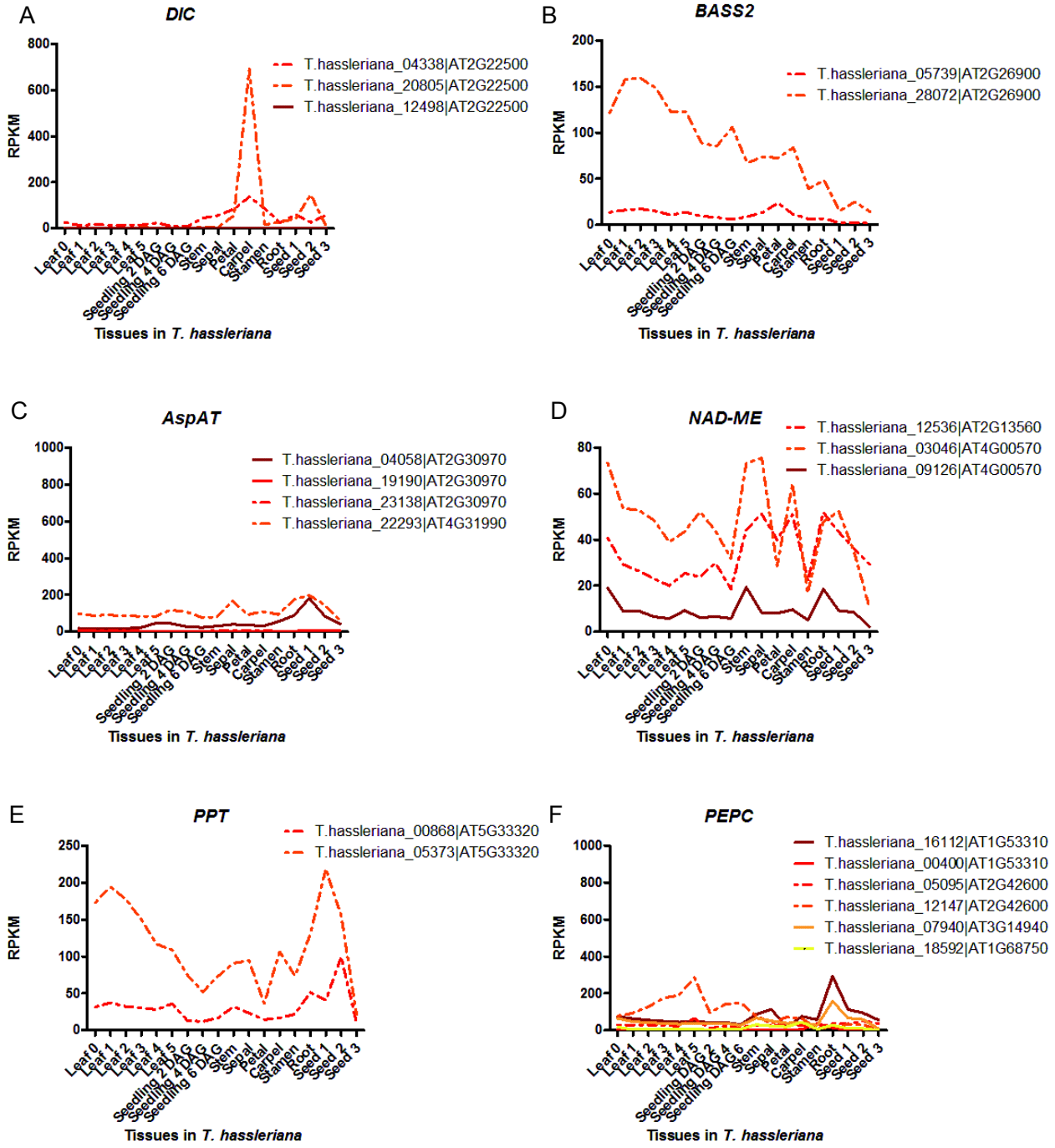
Supplemental Figure 11.2. Transcriptional investment at secondary Mapman category of each tissue compared in both species (Part 2). Distribution of the Mapman categories in each tissue in *G. gynandra* and *T. hassleriana*. Plot shows percent of average RPKMs of the 12 customized secondary Mapman bins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



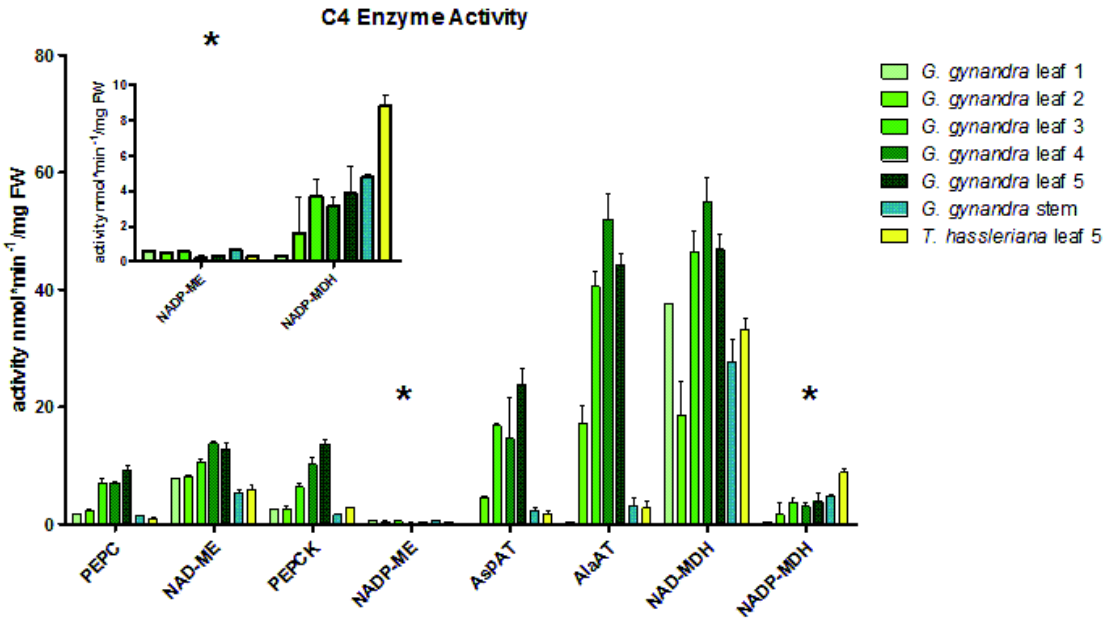
Supplemental Figure 12. Comparison of gene expression dynamics within the leaf gradient of both species.
(A-F) Average expression pattern of highest abundant putative ortholog of C_4 cycle genes (*NHD, PPDK, PPT, AlaAT, BASS2, PEPC*) in photo- and heterotrophic tissues in *G. gynandra* (light grey) and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



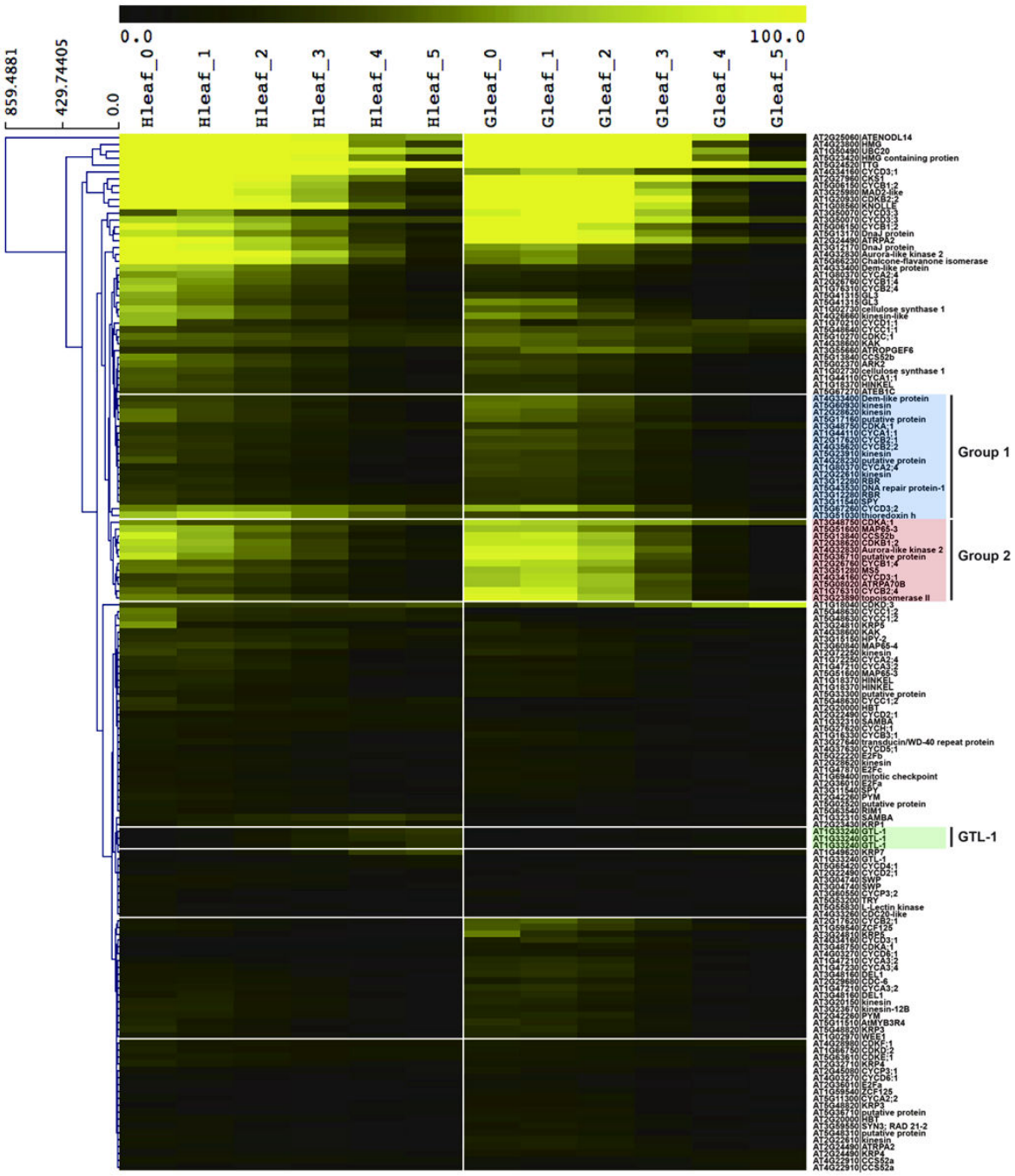
Supplemental Figure online 13. Plot of all C_4 gene putative orthologs expression pattern (RPKM) in *G. gynandra,* that were annotated as C_4 genes with AGI identifier and respective *T. hassleriana* ID. **(A-F)** Average expression pattern of putative ortholog of C_4 cycle genes (*DIC, BASS2, AspAT,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



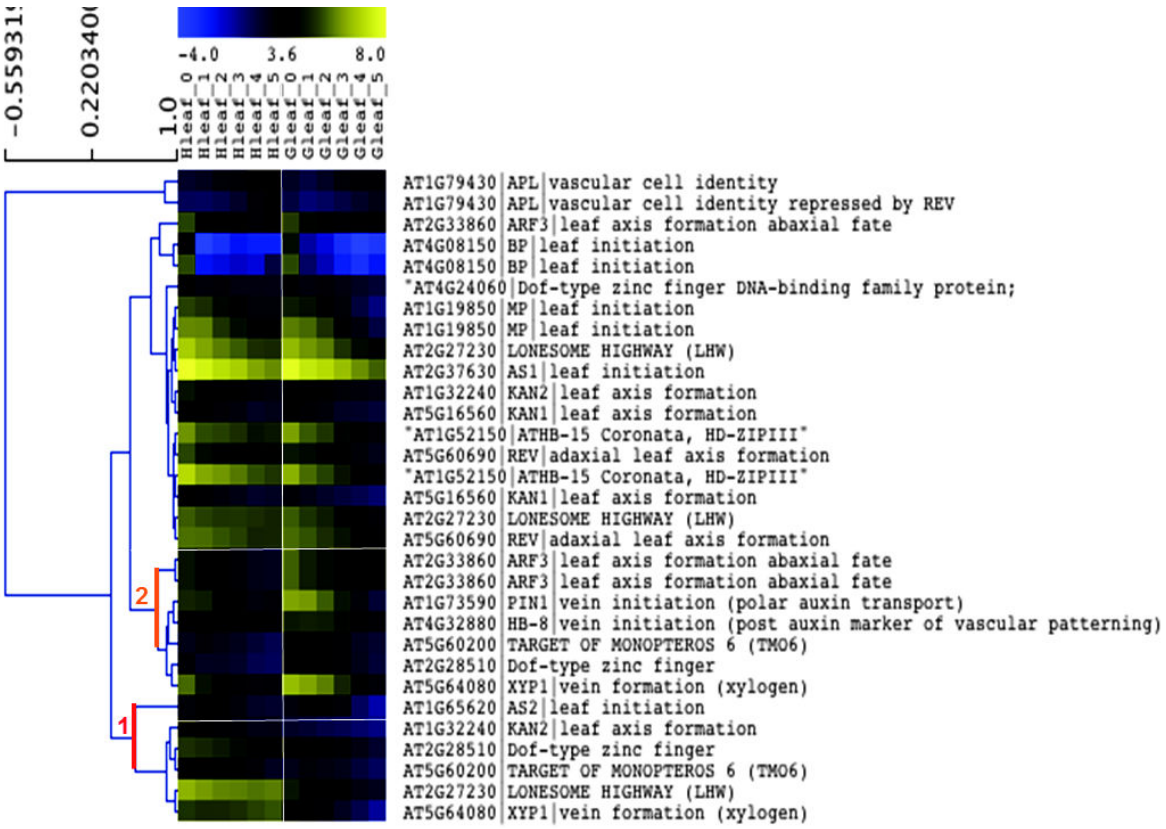
Supplemental Figure online 14. Plot of all C_4 gene putative orthologs expression pattern (RPKM) in *T. hassleriana,* that were annotated as C_4 genes with AGI identifier and respective *T. hassleriana* ID. **(A-F)** Average expression pattern of putative ortholog of C_4 cycle genes (*DIC, BASS2, AspAT,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



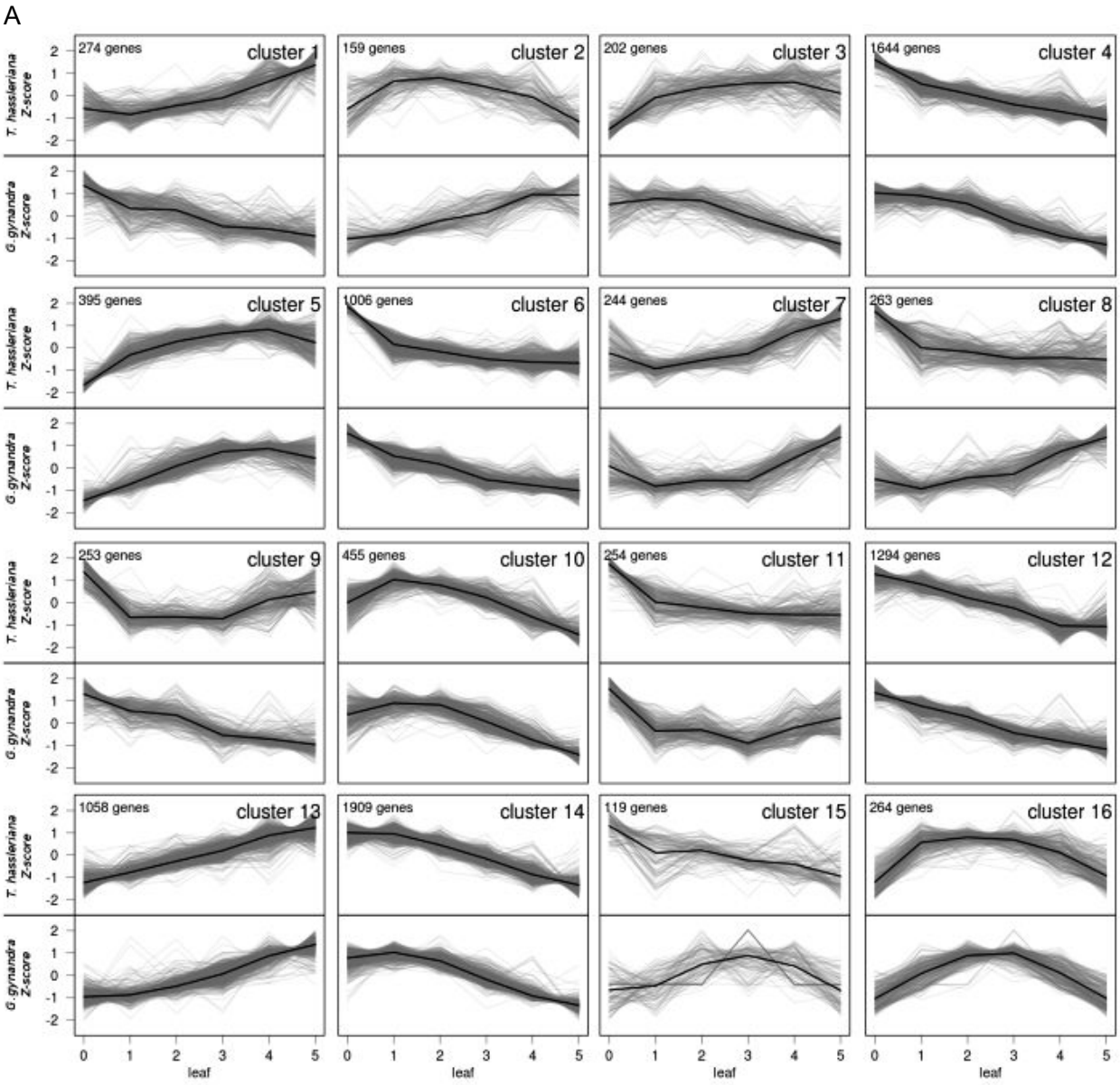
Supplemental Figure 15. Enzyme activity measurement of soluble C_4 cycle enzymes. Enzyme activities of PEPC, NAD-ME, PEPCK, NADP-ME, AspAT, AlaAT, NAD-MDH and NADP-MDH were measured along the developing *G. gynandra* leaf (stage 1-5) with the mature *T. hassleriana* leaf (stage 5) as C_3 control. (FW: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



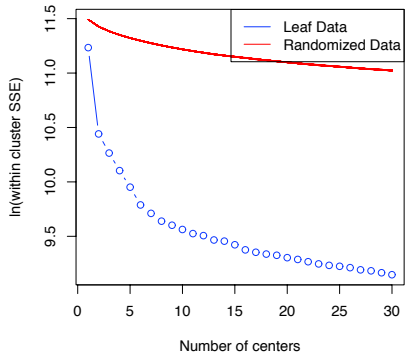
Supplemental Figure 16. Hierarchical clustering of average RPKM with Euclidean distance of core cell cycle genes in *T. hassleriana* and *G. gynandra*. Core cell cycle genes were extracted from (Vandepoele et al., 2002; Beemster et al., 2005). Deregulated cluster of interest are marked with blue and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



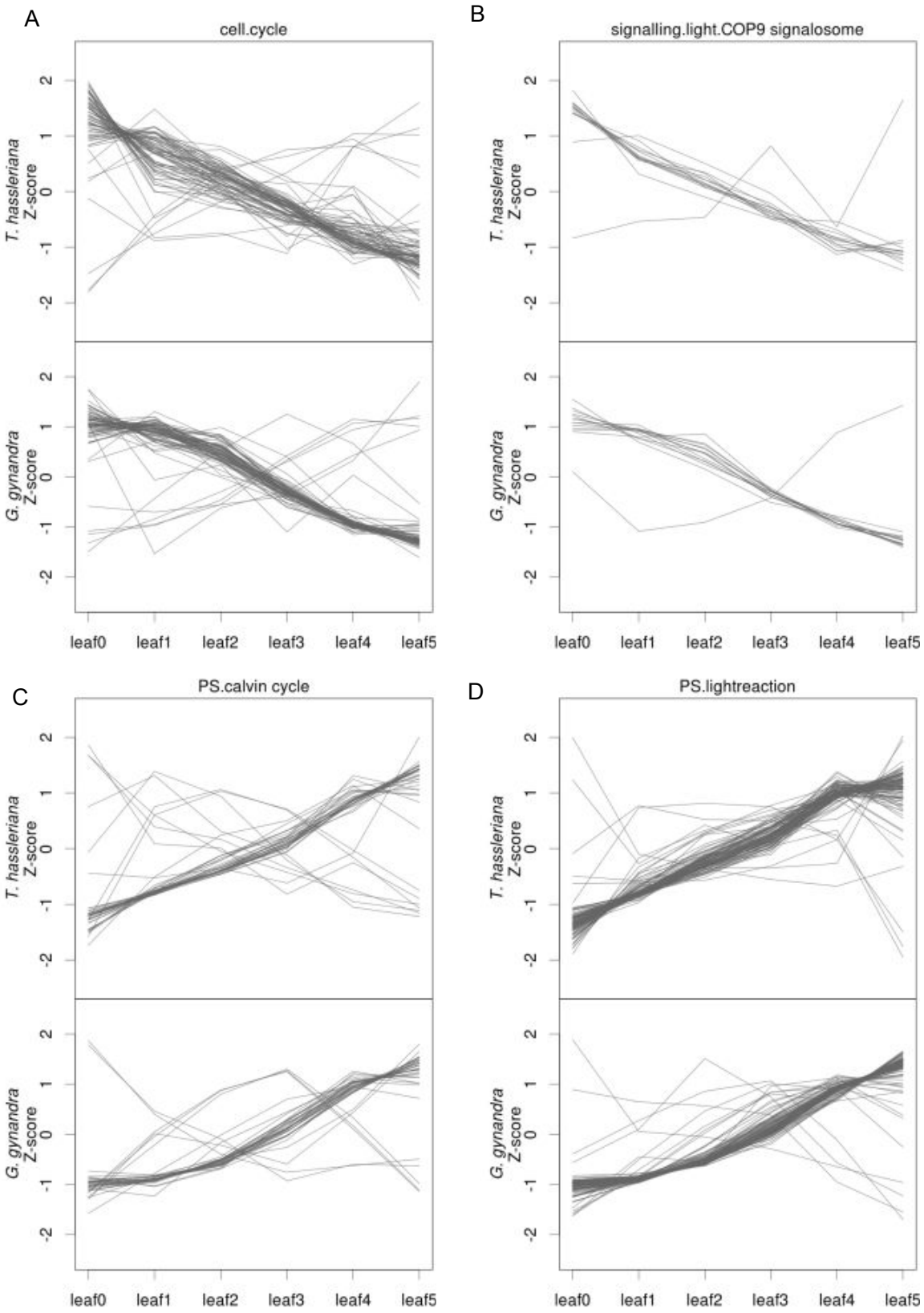
Supplemental Figure 17. Hierarchical clustering with Pearson's correlation of leaf developmental factors. Averaged transcript abundances (RPKM) of leaf gradient sample of transcriptional regulators involved in axial and vasculature fate determination were clustered. Group 1 (orange) and group 2 (red)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



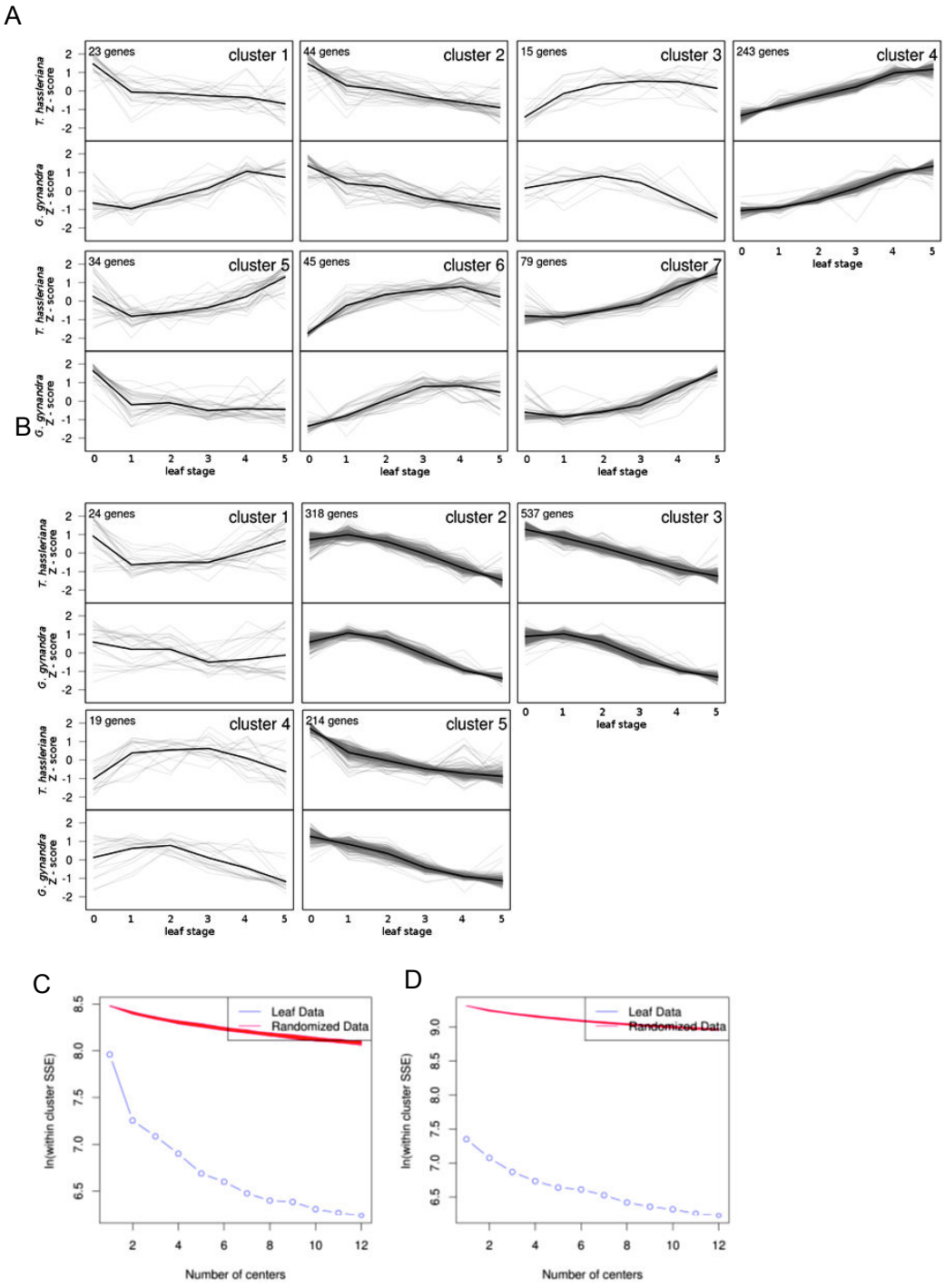
B



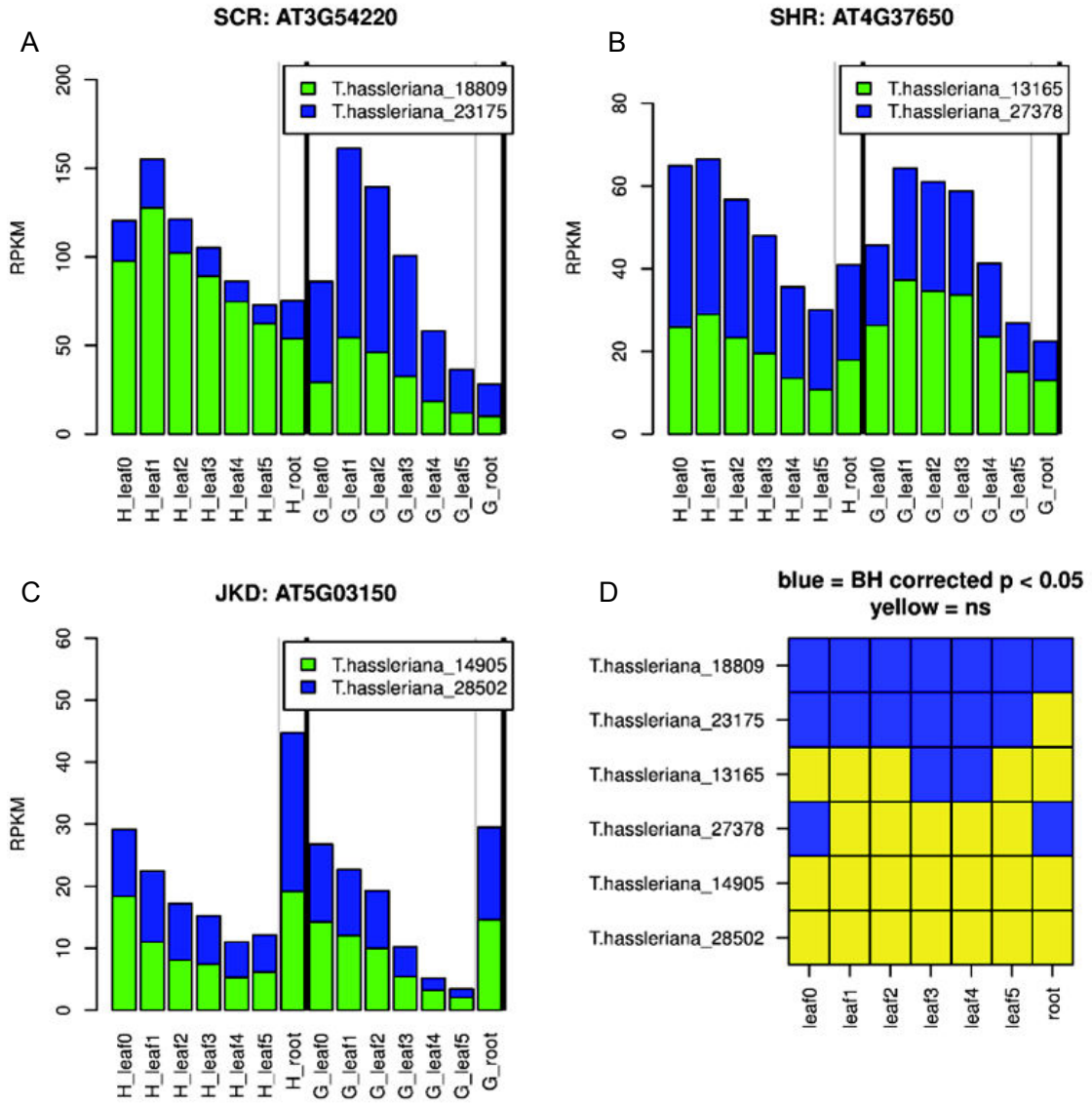
Supplemental Figure 18. *K*-means clustering of leaf gradient expression data and quality assessment. **(A)** *K*-means clustering of transcript abundances (RPKM) of leaf stage averages (*n*=3) between *T. hassleriana* and *G. gynandra* shown as species-scaled Z-scores. Size of each cluster is indicated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



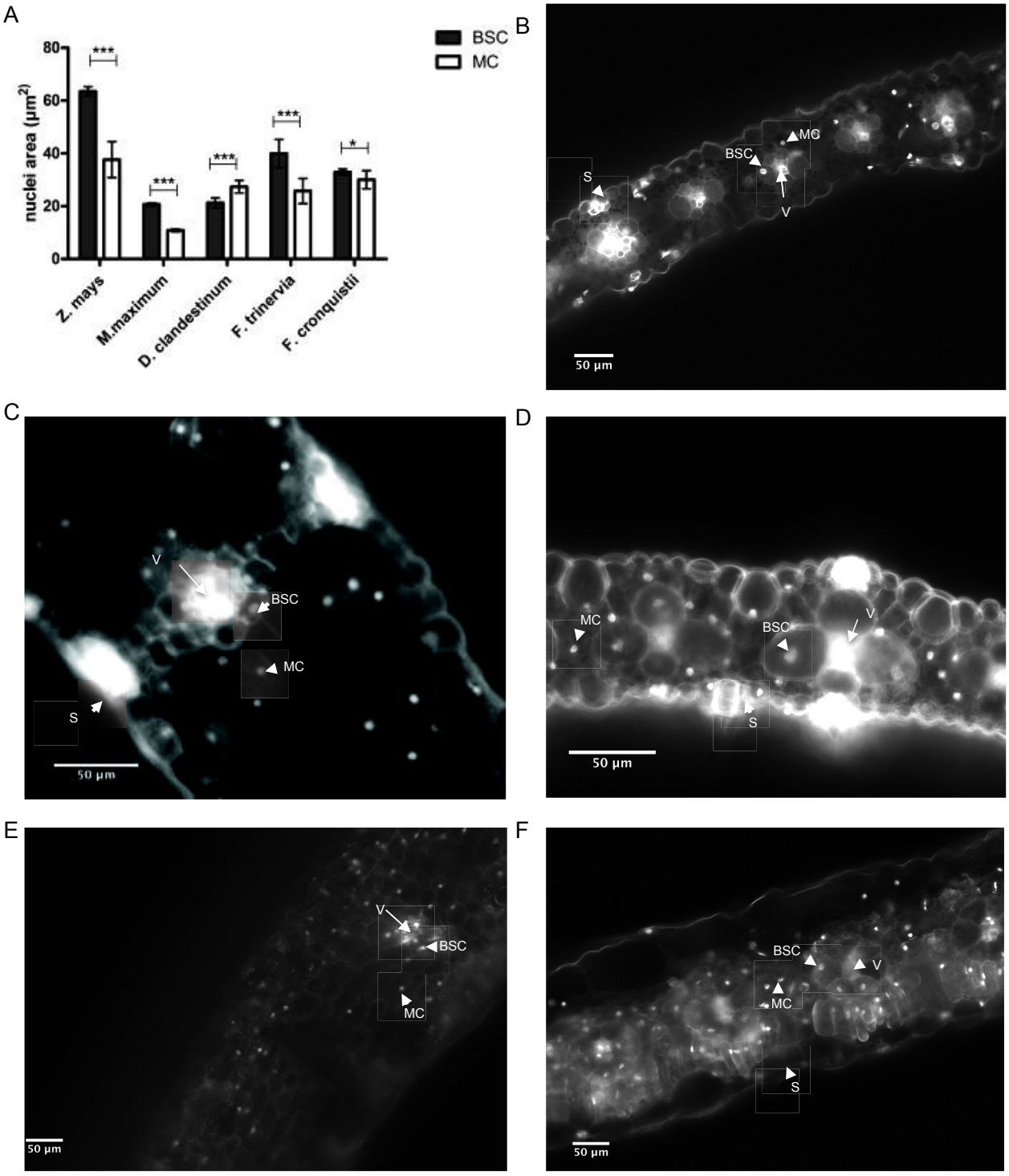
Supplemental Figure 19. Z-score plots of enriched mapman categories in the shifted clusters. Species scaled Z-scores from averaged transcript abundances (RPKM) for each leaf stage per species (n=3). **(A,B)** shifted enriched categories from cluster 4. **(C,D)** shifted enriched categories from cluster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



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



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



Supplemental Figure 22. Nuclei area and images of C_4 and C_3 species.
(A) Quantification of BSC and MC nuclei area of mature leaves of monocotyledonous (*Zea mays*; *Megathyrsus maximus*; *Dichantelium clandestinum*) and dicotyledonous (*Flaveria trinervia*; *Flaveria cronquistii*) C_4 and C_3 species cross sections (error bars ±SD; n=3). Area of nuclei is given as µm^2 with at least 100 nuclei analyzed per cell type per species. Asterisks indicate statistically significant differences between BSC and MC (*** p-value<0.001; * p-value<0.05). **(B-F)** Microscopic fluorescence images of propidium iodide stained mature leaf cross sections of *Zea mays*,C_4 (**B**); *Dichantelium clandestinum*;C_3 (**C**); *Megathyrsus maximus*, C_4 (**D**); *Flaveria cronquistii*, C_3 (**E**); *Flaveria trinervia,* C_4 (**F**). Scale bar: 50 µm; closed arrows pointing to nuclei of indicated cell type. BSC: bundle sheath cell; MC: mesophyll cell; V: vein; S: stomata.

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

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

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

A

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

B

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

C

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

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

A

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

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

B

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

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

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

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

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

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

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

Supplemental Methods

Leaf clearings and safranine staining (Supplemental Figure 1)

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

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

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

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

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

Principal component analysis (Supplemental Figure 8)

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

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

Enzyme Assays (Supplemental Figure 15)

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

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

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

Supplemental References

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

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

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

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

1098.

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

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

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

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

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

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

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