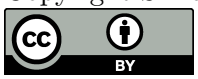
Functional Analysis of C_4-Photosynthesis Based on Next-Generation Sequencing

Simon Schliesky



1 Main Findings

1.1 Technical aspects of next-generation sequencing

1.1.1 Critical Assessment of Assembly Strategies ?

In this study, we tested six assembly algorithms[1] for quality in *de novo* assembly of 454 data. We could show that CAP3 and TGICL are more robust against single base-changes, which we introduced to simulate sequencing errors, as well as biological variance. TGICL and the commercial CLC *de novo* assembly performed best in terms of contig length, redundancy reduction, and chimeric contigs. Furthermore, we showed that the tested graph-based algorithms have difficulties assembling full-length transcripts, even when a high number of reads is available. In contrast, the OLC-based assemblers and the proprietary algorithm by CLC produced mostly full-length transcripts with read numbers above 100.

1.1.2 RNASeq Assembly - Are We There Yet? ?

The increase of read length and the decrease of cost led to a replacement of 454 with Illumina sequencing in many experiments. However, the validation of this method in *de novo* assembly of non-model sequences has not been validated. Here, we showed that the problems identified with 454 reads in our previous study **?** persist. Thus, we suggest that in order to make data comparable across studies, a set of quality control (QC) parameters need to be published along with the data. These QC parameters allow for a judgement of the reliability of a dataset. The major issue with *de novo* assembly of non-model plants is the lack of control over chimeric contigs, for there is no means of detecting them independently of a reference, yet.

1.2 Research on demand (aka collaborations)

1.2.1 The Protein Composition of the Digestive Fluid from the Venus Flytrap Sheds Light on Prey Digestion Mechanisms. ?

The venus flytrap is a carnivorous plant, which can digest insects and small spiders to assimilate nutrients. In this study, the transcriptome and the proteome were sequenced to identify the molecular mechanisms of the prey-response. Proteome data can only be itnerpreted based on a reliable reference sequence. To this end, an RNASeq experiment with *de novo* assembly of the sequences was conducted. The regulation patterns in

[1] Graph-based: SOAP, Velvet, MIRA; OLC-based: CAP3, TGICL; proprietary: CLC

both, transcriptome and proteome, suggest that the digestion system has evolved from defense-related genes.

1.2.2 Impact of SO_2 on Arabidopsis thaliana transcriptome in wildtype and sulfite oxidase knockout plants analyzed by RNA deep sequencing.?

In plants sulfur dioxide acts as an abiotic stress molecule. In this study, the effect of this stress on transcriptional regulation, especially the role of sulfite oxidase (SO) in detoxification, was investigated. New candidates, i.e. an apoplastic peroxidase and defensins, were identified to be coregulated with APS reductase and are most-likely involved in SO_2-stress response.

1.2.3 Analysis of the floral transcriptome of Tarenaya hassleriana (Cleomaceae), a member of the sister group to the Brassicaceae: towards understanding the base of morphological diversity in Brassicales ?

Tarenaya hassleriana belongs to a sister clade of the core Brassicaceae. The morphological diversity in T. hassleriana is much higher than in the Brassicaceae. In an RNA sequencing experiment the key changes between the two lineages were investigated. 5600 transcripts were identified to be specific for the Cleomaceae clade, as there were no homologous genes in Brassicaceae, Brassicales, or Rosids at all. A Comparison of T. hassleriana transcriptome data with A. thaliana microarray data showed 351 differentially expressed genes in the flower transcript levels of both species.

1.3 Main research focus

1.3.1 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 ?

The focus of this study was explaining the PEP-CK subtype of C4-photosynthesis in grasses at a molecular level. A comparative transcriptomics approach was chosen to determine the regulatory differences of a closely related C3/C4-PEP-CK species pair. A reliable sequence resource was created for future experiments. From the quantitative comparison of Megathyrsus maximus (C4) and Dichanthelium clandestinum (C3) transcripts we concluded that the PEP-CK cycle from an engineer's point of view is the easiest C_4 cycle to create. Intracellular transport processes could be grouped to modular transport clusters with simplified net transport reactions. Furthermore, from physiological data we could estimate the extent of intercellular transport necessary for maintaining C4-photosynthesis.

PEP-CK	phosphoenolpyruvate carboxy kinase
RuBisCO	ribulose-1,5-bisphosphate carboxynase oxygenase
OLC	overlap consensus-based
DBG	DeBruijn graph-based
QC	quality control

2 Introduction

2.1 C_4 Photosynthesis & C_4 rice

Around one billion people in the world feed on rice. Rice is a cheap, non-perishable crop. However, with growing population sizes the demand for rice increases, as well. Now, C_4 photosynthesis is considered a very promising trait to cope with this demand by increasing yield.

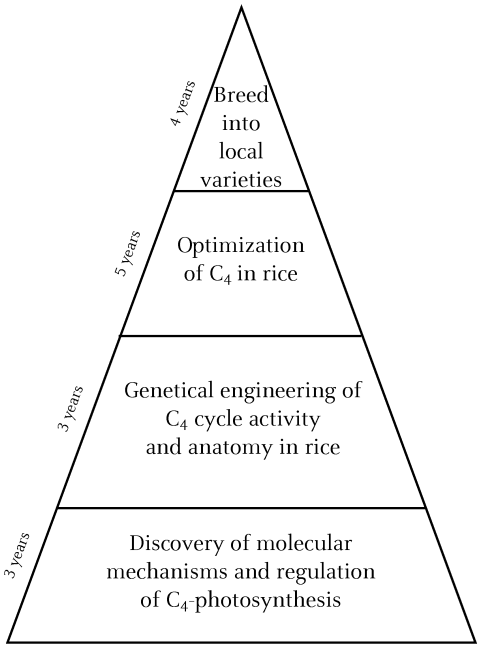


Figure 2.1: Roadmap for the C4-Rice Project (c4rice.irri.org)

2.1.1 C_3 Photosynthesis

Photosynthesis describes the conversion of CO_2 and light energy to sugars. The primary fixation of CO2 is catalysed by ribulose-1,5-bisphosphate carboxynase oxygenase (RuBisCO). CO_2 is attached to ribulose-1,5-bisphosphate (5C) yielding two molecules of phosphoglycerate (3C), hence the name C_3 photosynthesis. However, a side reaction to RuBisCO fixing CO_2 is the fixation of oxygen. The resulting molecule phosphoglycolate needs to be recycled at the loss of CO_2 and energy. This process, called photorespiration reduces the overall efficiency. Carbon fixation in plants performing only C_3 photosynthesis is, therefore, dependent on the CO_2 availability.

2.1.2 C_4 photosynthesis

C_4 photosynthesis is a trait that has evolved independently over 60 times throughout the plant kingdom. Within the *Poaceae* (grasses) the evolutionary origins of C_4 photosynthesis are confined within the PACMAD clade. Plants performing C_4 photosynthesis have evolved to avoid photorespiration and thus reduce energy loss. The mechanism to avoid photorespiration consists of changes in metabolism as well as cell architecture. A

spatial separation of primary carbon fixation and fixation by RuBisCO has evolved by confining the expression of RuBisCO to the bundle sheath cells. In parallel primary CO_2 fixation is catalysed by PEP carboxylase in the mesophyll cells. The transfer of fixed CO_2 from mesophyll to bundle sheath is carried out by transfer acids, such as malate or aspartate. After transfer to the bundle sheath, the fixed CO_2 is released again. Thus the two step fixation of C_4 photosynthesis increases the CO_2 concentration in the vicinity of RuBisCO and suppresses the need for photorespiration. As a consequence, less RuBisCO enzyme is needed. Therefore, C_4 plants present a higher efficiency.

2.1.3 C_4 rice

While the demand for food is growing the area accessible to agriculture is shrinking. One approach to induce a second green revolution is genetically engineering rice into a C_4 plant. Expectations are not only that the yield will increase in current environments, but also rice will become accessible to more climate regions. The International Rice Research Institute has laid out a roadmap to reach this goal. In general, the project can be summarised as: Understand, Imitate, Optimise, and Breed. That means, as depicted in figure 2.1, they suggest a coordinated research effort over the next 15 years. Starting with a detailed analysis of all molecular components involved in C_4. Followed by engineering efforts to establish a C_4 like metabolism in rice. Subsequently improving the cycle yield within rice transgenics. Finally crossing rice transgenics with existing rice cultivars to generate location-adapted C_4 rice plants. In terms of this roadmap, my work focuses on the analysis of the C_4-cycle within grasses. It aims at describing a set of parts and interconnections in order to build a blueprint for re-engineering C_4 photosynthesis. Thus the main driving force of my work was the question:

> *Which transcripts are involved in creating the difference between C_3 and C_4 photosynthesis in two closely related grass species?*

2.2 Next-Generation Sequencing

Sequencing DNA fragments has already been described in the late 70s. One chemical approach, as well as an approach based on PCR was presented. Despite optimisations, like fluorescent dyes and capillary electrophoresis, the throughput of these traditional approaches was rather low. The second generation of sequencing approaches includes the ones used in this work. Characteristic for these sequencing approaches is that they provide a high sequence throughput and in contrast to the first generation allow for a reliable estimation of sequence abundance in the sample. A new third generation of sequencing platforms has been introduced in recent years. As a major difference, third generation sequencing allows for sequencing of actual DNA molecules and therefore shows no bias of the amplification steps, some second generation sequencing methods suffered. The reliability of these platforms is still controversially discussed.

2.2.1 Sequencing platforms used during PhD project

Pyrosequencing (454)

Massively parallel pyrosequencing is an approach that has been commercially launched by 454 Life Sciences and was later acquired by Roche. DNA fragments are randomly distributed over a picotiter plate and subsequently amplified in the picoliter wells. This renders the DNA immobile and keeps the position of each fragment consistent during sequencing. Chemically the sequencing approach detects the pyrophosphate, which is released upon insertion of a nucleotide during DNA synthesis inside of each well. A combined sulfurylase:luciferase enzyme creates light emission for each pyrophosphate molecule released. Therefore, a CCD camera module can capture the amount of light generated per well. To sequence, the machine repeatedly adds dNTPs separately, so the light information collected is assigned to a certain nucleotide for each fragment. This sequencing approach suffers from long homopolymer stretches, because the light intensity detection becomes ambiguous. Read lengths of up to 800 bp are possible on GS FLX with Titanium chemicals. The number of reads is about a million per picotiter plate.

SOLiD

Sequencing by ligation has been introduced by Applied Biosystems (now Life Technologies) with the SOLiD platform. In this approach a fluorescent dye labeled oligonucleotide is binding two nucleotides of the template strand. The readout is a color value. One of the two bases is ligated and the process repeats with the position shifted by one. This way, each base is read twice, which increases the accuracy to 99.8%. Colorspace sequences can be converted to basespace by interpreting the sequence of colors as base-pairings. Even though this increases accuracy, one skipped readout is sufficient to invalidate the rest of the read. In other words, the information of each base is dependent on the information of the previous base. SOLiD sequencing can generate up to 30G bases with a maximum length of 85 bp.

Illumina

The use of modified ddNTPs to identify inserted bases during strand synthesis categorises Illumina sequencing as sequencing by synthesis technology. The provided nucleotides are enhanced with a cleavable fluorescent dye and a removable blocking group. The blocking group prevents multiple base insertions and leads to a one base at a time readout. A CCD camera captures the light emission of the fluorescent labels during laser excitation. Because of the blocking groups, Illumina sequencing has a fixed read length. Currently, the output of Illumina can reach up to 600G bases with a read length of up to 2x150 bp. Comparing the three sequencing technologies, Illumina has the highest rate of error with up to 2%. Working towards quantitative as well as qualitative sequence data:

> *What is the best sequencing approach to answer my experimental question?*

What is the best platform?

Choosing a sequencing platform is highly dependent on the experimental design. Sequencing *de novo*, i.e. without a reference genome, is considered easier if the read length is higher. Whereas, analysing statistical differences in gene regulation benefits from a high number of reads. Finally, the costs, accuracy, and availability of analysis software need to be considered. Therefore, there are different best platforms for a single use-case, but not a single best platform for all use-cases.

2.2.2 Assembly & mapping

High throughput sequencing technologies have arisen only recently. The rapid changes and improvements in sequencing technologies are accompanied by changes in big data analytics, as well. Therefore, the computational methods available for traditional sequencing need to be re-validated. Part of the work presented here focused on this validation especially towards *de novo* transcriptome sequencing, where the outcome is unpredictable.

De novo assembly

In contrast to reference-based sequence assembly, where the read sequences are matched to a known genome, in *de novo* assembly, the full-length information for a gene must be extracted from the read information alone. Two conceptually different solutions are currently known: overlap consensus-based (OLC) and DeBruijn graph-based (DBG) assembly.

In OLC assemblers all reads are compared to each other (pairwise or grouped) and assembled into contigs, where overlaps exist. This method is optimised for few and long sequences. The number of comparisons, hence the runtime, as well as the memory consumption increase more than linear with read numbers. Algorithms based on the OLC principle differ in the pre-selection of reads to compare or in the error tolerance.

DBG assemblers employ graph theory to solve the overlap problem. That means, a read is represented by a sequence of k-mer nodes, for which every node overlaps in k-1 positions with the neighboring nodes. This approach transfers the problem of finding overlaps in all against all reads to finding supported paths through the graph. DBG assemblers can handle much higher read numbers because memory consumption and computational power required are dependent on the number of unique k-mers used and not necessarily the number of reads. With many algorithms at hand:

> *What is the best way to assemble the generated 454 & Illumina reads to get reliable transcript sequences?*

Read mapping

In order to get quantitative information about gene expression, the sequenced reads need to be assigned to genes. Multiple approaches have been developed for read mapping to a

reference genome. One of the most commonly used mapping algorithms to achieve this is bowtie. Bowtie is optimised for mapping speed of nearly perfect sequences. However, this algorithm does not work on *de novo* sequencing data. In *de novo* approaches a reference genome is not available or not used on purpose. Thus, the read mapping is transferred to a reference sequence of a species that is closely related to the sequenced species. Even short evolutionary distances between species can account for many exchanges in the nucleotide sequence. Thus, mapping software like bowtie is unable to match many reads. Traditional programs, such as BLAST or BLAT, which are not optimised for speed, allow for mapping in proteinspace, i.e. translated nucleotide sequence mapped in all possible frames. Since there is redundancy in codon to amino acid translation, proteinspace mapping is not affected by synonymous base changes, and thus lenient about the actual nucleotide sequence. Hence, the question:

What is the best mapping approach for this study?

2.2.3 Statistics

One of the main challenges in statistics on transcriptome data is the dynamic range of read abundance. This abundance can range over five orders of magnitude. Many of the published Statistics for next-generation sequencing were initially designed for micro-array experiments. In these experiments the differential expression of genes was determined by a fold-change comparison between sampe and control. However, the dynamic range and the error types are different. Thus, the statistical methods to evaluate differential gene expression need to be adapted for RNA-Seq. Throughout this study, the statistical methods were subject to many changes and adaptations.

Under the assumption that each read is randomly drawn from a population of reads, **?** showed that this process can be explained by a binomial distribution and approximated by a Poisson distribution. This approach was provided in an R package called "'DEGSeq"'. Later, **?** could show, that modelling the stochastics of read sequencing is more accurate when using a negative binomial distribution. This approach was also provided in an R package called "'DESeq"'. None of these broadly accepted approaches for RNA-Seq statistics have found the golden solution, yet. The development and improvement of the statistics methods is still an ongoing process. Therefore an important decision was:

Which statistic model is the most reliable to detect differential expression in RNA-Seq approaches?

2.3 Personal Motivation

Main driver of my research was the urge to understand the C4 cycle in a depth that allows for building such a metabolic unit from scratch. So, with the fundholders' overall goal

of re-engineering C_4 crops in mind, during my research, I aimed at understanding the underlying molecular mechanisms in the C_4 grass *Megathyrsus maximus*. At the time of experimental design, the method of choice to achieve this was next-generation sequencing It was the most promising approach to provide a comparative and comprehensive insight into all involved genes in the phosphoenolpyruvate carboxy kinase (PEP-CK) type C_4 cycle as opposed to the C_3 cycle. Furthermore, no C_4 grass of the PEP-CK type has been investigated in such a comparative high-throughput experiment, before. Thus, my research was driven by the question:

> *Which transcripts are responsible for the functional mechanisms in Megathyrsus maximus that induce a C_4 cycle in contrast to a C_3 cycle in Dichanthelium clandestinum?*

As a side project I investigated, which sequencing techniques in combination with which processing approaches are suitable for answering this question. In the end, the prospect of an extensive automation and data processing approach tempted me into working on this project.

3 Conclusion

3.1 Next-Generation Sequencing

A variety of sequencing platforms, software, and use-cases is present in sequencing experiments. In our study, we wanted to evaluate the transcriptomes of non-model species (i.e. species without a sequenced genome). At the time of the initial experimental design, 454 sequencing was the commonly favorited approach for sequencing non-model species. However, it remained unclear, whether OLC or DBG assemblers were more successful in assembling full-length contigs for each transcript. To find an answer to this question, we used the *Arabidopsis thaliana* genome as a well known reference. Besides, we were interested in evaluating the assumption[1]: More reads lead to a better (i.e. more complete) assembly. From this reference we generated simulated 454 reads, based on actual sequencing data from former studies in *Cleome* **?**. These simulated reads were perfect in terms of sequencing errors. Therefore, we additionally created read libraries with artificial 1%, 3%, and 5% base changes. With the 4 read libraries as input sequences, we used the assembly software: **SOAP!** (**SOAP!**)**?**, **Velvet!** (**Velvet!**)**?**, **MIRA!** (**MIRA!**)**?**, **CAP3!** (**CAP3!**)**?**, **TGICL!** (**TGICL!**)**?**, and **CLC!** (**CLC!**)**?**. Traditional quality parameters to assess a *de novo* transcriptome assembly were taken from genome assembly. These metrics, namely N50, maximum contig length, number of contigs, describe the size and distribution of the contig library. As a rule of thumb, in genome assembly larger contigs represent a better assembly. In transcriptome assembly, however, this rule does not apply. A very long contig can result from the assembly of multiple genes. To describe the quality of an assembly in a less biased way, we used percentage of full-length transcript, i.e. the relative length of a contig to it's best hit reference gene's length. As a visualization we plotted number of reads used for the assembly against percentage of full-length transcript.

With perfect reads in all assembly software the contigs with ¿200 reads were mostly 100% assembled. However, in DBG assemblers several transcripts did not get assembled more than 20% - 60% of the full-length transcript. We did not find this phenomenon in OLCassemblers. Thus, we can conclude, that more reads make an assembly even worse, if using DBG assemblers, but have no negative effect on OLCassemblies. With increasing percentage of simulated sequencing errors (i.e. random base changes) the difference becomes less significant.

When using cheaper and less error prone Illumina sequencing, the numbers of reads one sequencing run yields increase drastically. Thus, OLC assemblers are not feasible if at all able to assemble the reads in terms of computational power. Yet, due to decreasing costs,

[1] It really was just a guess, but I might need to call it hypothesis in my final version

the number of transcriptome sequencing experiments in the public databases and the variety of used assembly software are increasing. As was shown before, RNA sequencing experiments are sensitive to uncontrolled experimental parameters. On top of that, many different assembly algorithms are available and used. Therefore, comparability of public data resources is not ensured. To cope with this, we suggested a set of standard quality measures, that can be obtained regardless of sequencing and assembly technology. The aim was making public datasets more accessible by providing additional information along with the sequences. These information do not alter nor improve the assembly quality, but they allow for a reliable quality estimate. Hence, the comparability of sequencing experiments can be determined based on the quality parameters.

The quality assessment parameters we suggest:

- number of contigs
- N50 of contigs
- Venn diagram of contigs mapping and reads mapping to reference gene
- percentage of contigs mapping to reference
- number of hybrid/chimeric contigs
- type of hybrids
  - read-through reads of neighbouring genes
  - fusion genes

We tested all these parameters in an actual sequencing of Arabidopsis thaliana transcriptome against the Arabidopsis thaliana reference genome (TAIR10). Compared with the theoretical values for the aforementioned parameters the experiments showed us that we are still far from ideal assemblies. Yet, the suggested parameters allow for a rating of the completeness and reliability of the assembled transcriptome. Until now, there is no approach to detect chimeric contigs in transcriptome assemblies, that is independent of a reference genome. Thus, detecting chimeric contigs in a non-model species without a reference is not possible so far. Having chimera in the set of contigs leads to an unpredictable shift of expression values in all involved transcripts. The problem of dissecting chimeric contigs into their underlying transcripts still needs to be solved. Only then, expression levels can be unambiguously determined genome wide.

An approach using expression information to identify the fusion positions did not succeed (figure 3.1). We tried to identify the fusion sites based on expression coverage of the underlying sequences. Unfortunately, we were not able to find a suitable mathematical model to predict these sites reliably. As a consequence we were not able to cut fused sequences into their originating transcripts without a priori knowledge about the genome.

RNA sequencing is a rapidly evolving technology, so a perfect assembly of the actual transcriptome might be possible with single molecule sequencing techniques. However,

as of now, these new third generation sequencing technologies are not broadly available, yet. Thus, most experimentalists continue to use short-read technologies and depend on reliable assembly algorithms.

3.2 Application of NGS to single gene or gene family research

3.2.1 Sulfite oxidase in Arabidopsis

3.2.2 Floral timing in Cleome

3.3 Genome wide transcript analysis

3.3.1 Dionaea

3.3.2 Megathyrsus maximus

- PEP-CK/NAD-ME cycle balance
- chloroplast transport balancing
- ATP for PEP-CK reaction provided through NADPH
- Preventing metabolites from leakage to secondary metabolism
- EC and Pfam-Domain combinations
- intercellular transport

Figure 3.1: The read expression across chimeric genes does not reveal the positions of gene fusion marked by dashed lines (determined by mapping to reference genome). (This figure was presented as part of a poster on BioComp2012)

4 Appendix

4.1 Publications

4.1.1 Copyright information

Following are hard-copies of the publications I worked on during my PhD. Besides the legal code required by the journals please find the author attributions in the Bibliography section of this thesis.

?: Permission for reproduction in my thesis was granted by Oxford University Press through Copyright Clearance Center. You can request a copy of the license document by emailing to simon.schliesky@hhu.de

?: MCP grants authors the right to reuse the publications with the following attribution: This research was originally published in Molecular & Cellular Proteomics under DOI 10.1074mcp.M112.021006

? New Phytologist grants authors the right to reproduce the publication for use in theses.

?, **?**, **?**: These publications are open access and thus published under the Creative Commons license

4.1.2 Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C_3 and C_4 species

Journal of Experimental Botany, Vol. 62, No. 9, pp. 3093–3102, 2011
doi:10.1093/jxb/err029 Advance Access publication 11 March, 2011

Journal of Experimental Botany
www.jxb.oxfordjournals.org

RESEARCH PAPER

Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C_3 and C_4 species

Andrea Bräutigam, Thomas Mullick, Simon Schliesky and Andreas P. M. Weber*

Plant Biochemistry, Universitätsstrasse 1, Heinrich-Heine-University, D-40225 Düsseldorf, Germany

* To whom the correspondence should be addressed. E-mail: andreas.weber@uni-duesseldorf.de

Received 5 November 2010; Revised 17 January 2011; Accepted 19 January 2011

Abstract

Next-generation sequencing enables the study of species without a sequenced genome at the 'omics' level. Custom transcriptome databases are generated and global expression profiles can be compared. However, the assembly of transcriptome sequence reads into contigs remains a daunting task. In this study, five different assembly programs, both traditional overlap-based, 'read-centric' assemblers and de Bruijn graph data structure-based assemblers, were compared. To this end, artificial read libraries with and without simulated sequencing errors were constructed from *Arabidopsis thaliana*, based on quantitative profiles of mature leaf tissue. The open source TGICL pipeline and the commercial CLC bio genomics workbench produced the best assemblies in terms of contig length, hybrid assemblies, redundancy reduction, and error tolerance. The mature leaf transcriptomes of the C_3 species *Cleome spinosa* and the C_4 species *Cleome gynandra* were assembled and analysed. The pathways and cellular processes tagged in the transcriptome assemblies reflect processes of a mature leaf. The databases are useful for extracting transcripts related to C_4 processes as full-length or nearly full-length sequences.

Key words: Assembly, C_4, next-generation sequencing, transcriptome.

Introduction

Sequence information, both qualitative (the sequence itself) and quantitative (how much each transcript is expressed), is important for the analysis of any trait at the molecular level. Next-generation sequencing (NGS) technologies have recently become widely available, and the creation of custom transcriptomes a possibility (Weber *et al.*, 2007; Novaes *et al.*, 2008; Alagna *et al.*, 2009; Barakat *et al.*, 2009; Dassanayake *et al.*, 2009; Wang *et al.*, 2009; Kumar and Blaxter, 2010). Using NGS to study the transcriptome of a species without a sequenced genome, such as a C_4 species other than *Zea mays* or *Sorghum bicolor*, simultaneously produces a transcriptome database for the tissue sampled as well as an expression profile of the tissue (Bräutigam *et al.*, 2011). It is even possible to compare the expression profiles of two different species, for example a C_3 and a C_4 species, with one another (Bräutigam and Gowik, 2010; Bräutigam *et al.*, 2011).

Different NGS technologies are currently available commercially (Metzker, 2010). Common to all available NGS technologies is that they produce much more sequence information compared with traditional Sanger sequencing at a much lower cost. However, there is no free lunch (yet), and with current technologies the payment is in short reads, from 36 bases (with Illumina technology; longer reads of up to 100 bases are possible at increased cost), over 75 bases (with SOLiD technology), and to ~450 bases (with Roche/454 technology). Roche/454 technology will produce fewer reads per run than Illumina and SOLiD, though. Although the sequence reads themselves are longer, the total sequence output is only about one-tenth of Illumina's output and ~1/100th of SOLiD's output. The possible applications of the long read and short read technologies in the context of plant research and their advantages and disadvantages have been reviewed in detail (Bräutigam and Gowik, 2010).

Briefly, in species without a sequenced genome, longer reads will facilitate both the assembly of a transcriptome database and the reliable quantification of expression. Therefore, the only technology which gives us reads >200 bases is currently used: Roche/454.

If transcriptome sequence information is generated for a species without a sequenced genome, two analyses are possible: the quantification of expression by aligning (also referred to as mapping) the reads to a related reference genome as explored by Bräutigam *et al.* (2011) and the assembly of the transcriptome to provide qualitative sequence information. A read mapping is prone to errors if the reads do not exactly match the reference (Palmieri and Schlotterer, 2009), a caveat certainly true for mapping only to a related reference genome and not the genome itself. Longer reads ensure a more accurate mapping (Palmieri and Schlotterer, 2009). BLAT has been shown to map reads reliably when the method was applied to compare the expression profiles of a C_3 and a C_4 species (Bräutigam *et al.*, 2011). In the quantitative comparison of the *Cleome* species *C. gynandra* and *C. spinosa*, the steady-state transcript levels of genes associated with C_4 photosynthesis were increased between 20- and 250-fold in the C_4 species, with the exception of malate dehydrogenase. This global approach also identified candidate genes for C_4-related processes, such as intra- and intercellular metabolite transport, as well as candidates for regulators, which maintain the C_4 state in mature leaves. Moreover, genes for protein biosynthesis, such as genes encoding ribosomal proteins, were more frequently down-regulated in the C_4 species, as were many of the genes encoding Calvin–Benson cycle and photorespiratory enzymes, indicating that down-regulation of ribosomal proteins in the cytosol and chloroplast may be contributing to nitrogen efficiency in some C_4 species.

However, the assembly of NGS reads remains a challenge. This is especially true for highly dynamic transcriptome read libraries. In principle, two different types of assemblers are available: a 'read-centric' overlap-based assembler, which has been used for assembling Sanger sequences, and an assembler specifically developed for handling the large amounts of reads provided by NGS, which is based on de Bruijn graph data structures (Flicek and Birney, 2009). However, the new type of assemblers have been developed for assembling genomic rather than transcriptomic sequence libraries (Flicek and Birney, 2009). While transcriptome libraries have dynamic ranges of 5–6 orders of magnitude between the highest abundant and the lowest abundant transcripts and their reads (Bräutigam *et al.*, 2011), genomic libraries ideally have no dynamic range. Traditional assemblers include CAP3 (Huang and Madan, 1999), TGICL, which is a pipeline of a megablast-like tool connected to the CAP3s clustering algorithm (Pertea *et al.*, 2003), and MIRA (Chevreux *et al.*, 2004). New assemblers of the de Bruijn graph type are, for example, SOAPdenovo (http://soap.genomics.org.cn/soapdenovo.html) and Velvet (Zerbino and Birney, 2008). Commercial programs such as the CLC bio genomics workbench do not fully disclose the type of assembler integrated into the program. In plants, several attempts to reconstruct a plant transcriptome from NGS reads have been published with different assembly programs without any tests of whether one assembler outperformed any other (Novaes *et al.*, 2008; Alagna *et al.*, 2009; Barakat *et al.*, 2009; Dassanayake *et al.*, 2009; Wang *et al.*, 2009). Such a critical assessment of assembler performance and suitability is conducted in this study.

The C_4 syndrome, a complex trait evolved to concentrate carbon in the vicinity of RubisCO, has originated in >45 plant lineages in a striking example of convergent evolution (Sage, 2004). It serves to minimize the oxygenation reaction of RubisCO while maximizing CO_2 fixation. C_4 plants thus are either able to accumulate biomass much faster than plants without this carbon concentration mechanism (e.g. *Z. mays* or *S. bicolor*) or able to live in adverse conditions that minimize CO_2 availability to and fixation by RubisCO, such as water limitation, heat, or poor soil conditions (Sage, 2004).

C_4 plants have a biochemical CO_2 pump: CO_2 is fixed by phospho*enol*pyruvate carboxylase (PEPC), an enzyme insensitive to O_2 but with a higher affinity for HCO_3^-, in the mesophyll cells, whilst RuBisCO is localized to the bundle sheath cells. Several different enzymes and transport proteins transfer the CO_2 as an acid with four carbon atoms, decarboxylate it to release the CO_2 in the vicinity of RubisCO, return a C_3 acid to the site of PEPC, and regenerate the CO_2 acceptor (Hatch, 1987). While the fixation of CO_2 using the acceptor phospho*enol*pyruvate (PEP) is always accomplished by the same enzyme, the C_4 transfer acids can be malate and/or aspartate and the C_3 transfer acid can be pyruvate, alanine, or PEP. Three different decarboxylation enzymes liberate the CO_2, NAD-dependent malic enzyme (NAD-ME), NADP-dependent malic enzyme (NADP-ME), and PEP carboxykinase (PEP-CK) (Hatch, 1987). The spatial separation of initial and final carbon fixation may be concomitant with the spatial separation of other anabolic pathways such as nitrogen and sulphur assimilation (Majeran *et al.*, 2005) as well as limited oxygen production at the site of RubisCO and increased ATP production (Meierhoff and Westhoff, 1993). In addition to this biochemical CO_2 pump, several adaptations on the cellular and tissue levels are necessary (Hatch, 1987). The majority of C_4 species, in terms of both species number and contribution to global biomass production, spatially separate RubisCO and PEPC in two different cell types called the mesophyll and the bundle sheath. For efficient C_4 photosynthesis, CO_2 released in the vicinity of RubisCO must not leak out of cells. A barrier is established either by a cell wall reinforced with lignin and/or suberin (Evert *et al.*, 1977) or by positioning the RubisCO-containing chloroplasts 'in the way' of loss by diffusion (Muhaidat *et al.*, 2007). While diffusion of CO_2 must be prevented, diffusion of the transfer acids must not only be allowed, but must also be very effective to accommodate the immense metabolite flux through the C_4 cycle, which operates at or above the speed of carbon fixation (Laisk and Edwards, 2000; Weber and von Caemmerer, 2010). Although all enzymes required for the C_4 cycle are characterized at least

at the biochemical level and many also at the molecular level (Hatch, 1987), the majority of molecular changes underlying the tissue and cellular adaptations, the intra- (Bräutigam *et al.*, 2008; Majeran *et al.*, 2008) and intercellular transport processes, such as plasmodesmatal regulation (Botha, 1992; Sowinski *et al.*, 2008), and most of the regulatory changes are unknown.

Forty-five origins of C_4 photosynthesis (Sage, 2004) provide at least 45 possible contrasting pairs of C_3 and C_4 plants to study. This study focuses on the C_4 plant *C. gynandra* (spider wisp, also known as 'African cabbage') and a C_3 relative, *C. spinosa* (spider plant). *Cleome gynandra* is currently known as the most closely related C_4 plant to *Arabidopsis thaliana* (thale cress) (Brown *et al.*, 2005). It is a leafy annual plant from the African continent and forms part of the diet in African countries (van Rensburg *et al.*, 2004). Within the genus Cleomaceae, there are other species that show characteristics of C_4 plants, such as carbon isotope discrimination in, for example, *C. angustifolia* and *C. oxalidae* (Marshall *et al.*, 2007). Since *C. gynandra* is easy to cultivate (it is considered an invasive weed in the USA; http://plants.usda.gov/java/profile?symbol=CLGY) and since it is a food plant for which seeds can be obtained in quantity from retailers, it is an attractive choice as an experimental organism. Transformation has recently been achieved (Newell *et al.*, 2010). *Cleome gynandra* is an old world plant and probably a basal branch within the Cleomaceae, although these basal branches are not well supported in phylogenetic trees (Inda *et al.*, 2008). A genome duplication event in the lineage of *C. gynandra*, which has 16 or 17 chromosomes, has been speculated about based solely on chromosome number (Inda *et al.*, 2008). The choice for the C_3 plant in the comparison pair is not obvious. The basal branch of *C. gynandra* does not contain a C_3 relative and, based on the evolutionary trees available (most recent in Inda *et al.*, 2008), the C_3 relatives are equidistant. The spider flower *C. spinosa* was chosen for its ease of cultivation and availability. *Cleome spinosa* is an ornamental plant which originated from South America. It is a member of the new world Cleomaceae with a chromosome number of $x=8$ or 9 (Inda *et al.*, 2008). Both plants can be cultivated alongside each other in a glass house or growth chamber. Under identical conditions in well-watered, rich soil, the C_3 plant will outgrow the C_4 plant. Extensive sequence information was not previously available for both species (Bräutigam *et al.*, 2011).

To provide not only the quantitative information (Bräutigam *et al.*, 2011) but also the best possible qualitative sequence information for the C_4 model *C. gynandra*, several different assemblers were tested in this study. To identify the most suitable assembler, a gold standard is needed against which the performance of the assemblers can be benchmarked. A simulated and therefore artificial read library, which represents the dynamics of a mature leaf transcriptome, was modelled based on the known *A. thaliana* genome. Assemblies were compared given that the ultimate best possible outcome, the real transcriptome, is known in this case. After the best assembler was determined, the sequence reads from both *Cleome* species were assembled, single nucleotide polymorphisms (SNPs) were annotated, and the transcript representation was analysed. The results of this study together with the recently published results of the transcriptome quantification (Bräutigam *et al.*, 2011) enable the identification and study of transcripts involved in maintaining C_4 tissue and cell architecture, and regulating and executing C_4 photosynthesis in mature leaves.

Materials and methods

Sequence read generation

The sequence reads from *C. gynandra* and *C. spinosa* were generated as described in Bräutigam *et al.* (2011). They are available at NCBI's short read archive under the accession numbers SRS002473 and SRS002474.

Generation of the simulated read database

Testing an assembly from sequencing reads that were generated from a species without a sequenced genome is problematic since the solution to the assembly, the correct transcriptome, is unknown. On the other hand, producing sequencing reads from a species with a sequenced genome for the express purpose of testing assembly programs is prohibitively expensive. To overcome these limitations, the quantitative information generated in Bräutigam *et al.* (2011) was used. If a Perl script draws the number of reads determined in Bräutigam *et al.* (2011) from each transcript and randomly distributes the reads along the length of the transcript, a simulated read library which reflects a C_4 transcriptome read library will be generated. This method does not take any 5' or 3' bias into account since it is not known whether the *Cleome* read libraries are indeed biased. The script used in this study is given in Supplementary Fig. S3 available at *JXB* online. In a second step, sequence variation was introduced. The most common sequencing error with 454 technology is miscalled homopolymer stretches; in other words, if the same nucleotide occurs multiple times in a row, the software may miss or add a nucleotide to the stretch. A coding sequence, the target of a transcriptome project, very rarely has homopolymer stretches. To determine the pattern of sequence variation, which may have resulted from either genetic variation in the sample or incorrect base calls during sequencing, reads from *C. gynandra* were mapped onto the *C. gynandra* unigenes generated in Bräutigam *et al.* (2011), and for each position nucleotide differences were annotated (Supplementary Fig. S3, three examples shown). The nucleotide substitutions appeared random (Supplementary Fig. S3, three examples shown). Therefore, to generate simulated read libraries with sequence variation present, error rates of 1, 3, and 5% were introduced at random positions in the read library with a Perl script. This script is also available in Supplementary Fig. S3.

Assembly

The source code for all assemblers except CLC is publicly available and was downloaded from public repositories. A free CLC trial version was downloaded from the company's website. All assemblers were run on a Linux Workstation with Dual Core CPU and 8 Gb of RAM. For both SOAP and Velvet, the k-mer size for the assembly was set to 19. MIRA was started with recommended parameters for expressed sequence tag (EST) assembly, using default settings otherwise: denovo, est, accurate, 454. CAP3 was run with default parameters and therefore an overlap identity requirement of 75% and minimal overlap length of 30 bases.

Table 1. *Assembly results of the artificial libraries (perfect, 1% error rate, 3% error rate, and 5% error rate) with five different assembly programs*

		SOAP	Velvet	MIRA	CAP3	TGICL	CLC
Perfect	Remaining sequence length in X%	14	10	11	10	11	11
	No. of contigs	27 315	21 059	14 064	12 518	12 277	11 787
	N25	892	840	984	1060	1110	1162
	N50	476	487	604	644	688	732
	No. of hybrids	434	29	190	180	190	184
1% error	Remaining sequence length in X%	40	12	18	10	11	10
	No. of contigs	646 735	66 298	31 027	12 747	12 338	11 707
	N25	147	431	676	1020	1101	1137
	N50	39	233	413	616	678	718
3% error	Remaining sequence length in X%	75	15	36	12	10	9
	No. of contigs	1 473 433	134 624	79 863	17 971	12 889	10 746
	N25	59	206	440	658	1003	1077
	N50	39	86	250	458	611	689
5% error	Remaining sequence length in X%	98	22	42	18	23	8
	No. of contigs	1 984 011	236 302	98 720	33 505	38 364	9837
	N25	61	95	392	475	536	890
	N50	39	66	250	383	418	593

TGICL is a pipeline consisting of a megablast-like tool, which was run with default parameters (overlap of at least 40 bases and 94% identity) which pre-clusters the reads into bins each of which contains only reads that overlap at least partially. After clustering, CAP3 addresses each cluster separately for assembly (parameters identical to previous assembly). CLC was run with default settings. After the assemblies, assembly parameters were calculated with Perl scripts and Linux commands.

SNP identification

Reads were aligned to the reference contig sequences produced by CLC using proprietary tools integrated into the CLC genomics workbench software suite, and an SNP was called if at least six reads covered the position, if at least two different nucleotides were present with at least two different reads each, and if 40% of the detected variation was one of the nucleotides. After automatic detection, the list was manually curated to include only SNPs with frequencies between 30% and 70%.

Mapping to Arabidopsis *and quantification*

Contigs from databases were mapped to *Arabidopsis* in protein space with BlastX. Annotation parameters were extracted using Perl scripts and Linux commands.

Results and Discussion

Determination of the most suitable assembly program

To produce the best assembly possible with non-normalized read libraries from the C_4 plant *C. gynandra* and its C_3 relative *C. spinosa*, different assembly programs were tested. If the assemblers are used with reads that have no reference genome available as a gold standard; that is, the *Cleome* reads, it is impossible to test the assemblies efficiently. Hybrid contigs, which are assembled from two different transcripts, and transcript coverage cannot be assessed since the transcripts from which the reads originate are unknown. To overcome this difficulty, four artificial read libraries were created. To that end, the quantitative results from Bräutigam *et al.* (2011) were used to produce a simulated read database from 11 889 *Arabidopsis* transcripts as described in detail in the Materials and methods. In a second step, 'errors', namely random base changes, were introduced at levels of 1, 3, and 5% to mimic both sequencing errors and genetic variation in the libraries. The artificial read libraries were assembled with the open source programs SOAPdenovo (SOAP), Velvet, MIRA, CAP3, and TGICL (Table 1) as well as the commercial program CLC bio genomics workbench (CLC bio, Aarhus, Denmark). Three different measures were used to test the assembly programs. The reduction in total sequence length of the initial libraries compared with the assembled contigs is one measure to test for reduced redundancy. A second measure is the number of contigs which are returned by the assembly. The length distribution is tested with the N25 and N50. If the contigs are sorted by length and one checks contigs until 25% of the sequence information (i.e. the total base count) is covered, the N25 is the length of the shortest contig in the list. In analogy, the N50 is the length of the shortest contig if one carries on until 50% of all sequence information is contained. With the artificial library and without sequencing errors, all assemblers reduce the redundancy by ~90%, indicating that a large proportion of reads are assembled into contigs. The programs produce between 27 315 contigs for SOAP and 12 277 and 11 787 contigs for TGICL and CLC, respectively, with N50s between 476 nucleotides for SOAP and 732 nucleotides for CLC (Table 1). Since 2128 of the transcripts are only represented by one read (Bräutigam *et al.*, 2011), all programs retain reads as singletons. The difference in contig number shows that not all assemblers come close to reducing the contig number back to the number of 11 889 transcripts from which reads were drawn. If SOAP was used, the number of transcripts from which the reads were

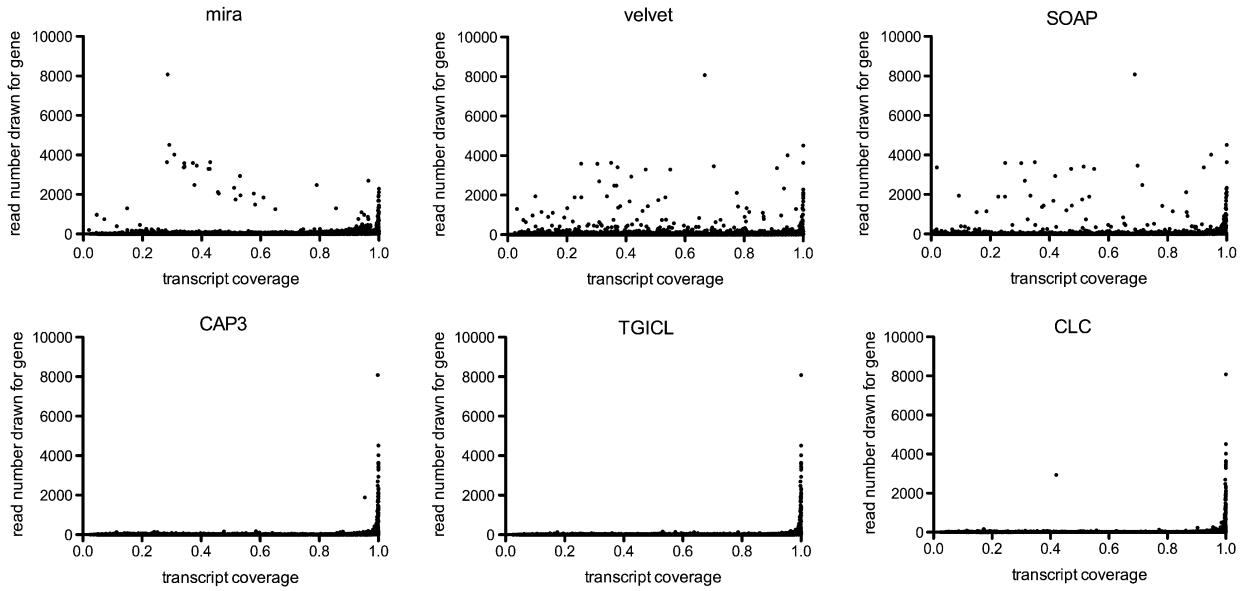


Fig. 1. Transcript coverage for the artificial library assemblies with perfect reads. For each contig, the corresponding *Arabidopsis* transcript was determined and the coverage, i.e. the percentage of bases from the transcript covered by the contig, was determined. For each transcript, the coverage was plotted against the number of reads that were drawn.

generated was overestimated at least 2.3-fold, whereas, if CLC was used, the number would be close to correct. In contrast to artificial reads, 'real world' sequencing data are never perfect as sequencing instruments make errors and natural populations contain genetic variation. The introduction of an error rate of 1% into the artificial read library causes a drop in redundancy reduction in SOAP to 60% while the other programs maintain numbers comparable with the assembly of the perfect library (Table 1). Both Velvet and MIRA as well as SOAP have dropped the N50 from between 434 nucleotides and 604 nucleotides to between 39 nucleotides and 413 nucleotides, while CAP3, TGICL, and CLC overall produce assemblies very similar to that of the perfect library. The number of contigs generated remains similar for CLC, TGICL, and CAP3, while it doubles in MIRA, triples in Velvet, and is 20-fold in SOAP (Table 1). As these artificial reads are closer to reality, depending on the assembler, the number of transcripts would be estimated as close to correct if TGICL, CLC, or CAP3 were used, to a 60-fold overestimate if SOAP was used. The introduction of an even higher rate of 3% variation exacerbates the problems of SOAP, Velvet, and MIRA with losses in redundancy reduction and shorter contigs. At an error rate of 3% in the library, CAP3 starts to lose some reduction in redundancy and has marked losses in long contigs, with N50 dropping from 616 bases at 1% to 458 bases at 3%. Both TGICL and CLC have minor losses in contig length, with the N50s dropping from 678 nucleotides and 718 nucleotides (at 1%) to 611 nucleotides and 689 nucleotides (at 3%). Contig numbers increase slightly for these programs (Table 1). Finally, at an error rate of 5%, all programs are unable to assemble contigs efficiently from the artificial read library (Table 1). The rate of sequence variation in real world data is currently not known. The rate is the sum of the genetic variation in the sampled population and the error rate of transcriptome sequencing.

It is a naïve assumption that long contigs mean 'best' assembly. It is critical that the assemblers do not produce hybrid contigs that have joined two different genes together. For the perfect artificial read library, the number of hybrids produced by each assembler was tested: the number of hybrids is ~190 for MIRA, CAP3, TGICL, and CLC, while SOAP produces 434 hybrids and Velvet only produces 29 (Table 1). While TGICL and CLC produce more hybrids than Velvet, they vastly exceed Velvet's abilities in assembling long contigs. Based on the criteria of contig length distribution and reduction in redundancy (Table 1), among those tested in this study, TGICL, CAP3, and CLC are the most suitable assemblers for non-model transcriptome data. TGICL and CLC are particularly resistant to sequence variation (Table 1).

To learn more about the reason why different assemblers produce such different results (Table 1), for each transcript, the number of reads drawn from the transcripts was plotted against the coverage achieved by the assemblers. The points representing those transcripts with a high number of reads drawn should form a line at the 100% coverage mark, while transcripts with fewer reads drawn may be distributed along the coverage gradient based on transcript length. For the CAP3-based assembly, the TCICL-based assembly, and the CLC-based assembly, the points form a line at the 100% coverage mark while points representing transcripts from which fewer reads were drawn are distributed along the coverage gradient (Fig. 1). However, for the assemblies performed with MIRA, Velvet, and SOAP, the points which represent transcripts from which many reads were drawn form a cloud above those from which few transcripts were

Table 2. *Assembly results of the C. gynandra library with five different assemblers and the results of the CLC assembly of the C. spinosa read library*

For *C. gynandra* 368 333 reads with 85 681 233 bases and for *C. spinosa* 284 318 reads with 65 525 139 bases were assembled. Remaining sequence length is the number of bases in the contigs after the assembly compared with the number of bases in the original sequence reads.

	***C. gynandra* SOAP**	**Velvet**	**MIRA**	**CAP3**	**TGICL**	**CLC**	***C. spinosa* CLC**
Remaining sequence length in X%	26	13	15	11	11	11	12
No. of contigs	383 907	92 149	30 785	20 259	19 019	17 851	16 770
N25	245	288	719	821	885	968	859
N50	106	173	434	496	529	596	521
N75	37	84	290	344	357	379	337

drawn (Fig. 1). This means that transcripts covered by a large number of reads cannot be assembled completely by a subset of assemblers. If errors are introduced, only TGICL and CLC remain capable of assembling high coverage transcripts efficiently (data not shown). In other words, the naïve assumption 'many reads from transcripts mean good assemblies' is not met for all assemblers, but is only true for TGICL, CLC, and CAP3 (with the caveat that CAP3 is not as tolerant of sequence variation). The reason for the problems of MIRA, Velvet, and SOAP in assembling 'perfect' reads into contigs is unknown, but it is suspected that the high number causes problems with the assembly algorithm. Like the question about hybrids, this analysis can only be performed with a simulated artificial read library and not with real sequence data.

The last deciding factor in choosing a good assembler is the time it takes to complete an assembly. The runtime of the assemblers is in the order of minutes for SOAP, Velvet, and CLC, while MIRA completes the assembly within hours, and CAP3 and TGICL take between 1 d and 2 d. Taken together, all results point to TGICL and CLC as the best assemblers among the tested programs, with TGICL taking a much longer time. However, TGICL is open source and the user has full control over the parameters. TGICL assemblies of libraries 10-fold larger than those tested here are possible if the RAM is scaled up from 8 Gb to 100 Gb (data not shown). CLC is a much quicker, but commercial program. Scaling has not yet been tested. Although MIRA has been adapted for assembling EST sequences (Chevreux *et al.*, 2004), it is not as capable, at least under the conditions of the present study. In particular, read libraries that include variation such as the artificial read libraries with errors cannot be assembled well. MIRA is designed to be conservative and thus may be unable to join reads with slight differences into contigs (Chevreux, 2006). Based on the short runtime, it is likely that CLC relies on a de Bruijn graph data structure (the algorithm is proprietary and thus unknown to the end-user), but is vastly more efficient compared with SOAP and Velvet. Recently, Velvet (Zerbino and Birney, 2008) has been extended with OASES with the goal of improving transcript assembly (http://www.ebi.ac.uk/~zerbino/oases/). The program is currently still a beta version and the corresponding publication has not been released. Taken together, all the results of the assembler tests point to TGICL as the oldest but most efficient open source program while the commercial program CLC produces results comparable with or better than TGICL at a much quicker pace.

Fig. 2. Venn diagram of the mapping results of the contigs from *C. spinosa* and *C. gynandra* libraries. After the assembly of the *C. gynandra* and *C. spinosa* reads, for each contig the corresponding *Arabidopsis* transcript was determined. A total of 6211 *Arabidopsis* transcripts were matched by contigs from both species.

Assembling the Cleome *transcriptomes*

After determining the best assembly program in terms of producing long contigs, few hybrids, and capability of assembling high coverage contigs, the results were confirmed with the *C. gynandra* read database. Only contig length and reduction in redundancy can be assessed, but not the number of hybrid contigs and the transcript coverage, since the true transcriptome of *C. gynandra* is unknown. The results of the *C. gynandra* library assembly mirror those of the artificial library assembly. SOAP produces the largest number of contigs with the smallest N50, while both TGICL and CLC produce the best results (Table 2). The contig number varies between ~18 000 and ~400 000, and the reduction in redundancy is between 74% and 89% (Table 2). CLC produces the longest contigs, with an N50 of 596 bases compared with TGICL's second best of 529 bases. Since CLC is even more capable than TGICL with the original *C. gynandra* library, it was also used to assemble the sequence read library from *C. spinosa*. The number of hybrid contigs cannot be determined. However,

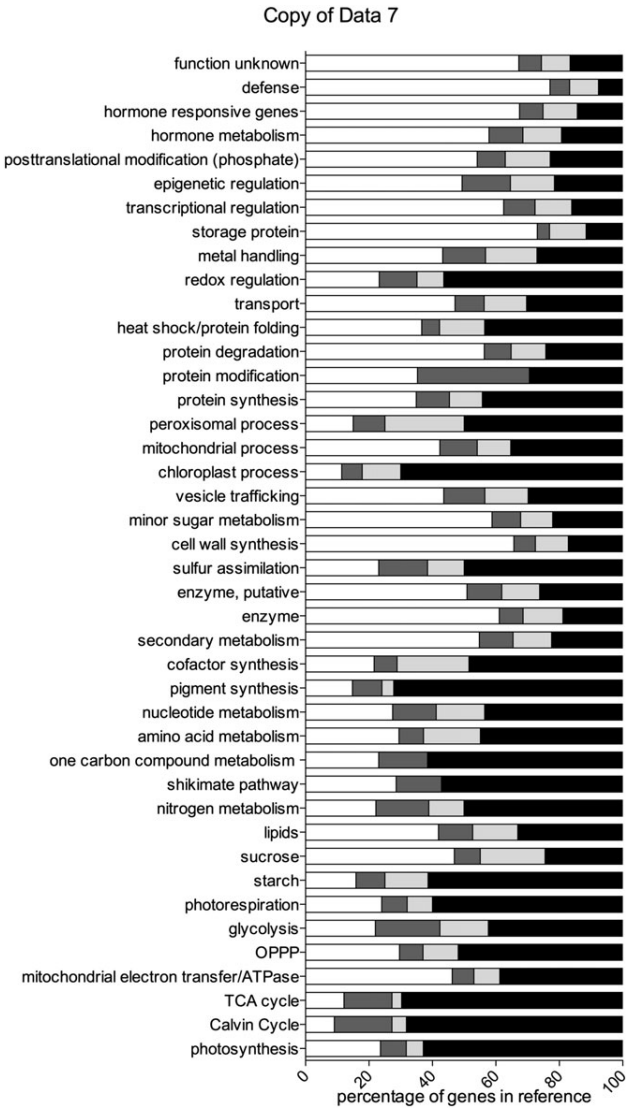


Fig. 3. Pathway representation analysis of the contig mappings to *A. thaliana*; white, not detected in either species; dark grey, detected in *C. spinosa*; light grey, detected in *C. gynandra*; black, detected in both species. For each *Arabidopsis* protein-coding transcript it was determined whether at least one contig from either *Cleome* species matches.

CLC produced 1.6% hybrid contigs with the artificial library (Table 1). In consequence, 1.6% hybrid contigs were considered as the lowest estimate for hybrids in the *C. gynandra* assembly. The new assemblies generated with CLC (Table 2) are better than the initial assemblies (benchmarks from Bräutigam *et al.*, 2011). For example, the N50 increased from 509 bases to 596 bases. When mapping the *C. gynandra* reads back to the contigs, only 6.5% or 24 068 of the 368 333 reads did not match a contig and thus represent singletons. From *C. spinosa*, only 7.7% or 21 870 of the 284 318 reads did not match a contig. A second strategy apart from using different assemblers to optimize transcriptome assemblies has been published recently. This work is based on combining two assemblers and producing an assembly of assemblies in a second step (Kumar and Blaxter, 2010). This strategy was not followed for the *Cleome* assemblies since the artificial read assemblies pointed to the production of hybrid contigs already in the first pass assembly and this problem was expected to be exacerbated in a second pass assembly.

Both *Cleome* plants were only partially inbred species. All reads were aligned to the consensus CLC contig sequences and SNPs were called as described in the Materials and methods. Based on these stringent criteria, 2323 SNPs were detected in the contigs of *C. spinosa*, of which eight are complex SNPs with three different possible nucleotides, and 2367 SNPs were detected in the contigs of *C. gynandra*, of which seven are complex SNPs. The SNPs annotated with this publication do not contain deletions or insertions of nucleotides and are thus unaffected by 454s technology's inability to read homopolymer stretches correctly. The SNP detection tables are available with this publication (Supplementary Table S1 at *JXB* online). Short read technologies such as Illumina, SOLiD, or HeliScope could be used to extend the SNP list by sampling and sequencing more individuals. Based on the SNP profile in the transcribed sequences alone, both *Cleome* species could be developed into inbred lines, of which one could be used for mutagenesis and the other for providing markers to map the mutation.

Comparison of the Cleome *assemblies*

The final assemblies from both *Cleome* species were compared with each other. Of the contigs, 86% and 87% could be annotated with *A. thaliana* using BlastX with a cut-off value of e^{-4}. A similar percentage of reads could be mapped to the *Arabidopsis* transcriptome (Bräutigam *et al.*, 2011). While *C. gynandra* contigs mapped to 9203 unique *Arabidopsis* transcripts, *C. spinosa* contigs mapped to 8598 unique *Arabidopsis* transcripts. Assuming a plant has ~30 000–40 000 protein-coding genes (Swarbreck *et al.*, 2008; Paterson *et al.*, 2009; Schnable *et al.*, 2009) and about half of these genes are expressed in leaves (Schmid *et al.*, 2005), one would expect ~15 000–20 000 transcripts expressed in leaf tissue of a plant. While the number of contigs is well within the estimated number of transcripts, the number of *Arabidopsis* transcripts matching the *Cleome* transcripts clearly is not. Two conclusions can be drawn based on this observation. First, the coverage of leaf transcripts is probably not complete. Even if the species *Cleome* had more genes (i.e. due to a genome duplication event), the genes expressed in leaves should roughly match half of the *Arabidopsis* transcripts, or ~15 000. Secondly, the number of contigs exceeds the number of matching *Arabidopsis* transcripts by a factor of ~2. The contig assembly thus probably retains an ~2-fold redundancy. This redundancy is probably due both to imperfect assemblies, for example due to sequence variation introduced by >2000 SNPs in the species (Supplementary Table S1 at *JXB* online), and to non-overlapping sequence reads, given that

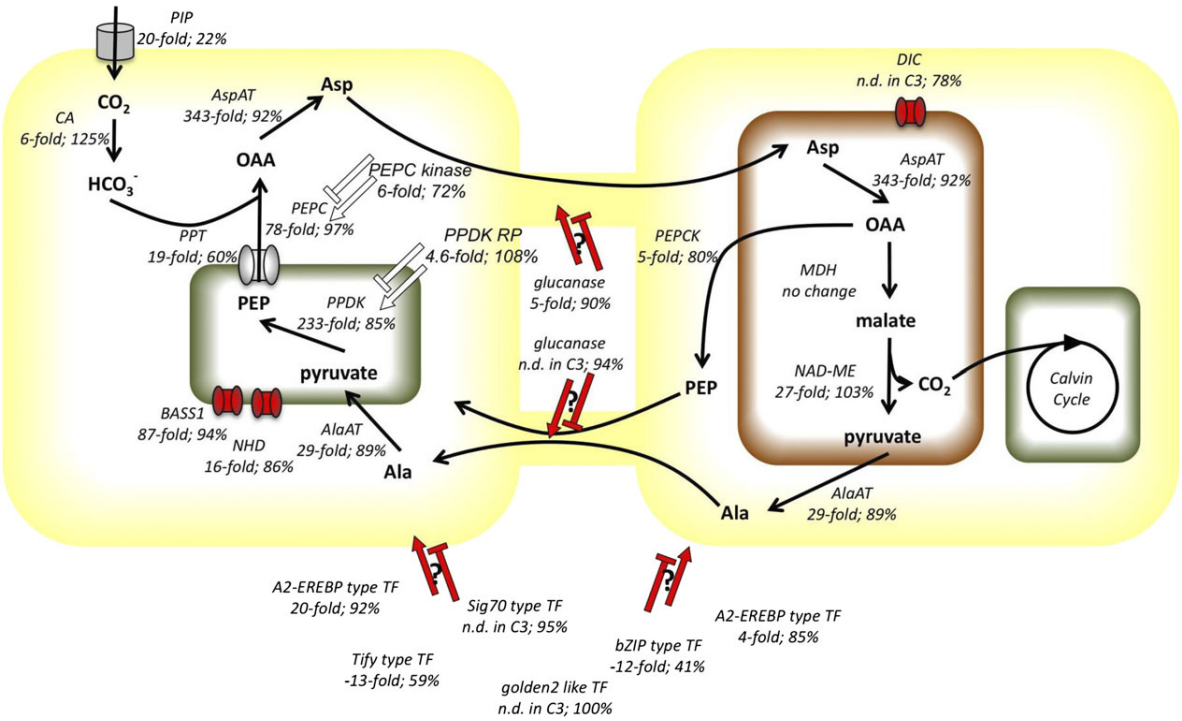


Fig. 4. Schematic representation of the C_4 cycle in *C. gynandra*. For each C_4 gene, the expression fold change compared with *C. spinosa* was extracted from Bräutigam *et al.* (2011) and compared with the contig coverage relative to the closest *Arabidopsis* homologue (in %). Organelles are colour-coded, mitochondria in brown, chloroplasts in green; candidate processes are marked in red. Abbreviations of enzymes: CA, carbonic anhydrase; PEPC, phosphoenolypyruvate carboxylase; AspAT, aspartate aminotransferase; MDH, malate dehydrogenase; NAD-ME, NAD-dependent malic enzyme; AlaAT, alanine aminotransferase; PEP-CK phosphoenolpyruvate carboxykinase; PPDK, phosphoenolpyruvate phosphate dikinase. Abbreviations of transport proteins: PIP, plasma membrane intrinsic protein; PPT, phosphoenolpyruvate phosphate translocator; BASS1, bile acid sodium symporter 1; DIC, dicarboxylic acic carrier; NHD, sodium–proton exchanger. Abbreviation of regulatory genes: PPDK RP, phosphoenolpyruvate phosphate dikinase; TF, transcription factor

the reads libraries are highly dynamic, spanning five orders of magnitude (Bräutigam *et al.*, 2011).

Of 9203 and 8598 contigs, in *C. spinosa* and *C. gynandra*, respectively, roughly two-thirds or 6211 were shared among both species, and 2980 and 2373 were unique to either species (Fig. 2). When only reads and not contigs are mapped to *Arabidopsis*, about half of the transcripts in the reference are identified (Bräutigam *et al.*, 2011) compared with about a third with contig mapping. There are two possible explanations. On the one hand, reads are shorter than contigs, so the mapping may not have been as precise and more different *Arabidopsis* transcripts were tagged. On the other hand, the reads were mapped to a minimal genome devoid of transcripts resulting from a whole-genome duplication and subsequent tandem duplications (Bräutigam *et al.*, 2011). Assuming equal mapping accuracies of reads and contigs, one may argue that the *Cleome* species lack a larger proportion of transcripts resulting from duplications.

To study the end-point of differentiation into C_4 leaf tissue, mature leaves were sampled from a C_3 and a C_4 species and the sequence libraries were not normalized prior to sequencing. These strategic decisions limit the qualitative sequence information: the contig database represents only transcripts expressed in mature leaf tissue and the depth of sequencing is relatively shallow, compared with normalized libraries. Read mapping to *Arabidopsis* indicated that primary metabolism as well as categories related to leaf functions were well represented in the *Cleome* libraries, while categories such as regulation were under-represented (Bräutigam *et al.*, 2011). Representation was tested using the contigs as well (Fig. 3). The results essentially mirror those obtained with reads alone (Fig. 3; Bräutigam *et al.*, 2011) The use of only mature leaf tissue clearly limits the number of transcripts which can be identified and it hinders the even distribution throughout functional categories. The sequencing of additional libraries from developing leaves and/or other plant tissues will increase both the number of transcripts and their coverage, as well as evening out the category representation. Based on the limitations in the contig database the question arises as to what degree the transcriptomes can be used to study the end-point of differentiation to a fully mature C_4 leaf.

To answer this question, the transcript coverage for transcripts known to be involved in C_4 photosynthesis was extracted and visualized (Fig. 4). Candidate transcripts for

regulatory processes were randomly chosen from Bräutigam *et al.* (2011) and also visualized (Fig. 4). By matching transcript sequences from genes overexpressed in C_4 tissue (Bräutigam *et al.*, 2011) with their contig sequences (this work), studies at the molecular level can indeed be initiated. For example, the full-length transcripts for the C_4 cycle enzymes of *C. gynandra* as well as the transport protein PPT known to be involved in C_4 photosynthesis can be extracted from the *C. gynandra* contig file (Fig. 4; Supplementary Fig. S1 at *JXB* online). Additional transcripts probably involved in the C_4 cycle such as those of candidate transport proteins like the plastidic BASS1, a member of the bile acid:sodium symporter family, or DIC, a dicarboxylate carrier at the mitochondrial membrane, can be extracted and studied. When the contig length of these transcripts is compared with the length of the *Arabidopsis* representative transcript model, the contigs achieve ⩾85% coverage. Known regulators of C_4 photosynthesis, which are more highly expressed in the C_4 species, are PEPC kinase (72% coverage), PPDK regulatory protein (108%), and a golden2-like transcription factor (100%). Candidate regulators, such as glucanases, potentially involved in increasing the open probability of plasmodesmata (Bräutigam *et al.*, 2011) are covered to 90% and 94%, respectively. Similar numbers are achieved for candidate transcription factors up-regulated in C_4 (Fig. 4). Since known and candidate transcripts for C_4-related processes are covered in full or nearly so, the transcriptome databases are suitable to study the end-point of C_4 differentiation. There may also be regulatory transcripts which need to be down-regulated in mature C_4 tissue and hence less abundant in the transcriptome database. For the two transcription factors down-regulated in the C_4 species which were tested, approximately half-length transcripts were assembled, indicating that the coverage is good enough to initiate further analysis. The NGS project for *Cleome* succeeded in both the comparative quantification of gene expression between a C_4 and a C_3 species (Bräutigam *et al.*, 2011) and in providing a sequence resource for further research.

Conclusion

The artificial read libraries constructed based on quantitative information enabled the study of different assembly programs and identified two useful assemblers for next-generation mRNA-Seq data: TGICL and CLC bio genomics workbench. The application of the assemblers to real world data of *C. gynandra* confirmed the results of artificial library assembly. The contig databases for mature C_3 and C_4 leaves represent the pathways of mature leaves well and enable the study of the end-point of C_4 photosynthetic differentiation.

Supplementary data

Supplementary data are available at *JXB* online.

Figure S1. The contig database of *C. gynandra.*

Figure S2. The contig database of *C. spinosa.*

Figure S3. Scripts for producing the simulated read libraries.

Figure S4. The sequence variation detected in the *C. gynandra* read library exemplified by three contigs with annotated sequence variation.

Table S1. Single nucleotide polymorphism tables for *C. gynandra* and *C. spinosa.*

Acknowledgements

This work was supported by grants of the German Research Council (DFG) to APMW (SFB 590, SFB TR1, WE2231/4-1). The authors would like to thank Lazar Pavlovic for language editing of the manuscript and Janina Maß for bioinformatic support.

References

Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, Pietrella M, Giuliano G, Chiusano ML, Baldoni L, Perrotta G. 2009. Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* **10,** 15.

Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, Powell WA, Wheeler N, Sederoff R, Carlson JE. 2009. Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biology* **9,** 11.

Botha CEJ. 1992. Plasmodesmatal distribution, structure and frequency in relation to assimilation in C3 and C4 grasses in Southern Africa. *Planta* **187,** 348–358.

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

Bräutigam A, Hoffmann-Benning S, Weber APM. 2008. Comparative proteomics of chloroplast envelopes from C3 and C4 plants reveals specific adaptations of the plastid envelope to C4 photosynthesis and candidate proteins required for maintaining C4 metabolite fluxes. *Plant Physiology* **148,** 568–579.

Bräutigam A, Kajala K, Wullenweber J, *et al.* 2011. An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiology* **155,** 142–156.

Brown NJ, Parsley K, Hibberd JM. 2005. The futured C-4 research—maize, Flaveria or Cleome? *Trends in Plant Science* **10,** 215–221.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WEG, Wetter T, Suhai S. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research* **14,** 1147–1159.

Chevreux C. 2006. MIRA: an automated genome and EST assembler. PhD Thesis.

Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM. 2009. Shedding light on an extremophile lifestyle through transcriptomics. *New Phytologist* **183,** 764–775.

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.

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

Hatch MD. 1987. C-4 photosynthesis—a unique blend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta* **895,** 81–106.

Huang XQ, Madan A. 1999. CAP3: a DNA sequence assembly program. *Genome Research* **9,** 868–877.

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

Kumar S, Blaxter ML. 2010. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* **11,** 571.

Laisk A, Edwards GE. 2000. A mathematical model of C-4 photosynthesis: the mechanism of concentrating CO2 in NADP-malic enzyme type species. *Photosynthesis Research* **66,** 199–224.

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, Zybailov B, Ytterberg AJ, Dunsmore J, Sun Q, van Wijk KJ. 2008. Consequences of C-4 differentiation for chloroplast membrane proteomes in maize mesophyll and bundle sheath cells. *Molecular and Cellular Proteomics* **7,** 1609–1638.

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 photosynthesism. *The Plant Journal* **51,** 886–896.

Meierhoff K, Westhoff P. 1993. Differential biogenesis of photosystem II in mesophyll and bundle-sheath cells of monocotyledonous NADP-malic enzyme-type C-4 plants—the nonstoichiometric abundance of the subunits of photosystem-II in the bundle-sheath chloroplasts and the translational activity of the plastome-encoded genes. *Planta* **191,** 23–33.

Metzker ML. 2010. Applications of next generation sequencing: sequencing technologies—the next generation. *Nature Reviews Genetics* **11,** 31–46.

Muhaidat R, Sage RF, Dengler NG. 2007. Diversity of Kranz anatomy and biochemistry in C-4 eudicots. *American Journal of Botany* **94,** 362–381.

Newell CA, Brown NJ, Liu Z, Pflug A, Gowik U, Westhoff P, Hibberd JM. 2010. *Agrobacterium tumefaciens*-mediated transformation of *Cleome gynandra* L., a C-4 dicotyledon that is closely related to Arabidopsis thaliana. *Journal of Experimental Botany* **61,** 1311–1319.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *BMC Genomics* **9,** 14.

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.

Paterson AH, Bowers JE, Bruggmann R, *et al*. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457,** 551–556.

Pertea G, Huang X, Liang F, *et al*. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19,** 651–652.

Sage RF. 2004. The evolution of C-4 photosynthesis. *New Phytologist* **161,** 341–370.

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. *Nature Genetics* **37,** 501–506.

Schnable PS, Ware D, Fulton RS, *et al*. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–1115.

Sowinski P, Szczepanik J, Minchin PEH. 2008. On the mechanism of C4 photosynthesis intermediate exchange between Kranz mesophyll and bundle sheath cells in grasses. *Journal of Experimental Botany* **59,** 1137–1147.

Swarbreck D, Wilks C, Lamesch P, *et al*. 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research* **36,** D1009–D1014.

van Rensburg WJ, Venter SL, Netshiluvhi TR, van den Heever E, Vorster HJ, de Ronde JA. 2004. Role of indigenous leafy vegetables in combating hunger and malnutrition. *South African Journal of Botany* **70,** 52–59.

Wang W, Wang YJ, Zhang Q, Qi Y, Guo DJ. 2009. Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* **10,** 10.

Weber APM, von Caemmerer S. 2010. Plastid transport and metabolism of C3 and C4 plants—comparative analysis and possible biotechnological exploitation. *Current Opinion in Plant Biology* **13,** 256–264.

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

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18,** 821–829.

4.1.3 The Protein Composition of the Digestive Fluid from the Venus Flytrap Sheds Light on Prey Digestion Mechanisms

The Protein Composition of the Digestive Fluid from the Venus Flytrap Sheds Light on Prey Digestion Mechanisms*[S]

Waltraud X. Schulze‡[a], Kristian W. Sanggaard§[a], Ines Kreuzer¶, Anders D. Knudsen§, Felix Bemm‖, Ida B. Thøgersen§, Andrea Bräutigam‡‡, Line R. Thomsen§, Simon Schliesky‡‡, Thomas F. Dyrlund§, Maria Escalante-Perez¶, Dirk Becker¶, Jörg Schultz‖, Henrik Karring§§, Andreas Weber‡‡, Peter Højrup¶¶, Rainer Hedrich¶‖‖, and Jan J. Enghild§****

The Venus flytrap (*Dionaea muscipula*) is one of the most well-known carnivorous plants because of its unique ability to capture small animals, usually insects or spiders, through a unique snap-trapping mechanism. The animals are subsequently killed and digested so that the plants can assimilate nutrients, as they grow in mineral-deficient soils. We deep sequenced the cDNA from *Dionaea* traps to obtain transcript libraries, which were used in the mass spectrometry-based identification of the proteins secreted during digestion. The identified proteins consisted of peroxidases, nucleases, phosphatases, phospholipases, a glucanase, chitinases, and proteolytic enzymes, including four cysteine proteases, two aspartic proteases, and a serine carboxypeptidase. The majority of the most abundant proteins were categorized as pathogenesis-related proteins, suggesting that the plant's digestive system evolved from defense-related processes. This in-depth characterization of a highly specialized secreted fluid from a carnivorous plant provides new information about the plant's prey digestion mechanism and the evolutionary processes driving its defense pathways and nutrient acquisition. ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M112.021006, 1306–1319, 2012.***

From the ‡Max Planck Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Potsdam, Germany; §Department of Molecular Biology and Genetics, Aarhus University, Gustav Wiedsvej 10C, 8000 Aarhus C, Denmark; ¶Department of Molecular Plant Physiology & Biophysics, Universität Würzburg, Julius-von-Sachs-Platz 2, 97082 Würzburg, Germany; ‖Department of Bioinformatics, Biozentrum, Am Hubland, Universität Würzburg, D-97074 Wuerzburg, Germany; ‡‡Department of Plant Biochemistry, Heinrich-Heine-Universitaet Duesseldorf, Universitaetsstrasse 1, 40225 Duesseldorf, Germany; §§University of Southern Denmark, Institute of Chemical Engineering, Biotechnology and Environmental Technology, Niels Bohrs Allé 1, 5230 Odense M, Denmark; ¶¶Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark; ‖‖Zoology Department, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia

Received June 5, 2012, and in revised form, July 26, 2012

Published, MCP Papers in Press, August 12, 2012, DOI 10.1074/mcp.M112.021006

Carnivorous plants capture, digest, and "eat" animals using four different types of trapping strategies: (i) flypaper or adhesive traps (*e.g. Drosera*, also known as sundews, and *Pinguicula*, also known as butterworts), (ii) sucking bladder traps (*e.g. Utricularia*, also known as bladderworts), (iii) pitfall traps (*e.g. Nepenthes*), and (iv) snap traps (*e.g. Dionaea muscipula*, also known as the Venus flytrap). These plants fascinated Charles Darwin. The Venus flytrap, in particular, attracted his attention, and he described the plant as "one of the most wonderful in the world" (1). The snap trap most likely evolved from the adhesive trap, because its ability to capture larger prey than the adhesive traps gives it an evolutionary advantage (2).

The trapping motion of *Dionaea muscipula* is among the fastest movements in the plant kingdom, and its mechanism has been described in detail, starting with Charles Darwin's work from ~150 years ago (3–6). The plant's leaves employ turgor pressure and hydrodynamic flow to close the trap (3). The closing is initiated by the mechanical stimulation of trigger hairs, eliciting an action potential to close the trap, which seals the fate of the animal inside (1). Then "touch" hormones such as 12-oxophytodienoic acid, which is a precursor of the phytohormone jasmonic acid, probably induce the secretion of digestive fluid (7). Touch hormones are likely to be released in response to the continuous mechanical stimulation of the trigger hairs by the prey as it struggles to escape (7). The trap may also be closed artificially by direct electrical stimulation or by the application of the bacterial phytotoxine coronatine (5, 7, 8).

The largest classes of Venus flytrap prey are spiders and flies. Highly active fliers, such as bees and wasps, are rarely caught (9). The trapped animal faces a slow death, and experiments with ants demonstrate that the prey are alive and capable of stimulating the trigger hairs up to 8 h after being caught (10). The nutrients obtained from the digestion of the different prey are important for the Venus flytrap. Among carnivorous plants in their natural habitats, the Venus flytrap appears to be the most dependent on the nitrogen obtained

from its digested prey (11). The nutrients from insects and spiders give the plants a competitive advantage in their natural low-nutrient soil habitats (12).

In contrast to its trapping mechanism, only a few studies have focused on the digestion process of the Venus flytrap, and none of the involved enzymes has been purified. However, the pH during Venus flytrap digestion has been studied. The pH of the digestive fluid is 4.3, and during the secretion phase the external "stomach" is further acidified to pH 3.4 (7, 13). The optimum pH for protease activity in the fluid has been analyzed in different studies, and the resulting values range from pH 3.0 to pH 7.0 (13–16). This discrepancy is likely due to differences in the assay conditions and the substrates used as targets during the analyses (13, 16).

In our work, we have determined the protein composition of the digestive fluid of the Venus flytrap. The protein identifications were based on a two-step approach involving (i) deep sequencing of the cDNA from stimulated leaves (RNA-seq) and (ii) subsequent mass spectrometric (MS)[1] analyses of the proteins in the collected digestive fluids (Fig. 1). Both the RNA-seq analyses and the digestion fluid proteomics were performed on independent samples using complementary approaches. The obtained mass spectra were searched against the two generated transcriptome databases, and the identified proteins in the secretome were abundance-ranked based on their intensity sums. Our results provide insights into the complex composition of the Venus flytrap's digestive fluid, which has vital functions in defense and nutrient digestion.

EXPERIMENTAL PROCEDURES

Plant Material for 454 Sequencing and Sampling by Filter Paper Stimulation—Dionaea muscipula plants were purchased from CRESCO Carnivora (De Kwakel, The Netherlands) and grown in plastic pots at 22 °C in a 16:8 h light:dark photoperiod. Three stimulation methods were used for the transcriptomic approach: (i) the plants were fed with ants, and the traps were collected after 24 h; (ii) the plants were sprayed with 100 μM coronatine, and the traps were harvested after 24 h; and (iii) the plants were stimulated by the placement of filter paper soaked with either 30 mM urea, 30 mM chitin, or water into the trap, and trap tissue was collected 1 and 8 h after stimulation. The material for the transcriptome analyses was harvested as follows: traps and excised trigger hairs were frozen in liquid nitrogen. Additionally, secretory cells were isolated from the inner trap surface by gently abrading the gland complexes using a razor blade. RNA was separately isolated from each sample, and for cDNA synthesis, the RNA from different tissues was pooled.

To stimulate fluid secretion for the protein analyses, the closure of healthy mature traps was initiated by tickling the trigger hairs within the trap. Because secretion does not begin without further stimulation of the trigger hairs while the trap is in the closed state, a fine piece of filter paper soaked with water was trapped in the closing snap trap, allowing for the induced movement of the trigger hairs by slight movements of the filter paper. Secreted liquid was then carefully sampled from the closed trap using a pipette tip that was inserted between the closed trap lobes.

[1] The abbreviations used are: DTT, dithiothreitol; LC, liquid chromatography; MS, mass spectrometric.

Plant Material for Illumina Sequencing and Sampling by Magnet-based Stimulation—Plants were purchased at the Lammehave nursery (Ringe, Denmark) and grown in a walk-in plant growth chamber at 26 °C in a 12:12 h light:dark photoperiod. All of the experiments were performed on healthy mature plants. For the transcriptomics analyses, the digestion process was initiated by feeding the plants yellow mealworm beetles (*Tenebrio molitor*). After 3, 8, 24, 48, and 72 h, the leaves were harvested, rinsed with water to remove the partially digested beetle and beetle fragments, snap frozen in liquid nitrogen, and stored at −80 °C. For each time point, two stimulated traps were harvested.

For the protein analyses, magnet-based stimulation of the leaf was used to induce the secretion of the digestive fluid. A small stick-magnet was positioned between the trap leaves, and 100 μl of the cysteine protease inhibitor trans-epoxysuccinyl-l-leucylamido-(4-guanidino) butane (E-64; 50 μM) was added to the trap to reduce the level of adventitious proteolysis (14, 16). Then a larger magnet was applied to move the smaller magnet inside the trap, stimulating the trigger hairs and resulting in complete closure. After 48 and/or 72 h, the secreted fluid (up to 200 μl) was aspirated using a pipette that was forced in between the leaf lobes of the trap. The collected material was centrifuged, and the supernatant was used for further analyses. If not used immediately, the fluid was stored at −20 °C.

Transcriptome Sequencing and Assembly—The leaves from the stimulated traps were pooled and homogenized before RNA was extracted using a previously described hot borate buffer protocol (17). Poly-A transcripts were enriched from 3.5 μg of total RNA, and the transcripts were fractionated in the presence of Zn^{2+}. Subsequently, double-stranded cDNA synthesis was performed using random primers and RNase H. After end repair and purification, the fragments were ligated with bar-coded paired-end adapters, and fragments with insert sizes of ~150 to 250 bp were isolated from an agarose gel. Half of the library was normalized, and the other half was amplified via PCR and purified from a gel. The library quality was assessed using capillary sequencing of randomly selected clones. The high-throughput sequencing of the cDNA samples from the beetle-stimulated traps was performed on an Illumina HiSeq 2000 instrument using a paired-end run with 2 × 50 bp. The cDNA samples isolated from the traps from the other combined stimulation approach were sequenced using a 454 GS FLX Titanium platform. The filtered reads from the two transcriptomic datasets were assembled using Oases software (56) on top of Velvet and Mira, respectively. The minimum size for the assembly was set to either 50 or 100 bases. Several parameter sets (*e.g.* Burrows-Wheeler Alignment) were tested to optimize the assemblies.

Sampling Procedures for Proteomics—For the filter paper stimulation method, the secreted fluid was collected into three independent pools from 15 to 20 traps stimulated for 68 h. The protein was precipitated using ice-cold acetone. Protein pellets were resuspended in 6 M urea and 2 M thiourea (pH 8). After the reduction of disulfide bonds with dithiothreitol (DTT), free cysteine residues were alkylated using iodacetamide. Proteins were predigested with Lys-C for 3 h before the dilution of the sample with four volumes of 10 mM Tris-HCl (pH 8). Trypsin was added (1 μg trypsin per 50 μg protein), and the digestions were performed at room temperature for 16 h. After acidification, the peptides were desalted over a C18 matrix prior to the MS analysis.

For the magnet-based method, digestive fluid was collected 48 h after the stimulation of 18 traps. Subsequently, three pools were prepared using the digestive fluid from five to eight plants. From these samples, 35 μl were withdrawn, and the pH was adjusted to 8.5. Subsequently, DTT was added, the samples were boiled for 5 min, and iodoacetamide was then added to alkylate free cysteine residues. After 15 min of incubation, trypsin was added (1:20 ratio), and the digestions were performed at 37 °C for 16 h.

Gel-free Proteomics of the Digestive Fluid (from Both Sampling Procedures)—The resulting peptides from each of the digests were separated using an EasyLC nanoflow HPLC system (Proxeon Biosystems, Odense, Denmark) connected to an LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific) equipped with a nanoESI ion source (Proxeon Biosystems, Odense, Denmark). The chromatographic separation was performed on a 15-cm fused silica emitter (100 μm i.d.) that was in-house packed with RP ReproSil-Pur C18-AQ 3 μm resin (Dr. Marisch GmbH, Ammerbuch-Entringen, Germany). The peptides were eluted using an acidic acetonitrile gradient at a flow rate of 250 nl min^{-1}, as described elsewhere (18, 19).

MS scans (300–1800 m/z) were recorded using an Orbitrap mass analyzer at a resolution of 60,000 at 400 m/z, with 1×10^6 automatic gain control target ions and a 500-ms maximum ion injection time. The MS scans were followed by data-dependent collision-induced dissociation MS/MS scans of the five most intense multiply charged ions in the mass spectrometer at a 15,000 signal threshold, 30,000 automatic gain control target, 300-ms maximum ion injection time, 2.5-m/z isolation width, 30-ms activation time at 35 normalized collision energy, and dynamic exclusion enabled for 30 s with a repeat count of 1. Peak picking was performed using either MaxQuant 1.114 (Max Planck Institute of Biochemistry, Martinsried, Germany) or Xcalibur 2.0 (Thermo Fisher Scientific Inc., Waltham, MA). The raw data files of the in-solution digestion of the Venus flytrap secretion fluid from both sampling methods have been deposited at the Tranche database (proteomecommons.org) under the following hash key: tzwFTnv4Y04ujdmGC1tVaULYKS3OQ/0i3pmQ2P0vvvdHe6+5vX09E6zW4OzKILOJJDZ9OTXzOB8N66+5czMOCv2MORAAAAAAAAAE/A==.

Identification and Quantification of the Secretome Proteins Using the 454 Transcriptome—The acquired raw data files were searched against the 6-frame translation of the 454 transcriptome (in total, 227,604 protein entries) using MaxQuant 1.114 (20). The carbamidomethylation of cysteine residues was set as a fixed modification, and the oxidation of methionine residues was set as a variable modification. Two missed cleavages were allowed. The mass tolerance for the first search was set to 10 ppm, and the fragment mass tolerance was set to 0.5 Da. The peptide and protein false discovery rates were set to 0.01, and the identified peptides were required to have a minimum length of six amino acid residues. The assignment of the identified peptides to translated proteins was primarily based on proteotypic peptides. Peptides with more than one protein match were assigned to protein groups consisting of all of the proteins with their respective peptide matches. The retention time alignment of the precursor ions was used to extract intensity information from the peaks with matching *m/z* values from samples in which these peaks were not selected for data-dependent fragmentation.

Identification and Quantification of the Secretome Proteins Using the Illumina Transcriptome—For protein identification, the raw data from both sampling procedures were searched against the 6-frame translation of the Illumina transcriptome (in total, 97,728 protein entries) using Mascot 2.3.02 (Matrix Science, London, UK) (21). The searches were performed with up to one missed cleavage allowed, carbamidomethyl (C) as a fixed modification, methionine oxidation as a variable modification, a peptide mass tolerance offset of 10 ppm, a fragment mass tolerance of 0.5 Dalton, and an ion score cutoff at 20. Peptide identification was defined as peptides with scores above Mascot's homology threshold and at a significance threshold (*p*) of 0.01. Peptide assignments to proteins were performed according to the default Mascot settings, *i.e.* each redundant peptide was primarily assigned to the highest scoring protein. The described settings resulted in an average false discovery rate of 3.3%. However, additional stringent criteria for protein and peptide acceptances were applied (see below). Proteins that did not meet the quantitative thresholds and that were not identified and quantified in at least two of the six samples were rejected. To extract the quantitative information, Mascot Distiller 2.4.2.0 (Matrix science) was applied using fraction and correlation thresholds of 0.7 and 0.9, respectively. The data were parsed using MS Data Miner 1.0, which is in-house-developed software (57).

Gel-based Proteomics of the Digestive Fluid—A total of 100 μl of digestive fluid from the magnet-stimulated traps was lyophilized and subsequently dissolved in SDS sample buffer containing 30 mM DTT. The proteins were resolved in 5% to 15% acrylamide gradient gels (23). Subsequently, the gel was silver-stained, and all of the visible bands were excised and digested with trypsin (24, 25). Tryptic peptides were purified using a C18 stage tip (Proxeon Biosystem A/S part of Thermo Fisher Scientific Inc., Odense, Denmark) and were subsequently analyzed via liquid chromatography (LC)-MS/MS using an EASY-nLC (Proxeon Biosystems) connected to a Q-TOF Ultima API (Micromass/Waters, Milford, MA) mass spectrometer. As described above, the obtained mass spectra were analyzed using Mascot; however, the significant threshold value (*p* value) was set at 0.05 because of the lower complexity of the samples and the lower sensitivity of the instrument. In addition, the peptide mass tolerance was set at 1 Dalton because of the lower accuracy of this instrument. If the identification was based on a single peptide, then it was accepted only if the protein was also observed during the gel-free analyses using the more sensitive Orbitrap mass spectrometers.

Verification of the Peptide and Protein Assignments—We manually verified that the peptide hits corresponding to the same transcript corresponded to the same reading frame. If the peptides were identified from different reading frames, then we evaluated whether the different reading frames were likely explained by missing regions in the sequence of interest, which could have led to a frameshift. If this was the case, then the reading frames that resulted in the identification of peptides were merged in the part of the transcript that was missing (see "X" in the sequences). If the identification was based on one unique peptide, then the MS/MS data were manually validated using the following criteria: the assignment of major peaks and the occurrence of uninterrupted y- or b-ion series of at least three amino acid residues. The full list of identified peptides and proteins can be found in supplemental Tables S1 (454 transcriptome) and S2 (Illumina transcriptome). The spectra for proteins with single-peptide identifications can be found in supplemental Figs. S2 (454 transcriptome, in-solution analysis), S3 (Illumina transcriptome, in-solution analysis), and S4 (Illumina transcriptome, gel analysis). The majority of the proteins were identified and quantified by means of both the Mascot/Illumina transcriptome approach and the MaxQuant/454 transcriptome approach, which strengthens our data.

Determination of the Proteins' Relative Abundances—To calculate the relative abundances, we initially calculated the sum of the ion intensities from the extracted ion chromatograms for all of the identified peptides in that particular analysis (*i.e.* the total ion intensity). Subsequently, the individual samples were normalized. Then, the relative abundance (or fraction of the total) of a particular protein (weighted by the protein's molecular weight) was derived using the following formula: (MS intensity sum for the peptides belonging to a specific protein/total ion intensity of the sample)/molecular weight of the transcript. Subsequently, the average value of the relative abundances was calculated based on samples from the same stimulation method when the protein was present. The abundance ranking performed here was similar to the emPAI calculation (26); however, we based our rankings on summed ion intensities rather than peptide counts, which is analogous to the iBAQ quantification (27). Transcript molecular weight was used for protein size normalization. If only one or two peptides were identified and quantifiable for a single protein, then those peptides were used for quantitation in any case. Proteins

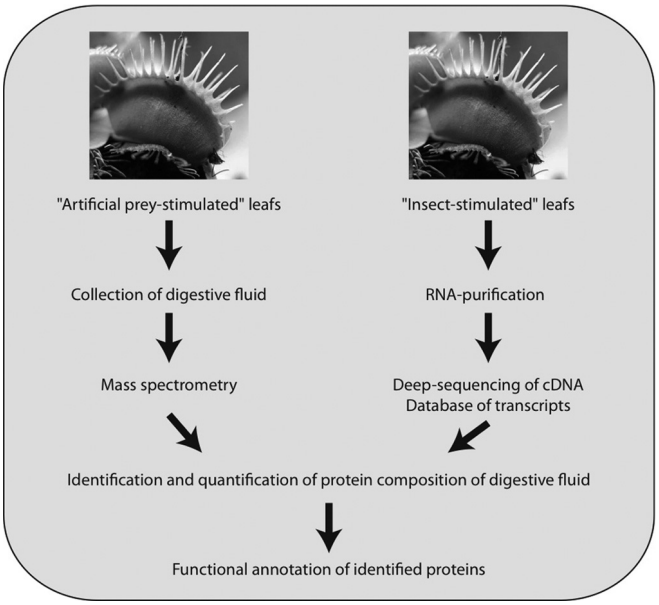


FIG. 1. **Workflow of the Venus flytrap digestive fluid analysis.**

were excluded if they were identified and quantifiable in only one of the six LC-MS/MS samples. We focused on the identified proteases in the Results section, and we searched specifically for proteases in the obtained data. Consequently, if the identified protein was a protease, then we carefully searched the peptide spectra, and if acceptable (see supplemental Fig. S2) the protein was included in Table II even if it was identified and quantified in only one of the six samples. Based on this screening, only contig 18374 was included.

Functional Annotation and Alignment Analyses—Each identified protein was functionally annotated via comparison with *Arabidopsis thaliana* using the TAIR BLAST 2.2.8 tool with the default settings (www.arabidopsis.org). The highest scoring hit was used for the annotation. If a homologous protein was not identified using this approach, a comparison with the NCBInr database was performed using the NCBI BLAST tool. The *MEROPS* peptidase database (merops-.sanger.ac.uk) and the Biology Workbench software from the San Diego Supercomputer Center (workbench.sdsc.edu) were used for sequence analyses and alignments of the identified proteolytic enzymes (28).

RESULTS

Deep Sequencing of the cDNA from Stimulated Venus Flytrap Snap Trap Leaf Lobes—In the present study, we used transcriptomics-generated databases to facilitate the subsequent proteomics-based identification and quantification of the proteins in Venus flytrap digestive fluid (Fig. 1). To increase the likelihood of including all of the relevant transcripts, we employed two different deep sequencing RNA-seq technologies to generate two comprehensive cDNA databases. After the raw data were cleaned, the 454 transcriptome was assembled from 3.5 million reads. In total, 37,934 contigs with matching homologies in the plant kingdom were assembled with an average length of 550 (Table I). In the Illumina transcriptome, ~41 and 120 million paired-end reads of 50 bp lengths were obtained for the non-normalized sample and the normalized sample, respectively. The data from these two

TABLE I
Transcriptome properties

	Illumina	454
Number of contigs	16,288	37,934
Sum of the contig lengths	12,177,595	21,138,288
Cut-off contig length	100	50
Average length of the contigs	747	550

sequencing runs were combined and used for contig and transcript assemblies. A hash value of 41 was selected for the contig and transcript assembly, and 16,288 transcripts were assembled with an average length of 747 (Table I).

The Basic Pattern of the Protein Composition of the Venus Flytrap Digestive Fluid—If the Venus flytrap plants had been stimulated to secrete digestive fluid by a natural prey, then the peptides derived from "prey proteins" would have influenced the mass spectrometry analyses, likely affecting the sensitivity of the analyses, as well as the quantification and identification of the secreted Venus flytrap proteins. A standardized protocol for digestive fluid harvesting would have been difficult to implement if natural prey had been used. Instead, we developed two methods to stimulate the secretion of the digestive fluid without adding real prey.

Upon magnet-based stimulation, the secretion process was initially monitored over 3 days. After 24 h, moisture was detected in the traps, but the amounts were too small for sampling. Up to 200 μl of digestive fluid was collected 48 h (10 replicates) and 72 h (three replicates) after stimulation. The traps that were emptied after 48 h were allowed to continue to secrete for an additional 24 h after stimulation (three replicates). The proteins in the collected samples were separated by SDS-PAGE (supplemental Fig. S1), resulting in the same pattern of protein bands. This suggested that the variation over time and from plant to plant in the overall protein composition of the digestive fluid was low under the conditions employed.

Identification and Abundance Ranking of the Proteins in the Venus Flytrap Digestive Fluid—The secreted proteins from the magnet-based and filter-paper-based stimulations were analyzed by LC-MS/MS. Six datasets were searched against both of the generated transcriptomes and analyzed using MaxQuant and Mascot Distiller. Using this approach, we identified and quantified 76 proteins in the digestive fluid (Table II, supplemental Table S3), with 32 proteins detected in both transcripts. Furthermore, 34 were present in only the 454 transcriptome (supplemental Tables S1 and S4), and 10 proteins were present in only the Illumina transcriptome (supplemental Tables S2 and S5). In total, 66 proteins were identified in the secreted fluid upon filter-paper-based stimulation, and 42 proteins were identified in the secretion fluid after magnet-based stimulation (Table II). Of these 42 proteins, 30 proteins were also identified upon filter-paper-based stimulation, demonstrating large overlap (71%) between the sampling proce-

TABLE II

List of the identified and quantified proteins from the secreted fluid after two different stimulation methods (paper-based and magnet-based) and after matching against two different transcriptomes (454 and Illumina); the abundance ranking within each sample is indicated

Identifier (454-transcript)	Identifier (Illumina-transcript)	Accession (A. thaliana)	Putative function	Average rank in both transcripts	
				Paper-based stimulation	Magnet-based stimulation
	Locus_8322_Transcript_1/1_Confidence_1.000		No BLAST hit with E-value < 1	5	1
DM_TRA02_REP_contig53074	Locus_610_Transcript_2/2_Confidence_1.000	AT3G12500	Chitinase	5	2
NG-5590_Gland_cleanedcontig148924	Locus_223_Transcript_110/117_Confidence_0.063	AT5G45890	Cysteine protease (dionain-3)	6	3
DM_TRA02_REP_contig82037			No BLAST hit with E-value < 1	16	4
DM_TRA02_REP_contig53027	Locus_223_Transcript_111/117_Confidence_0.063	AT4G33720	PR (pathogenesis-related) protein	5	5
DM_TRA02_contig15221	Locus_369_Transcript_1/1_Confidence_1.000	AT4G33355	PR (pathogenesis-related) protein	20	5
DM_TRA02_REP_contig50549	Locus_270_Transcript_1/2_Confidence_1.000	AT4G11650	Osmotin-like protein	29	6
DM_TRA02_contig105872		AT1G11905	Protein of unknown function	22	6
DM_TRA02_REP_contig53296		AT4G33355	PR (pathogenesis-related) protein		7
NG-5590_Gland_cleanedcontig74818&contig124417	Locus_21_Transcript_5/6_Confidence_0.188	AT5G45890; AT1G	Cysteine protease (dionain-2)	54	9
NG-5590_Gland_cleanedcontig110078		AT1G47128	Cysteine proteinase (Dionain 4)	10	9
DM_TRA02_contig14398	Locus_326_Transcript_1/1_Confidence_1.000	AT2G38530	Lipid transfer protein	24	10
DM_TRA02_contig352	Locus_1885_Transcript_2/2_Confidence_1.000	AT5G06860	Polygalacturonase inhibiting protein	8	11
NG-5590_Gland_cleanedcontig146186	Locus_2155_Transcript_2/2_Confidence_1.000	AT3G10410	Serine carboxypeptidase-like 49	13	11
	Locus_3837_Transcript_1/1_Confidence_1.000	AT2G02990	Ribonuclease T2 family	15	13
DM_TRA02_contig19199	Locus_673_Transcript_1/1_Confidence_1.000	AT3G57260	Beta 1,3-glucanase	1	14
DM_TRA02_contig16330	Locus_52_Transcript_2/5_Confidence_0.385	AT1G14540	Peroxidase superfamily protein	2	15
NG-5590_Gland_cleanedcontig2880	Locus_448_Transcript_1/1_Confidence_1.000	AT5G54370	Late embryogenesis abundant protein		15
DM_TRA02_contig33225			No BLAST hit with E-value < 1		16
DM_TRA02_REP_contig51333	Locus_3455_Transcript_1/1_Confidence_1.000	AT5G59970	Histone superfamily protein	20	17
DM_TRA02_contig18767		AT4G35790	Protein with phospholipase D activity		17
DM_TRA02_REP_contig85615	Locus_462_Transcript_4/6_Confidence_0.200	AT1G53130	PR (pathogenesis-related) protein		18
DM_TRA02_REP_contig55271	Locus_522_Transcript_1/2_Confidence_1.000	AT1G51060	A histone H2A protein	25	18
	Locus_21_Transcript_3/6_Confidence_0.562	AT5G45890	Cysteine proteinase (dionain-1)		19
DM_TRA02_contig14593		XP_002510033	Branched Chain amino acid pathway		19
	Locus_462_Transcript_2/6_Confidence_0.300	AT4G26880	Stigma-specific Stig1 family protein		20
	Locus_5180_Transcript_1/1_Confidence_1.000	AT1G79820	Suppressor of G protein beta1 (SGB1)	27	21

TABLE II—*continued*

Identifier (454-transcript)	Identifier (Illumina-transcript)	Accession (A. thaliana)	Putative function	Average rank in both transcripts	
				Paper-based stimulation	Magnet-based stimulation
DM_TRA02_contig12450	Locus_5610_Transcript_1/1_Confidence_1.000	AT5G05340	Peroxidase superfamily protein	19	22
DM_TRA02_REP_contig78697			No BLAST hit with E-value < 1	15	22
	Locus_6940_Transcript_1/1_Confidence_1.000	AT5G05340	Peroxidase superfamily protein	21	22
DM_TRA02_contig53&contig6292	Locus_8046_Transcript_1/1_Confidence_1.000	AT5G50400	Purple acid phosphatase 27 (PAP27)	10	23
DM_TRA02_contig14693			No BLAST hit with E-value < 1	32	24
	Locus_5274_Transcript_1/2_Confidence_0.800	AT2G07698	ATPase, F1 complex, alpha subunit protein	33	25
NG-5590_Gland_cleanedcontig104489		AT4G35790	Encodes a protein with phospholipase D activity		25
DM_TRA02_contig7401	Locus_597_Transcript_3/3_Confidence_0.667	AT5G14780	NAD-dependent formate dehydrogenase	41	26
DM_TRA02_contig16324		AT3G04720	PR (pathogenesis-related) protein		26
DM_TRA02_contig24716			No BLAST hit with E-value < 1	18	27
DM_TRA02_REP_contig50397	Locus_864_Transcript_1/1_Confidence_1.000	AT5G06860	Polygalacturonase inhibiting protein	5	28
DM_TRA02_contig9524		AT2G27130	Lipid transfer protein (LTP) family protein		28
DM_TRA02_REP_contig87892	Locus_5734_Transcript_1/1_Confidence_1.000	AT2G17120	Lysm domain GPI-anchored protein 2 precursor	35	29
DM_TRA02_REP_contig55391		AT4G26880	stigma-specific Stig1 family protein		30
DM_TRA02_contig19291	Locus_187_Transcript_30/39_Confidence_0.191	AT4G39330	Cinnamyl alcohol dehydrogenase 9	41	31
DM_TRA02_REP_contig63924		VITISV_027092	Hypothetical protein, peroxidase-like	4	
DM_TRA02_contig6244	Locus_9868_Transcript_1/1_Confidence_1.000	AT5G45960	GDSL-like lipase	10	
DM_TRA02_REP_contig53347	Locus_5819_Transcript_1/1_Confidence_1.000	AT3G52780	Purple acid phosphatase 20	12	
NG-5590_Gland_cleanedcontig146154	Locus_258_Transcript_1/1_Confidence_1.000	AT5G48430	Aspartyl protease family protein-like	13	
	Locus_2276_Transcript_1/1_Confidence_1.000	AT5G09810	Actin	14	
	Locus_2818_Transcript_1/1_Confidence_1.000	AT1G68290	ENDO 2 (endonuclease 2)	17	
DM_TRA02_REP_contig79562		AT3G50990	Peroxidase superfamily protein	17	
DM_TRA02_contig126504		AT2G42840	PDF1 (protodermal factor 1)	21	
DM_TRA02_contig120869	Locus_1790_Transcript_1/2_Confidence_1.000	AT5G03240; AT3G	Ubiquitin extension protein	23	
DM_TRA02_contig14293	Locus_1849_Transcript_1/1_Confidence_1.000	AT1G71695	Peroxidase superfamily protein	23	
DM_TRA02_REP_contig72380		XP_003536264	Peroxidase superfamily protein	23	
DM_TRA02_REP_contig49581	Locus_468_Transcript_1/1_Confidence_1.000	AT4G28390	AAC3 (ADP/ATP carrier 3)	28	
DM_TRA02_REP_contig137982		AT1G71695	Peroxidase superfamily protein	28	
DM_TRA02_contig14688	Locus_1267_Transcript_1/1_Confidence_1.000	AT5G08680	ATP synthase beta-subunit	31	
NG-5590_Gland_cleanedcontig79407	Locus_3385_Transcript_1/2_Confidence_1.000	AT5G22850	Aspartyl protease family protein (dionaeasin-2)	31	

TABLE II—*continued*

Identifier (454-transcript)	Identifier (Illumina-transcript)	Accession (A. thaliana)	Putative function	Average rank in both transcripts	
				Paper-based stimulation	Magnet-based stimulation
DM_TRA02_REP_contig50840	Locus_175_Transcript_1/1_Confidence_1.000	AT1G78300	G-box binding factor GF14 omega	34	
	Locus_5247_Transcript_1/1_Confidence_1.000	AT3G12580	Heat shock protein 70	34	
DM_TRA02_REP_contig60010		AT3G54420	ATEP3; chitinase	34	
DM_TRA02_REP_contig56545		XP_003380422	superoxide dismutase [Trichinella spiralis]	35	
DM_TRA02_contig7075		AT5G08680	ATP synthase beta-subunit	36	
DM_TRA02_REP_contig48999		AT5G60390	elongation factor 1-alpha/EF-1-alpha	37	
DM_TRA02_contig12608		AT1G68290	ENDO 2 (endonuclease 2)	38	
DM_TRA02_contig1301		AT1G13440	Glyceraldehyde-3-phosphate dehydrogenase C2 (GAPC2)	39	
DM_TRA02_contig16247	Locus_1944_Transcript_1/1_Confidence_1.000	AT4G37530	Peroxidase superfamily protein	40	
DM_TRA02_contig13240		AT3G54420	ATEP3; chitinase	41	
DM_TRA02_contig347	Locus_4642_Transcript_1/1_Confidence_1.000	AT5G67130	Phosphodiesterase superfamily protein	42	
NG-5590_Gland_cleanedcontig105066		AT1G03220	Aspartyl protease family protein-like	42	
DM_TRA02_REP_contig50509		AT5G02500	HSC70–1 (heat shock cognate protein 70–1)	44	
DM_TRA02_contig28745			No BLAST hit with E-value < 1	45	
DM_TRA02_REP_contig129322		AT5G41210	Glutathione transferase	48	
DM_TRA02_contig16780		AT2G36460	Fructose-bisphosphate aldolase, putative	49	
DM_TRA02_REP_contig48954		AT2G44490	Thioglucosidase	50	
DM_TRA02_contig14440		AT5G17920	Methionine synthase	54	
DM_TRA02_contig18374		AT2G03200	Eukaryotic aspartyl protease family protein (dionaeasin-1)	55	

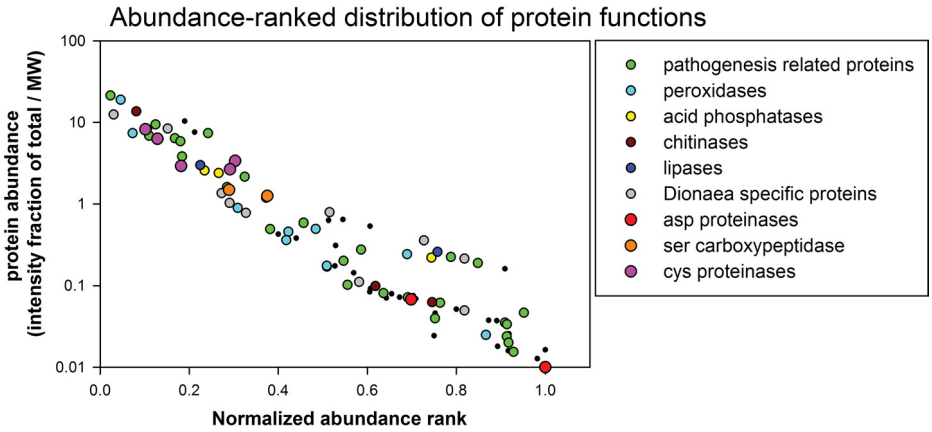


FIG. 2. **The summed ion intensity fraction of the totals for the proteins identified in the two transcriptome versions plotted against the normalized abundance rank within the respective sample.** Different colors indicate proteins of a similar function.

dures (Table II). However, the abundance ranking displayed differences between the two sampling procedures, emphasizing the importance of using complementary methods. The rankings based on MaxQuant and Mascot Distiller were generally in good agreement (supplemental Table S3). The 32 proteins identified from both transcriptome databases displayed significant identical overlaps over long, continuous stretches. Frequently, the shorter contig from one transcript was completely contained within another, longer transcript (supplemental Table S6). This finding clearly supports the value of the complementary approach utilized in this analysis of the Venus flytrap secretome.

In order to annotate the identified proteins based on function, the corresponding translated transcripts were homology searched against the well-annotated *Arabidopsis thaliana* proteome (supplemental Tables S4 and S5). Among the most abundant proteins in the Venus flytrap secretome were homologs of proteases, chitinases, osmotin-like protein, pathogenesis-related proteins, lipid transfer proteins, peroxidases, and beta-1,3-glucanase, which all belong to families of pathogenesis-related proteins (29). The presence of defense-related proteins has also been observed in the digestive fluid of *Nepenthes* (30), indicating that the digestive process in these two plants is functionally similar, although the plants belong to two different families, namely, *Droseraceae* and *Nepenthaceae*. Nine of the transcripts were homologous to different proteases, showing that the proteases are one of the two largest protein families in the secretome. This result correlates with the degrading function of the fluid. The sequence of one of the identified proteases was identical to the peptide sequence of a Venus flytrap digestive fluid cysteine protease from a previous study, except for one amino acid residue (16). In that particular study, the protease was termed "dionain." Here, we adhered to this nomenclature and named the cysteine proteases dionain-1, dionain-2, dionain-3, and dionain-4. Dionain-1 is the protease that contains the previously sequenced peptides (Table II). Nine peroxidase homologs, including one of the top-ranking proteins from the filter-paper-based sampling procedure, were identified. These data indicate that peroxidases have an important role in the digestive fluid. For each identified protein, we plotted the summed ion intensity fraction of the total against the normalized abundance rank in the two transcriptomes. The normalized abundance ranks were calculated as the abundance rank divided by the total number of proteins identified in each sample. When performing this ranking, it became apparent that proteins with antimicrobial/defense functions were among the most abundant proteins in the Venus flytrap secretion fluid (Fig. 2). Cysteine proteinases were consistently the most abundant proteases, followed by serine carboxypeptidases and aspartic proteases.

To strengthen our findings on the in-solution-based abundance ranking and composition analysis of the Venus flytrap digestive fluid, we conducted a complementary gel study (supplemental Fig. S1). Digestive fluid was collected from magnet-stimulated traps, and the proteins were separated on a silver-stained polyacrylamide gel prior to analysis via LC-MS/MS. The obtained spectra were queried against the Illumina transcriptome (supplemental Table S7). The majority of the proteins identified in the gel-based analysis were among the average top-10 ranked proteins present in the in-solution analysis of the magnet-based stimulation. Although there is a general relationship between the summed peptide intensities and the amount of protein present (27), outliers with very few tryptic peptides compatible with an MS analysis can also occur. This is likely to be the case for the relatively low ranking of the dionain-1 protein. This protein has been described previously as the major protein in the digestive fluid (16), and we identified this protein in four of the bands from the gel, which indicates the prominence of this protease. However, this protein was quantified by only one peptide, so its ranking is relatively low (supplemental Table S7).

The Proteolytic Enzymes of the Venus Flytrap Digestive Fluid—The protein sequences of four identified cysteine proteases were compared with those of other proteases using the *MEROPS* BLAST service (28). Dionains 1–3 and dionain-4 most resembled the cysteine proteases SPG31-like peptidase (31) and pseudotzain (32), respectively. All four proteases

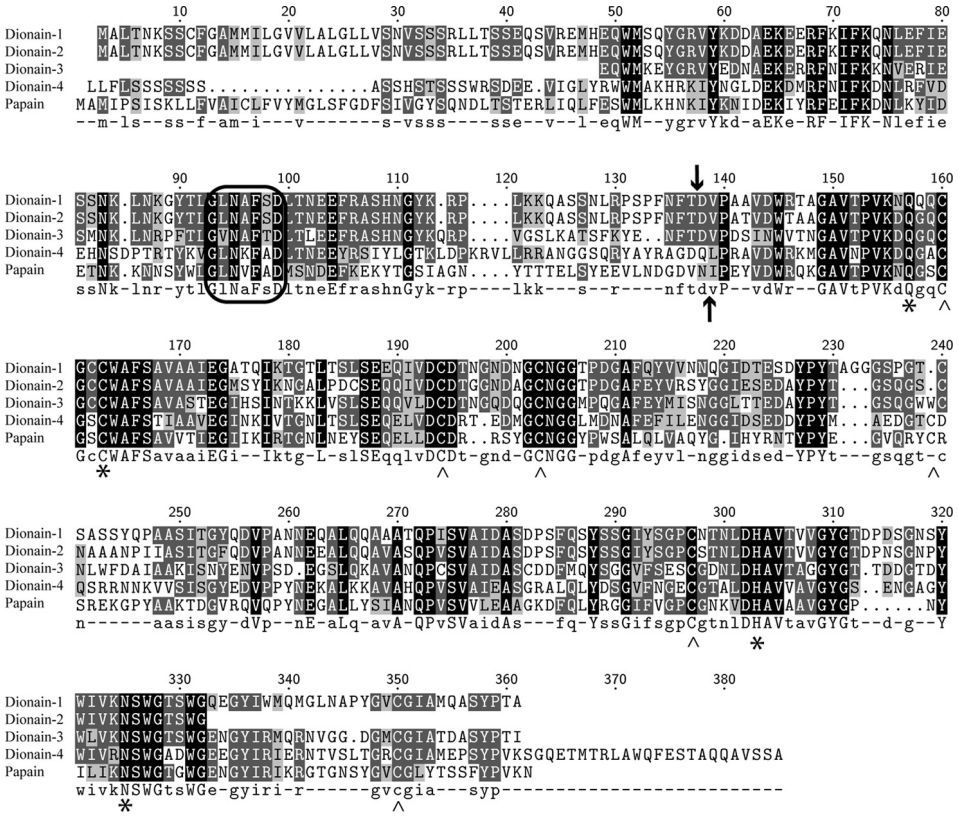


FIG. 3. **The alignment of the open reading frames of the Venus flytrap cysteine proteases with papain.** The encircled area represents a conserved motif that is essential for the conversion from zymogen to the mature protease in papain and cathepsin L. The upward-pointing arrow (↑) indicates the cleavage site between the activation peptide and the mature protease in papain, and the downward-pointing arrow (↓) indicates the cleavage site between the activation peptide and the mature protease in dionain-1. The asterisks (*) represent the active site residues in papain, and the carets (^) represent the cysteine residues involved in the disulfide bridges. The alignment strongly suggests that the active site residues and disulfide bridges are conserved between the five different plant cysteine proteases.

belong to the subfamily C1A, represented by the cysteine protease papain and to which the identified sequences were aligned (Fig. 3). The alignment illustrated that all of the reactive site residues and disulfide bridge-forming cysteines in papain were conserved in all of the dionains. The similarity to papain was also apparent because the pH optima of dionains are expected to be acidic, which is also the case for papain (33). In addition, the alignment showed that an evolutionarily conserved motif (Gly-Xxx-Asn-Xxx-Phe-Xxx-Asp), pivotal for the pH-dependent autoactivation of cysteine proteases, was present in the pro-peptide of dionains (34, 35). These results suggest that autoactivation is a part of the dionain activation mechanism. Indeed, SDS-PAGE of the digestive fluid proteins (collected in the absence of E-64) followed by the Edman degradation of dionain-1 (data not shown; also Ref. 16) revealed that activation occurred due to the proteolysis of a peptide bond within the same region observed for papain (Fig. 3).

The other large group of proteolytic enzymes in the secretome was the aspartic proteases. To characterize these proteases, the sequences were analyzed using BLAST against the MEROPS database, which demonstrated that these proteases belong to subfamily A1B. This subfamily is represented by the nepenthesin from *Nepenthes gracilis* (36). The other large aspartic protease subfamily is A1A and is represented by pepsin, which is important for digestion processes in vertebrates. The main parts of the proteases in subfamilies A1A and A1B are most active at acidic pHs, but A1B differs from A1A in that it normally has six disulfide bridges, which are likely responsible for the remarkable stability of these proteins (37). Two of the identified aspartic proteases (contig 146154 and contig 105066 when using the 454 transcriptome naming) did not contain the active site residues found in the nepenthesins (Fig. 4). In addition, the cysteine patterns were different. These data indicate that although these proteases belong to the A1B subfamily, they are unable to display catalytic activities. An NCBI-BLAST search of these sequences indicated that the two proteins might be xylanase and endoglucanase inhibitor proteins, which are involved in plant defense (38). Taken together, two of the putative aspartic proteases are likely not involved in the degradation of prey proteins in the Venus flytrap.

In contrast, the other two Venus flytrap aspartic proteases contained the active site residues and nearly all of the disul-

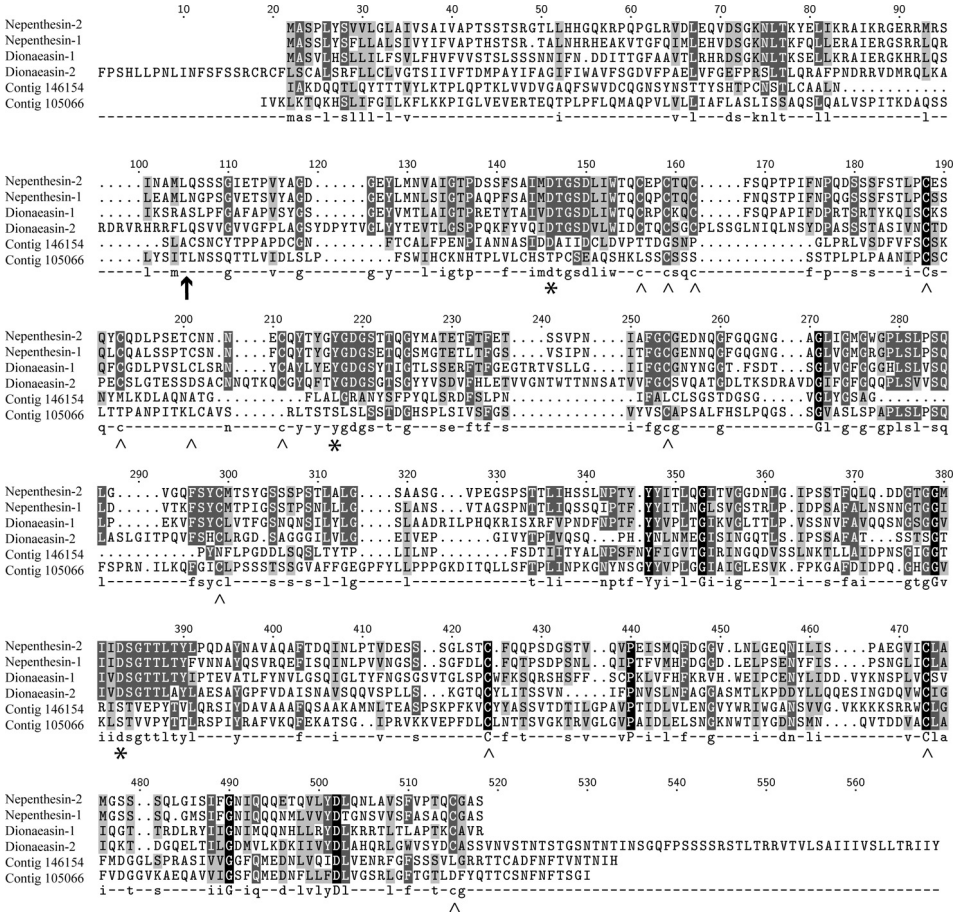


FIG. 4. **The alignment of the Venus flytrap aspartate proteases with nepenthesin 1 and nepenthesin 2 identified from *Nepenthes gracilis.*** The upward-pointing arrow (↑) indicates the cleavage site between the activation peptide and the mature nepenthesin proteases. The asterisks (*) represent the active site residues in the nepenthesins, and the carets (^) indicate the positions of the cysteine residues involved in the disulfide bridges in the nepenthesins. The alignment strongly suggests that two of the Venus flytrap aspartic proteases are catalytically active proteases (referred to as dionaeasins), in contrast to the two other identified proteins (contigs 146154 and 105066), which are likely not catalytically active proteases.

fide bridge-forming cysteine residues at the same positions as in the nepenthesins. Consequently, we named these sequences dionaeasin-1 (contig 18374) and dionaeasin-2 (contig 79407), which is analogous to the nepenthesins from *Nepenthes.* These dionaeasins are likely involved in prey digestion. However, these two proteases were less abundant in the digestive fluid than the cysteine proteases (Table II and Fig. 2), a finding also emphasized by the fact that dionaeasin-1 was found in only one of the six LC-MS/MS analyses. These results suggest that aspartic proteases have a minor role in the degradation process. This is in contrast to *Nepenthes*, in which the nepenthesins are the most prominent proteolytic enzymes.

Regardless of the sampling procedure, the identified serine carboxypeptidase was relatively abundant (Table II), and it was identified using the less sensitive gel-based approach (supplemental Table S7). These results suggest that this protease has an important role in the digestion process. Sequence analyses revealed that it belongs to the S10 family of serine proteases and to the plant serine carboxypeptidase III group (*MEROPS* ID S10.009) (39). The S10 family, represented by carboxypeptidase Y, is active only at acidic pHs, which makes it different from all of the other serine protease families (except for the S53 family). This activity at a low pH correlates with the finding of this protease in the acidic Venus flytrap digestive fluid.

DISCUSSION

Dionaea muscipula, the Venus flytrap plant, is not eaten by animals and is rarely infected by microbes due to its high content of defense metabolites and proteins (40). Instead, it actively traps, kills, and consumes animals. To gain insight into its digestive processes, we analyzed the composition of the fluid secreted by the plant. Most previous studies of the Venus flytrap's digestive fluid (13–15, 41) used an enzyme activity-based approach to characterize the composition of the fluid. In contrast to these indirect methods, we used a combination of transcriptomics and proteomics to identify the

major proteins present in the secreted fluid. Using the enzyme activity-based approach, protease, chitinase, peroxidase, phosphatase, and nuclease activities were detected (14, 15). With regard to protease activity, it has been suggested that the major protease in Venus flytrap digestive fluid is a cysteine protease and that carboxypeptidase activity is also present (14). With the present proteome study, we can explain these different enzymatic activities and correlate them with specific protein sequences.

Transcriptomics-facilitated Proteomics—A direct MS-based identification of the secreted proteins in Venus flytrap digestive fluid was complicated by the lack of the Venus flytrap genome. Recently, deep sequencing technologies have revolutionized the transcriptomics field (42). These methods are largely unbiased and high-throughput, and they provide a large dynamic range (43). To facilitate MS-based protein identifications and to obtain more sequence information, in contrast to a *de novo* MS approach, we used deep sequencing methods to produce a cDNA library of the transcribed mRNA sequences in stimulated Venus flytrap leaves (Fig. 1). The use of the assembled transcriptome as a database for protein identification was significantly more comprehensive and faster than *de novo* peptide sequencing. The rapid development of the "next-generation sequencing" field and the decreasing costs of this technology suggest that the approach applied here is generally applicable for proteome-based studies of organisms and systems for which genome sequence information is limited (44).

Secreted Proteins are Actively Synthesized in the Traps—The gel-based analysis of the secreted fluid confirmed that all of the major proteins (*i.e.* those visible on the silver-stained gel) in the digestive fluid were identified based on the transcriptome derived from the trap tissue. This analysis demonstrates that the cDNA sequences of the major secreted proteins are present in the transcriptome database. Indirectly, it shows that the mRNA coding for these proteins is present in the stimulated traps. Therefore, our results reveal that the Venus flytrap does not exclusively secrete proteins into the digestive fluid from preformed vesicles. Instead, the presence of mRNA indicates that the plant also synthesizes these proteins in the trap during the digestion process. These results are substantiated by the finding that only a limited amount of digestive fluid was present in the traps even after 24 h. If the proteins to be secreted had been present in storage vesicles and ready to be released, then we would have expected the digestive fluid at an earlier time point. These results are supported by previous findings showing that protein synthesis occurs during the secretory phase and that some of the synthesized protein is directly secreted (41).

Abundance Ranking of the Secreted Proteins—In total, 71% of the proteins identified upon magnet-based stimulation were also identified by filter-paper-based stimulation, demonstrating a large overlap between the two stimulation procedures. However, the abundance ranking displayed some differences between the two procedures, at least when the integer ranks were compared. Because more proteins were identified using the filter-paper-based method, these rank numbers contained larger values than those from the magnet-based sampling. When the rankings were standardized for the number of identified proteins (Fig. 2), the abundances were more similar.

The low number of identified peptides for some of the proteins and the rather low abundance ranks likely can be explained by mispredicted nucleotides in some of the transcripts, which is consistent with previous findings demonstrating that contigs obtained from Illumina RNA-seq contain ~5% mispredicted nucleotides (45). Furthermore, the complete open reading frames were not sequenced for all of the transcripts, which can be partly explained by the low amounts of the respective mRNAs. In addition, the assembly programs used, Velvet and Mira, do not perform perfectly during redundancy reduction even at an error rate of 1% in the transcriptome (46).

The cysteine proteases were difficult to quantify and rank properly. These proteases contain a large activation peptide, and the arginine and lysine contents of the mature proteases are low; as a result, only a few tryptic peptides are applicable for MS analyses. These proteins also contain several glycosylation sites, which hamper MS-based identification. In addition, many of the peptides contain cysteine residues, and in our work these were more difficult to detect even though the proteins were alkylated. Thus, the ranking of *e.g.* dionain-1, the cysteine protease that was previously found to be a major component of the digestive fluid (16), was likely too low relative to its actual abundance. The high numbers of cysteine residues and putative disulfide bridges (Figs. 3 and 4) indicated that the identified proteins were compact and stable. Consequently, these proteins are likely to be relatively resistant to the promiscuous protease degradations taking place during digestion. Similar to other plant proteases (47), potential *N*-linked glycosylation sites indicate that the identified proteases could be glycosylated, stabilizing the proteins with respect to proteolytic degradation.

The Composition of the Digestive Environment Sheds Light on the Prey Digestion Mechanism—The low pH of the Venus flytrap digestive fluid is similar to that of other carnivorous plants (*e.g. Nepenthes*) and to the digestive fluid pHs among vertebrates. However, in vertebrates and *Nepenthes*, the proteolytically active enzymes are predominantly aspartic proteases (36, 48). In contrast, our findings suggest that cysteine proteases are the most abundant class of proteases in the digestive fluid of the Venus flytrap, followed by a serine carboxypeptidase and aspartic proteases. This composition and diversity of proteases has not been observed in other digestive fluids. The enzyme composition that resembles our findings the most is the intestinal protein digestion cascade employed by some invertebrates (49–51), which similarly includes aspartic proteases and cysteine proteases from the

same protease families (the pepsin family (A1) and the papain family (C1)) observed in the Venus flytrap. In general, cysteine proteases have a neutral pH optimum. However, adaptations to acidic pH optima have been observed among the lysosomal cathepsins, which are primarily involved in unspecific bulk protein degradation (52). The protease composition of the Venus flytrap's digestive fluid, with three classes of peptidases, is likely a potent digestion system, emphasizing the strong dependence of *Dionaea* on the nutrients supplied through prey capture and digestion (11). Particularly as the pH of the digestive fluid changes over time (7), the different enzymes might reach their maximum activities at different digestion stages after the prey is captured. As previously mentioned, the natural habitat of Venus flytrap plants is low-nutrient soils, and the plants depend on nutrients obtained by digesting trapped prey. These identified proteases are likely involved in the release of nitrogen from the prey proteins. In addition to proteases, a number of other hydrolytic enzymes are present in the digestive fluid, and the fact that nucleases, phosphatases, and phospholipases were identified indicates that phosphate is similarly obtained from the prey's nucleic acids, proteins, and cell membranes.

Three chitinases were also identified, including one of the proteins found to be most abundant in the digestive fluid regardless of the stimulation method. These chitinases would be expected to degrade the exoskeletons of captured insects or spiders and thereby facilitate enzymatic access to the inner part of the prey. Furthermore, chitinases are pathogenesis-related proteins that might prevent microbial growth on the trapped prey during the digestion process.

It has been suggested that prey proteins in the Venus flytrap are initially oxidized in order to facilitate their subsequent proteolysis (53), and it has been demonstrated that *Nepenthes gracilis* uses free radicals during the digestion process (54). Plumbagin, a low-molecular-weight compound present in Venus flytrap digestive fluid, likely facilitates this oxidation (40). The identified peroxidases from the present study are likely involved in these oxidative processes. Thus, our findings support the hypothesis that the oxidation of prey molecules facilitates the digestion mechanisms of the Venus flytrap.

The functions of the hydrolytic enzymes in the digestive fluid are intuitively easy to envision. These enzymes are likely directly involved in prey digestion. The functions of some of the other proteins present in the fluid are more challenging to elucidate, and a functional annotation based on the name of the best match in a homology search (Table II) does not necessarily shed light on the *in vivo* role of the protein. The roles of these proteins in the digestion mechanism remain to be investigated.

The Digestive Fluid Proteome Suggests a Shift from Defense-related Processes to Digestion-related Processes among the Carnivorous Plants—The only previously characterized digestive fluid proteome from a carnivorous plant was derived from *Nepenthes* (30). The depth of that *de novo* sequencing-based study was lower than in the present study; however, aspartic proteases (nepenthesin I and II), a chitinase, a glucanase, a xylosidase, and a thaumatin-like protein were identified. These protein classes were, with the exception of the xylosidase, also identified in the present analysis of the Venus flytrap, indicating the conserved functions of the digestive fluid among carnivorous plant species. Similar to our results, the *Nepenthes* digestive proteins are also predominantly pathogenesis-related proteins. Higher plants express pathogenesis-related proteins as a response to an attack by pathogens, and consequently, many of these proteins possess hydrolytic activities that are potentially applicable to prey digestion in carnivorous plants. The identification of several defense-related proteins suggests that carnivorous plants have exploited the hydrolytic properties of these pathogenesis-related proteins (55). Many pathogenesis-related proteins are resistant to low pHs and to proteolytic degradation (29), making them functional in digestive fluids. During the evolution of carnivory in plants, there has likely been a shift from a pathogen-related response to a prey-related response and a shift from the hydrolysis and destruction of the pathogens to the hydrolysis and digestion of the prey. The defense-related proteins in digestive fluid likely still display antibacterial and antifungal effects, as in *e.g.* poplar extrafloral nectaries (22), in order to avoid pathogenic attacks during the digestion process.

CONCLUSION

The present characterization of Venus flytrap digestive fluid employed deep sequencing of the transcriptome followed by its assembly and subsequent use as a database during the proteomic analyses. This study demonstrates the use of high-throughput technologies in expanding molecular analyses to organisms for which the genome sequence is unknown. The Venus flytrap secretome reveals a unique diversity of hydrolytic enzymes, and the results shed light on the purpose and mechanisms of digestion. Furthermore, the *Dionaea* secretome contains a high proportion of pathogenesis-related proteins, suggesting that the capability of carnivorous plants to digest prey evolved from a plant defense system.

Acknowledgments—We thank Katharina Markmann (Aarhus, Denmark) and Tania A. Nielsen (Aarhus, Denmark) for help with RNA purification; Tom A. Mortensen (Aarhus, Denmark) and the nursery Lammehave (Denmark) for assistance and helpful suggestions regarding the cultivation of Venus flytrap plants; Fasteris SA (Switzerland) for library preparation, Illumina sequencing, and Illumina transcriptome assembly; Kerstin Zander (Golm) for help in sample preparation for mass spectrometry; and Brigitte Neumann for excellent technical assistance (Würzburg).

* This work was supported by a grant from the Danish Research Council for Strategic Research to J.J.E. and an ERC Advanced Grant to R.H.

S This article contains supplemental material.

[a] These authors contributed equally to this work.

** To whom correspondence should be addressed: Rainer Hedrich, E-mail: hedrich@botanik.uni-wuerzburg.de; and Jan Johannes Enghild, E-mail: jje@mb.au.dk.

REFERENCES

1. Darwin, C. (1875) *Insectivorous Plants*, Murray, London
2. Gibson, T. C., and Waller, D. M. (2009) Evolving Darwin's 'most wonderful' plant: ecological steps to a snap-trap. *New Phytol*. **183,** 575–587
3. Forterre, Y., Skotheim, J. M., Dumais, J., and Mahadevan, L. (2005) How the Venus flytrap snaps. *Nature* **433,** 421–425
4. Markin, V. S., Volkov, A. G., and Jovanov, E. (2008) Active movements in plants: mechanism of trap closure by Dionaea muscipula Ellis. *Plant Signal Behav*. **3,** 778–783
5. Ueda, M., Tokunaga, T., Okada, M., Nakamura, Y., Takada, N., Suzuki, R., and Kondo, K. (2010) Trap-closing chemical factors of the Venus flytrap (Dionaea muscipulla Ellis). *Chembiochem.* **11,** 2378–2383
6. Williams, S. E., and Bennett, A. B. (1982) Leaf closure in the Venus flytrap: an acid growth response. *Science* **218,** 1120–1122
7. Escalante-Perez, M., Krol, E., Stange, A., Geiger, D., Al-Rasheid, K. A., Hause, B., Neher, E., and Hedrich, R. (2011) A special pair of phytohormones controls excitability, slow closure, and external stomach formation in the Venus flytrap. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 15492–15497
8. Volkov, A. G., Carrell, H., Baldwin, A., and Markin, V. S. (2009) Electrical memory in Venus flytrap. *Bioelectrochemistry* **75,** 142–147
9. Griggs, R. F. (1935) Victims of the Venus flytrap. *Science* **81,** 7–8
10. Lichtner, F. T., and Williams, S. E. (1977) Prey capture and factors controlling trap narrowing in Dionaea (Droseraceae). *Am. J. Bot*. **64,** 881–886
11. Schulze, W., Schulze, E. D., Schulze, I., and Oren, R. (2001) Quantification of insect nitrogen utilization by the Venus fly trap Dionaea muscipula catching prey with highly variable isotope signatures. *J. Exp. Bot*. **52,** 1041–1049
12. Adamec, L. (1997) Mineral nutrition of carnivorous plants: a review. *Bot. Rev.* **63,** 273–299
13. Takahashi, K., Matsumoto, K., Nishi, W., Muramatsu, M., and Kubota, K. (2009) Comparative studies on the acid proteinase activiteis in the digestive fluids of *NEPENTHES, CEPHALOTOUS, DIONAEA,* and *DROSERA. Carnivorous Plant Newsletter* **38,** 75–82
14. Robins, R. I., and Juniper, B. E. (1980) The secretory cycle of Dionaeamuscipula ellis. IV. The enzymology of the secretion. *New Phytol*. **86,** 401–412
15. Scala, J., Iott, K., Schwab, D. W., and Semersky, F. E. (1969) Digestive secretion of Dionaea muscipula (Venus's-flytrap). *Plant Physiol*. **44,** 367–371
16. Takahashi, K., Suzuki, T., Nishii, W., Kubota, K., Shibata, C., Isobe, T., and Dohmae, N. (2011) A cysteine endopeptidase ("dionain") is involved in the digestive fluid of Dionaea muscipula (Venus's fly-trap). *Biosci. Biotech. Bioch.* **75,** 346–348
17. Hogslund, N., Radutoiu, S., Krusell, L., Voroshilova, V., Hannah, M. A., Goffard, N., Sanchez, D. H., Lippold, F., Ott, T., Sato, S., Tabata, S., Liboriussen, P., Lohmann, G. V., Schauser, L., Weiller, G. F., Udvardi, M. K., and Stougaard, J. (2009) Dissection of symbiosis and organ development by integrated transcriptome analysis of lotus japonicus mutant and wild-type plants. *PLoS One* **4,** e6556
18. Engelsberger, W. R., and Schulze, W. X. (2012) Nitrate and ammonium lead to distinct global dynamic phosphorylation patterns when resupplied to nitrogen-starved Arabidopsis seedlings. *Plant J*. **69,** 978–995
19. Rasmussen, M. I., Refsgaard, J. C., Peng, L., Houen, G., and Hojrup, P. (2011) CrossWork: software-assisted identification of cross-linked peptides. *J. Proteomics* **74,** 1871–1883
20. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol*. **26,** 1367–1372
21. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
22. Escalante-Perez, M., Jaborsky, M., Lautner, S., Fromm, J., Muller, T., Dittrich, M., Kunert, M., Boland, W., Hedrich, R., and Ache, P. (2012) Poplar extrafloral nectaries: two types, two strategies of indirect defenses against herbivores. *Plant Physiol*. **159,** 1176–1191
23. Bury, A. F. (1981) Analysis of protein and peptide mixtures—evaluation of three sodium dodecyl sulfate-polyacrylamide gel-electrophoresis buffer systems. *J. Chromatogr* **213,** 491–500
24. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1,** 2856–2860
25. Sanggaard, K. W., Karring, H., Valnickova, Z., Thogersen, I. B., and Enghild, J. J. (2005) The TSG-6 and Ialphal interaction promotes a transesterification cleaving the protein-glycosaminoglycan-protein (PGP) cross-link. *J. Biol. Chem.* **280,** 11936–11942
26. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005) Exponentially Modified Protein Abundance Index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4,** 1265–1272
27. Schwannhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473,** 337–473
28. Rawlings, N. D., Barrett, A. J., and Bateman, A. (2010) MEROPS: the peptidase database. *Nucleic Acids Res.* **38,** D227–D233
29. van Loon, L. C., Rep, M., and Pieterse, C. M. (2006) Significance of inducible defense-related proteins in infected plants. *Annu. Rev. Phytopathol*. **44,** 135–162
30. Hatano, N., and Hamada, T. (2008) Proteome analysis of pitcher fluid of the carnivorous plant Nepenthes alata. *J. Proteome Res*. **7,** 809–816
31. Chen, G. H., Huang, L. T., Yap, M. N., Lee, R. H., Huang, Y. J., Cheng, M. C., and Chen, S. C. (2002) Molecular characterization of a senescence-associated gene encoding cysteine proteinase and its gene expression during leaf senescence in sweet potato. *Plant Cell Physiol*. **43,** 984–991
32. Tranbarger, T. J., and Misra, S. (1996) Structure and expression of a developmentally regulated cDNA encoding a cysteine protease (pseudotzain) from Douglas fir. *Gene* **172,** 221–226
33. Hoover, S. R., and Kokes, E. L. (1947) Effect of pH upon proteolysis by papain. *J. Biol. Chem.* **167,** 199–207
34. Vernet, T., Berti, P. J., de Montigny, C., Musil, R., Tessier, D. C., Menard, R., Magny, M. C., Storer, A. C., and Thomas, D. Y. (1995) Processing of the papain precursor. The ionization state of a conserved amino acid motif within the Pro region participates in the regulation of intramolecular processing. *J. Biol. Chem.* **270,** 10838–10846
35. Collins, P. R., Stack, C. M., O'Neill, S. M., Doyle, S., Ryan, T., Brennan, G. P., Mousley, A., Stewart, M., Maule, A. G., Dalton, J. P., and Donnelly, S. (2004) Cathepsin L1, the major protease involved in liver fluke (Fasciola hepatica) virulence: propetide cleavage sites and autoactivation of the zymogen secreted from gastrodermal cells. *J. Biol. Chem.* **279,** 17038–17046
36. Athauda, S. P. B., Matsumoto, K., Rajapakshe, S., Kuribayashi, M., Kojima, M., Kubomura-Yoshida, N., Iwamatsu, A., Shibata, C., Inoue, H., and Takahashi, K. (2004) Enzymic and structural characterization of nepenthesin, a unique member of a novel subfamily of aspartic proteinases. *Biochem. J.* **381,** 295–306; Erratum (2004) *Biochem. J.* **382,** 1039
37. Takahashi, K., Athauda, S. B. P., Matsumoto, K., Rajapakshe, S., Kuribayashi, M., Kojima, M., Kubomura-Yoshida, N., Iwamatsu, A., Shibata, C., and Inoue, H. (2005) Nepenthesin, a unique member of a novel subfamily of aspartic proteinases: enzymatic and structural characteristics. *Curr. Protein Pept. Sci*. **6,** 513–525
38. Sansen, S., De Ranter, C. J., Gebruers, K., Brijs, K., Courtin, C. M., Delcour, J. A., and Rabijns, A. (2004) Structural basis for inhibition of Aspergillus niger xylanase by triticum aestivum xylanase inhibitor-I. J. *Biol. Chem.* **279,** 36022–36028
39. Drzymala, A., and Bielawski, W. (2009) Isolation and characterization of carboxypeptidase III from germinating triticale grains. *Acta Biochim. Biophys. Sin.* **41,** 69–78
40. Tokunaga, T., Takada, N., and Ueda, M. (2004) Mechanism of antifeedant activity of plumbagin, a compound concerning the chemical defense in carnivorous plant. *Tetrahedron Lett.* **45,** 7115–7119
41. Robins, R. J., and Juniper, B. E. (1980) The secretory cycle of Dionaeamuscipula ellis. II. Storage and synthesis of the secretory proteins. *New Phytol.* **86,** 297–311
42. Shendure, J., and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol*. **26,** 1134–1145
43. Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*. **10,** 57–63
44. Brautigam, A., Shrestha, R. P., Whitten, D., Wilkerson, C. G., Carr, K. M., Froehlich, J. E., and Weber, A. P. (2008) Low-coverage massively parallel pyrosequencing of cDNAs enables proteomics in non-model species: comparison of a species-specific database generated by pyrosequenc-

ing with databases from related species for proteome analysis of pea chloroplast envelopes. *J. Biotechnol*. **136,** 44–53
45. Wall, C. E., Cozza, S., Riquelme, C. A., McCombie, W. R., Heimiller, J. K., Marr, T. G., and Leinwand, L. A. (2011) Whole transcriptome analysis of the fasting and fed Burmese python heart: insights into extreme physiological cardiac adaptation. *Physiol. Genomics* **43,** 69–76
46. Bräutigam, A., Mullick, T., Schliesky, S., and Weber, A. P. (2011) Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C(3) and C(4) species. *J. Exp. Bot*. **62,** 3093–3102
47. Choi, K. H., and Laursen, R. A. (2000) Amino-acid sequence and glycan structures of cysteine proteases with proline specificity from ginger rhizome Zingiber officinale. *Eur. J. Biochem.* **267,** 1516–1526
48. Kageyama, T. (2002) Pepsinogens, progastricsins, and prochymosins: structure, function, evolution, and development. *Cell Mol. Life Sci*. **59,** 288–306
49. Williamson, A. L., Brindley, P. J., Knox, D. P., Hotez, P. J., and Loukas, A. (2003) Digestive proteases of blood-feeding nematodes. *Trends Parasitol*. **19,** 417–423
50. Delcroix, M., Sajid, M., Caffrey, C. R., Lim, K. C., Dvorak, J., Hsieh, I., Bahgat, M., Dissous, C., and McKerrow, J. H. (2006) A multienzyme network functions in intestinal protein digestion by a platyhelminth parasite. *J. Biol. Chem.* **281,** 39316–39329
51. Knox, D. (2011) Proteases in blood-feeding nematodes and their potential as vaccine candidates. *Adv. Exp. Med. Biol*. **712,** 155–176
52. Barrett, A. J. (1992) Cellular proteolysis. An overview. *Ann. N.Y. Acad. Sci.* **674,** 1–15
53. Galek, H., Osswald, W. F., and Elstner, E. F. (1990) Oxidative protein modification as predigestive mechanism of the carnivorous plant Dionaea muscipula: an hypothesis based on in vitro experiments. *Free Radic. Biol. Med.* **9,** 427–434
54. Chia, T. F., Aung, H. H., Osipov, A. N., Goh, N. K., and Chia, L. S. (2004) Carnivorous pitcher plant uses free radicals in the digestion of prey. *Redox Rep*. **9,** 255–261
55. Mithofer, A. (2011) Carnivorous pitcher plants: insights in an old topic. *Phytochemistry* **72,** 1678–1682
56. 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* DOI: 10.1093/bioinformatics/bts094
57. Dyrlund, T. F., Poulsen, E. T., Scavenius, C., Sanggaard, K. W., and Enghild, J. J. (2012) MS Data Miner: a web-based software tool to analyze, compare and share mass spectrometry protein identifications. *Proteomics*, in press, DOI: 10.1002/pmic.201200109

4.1.4 Impact of SO2 on Arabidopsis thaliana transcriptome in wildtype and sulfite oxidase knockout plants analyzed by RNA deep sequencing

Research

New Phytologist

Impact of SO_2 on *Arabidopsis thaliana* transcriptome in wildtype and sulfite oxidase knockout plants analyzed by RNA deep sequencing

Domenica Hamisch[1], Dörte Randewig[2], Simon Schliesky[3], Andrea Bräutigam[3], Andreas P. M. Weber[3], Robert Geffers[4], Cornelia Herschbach[2], Heinz Rennenberg[2,5], Ralf R. Mendel[1] and Robert Hänsch[1]

[1]Institut für Pflanzenbiologie, Technische Universität Braunschweig, Humboldtstraße 1, D-38106, Braunschweig, Germany; [2]Institut für Forstbotanik und Baumphysiologie, Professur für Baumphysiologie, Albert-Ludwigs-Universität Freiburg, Georges-Köhler Allee 53/54, D-79085, Freiburg, Germany; [3]Institut für Biochemie der Pflanzen, Heinrich-Heine-Universität, Universitätsstraße 1, D-40225, Düsseldorf, Germany; [4]Genome Analytics, Helmholtz Centre for Infection Research, Inhoffenstrasse 7, D-38124, Braunschweig, Germany; [5]King Saud University, PO Box 2454, Riyadh, 11451, Saudi Arabia

Author for correspondence:
Robert Hänsch
Tel: +49 531 391 5867
Email: r.haensch@tu-bs.de

Received: *29 May 2012*
Accepted: *9 August 2012*

New Phytologist (2012) **196**: 1074–1085
doi: 10.1111/j.1469-8137.2012.04331.x

Key words: Arabidopsis knockout mutants, cluster analyses, gene ontology, RNA-deep-sequencing, SO_2 fumigation, sulfate assimilation, sulfite detoxification, sulfite oxidase (SO).

Summary

- High concentrations of sulfur dioxide (SO_2) as an air pollutant, and its derivative sulfite, cause abiotic stress that can lead to cell death. It is currently unknown to what extent plant fumigation triggers specific transcriptional responses.
- To address this question, and to test the hypothesis that sulfite oxidase (SO) is acting in SO_2 detoxification, we compared Arabidopsis wildtype (WT) and SO knockout lines (SO-KO) facing the impact of 600 nl l^{-1} SO_2, using RNAseq to quantify absolute transcript abundances. These transcriptome data were correlated to sulfur metabolism-related enzyme activities and metabolites obtained from identical samples in a previous study.
- SO-KO plants exhibited remarkable and broad regulative responses at the mRNA level, especially in transcripts related to sulfur metabolism enzymes, but also in those related to stress response and senescence. Focusing on *SO* regulation, no alterations were detectable in the WT, whereas in SO-KO plants we found up-regulation of two splice variants of the *SO* gene, although this gene is not functional in this line.
- Our data provide evidence for the highly specific coregulation between SO and sulfur-related enzymes like APS reductase, and suggest two novel candidates for involvement in SO_2 detoxification: an apoplastic peroxidase, and defensins as putative cysteine mass storages.

Introduction

Sulfur is an essential nutrient for plant growth. Assimilatory reduction of soil-available sulfate is the main pathway of sulfur acquisition (Rennenberg, 1984), but plants can also use atmospheric sulfur dioxide (SO_2) gas as additional sulfur source (De Kok *et al.*, 2007). If, however, atmospheric SO_2 exceeds a critical threshold concentration, it becomes toxic for the plant and causes irreversible injury. Toxicity of sulfite strongly depends on dosages of SO_2, susceptibility of the plant species, and physiological and environmental factors (Bell, 1980; Ayazloo & Bell, 1981; Rennenberg, 1984; Alscher *et al.*, 1987; De Kok, 1990). Plants as sessile organisms have evolved several protection mechanisms: (1) the cuticle, which functions as the first barrier for toxic gases, largely restricting the pathway for influx to the stomata (Tamm & Cowling, 1977); (2) active stomatal closure, reducing SO_2 uptake (Rao & Anderson, 1983), and mesophyll resistances to SO_2 flux, mainly determined by metabolism of sulfite, which adjust SO_2 flux into leaves (Pfanz *et al.*, 1987); and (3) active detoxification of sulfite or bisulfite. These ions are metabolized within the plant either by feeding into sulfur assimilation, to form cysteine and other sulfur compounds (Filner *et al.*, 1984; Heber & Hüve, 1998), or by oxidation to sulfate by nonenzymatic (Rennenberg, 1984) or enzymatic processes (Pfanz *et al.*, 1990; Eilers *et al.*, 2001). Sulfite conversion to sulfate is catalyzed by the enzyme sulfite oxidase (SO) (Eilers *et al.*, 2001; Hänsch *et al.*, 2006). Loss of SO activity impairs the plant's ability to survive upon SO_2 exposure; conversely, overexpression of SO helps the plants to withstand even toxic SO_2 concentrations (Brychkova *et al.*, 2007; Lang *et al.*, 2007; Randewig *et al.*, 2012).

Recently, we used *Arabidopsis thaliana* wildtype (WT) and SO knockout (SO-KO) plants to decipher in detail responses to SO_2 fumigation in leaf rosettes (Randewig *et al.*, 2012). We identified the significance of SO for the overall shoot response to SO_2 in relation to alterations in plant phenology and physiology (gas exchange, metabolites and enzyme activities involved in assimilatory sulfate reduction). SO-KO and WT plants were exposed to SO_2 dosages that are known to be nontoxic to WT plants (Randewig *et al.*, 2012). Effects on sulfite detoxification and sulfur assimilation, particularly metabolic coregulation of enzymes

involved in sulfur assimilation, were compared. SO_2 exposure caused a significant increase in sulfate and glutathione (GSH) pool in wildtype Arabidopsis. Conversely, in KO plants the sulfate pool was kept constant, but thiol concentrations were strongly increased (14-fold for cysteine). Moreover, these metabolic changes were connected with a strong regulation of adenosine 5′-phosphosulfate reductase (APR) activity, the key enzyme of sulfate assimilation (Kopriva & Rennenberg, 2004). Based on these results we suggested a tight coregulation of SO and APR, thus controlling the sulfate assimilation pathway and stabilizing sulfite distribution.

Next, we conducted transcriptome analyses and followed a twofold strategy: the comparison of WT vs SO-KO plants before and after SO_2 fumigation would permit (1) a comprehensive analysis of the transcriptional regulation of sulfur metabolism, and (2) the deciphering of more complex and far-reaching reactions of the plant beyond sulfur metabolism. We hypothesized that sulfur metabolism in response to SO_2 is at least partially regulated at the transcriptional level and that an unbiased transcriptome analysis would permit the identification of unknown genes involved in the SO_2 response. For transcriptional profiling, sequencing-based techniques (RNA deep sequencing, RNAseq) offer numerous advantages over microarrays, such as: (1) a larger and more quantitative dynamic range of the experiment; (2) the ability to estimate absolute transcript numbers; and, therefore, (3) the opportunity to perform more accurate quantification of relative changes in transcript numbers. In the present paper we provide a detailed analysis of consequences of fumigation with *c.* 600 nl l^{-1} SO_2 for 60 h – a nontoxic dosage for *A. thaliana* wildtype plants – and compare the effect on WT and SO-KO plants. Fortunately, we were able to use plant samples that had already been analyzed in a previous study (Randewig *et al.*, 2012), permitting us to compare the transcriptome of WT and SO-KO plants under ambient and elevated SO_2 conditions with sulfur metabolite concentrations, a set of enzyme activities, and physiological data.

Materials and Methods

Plant material

For RNAseq experiments *Arabidopsis thaliana* plant samples of two different genotypes were used: *A. thaliana* (L.) Heynh. ecotype Columbia (WT plants) and transgenic SO knockout plants (SO-KO, GABI-Kat T-DNA insertion line (850B05) generated within the GABI-Kat program (Rosso *et al.*, 2003) kindly provided by Bernd Weisshaar (MPI for Breeding Research, Cologne, Germany)). Plantlets were grown in 500 ml plastic boxes at 22 : 20°C, day : night (8 h photoperiod) in controlled environmental chambers (HPS 1500, Voetsch Industrietechnik GmbH, Balingen, Germany). Eight-week-old plants were used for fumigation with 600 ± 15 nl l^{-1} SO_2. Four pots, each with four plants (WT and SO-KO), were placed separately into a single enclosure for 86 h. Three hours after the beginning of the dark period during the second night, SO_2 exposure was started and finished after 60 h. This treatment was reproduced with a new set of plants at least three times (for more details, see Randewig *et al.*, 2012).

Total RNA extraction and mRNA purification

Total RNA for WT, fumigated WT (WT[+]), SO-KO and fumigated SO-KO (SO-KO[+]) plants was isolated using the NucleoSpin® RNA Kit (Macherey-Nagel, Düren, Germany). For each genotype/treatment, 10 samples (each consisting of two plants randomly chosen from the three independent fumigation experiments) of 100 mg powdered plant tissue were separately used for total RNA isolation according to the manual (an exception to this was that elution was performed using two times 20 µl of RNAse-free H_2O). Total RNA preparations of 10 samples per probe set (WT, WT[+], SO-KO and SO-KO[+]) were pooled. Dynabeads® Plant Oligo$(dT)_{25}$ (Invitrogen, Darmstadt, Germany) were used for final mRNA purification according to the manufacturer's instructions. The quality of total RNA and isolated mRNA was checked using the Agilent 2100 Bioanalyzer RNA chip (Agilent Technologies, Böblingen, Germany). A sequencing library for RNAseq was created from 3 µg of mRNA using the SOLiD Whole Transcriptome Analysis Kit (Applied Biosystems, Carlsbad, California, USA). Thereafter, emulsion PCR was performed using SOLiD EZ bead kits. The resulting bead library was divided into three aliquots, loaded in separate flow cells and sequenced for 50 bp on an ABI SOLiD 5500XL system (Applied Biosystems). Using CLC workbench (CLC bio, Mühltal, Germany), transcriptome reads were aligned to whole genome sequences from the TAIR10 *A. thaliana* database (www.arabidopsis.org).

RNAseq data analyses

Reads were exported as color-space FASTA (filename.csfasta) and the associated quality (filename.qual) files and afterwards imported to CLC Genomics Workbench (CLC bio) using the NGS import function. Thus erroneous reads were cropped at the position of the error. Alignment and expression values in reads per kilobase of exon model per million mapped reads (RPKM) were obtained using the 'RNAseq Analysis' feature of CLC Genomics Workbench. RPKM are defined as follows:

$$\mathrm{RPKM} = \frac{\text{total exon reads}}{\text{mapped reads (millions)} \cdot \text{exonlength (kb)}}.$$

All four libraries were analyzed separately using standard parameters, that is, minimum length = 90%, minimum similarity = 80%, maximum number of hits for a read = 10, use color space = yes, type of organism = eukaryote. The reference was set to annotated *A. thaliana* chromosomes from the TAIR10 release 20 June 2009. The gene expression values were exported for further analysis.

DEGseq

Identification of differentially expressed genes was done using the R package 'DEGseq' (Wang *et al.*, 2010). This package allowed statistical analysis despite the lack of technical replicates. The underlying algorithm projects a random sampling model to the expression data to estimate the variance and calculates *P* values

based on this estimation. As input, RPKM for each gene were provided. The parameters were set to nondefault values: method = 'MARS', normal method = none (Supporting Information, Table S5). For each pair of input files, DEGseq provides a list of *P* values to determine significantly differentially expressed genes. Expression was considered significant if the uncorrected *P* value was < 0.001 (corresponding to Benjamini–Hochberg false discovery rate-corrected *P*-values, $P < 0.014$ for SO-KO control vs treated, < 0.017 for WT-treated vs SO-KO-treated, < 0.029 for WT control vs SO-KO control and below 0.087 for WT control vs WT treated). These genes were marked using the verbalization 'TRUE'; those remaining were tagged as 'FALSE'.

GeneSpring GX

Gene expression values from the RNAseq experiment were used within the GeneSpring GX software, version 11.5 (Agilent Technologies, Waldbronn, Germany). There were two data sets used within the studies: (1) absolute expression data and (2) log-scaled data.

Absolute, normalized data, which were not log-scaled and not processed using baseline transformation to the median of all sample data, were necessary to get a more detailed view of the transcription amounts of different genes. For this purpose, raw data were prepared in GeneSpring GX 11.5, pointing out that they were already log-scaled, which resulted in cutting extremely low data values using the 20th to 100th percentile normalization but no baseline transformation. The resulting data were absolute expression data and used for detailed analysis of transcript abundances involved in the sulfur metabolism. The detailed view of absolute transcript data is applicable for each requested gene and its associated splice variants, which is the primary advantage of using RNAseq data compared with microarrays where transcript abundances are always given in relation to several control genes and not taken individually.

To prepare **log-scaled data**, RPKM values were processed using the GeneSpring GX 11.5 generic single-color experiment according to the manual. After the normalization step, experimental data were grouped as genotype × treatment : WT/SO-KO × control/treated with 600 nl l^{-1} SO_2. To identify genes that show differences between treated and control samples or samples with different genotypes, expression ratios (**fold-changes**) were calculated in the following way:

$$\text{Expression ratio} = \text{Fold change} = \frac{\text{Condition 1}}{\textit{Condition2}}.$$

Ratios below or above a determined cutoff show that these genes are *x*-fold up- or down-regulated. Four different pairs of conditions were used within the fold-change analyses: WT vs WT[+]; SO-KO vs SO-KO[+]; WT vs SO-KO; and WT[+]vs SO-KO[+]. Data that were fivefold up- or down-regulated and DEGseq-verified data were used to obtain deeper insights into regulation of other processes beyond sulfur metabolism.

Using the GeneSpring GX 11.5 **cluster analyses** tool, hierarchical clustering was performed for data with a fivefold-change threshold verified with DEGseq. Hierarchical clustering was carried out on entities (differentially regulated genes) and conditions (different genotypes and treatments) using combined trees. Merging of entities in different clusters is controlled by applying a certain linkage rule; here we used 'complete'. Cluster entities were colored according to the numeric values of the normalized, log_2-scaled data. Expression profiles from each of the eight identified clusters were generated; transcripts belonging to the different clusters were exported and used for Gene Ontology (GO) analyses.

The GO database (www.geneontology.org) describes connections between gene expression data and defined GO terms. Using the GeneSpring GX 11.5 GO analysis tool, entities of interest obtained from one experiment can be explored, finding matching GO terms. The output of GO analysis is a tree containing GO terms enriched with a *P*-value cutoff of 0.1. Transcripts belonging to the different clusters (I to VIII) defined in the cluster analyses were used for GO analysis.

Results and Discussion

General view of the Arabidopsis transcriptome under SO_2 fumigation

A total number of 22 130 genes, including their different splice variants, were identified for WT plants in this experiment: 23 232 for WT[+], 22 424 for SO-KO, and 22 255 for SO-KO[+]. Quantile-normalized, log_2-scaled and nonbaseline-transformed RPKM of these transcripts were widely spread (Fig. S1). Each analyzed sample consisted of 10 independently prepared RNAs from a total of 20 treated plants. The biggest spread of RPKM, and therefore the greatest change in transcripts, was detected for WT vs SO-KO[+], followed by SO-KO vs SO-KO[+]. A narrower distribution, indicating a weaker response to SO_2 or the genotypic modification, was identified in WT vs WT[+], WT vs SO-KO and SO-KO vs WT[+]. Fig. 1(a,b) present the amounts of differentially expressed genes in relation to the total number of transcripts (different splice variants included) using data within a fivefold-change cutoff, verified with DEGseq (Table S1). With the fivefold-change threshold, between 0.4 and 1.6% of genes were differentially expressed for all condition pairs. Approx. 60% of the **differentially expressed genes** were up-regulated (Fig. 1a), whereas *c.* 40% were down-regulated (Fig. 1b).

Venn diagrams (Fig. 1c,d) were created to investigate several hypotheses concerning the biological evidence of the genotypic variation in SO-KO and the effects caused by SO_2 treatment. The Venn diagrams show the number of fivefold up- (Fig. 1c) or down-regulated (Fig. 1d) transcripts found in different treatment–genotype combinations and which of those genes were differentially expressed in different condition pairs.

First, we asked if knocking out the *SO* leads to the same change in transcripts and transcript numbers as does fumigation of the WT. If this were the case there would be more transcripts in the overlap of WT vs WT[+] and WT vs SO-KO and fewer in the section of solely regulated transcripts. We identified 11 transcripts for up-regulation and 10 for down-regulation in the overlapping section. For the genotypic comparison, we found 102 transcripts up-regulated

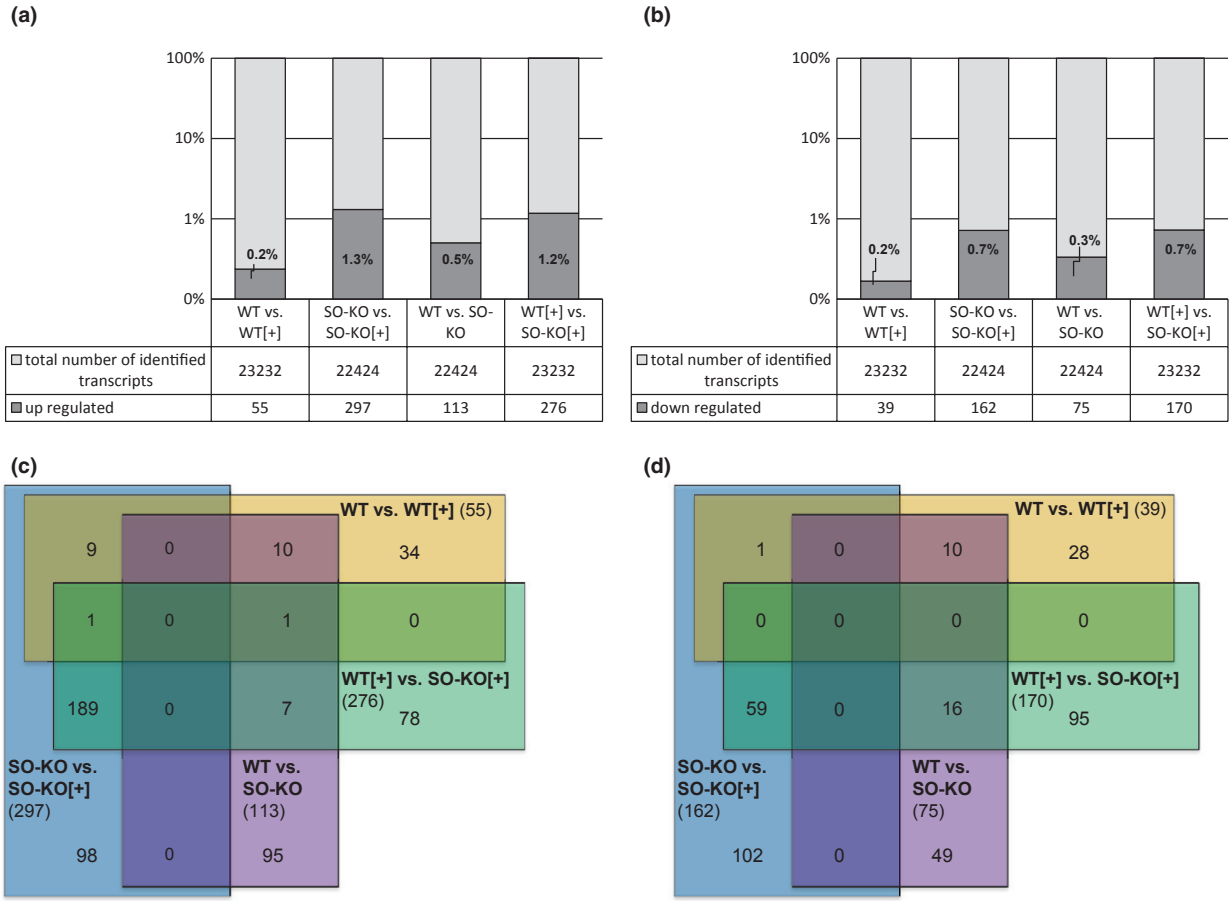


Fig. 1 Differentially expressed transcripts and Venn diagrams presenting intersections after combining the four different *Arabidopsis thaliana* comparison pairs: wildtype (WT) vs SO_2-fumigated WT (WT[+]), sulfite oxidase knockout (SO-KO) vs SO-KO[+], WT vs SO-KO and WT[+] vs SO-KO[+]. The numbers and percentages of differentially expressed transcripts for fivefold up-regulated (a) and down-regulated (b) genes presented the highest abundance of up- and down-regulated transcripts in SO-KO vs SO-KO[+]; the lowest number was observed for WT vs WT[+]. Venn diagrams represent numbers of fivefold up-regulated (c) and down-regulated (d) transcripts when different comparisons overlap. The Venn diagram for up-regulation showed two transcripts which are solely unregulated in SO-KO vs SO-KO[+] and WT vs SO-KO. [+] indicates SO_2 fumigation.

and 65 transcripts solely down-regulated, and there were 44 up-regulated transcripts and 29 down-regulated transcripts for WT vs WT [+]. These findings ran counter to our hypothesis that the gene expression changes caused by SO knockout were similar to those caused by SO_2 fumigation. This result was, however, in line with the second assumption that we would not find any intersections for WT vs SO-KO and SO-KO vs SO-KO[+], because *SO* gene knockout and fumigation of SO-KO would not have any regulated transcripts in common, and interpretation of our data validated this expectation. Thirdly, SO is predicted to play a key role in SO_2 detoxification, and we therefore hypothesized a higher number of regulated transcripts in SO-KO[+] than in WT[+], since the knockout of *SO* inhibits SO-mediated SO_2 protection. This may further induce other processes. Our findings of 45 up- and 38 down-regulated transcripts in WT vs WT[+] and 287 up- and 161 down- regulated transcripts in SO-KO vs SO-KO[+] confirmed this hypothesis. Only 10 transcripts for up-regulation and one for down-regulation were identified as commonly regulated. Fourthly, we expected that SO-KO plants would have to use different defense mechanisms to detoxify SO_2 than those used by WT plants. If this were the case, we would therefore find only very few transcripts that were commonly up- or down-regulated in WT vs WT[+] and WT[+] vs SO-KO[+]. Transcripts in the overlap represented genes that were already highly regulated in WT[+] and even more highly regulated in SO-KO[+]. This was a small common transcript set resulting from the different transcript usages of WT[+] and SO-KO[+] during fumigation. For up-regulation we identified two transcripts: AT5G44420, which belongs to the plant defensin family (plant defensin 1.2), and AT3G44310, encoding the nitrilase 1. The genotypic comparison (WT vs SO-KO) marked the defensin as an unregulated transcript (1.19-fold), whereas the nitrilase was significantly regulated (3.97-fold) in SO-KO vs SO-KO[+], but this was not visible in the fivefold comparison. For down-regulation there was no transcript detectable in the intersection. Verification of this hypothesis led to the fifth expectation, that we would find more genes regulated in SO-KO vs SO-KO[+] than in WT[+] vs SO-KO[+], but overall a high number of commonly regulated transcripts. Counting the transcripts for SO-KO vs SO-KO[+] revealed 297 up-regulated and 162 down-regulated transcripts. For WT[+] vs SO-KO[+] we found 276 up-regulated and 170 down-regulated transcripts: 249 transcripts were commonly regulated in SO-KO vs SO-KO[+] and WT[+] vs

New Phytologist

SO-KO[+], and a total of 408 transcripts were solely regulated. These results confirm a high number of commonly regulated transcripts (50%), but calculations did not verify our hypothesis of a much higher percentage of regulated genes in SO-KO vs SO-KO [+] than in WT[+] vs SO-KO[+].

To obtain further insights into the molecular mechanisms affected by knocking out *SO*, we identified differentially expressed genes by comparing WT and SO-KO with and without SO_2 fumigation. Significantly regulated genes with a greater than fivefold transcriptional change were selected and significance was determined with the DEGseq tool. We applied **hierarchical clustering** (Fig. 2a) to further delineate associated gene groups with similar expression profiles (Fig. 2b). Most of the transcriptional changes were induced after SO_2 fumigation in the SO-KO mutant plants, represented by the largest clusters, IV and VII. In total we were able to identify eight individual gene expression clusters numbered from I to VIII. For further analyses we used the top 20 regulated transcripts and **GO** analyses for each identified cluster.

Application of SO_2 to WT plants should lead to several transcriptional changes, but because of the dosage of 600 nl l^{-1} used and because SO acts as a detoxificating enzyme, we hypothesized a smaller reaction than we would expect for SO-KO treatment. GO analysis of fivefold regulated transcripts in WT vs WT[+] was in line with this expectation by revealing the lowest number of GO terms. Transcripts could be assigned to the categories BIOLOGICAL PROCESS (16 genes, 36%) and CELLULAR COMPONENT (29 genes, 64%). We identified up-regulated

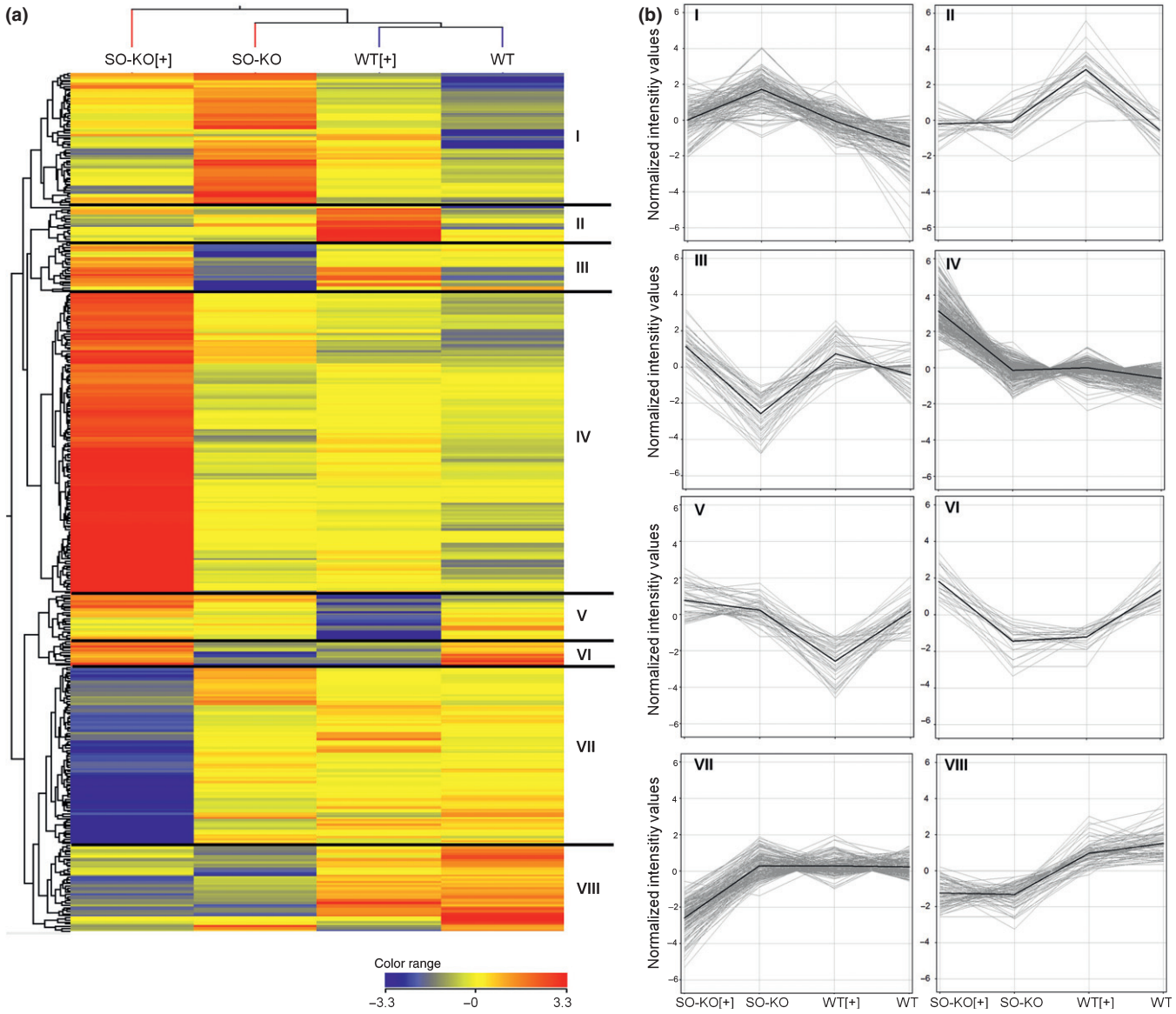


Fig. 2 Hierarchical clustering of fivefold regulated transcripts and profile plots of selected clusters. Colors of the cluster in (a) were assigned based on the normalized, log-scaled RPKM (reads per kilobase of exon model per million mapped reads) values (significance tested using DEGseq). In panel (b), expression profiles of transcripts involved in these eight clusters are depicted. Cluster IV shows transcripts which are solely up-regulated in the SO_2-fumigated *Arabidopsis thaliana* sulfite oxidase knockout (SO-KO[+]); cluster VII includes those which are mutually down-regulated in SO-KO[+]. Both clusters presented the highest number of involved transcripts.

transcripts for WT vs WT[+] in clusters I, II, and III; down-regulation was detected in clusters V, VI, and VIII. The top 20 transcripts in cluster II included four transcripts associated with ribosomal and translation processeses, which indicated an influence of SO_2 fumigation on mRNA synthesis regulation. Cluster II contained transcripts which showed their highest abundances in WT[+] only and which therefore revealed a moderate reaction of WT plants to SO_2 with transcriptional adaptation and thus up-regulation of transcripts belonging to ribosomal processes.

Hypothetically, treatment of SO-KO plants with SO_2 should lead to higher and different transcriptional responses compared with treated WT plants, as this was already verified by Venn diagrams. Additionally, GO analyses of fivefold regulated transcripts showed the highest numbers of significantly enriched transcripts ($P < 0.1$) in SO-KO vs SO-KO[+] (Fig. S2), which was also clear from the Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses (Fig. S3). GO terms enriched in SO-KO vs SO-KO[+] were principally associated with the terms MOLECULAR FUNCTION and BIOLOGICAL PROCESS. Up-regulation of transcripts was identified in clusters III, IV, V and VI, while down-regulation was identified in clusters I and VII. Transcript functions for up-regulation in SO-KO vs SO-KO[+] differed from those identified as up-regulated for WT vs WT[+]. The WT[+] top 20 included the greatest number of transcripts associated with transcriptional regulation (cluster I, II), whereas the SO-KO[+] top 20 included transcripts involved in defense processes (cluster IV) and peptidase activity (cluster VI). The top 20 of cluster IV contained nine transcripts associated with defense, including defensins (AT5G44420, AT5G44430, AT2G26020 and AT2G26010), GSTs (AT1G02930, AT1G02920, AT4G02520) and peroxidase CB (AT3G49120).

Transcriptional regulation of enzymes related to sulfur metabolism: effects resulting from SO_2 and/or genotypic variation of SO knockout

We recently reported enzyme activities and S-metabolites of *A. thaliana* WT and transgenic lines subjected to SO_2 fumigation (Randewig *et al.*, 2012). Aliquots of the same plant material were used in this study for the RNAseq experiment. This combination of transcriptome data with enzyme activities and metabolite concentrations provided new insights into the regulation of sulfate assimilation and related metabolic pathways. We hypothesized that excess SO_2, especially in the absence of SO, would lead to transcriptional down-regulation of the enzymes producing sulfite and transcriptional increases in at least some enzymes mediating reactions downstream of sulfite to sequester the excess organic sulfur produced. The following description and discussion of the results are summarized in Fig. 3, which presents the regulation of genes for SO-KO vs SO-KO[+]. Raw data and fold changes are presented in Table S2.

Sulfate is taken up by the root and transported via the xylem stream into the leaves for further assimilation. Required **sulfate transporters (SULTR)** are divided into subfamilies on the basis of their protein sequence similarities (Hawkesford, 2003; for review, see Davidian & Kopriva, 2010). With the exception of SULTR3;1, none of the sulfate transporters was significantly regulated in leaves as analyzed by the DEGseq tool. However, from the group two of the SULTR, responsible for long-distance transport and localized in xylem parenchyma cells, *SULTR2*;1 transcript abundances were similar in the nonfumigated plant material, but increased in WT[+] by 30% or decreased in SO-KO[+] by 50%. Moreover, the expression of *SULTR2*;2 was down-regulated threefold in SO-KO plants in the fumigation experiment, which possibly reflects a reduction in sulfate uptake and transport. From the SULTR of group four – suggested to function in vacuolar sulfate remobilization to the cytoplasm in roots and leaves (Kataoka *et al.*, 2004) – *SULTR4*;2 mRNA amounts were two- and 10-fold decreased during fumigation in WT and SO-KO rosettes, respectively. This led us to the hypothesis that, particularly in SO-KO, sulfite cannot be oxidized to sulfate and hence there is no requirement for sulfate to be introduced into the assimilatory stream via SULTR2 and SULTR4.

For assimilation, sulfate has to be activated by the **ATP sulfurylase (APS)**, which catalyzes the first step in this pathway. Determination of *APS* transcript abundances revealed that *APS1* was the most abundant (between 244 and 349 quantile-normalized RPKM; Table S2) and the only isoform that was significantly down-regulated (1.3-fold) in SO-KO[+] and when comparing WT[+] and SO-KO[+]. This supported the hypothesis that sulfate reduction is down-regulated transcriptionally if excess SO_2 is present. The activated sulfate is partly converted into PAPS (3′-phosphoadenosine 5′-phosphosulfate) by one of the four isoforms of **APS kinase (APK)**. Compared with all other samples, a significant decrease (two- to threefold) of *APK1-3* mRNA was detected only in SO-KO[+] plants. Plants possess large numbers of **sulfotransferases (SOTs)** that are responsible for sulfonation of small molecules by using PAPS, cysteine, or other reduced S-compounds, as an important component of plant stress responses (Klein & Papenbrock, 2004). SOTs thus can sequester organic sulfur. Fumigation of SO-KO led to an almost 13-fold increase of *SOT12* transcript amounts. *SOT12* is known to be stress-inducible and has been described to confer pathogen resistance in *A. thaliana* by sulfonation of salicylic acid (Baek *et al.*, 2010).

The majority of activated sulfate is metabolized further on by **APS reductase (APR)**. Three isoforms described in the literature are localized in the chloroplast. APR is known to be the key enzyme of the sulfate assimilation pathway (Kopriva & Rennenberg, 2004) and is regulated transcriptionally and post-translationally, respectively (Kopriva & Koprivova, 2004). Our data confirm these findings: in WT plants, 600 nl l^{-1} SO_2 did not change the transcript abundances of any APR isoform or splice variant. The enzyme activity decreased significantly (Randewig *et al.*, 2012), presumably as a result of feedback inhibition (Vauclare *et al.*, 2002). In SO-KO control plants, *APR1*, *APR2* and *APR3* transcripts were increased significantly compared with WT control samples (Table S2). Fumigation of these SO-KO plants led to a dramatic decrease in both transcript abundance (Table S2) and enzyme activity levels (Randewig *et al.*, 2012). The strong down-regulation of APR reflects a tight control of sulfite synthesis at the transcriptional level as well as at the post-translational level. Such a negative feedback inhibition of APR mediated by increasing

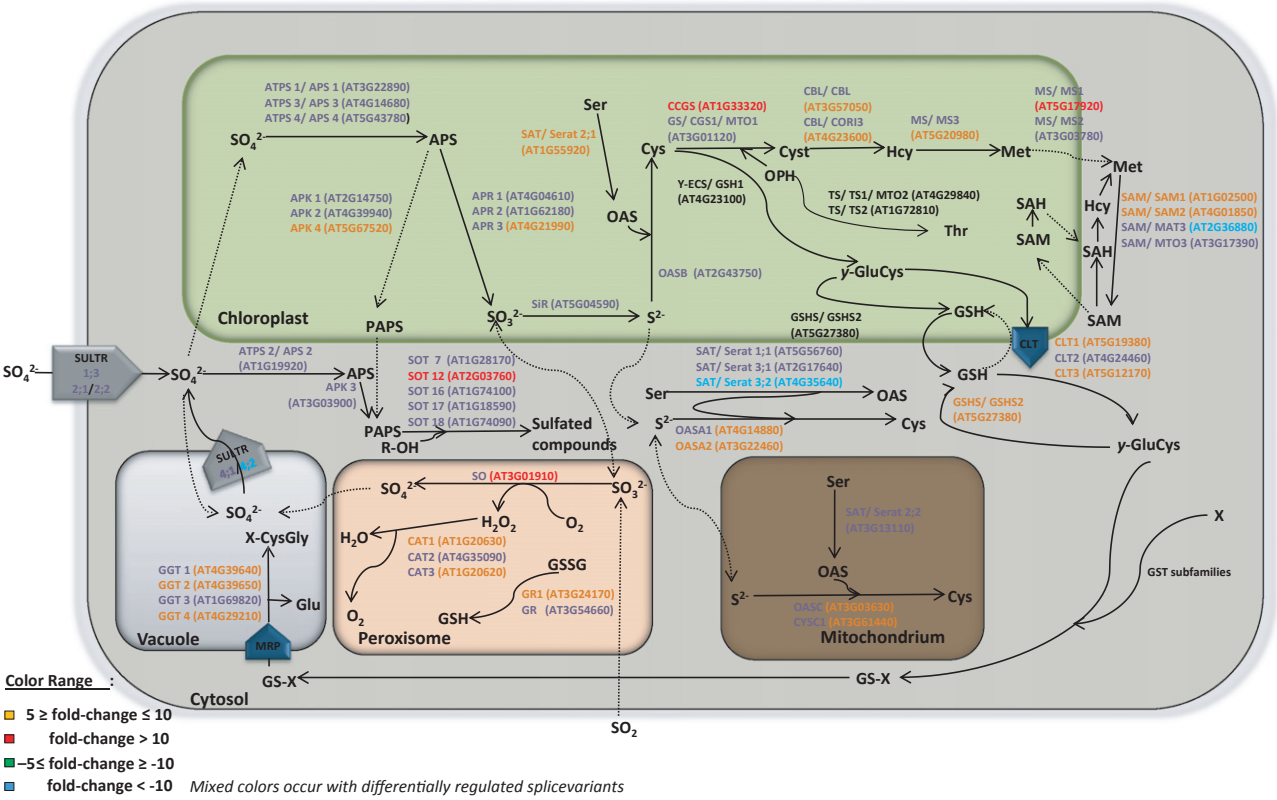


Fig. 3 Alterations in the sulfur metabolism for the *Arabidopsis thaliana* sulfite oxidase knockout (SO-KO) vs SO_2-fumigated SO-KO (SO-KO[+]) comparison. Regulation of sulfur metabolism-associated genes for SO-KO[+] is presented using different colors, as described in the color range. Gene name abbreviations are described in Table S2. To generate this scheme no fold change was applied to the expression values; mostly the glutathione S-transferases (GSTs) and the sulfotransferase (SOT) show a strong reaction in the SO-KO[+], in particular in the up-regulated transcripts.

amounts of thiols (Vauclare *et al.*, 2002) has been discussed previously (for a review, see Kopriva & Koprivova, 2004; Davidian & Kopriva, 2010). Moreover, GSH itself is involved in cell proliferation of cell cultures and lateral roots of Arabidopsis (Vivancos *et al.*, 2010) as well as in meristem development and embryo maturation of Brassica (Stasolla *et al.*, 2008) by changing the transcript abundance of definite genes. This has also been proposed by Szalai *et al.* (2009) for abiotic stress conditions either via H_2O_2 or GSH/ oxidized glutathione (GSSG) in general.

Sulfite reductase (SiR) converts sulfite into sulfide and is a single copy gene. **O-acetylserine(thiol)lyase (OASTL, OASx, CYSC1)** and **serine acetyltransferase (SAT, Seratx;x)** together form the cysteine synthase complex (Wirtz *et al.*, 2010), which produces organic sulfocompounds from sulfide (Fig. 3). For *SIR* transcripts, no significant regulation was observed for WT and SO-KO plants after fumigation. OASTL and SAT enzymes occur as different isoforms, which are located in several cellular compartments. OASB and Serat2;1 are localized in chloroplasts, Serat2;2, OASC and CYSC1 were described to act within the mitochondria, and OASA1, OASA2, Serat1;1, Serat3;1 and Serat3;2 are cytosolic enzymes (Jost *et al.*, 2000; Yamaguchi *et al.*, 2000; Kawashima *et al.*, 2005). We found that, after fumigation of SO-KO, only transcripts encoding the chloroplastidic *SERAT2;1* and mitochondrial *CYSC1* were significantly increased. This provokes the hypothesis that S-assimilation, and therefore cysteine synthesis, is possibly induced after fumigation. In summary, *OASTL* and *SAT* transcript abundances showed that additional sulfur was channeled into the direction of cysteine production after SO_2 fumigation, which held true for both fumigated WT and SO-KO plants.

Organic sulfur may flow towards methionine via cystathione or towards GSH. **Cysteine gamma-synthase (CGS/MTO1)** is involved in the conversion of cysteine into cystathionine. *CGS* sequencing data did not show any alterations after fumigation regarding WT and SO-KO samples. **Cystathionine beta-lyases (CBL and CORI3)** are involved in the conversion of cystathionine into homocysteine. For SO-KO and SO-KO[+], significantly lower transcript abundances (roughly fivefold) were found for all of the three *CORI3* splice variants. Moreover, after fumigation of WT plants, *CORI3* transcripts were up-regulated. *CGS* was expressed at a higher level than *CBL* (Table S2). RNAseq data showed only minor alterations in **methionine synthase (MS)** and **S-adenosyl-methionine synthetase (SAM-synthetase)** transcripts. MS converts homocysteine into methionine, which could be used by SAM-synthetase to generate S-adenosylmethionine as a methyl group donor in numerous transmethylation reactions (Peleman *et al.*, 1989). Only *MS2* displayed a twofold decrease in SO-KO[+] compared with SO-KO, and a threefold decrease in SO-KO[+] compared with WT [+]. Based on the transcriptional profile, the excess SO_2 did not flow towards methionine.

Gamma glutamylcysteine synthase (γ-ECS) catalyzes the first step in GSH biosynthesis. We identified a small increase in *γ-ECS* transcript amounts for SO-KO[+]. **Glutathione synthetase (GSHS)** produces GSH from γ-glutamylcysteine and glycine. *GSHS* transcripts showed an average abundance of 30 RPKM for WT, WT[+] and SO-KO. Transcript abundances for SO-KO[+] were significantly higher (2.6-fold) compared with other samples, indicating that the higher amounts of produced γ-glutamylcysteine are converted into GSH. **Glutathione reductase (GR)** reduces GSSG to GSH. GR1 had enhanced transcript abundance, especially in SO-KO[+]. Transcript abundances of SO-KO and SO-KO[+] were higher than those measured for WT and WT[+]. Higher amounts of *GR1* transcripts in SO-KO[+] samples may indicate that higher amounts of GSSG have to be reduced back to GSH during the SO_2 fumigation process. Based on the transcriptional profile, the excess SO_2 flowed towards GSH. These transcriptional up-regulations of several enzymes in GSH biosynthesis fitted well with the increased amount of γ-glutamylcysteine and GSH measured in these samples previously (Randewig *et al.*, 2012) and confirmed the hypothesis of an enhanced S-flux into the S-assimilation stream. However, accumulation of GSH above a specific threshold could be dangerous to plant cells, causing increased oxidative stress in tobacco (Creissen *et al.*, 1999) and affecting photosynthesis, growth and sulfur metabolism in poplar (Herschbach *et al.*, 2010). Moreover, GSH is demonstrated to be the sulfur donor in the biosynthesis of glucosinolates in Arabidopsis (Schlaeppi *et al.*, 2008; Geu-Flores *et al.*, 2011). However, in our investigation, transcript data of the key enzymes processing GSH conjugates into glucosinolate and camalexin pathways (γ-glutamyl peptidases GGP1 (AT4G30530) and GGP3 (AT4G30550)) were not altered or even slightly decreased (Table S1), suggesting that formation of glucosinolates was only a minor sink for excess sulfur, as was also shown in previous studies with *Arabidopsis* (Van der Kooij *et al.*, 1997). Therefore a supplemental mass storage for the reduced sulfur should be postulated.

Glutathione S-transferase (GST) proteins are arranged in different subfamilies, GSTF, GSTL, GSTT, GSTU and GSTZ (Frova, 2003). In all GST subfamilies, isoform transcripts seemed to be regulated after fumigation with SO_2, which was especially obvious for SO-KO[+], with an up-regulation of 10-fold and higher. *GST6* (AT1G02930) was 10.5-fold up-regulated in response to the fumigation stress.

In WT plants, excess sulfite can be detoxified by oxidation to sulfate (Fig. 3). **Sulfite oxidase (SO)** counteracting the APR is supposed to be the most effective tool within the plant cell for removing excess amounts of sulfite (Brychkova *et al.*, 2007; Lang *et al.*, 2007; Randewig *et al.*, 2012). As shown very recently using microarrays, fumigation of grape berries with 1–3 μl l^{-1} SO_2 surprisingly resulted in a decrease of *SO* transcripts (Giraud *et al.*, 2012). In our RNAseq experiment, transcript numbers of all three different SO splice variants were determined. SO splice variant 1 (*SO-1*) showed the highest transcript abundance for WT and WT[+] (60 and 57 normalized RPKM) in comparison to the two additional splice variants (0.07 and 0.16 normalized RPKM for *SO-2*, and 0.2 and 0.45 normalized RPKM for *SO-3*). SO-KO plants lacked detectable amounts of SO protein, as determined by immunoblot analysis, because of the T-DNA insertion within this gene (Lang *et al.*, 2007); consequently, no SO activity was detectable (Randewig *et al.*, 2012). Activity measurements applied in Randewig *et al.* (2012)showed no alterations in SO activity for WT[+]. RNAseq confirmed this result (Table S2). Surprisingly, RNAseq data showed that in SO-KO the *SO* transcripts were detectable. However, a closer look into the sequencing data (i.e. mapping of sequence reads to Arabidopsis mRNA sequences) showed that SO-KO did not produce a functional transcript. Although the reading frame for the transcript is disrupted as a result of the T-DNA insertion and the resulting mRNA is thus noncoding, the transcriptional response apparently attempts to enhance SO production (Figs S4, S5). For *SO*-1 we detected 2.5-fold (and for *SO*-2 *c.* 16-fold) increased transcript abundance after fumigation with 600 nl l^{-1} SO_2 for 60 h. At present, the physiological relevance of the different splice variants is unclear. The current interpretation of *SO*-1 and *SO*-2 abundances could only be alternative splicing as known from other eukaryotic systems (Graveley, 2005; Smith, 2005).

Transcriptional regulation of biological processes beyond sulfur metabolism

We hypothesized that additional lines of defense against SO_2 and additional consequences of SO_2 poisoning could be deduced from the global transcriptome analysis (see also Figs S3, S4). Therefore, RNAseq supported the development of a new regulation model (Fig. 4) based on transcript data, explaining plant reactions to excess SO_2. Investigations of regulatory mechanisms beyond sulfur metabolism were based on transcripts enriched in specific gene groups or processes for the 717 genes identified as fivefold regulated in at least one of the comparison pairs, WT vs WT[+], SO-KO vs SO-KO[+], WT vs SO-KO and WT[+] vs SO-KO[+].

Photosystem components Several studies describe the down-regulation of photosystem components after application of stresses, leading to an inhibition of energy production, increased oxidative stress (Chaves *et al.*, 2009) and the activation of catabolic processes. Comparing the genotypic changes between SO-KO and WT, different components of the photochemical apparatus are down-regulated, but here the limit is a twofold down-regulation. However, the analysis of SO-KO data after fumigation shows an even stronger down-regulation of these transcripts, reaching threefold changes and a higher number of regulated genes. The influence on photosynthesis was also stated in Randewig *et al.* (2012). The CO_2 assimilation rate was almost halved in SO-KO after SO_2 fumigation, and both the stomatal conductance ($g_{(H2O)}$) and the SO_2 uptake rate were reduced. SO_2 fumigation as well as inhibition of photosynthesis resulted in strong oxidative stress for the plant. As a consequence, genes associated with the oxidative stress response should also be a subject of regulation.

Senescence-associated genes Contact with SO_2 should lead to enhancements of the senescence processes depending on the dosage, a hypothesis we arrived at because of the initial phenotypic symptoms of injury, with small necrotic spots on the leaf surface

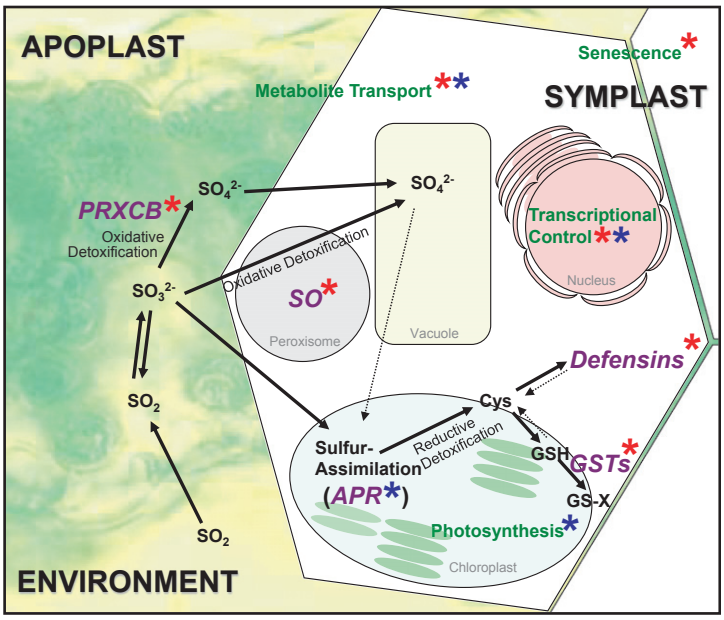


Fig. 4 Current working model derived from RNAseq data interpreting overall plant reaction and detoxification mechanisms in the face of excess SO_2. SO_2 enters the plant cell via the stomata, where it is converted into sulfite: sulfite can be detoxified by an apoplastic peroxidase (PRXCB) and sulfite oxidase (SO) in an oxidative reaction or fed into the sulfur assimilation stream (reductive detoxification). These and further specific transcripts (violet) and processes (green) were identified as being up-regulated (red star) or down-regulated (blue star); transcript isoforms involved in specific processes in some cases presented different regulations (red and blue star). In general, SO_2 fumigation leads to different reactions, including transcriptional control by regulation of transcription factors, changes of metabolite transport, induction of senescence and down-regulation of photosynthetic processes. This model hypothesizes the apoplastic peroxidase PRXCB as a protagonist for plant SO and uncovers plant defensins as a novel mass storage of reduced sulfur. Cys, cysteine; GSH, glutathione; GST, glutathione S-transferase.

detected in the fumigated plant material (Randewig *et al.*, 2012). RNAseq data confirmed these observations at the transcriptional level: SO-KO plants which are not able to remove SO_2 using SO should present higher numbers of genes associated with senescence. Confirming these expectations, we identified most up-regulated transcripts with the highest fold changes in SO-KO[+]: eight genes with greater than fivefold changes presenting a maximum at 18.2 normalized RPKM. For the genotypic comparison we found two genes more than fivefold up-regulated, which were further down-regulated or not regulated in SO-KO vs SO-KO[+]: the *senescence-associated gene 29* (*SAG29*, AT5G13170) was 7.5-fold up-regulated, and the *dark inducible 2* gene (*DIN2*, AT3G60140) was 5.2-fold up-regulated in SO-KO in the genotypic comparison to WT. Increases in transcript abundances for WT vs SO-KO samples showed that already the genotypic variation has an effect on senescence processes. Up- or down-regulation of senescence-associated genes was not observed in the WT after fumigation; no transcripts > fivefold regulated were detected.

Transcriptional regulation Cluster analyses revealed different regulation patterns of transcription factors after SO_2 fumigation and in the genotypic comparison. The highest enrichment of regulated transcription factors was expected for SO-KO vs SO-KO[+], followed by WT vs WT[+]. Overall, we identified 41 genes associated with transcriptional regulation out of 717 fivefold regulated genes. The highest fold-changes were detected for fumigated SO-KO (24 genes more than fivefold regulated), whereas WT vs WT[+] samples revealed three genes, and WT vs SO-KO 12 genes, that were more fivefold up- or down-regulated. Therefore we can assume that transcriptional regulation mainly plays a role in treated SO-KO plants.

Transporters With respect to the GO analysis, we hypothesized an enhanced regulation of transcripts associated with transport after fumigation with SO_2 in all comparisons. An enrichment of genes involved in transport was confirmed by 41 genes out of 717 fivefold regulated genes. For SO-KO vs SO-KO[+] we identified 29 genes regulated over fivefold, six for WT vs SO-KO and WT vs WT[+] samples. A large number of genes belonging to the multidrug and toxin extrusion (MATE) family of efflux pumps was identified to be up-regulated in SO-KO[+]. MATE efflux transporters were already identified in *A. thaliana* via microarray and showed an induced transcription in plants treated with high amounts of boron (Kasajima & Fujiwara, 2007). We found four MATE efflux transcripts (AT2G04050, AT2G04040, AT3G23550, AT2G04070) of as yet unknown function which showed 30- to 65-fold higher transcript abundance and were thus remarkably up-regulated in SO-KO[+] plants.

Oxidoreductases and response to oxidative stress The influence of SO_2 should lead to plant reactions, including several oxidative and reductive processes; therefore we expected and confirmed an enrichment of altered transcripts associated with oxidoreductase activity. We detected 57 genes out of 717 fivefold regulated genes that are involved in oxidoreductase processes. Forty-one genes

were identified as over fivefold regulated for SO-KO vs SO-KO[+], nine for the genotypic comparison and seven for WT vs WT[+]. These data confirmed the hypothesis of a higher regulation of oxidative and reductive processes in fumigated SO-KO plants. Oxidoreductases include the group of peroxidases that were as much as 54.8-fold up-regulated in SO-KO[+], compared with 10-fold up-regulation in WT[+] (Table S4). Peroxidase CB (AT3G49120), which belongs to the class III peroxidases, was fivefold up-regulated in SO-KO[+]. This peroxidase is localized in the apoplast (PeroxiBase, http://peroxibase.toulouse.inra.fr; Shah *et al.*, 2004) and is involved in cell wall elongation (Irshad *et al.*, 2008) and reactive oxygen species (ROS) generation under biotic stress reactions (Bindschedler *et al.*, 2006). Moreover, Pfanz and colleagues studied apoplastic peroxidases in response to SO_2 fumigation and suggested a role in SO_2 detoxification (Pfanz *et al.*, 1990; Pfanz & Oppmann, 1991). The present interpretation of these findings is that plants have two independent methods of SO_2 detoxification: one in the apoplastic space and the other at the cellular level, which will be in the focus of future work.

Defense Fumigation of *A. thaliana* resulted in higher expression of several defense-related genes. We identified 56 out of 717 fivefold regulated genes associated with defense processes. Seven genes in this group are plant defensins (PDFs) or plant defensin-like proteins. Defensins are small (4–6 kDa) peptides, whose three-dimensional structures are stabilized via eight disulfide-linked cysteines (Thomma *et al.*, 2002). These peptides represent 0.5% of the whole-plant protein content (Stotz *et al.*, 2009) and belong to the family of antimicrobial peptides (Kovaleva *et al.*, 2010). In WT[+] and SO-KO[+], the same PDFs were up-regulated: PDF1.2 (AT5G44420), PDF1.2b (AT2G26020), PDF1.2c (AT5G44430) and PDF1.3 (AT2G26010), a defensin-like protein (AT2G43510), as well as the low-molecular-weight cysteine-rich 67 protein (LCR67, AT1G75830). In general, defensin genes were four- to fivefold up-regulated in WT[+], but in SO-KO[+] these transcripts showed the most impressive and highest regulation found in this RNAseq experiment. Here 17.8- to 244.5-fold higher transcript abundances were measured. Moreover, for these defensins, the highest RPKM values (Table S3) were measured: 2936.8 RPKM for PDF1.2 in SO-KO [+]. Comparing genotypic changes between SO-KO and WT, no defensins were differentially expressed. One possible and logical reason could be the mass storage of reduced sulfur in these cysteine-rich peptides additional to GSH. In defensins, typically four to eight amino acids out of 45–54 amino acids are cysteine residues.

Conclusion

In the present study, SO_2 at a concentration of 600 nl l^{-1} was applied to Arabidopsis WT and SO-KO for 60 h, which represents neither fully acclimated plants nor immediate stress responses. Before the current investigations of mRNA alterations, S-metabolism-related enzyme activities and S-metabolite concentrations were determined using aliquots of the same plant material (Randewig *et al.*, 2012). Changes in S-metabolite concentrations of WT and SO-KO plants in response to SO_2 were related to enzyme activities and absolute transcript abundances of mRNA. Removal of excess sulfate by conversion into sulfur-containing compounds via the sulfur assimilatory stream in response to SO_2 exposure – as concluded from S-associated enzyme activities (Randewig *et al.*, 2012) – was supported by transcript data of the present investigations. These results make the hypothesis of a tight coregulation between SO and APR plausible, meaning there is a role in keeping the intracellular sulfite pool constant at a low concentration.

To prevent damage from atmospheric SO_2, additional mechanisms play a role *in planta* as well. In this RNAseq experiment, two other factors that are possibly involved in sulfite detoxification, and which therefore assist in coregulation of APR and SO, were identified: PDFs, a group of small cysteine-rich peptides, and a peroxidase which is localized in the apoplastic space (Shah *et al.*, 2004). Up-regulation of PDFs after fumigation seems to be a strikingly new response to excess SO_2 concentrations. Owing to the strong reaction of WT and SO-KO plants to SO_2, defensins may function in both processes: excess sulfur storage and sulfite detoxification. Another outstanding finding of this RNAseq analysis is the up-regulation of transcripts encoding a peroxidase (peroxidase CB, AT3G49120). As a result of the apoplastic localization, this enzyme may function as a first line of defense in the detoxification of SO_2. This hypothesis was advanced previously (Pfanz *et al.*, 1990; Pfanz & Oppmann, 1991), but was never tested at the molecular level. The up-regulation of peroxidase CB in SO-KO[+] may indicate it has some role in removing sulfite before it enters the cytoplasm. Both topics will be of great interest in our upcoming investigations.

Our transcript analyses exhibited a set of regulated genes amounting to *c.* 5% (cluster analyses). Giraud *et al.* (2012) reported a strong influence of 1–3 µl l^{-1} SO_2 on grape berries using microarray analyses. Although 600 nl l^{-1} SO_2 is a nontoxic dosage for Arabidopsis (Hänsch & Mendel, 2005), the first responses at the mRNA level were detected in WT[+], though to a lesser extent than detected in SO-KO [+]. WT plants showed a strong and fast reaction to SO_2 (Fig. 1), whereas SO-KO[+] presented the highest RPKM values of several transcripts and a reaction involving a much broader range of transcripts. In contrast to microarray approaches, RNAseq enabled us to obtain details on splice variant gene expression and therefore even allowed *SO* transcript observations for SO-KO[+]. Although the resulting protein products are not functional, as determined in SO enzyme activities for SO-KO and SO-KO[+] (Randewig *et al.*, 2012), plants seem to have a driving force that categorically tries to produce SO when SO_2 is present.

In conclusion, RNAseq of WT[+] and SO-KO[+] and their controls not only gave quantitative insights into the transcriptional response of Arabidopsis plants to SO_2 fumigation, but also permitted some first insights into novel putative mechanisms for SO_2 detoxification beyond SO activity and transportation of excess sulfite into the S-assimilation stream. Mainly based on SO-KO, we present in Fig. 4 our new working model for plant reactions to excess SO_2, including the hypothesized SO_2 detoxification mechanisms.

Acknowledgements

This work was supported by a grant from the Deutsche Forschungsgemeinschaft to R.H. under contract no. HA3107/4 and to H.R. under contract no. Re515/32. A.P.M.W. appreciates support from DFG grants IRTG 1525 and SFB 590. We are grateful to Dr Jürgen Kreuzwieser, Axel Schwietale and Michael Rienks for excellent technical help. We thank Prof. Rudolf Galensa, Prof. Norbert Käufer and Mariko Matsuda Alexander for discussing and critically reading the manuscript.

References

Alscher R, Bower JL, Zipfel W. 1987. The basis for different sensitivities of photosynthesis to SO_2 in two cultivars of pea. *Journal of Experimental Botany* **38**: 99–108.

Ayazloo M, Bell JNB. 1981. Studies on the tolerance to sulphur dioxide of grass populations in polluted areas. I. Identification of tolerant populations. *The New Phytologist* **88**: 203–222.

Baek D, Pathange P, Chung JS, Jiang J, Gao L, Oikawa A, Hirai MY, Saito K, Pare PW, Shi H. 2010. A stress-inducible sulphotransferase sulphonates salicylic acid and confers pathogen resistance in Arabidopsis. *Plant, Cell & Environment* **33**: 1383–1392.

Bell JNB. 1980. Response of plants to sulphur dioxide. *Nature* **284**: 399–400.

Bindschedler LV, Dewdney J, Blee KA, Stone JM, Asai T, Plotnikov J, Denoux C, Hayes T, Gerrish C, Davies DR *et al.* 2006. Peroxidase-dependent apoplastic oxidative burst in Arabidopsis required for pathogen resistance. *The Plant Journal* **47**: 851–863.

Brychkova G, Xia Z, Yang G, Yesbergenova Z, Zhang Z, Davydov O, Fluhr R, Sagi M. 2007. Sulfite oxidase protects plants against sulfur dioxide toxicity. *Plant Journal* **50**: 696–709.

Chaves MM, Flexas J, Pinheiro C. 2009. Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell. *Annals of Botany* **103**: 551–560.

Creissen G, Firmin J, Fryer M, Kular B, Leyland N, Reynolds H, Pastori G, Wellburn F, Baker N, Wellburn A *et al.* 1999. Elevated glutathione biosynthetic capacity in the chloroplasts of transgenic tobacco plants paradoxically causes increased oxidative stress. *Plant Cell* **11**: 1277–1292.

Davidian JC, Kopriva S. 2010. Regulation of sulfate uptake and assimilation – the same or not the same? *Molecular Plant* **3**: 314–325.

De Kok L. 1990. Sulfur metabolism in plants exposed to atmospheric sulfur. In: Rennenberg H, Brunold C, De Kok LJ, Stulen I, eds. *Sulfur nutrition and sulfur assimilation in higher plants.* The Hague, the Netherlands: SPB Academic Publishing, 111–130.

De Kok L, Durenkamp M, Yang L, Stulen I. 2007. Atmospheric sulfur. In: *Sulfur in plants - an ecological perspective. Plant Ecophysiology* **6**: 91–106.

Eilers T, Schwarz G, Brinkmann H, Witt C, Richter T, Nieder J, Koch B, Hille R, Hansch R, Mendel RR. 2001. Identification and biochemical characterization of *Arabidopsis thaliana* sulfite oxidase. A new player in plant sulfur metabolism. *Journal of Biological Chemistry* **276**: 46989–46994.

Filner P, Rennenberg H, Sekiya J, Bressan RA, Wilson LG, LeCureux L, Shimei T. 1984. Biosynthesis and emission of hydrogen sulfide by higher plants. In: Koziol MJ, Whatley FR, eds. *Gaseous Air Pollutants and Plant Metabolism.* London, UK: Butterworth, 291–312.

Frova C. 2003. The plant glutathione transferase gene family: genomic structure, functions, expression and evolution. *Physiologia Plantarum* **119**: 469–479.

Geu-Flores F, Moldrup ME, Bottcher C, Olsen CE, Scheel D, Halkier BA. 2011. Cytosolic gamma-glutamyl peptidases process glutathione conjugates in the biosynthesis of glucosinolates and camalexin in Arabidopsis. *Plant Cell* **23**: 2456–2469.

Giraud E, Ivanova A, Gordon CS, Whelan J, Considine MJ. 2012. Sulphur dioxide evokes a large scale reprogramming of the grape berry transcriptome associated with oxidative signalling and biotic defence responses. *Plant, Cell & Environment* **35**: 405–417.

Graveley BR. 2005. Mutually exclusive splicing of the insect dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell* **123**: 65–73.

Hänsch R, Lang C, Riebeseel E, Lindigkeit R, Gessler A, Rennenberg H, Mendel RR. 2006. Plant sulfite oxidase as novel producer of H_2O_2: combination of enzyme catalysis with a subsequent non-enzymatic reaction step. *Journal of Biological Chemistry* **281**: 6884–6888.

Hänsch R, Mendel RR. 2005. Sulfite oxidation in plant peroxisomes. *Photosynthesis Research* **86**: 337–343.

Hawkesford MJ. 2003. Transporter gene families in plants: the sulphate transporter gene family-redundancy or specialization? *Physiologia Plantarum* **117**: 155–163.

Heber U, Hüve K. 1998. Action of SO_2 on plants and metabolic detoxification of SO_2. *International Review of Cytology - A Survey of Cell Biology* **177**: 255–286.

Herschbach C, Rizzini L, Mult S, Hartmann T, Busch F, Peuke AD, Kopriva S, Ensminger I. 2010. Over-expression of bacterial gamma-glutamylcysteine synthetase (gsh1) in plastids affects photosynthesis, growth and sulphur metabolism in poplar (populus tremula x populus alba) dependent on the resulting gamma-glutamylcysteine and glutathione levels. *Plant, Cell & Environment* **33**: 1138–1151.

Irshad M, Canut H, Borderies G, Pont-Lezica R, Jamet E. 2008. A new picture of cell wall protein dynamics in elongating cells of *Arabidopsis thaliana*: confirmed actors and newcomers. *BMC Plant Biology* **8**: 94.

Jost R, Berkowitz O, Wirtz M, Hopkins L, Hawkesford MJ, Hell R. 2000. Genomic and functional characterization of the oas gene family encoding o-acetylserine (thiol) lyases, enzymes catalyzing the final step in cysteine biosynthesis in *Arabidopsis thaliana*. *Gene* **253**: 237–247.

Kasajima I, Fujiwara T. 2007. Identification of novel *Arabidopsis thaliana* genes which are induced by high levels of boron. *Plant Biotechnology* **24**: 355–360.

Kataoka T, Watanabe-Takahashi A, Hayashi N, Ohnishi M, Mimura T, Buchner P, Hawkesford MJ, Yamaya T, Takahashi H. 2004. Vacuolar sulfate transporters are essential determinants controlling internal distribution of sulfate in Arabidopsis. *Plant Cell* **16**: 2693–2704.

Kawashima CG, Berkowitz O, Hell R, Noji M, Saito K. 2005. Characterization and expression analysis of a serine acetyltransferase gene family involved in a key step of the sulfur assimilation pathway in Arabidopsis. *Plant Physiology* **137**: 220–230.

Klein M, Papenbrock J. 2004. The multi-protein family of Arabidopsis sulphotransferases and their relatives in other plant species. *Journal of Experimental Botany* **55**: 1809–1820.

Van der Kooij TAW, De Kok LJ, Haneklaus S, Schnug E. 1997. Uptake and metabolism of sulfur dioxide by *Arabidopsis thaliana*. *New Phytologist* **135**: 101–107.

Kopriva S, Koprivova A. 2004. Plant adenosine 5'-phosphosulphate reductase: the past, the present, and the future. *Journal of Experimental Botany* **55**: 1775–1783.

Kopriva S, Rennenberg H. 2004. Control of sulphate assimilation and glutathione synthesis: interaction with N and C metabolism. *Journal of Experimental Botany* **55**: 1831–1842.

Kovaleva V, Krynytskyy H, Gout I, Gout R. 2010. Recombinant expression, affinity purification and functional characterization of scots pine defensin 1. *Applied Microbiology and Biotechnology* **89**: 1093–1101.

Lang C, Popko J, Wirtz M, Hell R, Herschbach C, Kreuzwieser J, Rennenberg H, Mendel RR, Hänsch R. 2007. Sulphite oxidase as key enzyme for protecting plants against sulphur dioxide. *Plant, Cell & Environment* **30**: 447–455.

Peleman J, Boerjan W, Engler G, Seurinck J, Botterman J, Alliotte T, Van Montagu M, Inze D. 1989. Strong cellular preference in the expression of a housekeeping gene of *Arabidopsis thaliana* encoding s-adenosylmethionine synthetase. *Plant Cell* **1**: 81–93.

Pfanz H, Dietz K-J, Weinerth I, Oppmann B. 1990. Detoxification of sulfur dioxide by apoplastic peroxidases. In: Rennenberg H, Brunold C, De Kok LJ, Stulen I, eds. *Sulfur nutrition and sulfur assimilation in higher plants.* The Hague, the Netherlands: SPB Academic Publishing, 229–233.

Pfanz H, Martinoia E, Lange OL, Heber U. 1987. Mesophyll resistances to SO_2 fluxes into leaves. *Plant Physiology* **85**: 922–927.

Pfanz H, Oppmann B. 1991. The possible role of apoplastic peroxidases in detoxifying the air pollutant sulfur dioxide. In: Lobarzewski J, Greppin H, Penel C, Gaspar T, eds. *Biochemical, molecular and physiological aspects of plant peroxidases.* Geneva, Switzerland: University of Geneva, 401–417.

Randewig D, Hamisch D, Herschbach C, Eiblmeier M, Gehl C, Jurgeleit J, Skerra J, Mendel RR, Rennenberg H, Hänsch R. 2012. Sulfite oxidase controls sulfur metabolism under SO_2 exposure in *Arabidopsis thaliana*. *Plant, Cell & Environment* **35**: 100–115.

Rao IM, Anderson LE. 1983. Light and stomatal metabolism: II. Effects of sulfite and arsenite on stomatal opening and light modulation of enzymes in epidermis. *Plant Physiology* **71**: 456–459.

Rennenberg H. 1984. The fate of excess sulfur in higher plants. *Annual Review of Plant Physiology and Plant Molecular Biology* **35**: 121–153.

Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B. 2003. An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Molecular Biology* **53**: 247–259.

Schlaeppi K, Bodenhausen D, Buchala A, Mausch F, Reymond P. 2008. The glutathione-deficient mutant pad2-1 accumulates lower amounts of glucosinolates and is more susceptible to the insect herbivore *Spodoptera littoralis*. *Plant Journal* **55**: 774–786.

Shah K, Penel C, Gagnon J, Dunand C. 2004. Purification and identification of a Ca^{2+}-pectate binding peroxidase from Arabidopsis leaves. *Phytochemistry* **65**: 307–312.

Smith CW. 2005. Alternative splicing - when two's a crowd. *Cell* **123**: 1–3.

Stasolla C, Belmonte MF, Tahir M, Elhiti M, Khamiss K, Joosen R, Maliepaard C, Sharpe A, Gjetvaj B, Boutilier K. 2008. Buthionine sulfoximine (bso)-mediated improvement in cultured embryo quality in vitro entails changes in ascorbate metabolism, meristem development and embryo maturation. *Planta* **228**: 255–272.

Stotz HU, Thomson JG, Wang Y. 2009. Plant defensins: defense, development and application. *Plant Signaling and Behavior* **4**: 1010–1012.

Szalai G, Kellos T, Galiba G, Kocsy G. 2009. Glutathione as an antioxidant and regulatory molecule in plants under abiotic stress conditions. *Journal of Plant Growth Regulation* **28**: 66–80.

Tamm CO, Cowling EB. 1977. Acid precipitation and forest vegetation. *Water, Air, and Soil Pollution* **7**: 503–511.

Thomma BP, Cammue BP, Thevissen K. 2002. Plant defensins. *Planta* **216**: 193–202.

Vauclare P, Kopriva S, Fell D, Suter M, Sticher L, von Ballmoos P, Krahenbuhl U, den Camp RO, Brunold C. 2002. Flux control of sulphate assimilation in *Arabidopsis thaliana*: adenosine 5'-phosphosulphate reductase is more susceptible than ATP sulphurylase to negative control by thiols. *Plant Journal* **31**: 729–740.

Vivancos PD, Dong Y, Ziegler K, Markovic J, Pallardo FV, Pellny TK, Verrier PJ, Foyer CH. 2010. Recruitment of glutathione into the nucleus during cell proliferation adjusts whole-cell redox homeostasis in *Arabidopsis thaliana* and lowers the oxidative defence shield. *Plant Journal* **64**: 825–838.

Wang L, Feng Z, Wang X, Wang X, Zhang X. 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**: 136–138.

Wirtz M, Birke H, Heeg C, Muller C, Hosp F, Throm C, Konig S, Feldman-Salit A, Rippe K, Petersen G *et al.* 2010. Structure and function of the hetero-oligomeric cysteine synthase complex in plants. *Journal of Biological Chemistry* **285**: 32810–32817.

Yamaguchi Y, Nakamura T, Kusano T, Sano H. 2000. Three *Arabidopsis* genes encoding proteins with differential activities for cysteine synthase and beta-cyanoalanine synthase. *Plant and Cell Physiology* **41**: 465–476.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Scattered matrix plot presenting the distribution RPKM values for different comparisons between WT, WT[+], SO-KO and SO-KO[+].

Fig. S2 GO analysis of fivefold regulated transcripts.

Fig. S3 KEGG metabolic pathways of the four different comparison pairs using fivefold data.

Fig. S4 *SO* [At3G01910] splice variants and mRNA fragment mapping for WT, WT[+], SO-KO and SO-KO[+].

Fig. S5 *SO* [At3G01910] mRNA mapping for WT, WT[+], SO-KO and SO-KO[+].

Table S1 Transcript raw data

Table S2 Raw data for sulfur metabolism-associated genes

Table S3 RPKM values and fold-change calculations of highly regulated plant defensins (PDFs)

Table S4 RPKM values and fold-change calculations of highly regulated peroxidase genes

Table S5 Script for programming in *R* to execute DEGseq tool

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

4.1.5 RNA-seq assembly - are we there yet?

frontiers in
PLANT SCIENCE

REVIEW ARTICLE
published: 25 September 2012
doi: 10.3389/fpls.2012.00220

RNA-seq assembly – are we there yet?

***Simon Schliesky*[1], *Udo Gowik*[2], *Andreas P. M. Weber*[1] *and Andrea Bräutigam*[1]*

[1] Center of Excellence on Plant Sciences (CEPLAS), Institute for Plant Biochemistry, Heinrich Heine University, Düsseldorf, Germany
[2] Center of Excellence on Plant Sciences (CEPLAS), Institute for Plant Developmental and Molecular Biology, Heinrich Heine University, Düsseldorf, Germany

Edited by:
Bjoern Usadel, Rheinisch-Westfaelische Technische Hochschule Aachen University, Germany

Reviewed by:
Jose M. Jimenez-Gomez, Max Planck Institute for Plant Breeding, Germany
Marc Lohse, Max Planck Institute of Molecular Plant Physiology, Germany

***Correspondence:**
Andrea Bräutigam, Institute for Plant Biochemistry, 26.03.01.Room 32, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany.
e-mail: andrea.braeutigam@uni-duesseldorf.de

Transcriptomic sequence resources represent invaluable assets for research, in particular for non-model species without a sequenced genome. To date, the Next Generation Sequencing technologies 454/Roche and Illumina have been used to generate transcriptome sequence databases by mRNA-Seq for more than fifty different plant species. While some of the databases were successfully used for downstream applications, such as proteomics, the assembly parameters indicate that the assemblies do not yet accurately reflect the *actual* plant transcriptomes. Two different assembly strategies have been used, overlap consensus based assemblers for long reads and Eulerian path/de Bruijn graph assembler for short reads. In this review, we discuss the challenges and solutions to the transcriptome assembly problem. A list of quality control parameters and the necessary scripts to produce them are provided.

Keywords: RNA-seq, assembly, plant, NGS, next generation sequencing, transcriptome

INTRODUCTION

Access to a sequence database for a plant species of interest tremendously advances that plant species' potential use in research, as is evidenced by the success story of the small weed *Arabidopsis thaliana*. However, the complexities of many plants' genomes and prohibitive costs have precluded the sequencing of their genomes. Instead of the genome, the transcriptomes of tissues of interest for many important crop plants were sequenced[1]. The majority of those sequencing efforts were carried out with substantial funding and frequently in consortia. The advent of next generation sequencing (NGS) technologies has however marked a new era of transcriptomics (Metzker, 2010). Single laboratories are now enabled to produce a sequence resource for their species of choice, be it for commercial, medicinal, ecological, or any other reason. Since the initial proof of concept through the sequencing of the transcriptome of *Arabidopsis* seedlings (Weber et al., 2007), at least 60 additional plant transcriptomes have been sequenced *de novo*. Currently, the 1KP project aims for transcriptomic sequencing of 1,000 plant species[2].

The quest for a $1,000 human genome has driven the sequencing industries to formidable innovations. The gold rush started with the 454 platform (later acquired by Roche) and the 100 bases long reads that could be obtained on the initial GS20 instrument. Improvements to the platform lead to reads of 250 bases in length. The latest 454/Roche platform used for (plant) transcriptome sequencing is the GS FLX Titanium which allows read lengths of 400 bases (Glenn, 2011)[3]. While a typical 454/Roche sequencing run is finished within less than a day, it yields only 400 Mb per run. Illumina (formerly Solexa) employs a different technology platform. Initially reads were as short as 36 bases but improvements to the technology have led to increased read length of 100 bases (and if paired reads are used, 200 bases of the same transcript). In contrast to the 454/Roche platform, sequencing runs take from several days to more than 1 week but produce ∼600 Gb per run (Glenn, 2011)[4]. With respect to cost per base sequenced, Illumina will beat Roche/454 by a factor of more than 100. Both the 454/Roche and the Illumina platform have been used for transcriptome sequencing and assembly (**Table 1**). To our knowledge, the two other established NGS technologies, SOLiD and Ion Torrent, have not been used for published plant transcriptome projects (using the search words of RNA-seq, plant AND transcriptome, plant AND NGS at ISI Web of Knowledge).

TRANSCRIPTOME SEQUENCING AND ITS APPLICATIONS

The initial *de novo* plant transcriptome sequencing by mRNA-Seq was conducted on *Arabidopsis thaliana* (Weber et al., 2007). Only half a million reads of close to 100 bases in length were sequenced in this proof of concept approach. It was recognized already at this early stage that remapping the reads to the *Arabidopsis* genome tagged many more transcripts than could be assembled with Newbler, Phrap, or CAP3 (Emrich et al., 2007; Weber et al., 2007). Indeed, assembly was recognized as a future challenge.

Virtually all of the 454/Roche transcriptome sequencing projects following this initial work did have the generation of a transcriptome resource as one of their major objectives (**Table 1**). Many NGS experiments provide a resource of markers for molecular breeding, for example for eucalyptus, melon, and different legumes (Novaes et al., 2008; Guo et al., 2010; Blavet et al., 2011; Hiremath et al., 2011; Kaur et al., 2011).

[1] http://compbio.dfci.harvard.edu/tgi/plant.html
[2] http://www.onekp.com
[3] http://www.molecularecologist.com/next-gen-fieldguide/
[4] http://www.molecularecologist.com/next-gen-fieldguide/

Table 1 | Plant transcriptome sequencing projects until today (complete table available as Table S1 in Supplementary Material).

Reference	Plant	Type of reads
Weber et al. (2007)	*Arabidopsis thaliana*	454
Novaes et al. (2008)	*Eucalyptus grandis*	454
Barakat et al. (2009)	*Castanea dentata, C. mollissima*	454
Alagna et al. (2009)	*Olea europaea*	454
Dassanayake et al. (2009)	*Heritiera littoralis, Rhizophora* mangle	454
Wang et al. (2009)	*Artemisia annua*	454
Swarbreck et al. (2011)	*Avena barbata*	454
Guo et al. (2010)	*Cucumis sativus*	454
Riggins et al. (2010)	*Amaranthus uberculatus*	454
King et al. (2011)	*Jatropha curcas*	454
Hiremath et al. (2011)	*Cicer arietinum*	454
Troncoso-Ponce et al. (2011)	*Ricinus communis, Brassica napus, Eunonymus alatus, Tropaeolum majus*	454
Bräutigam et al. (2011a)	*Cleome gynandra, C. spinosa*	454
Cantu et al. (2011)	*Triticum aestivum*	454
Dai et al. (2011)	*Cucumis melo* (sweet melon)	454
Sun et al. (2011)	*Pinus sylvestris*	454
Der et al. (2011)	*Pteridium aquilinium*	454
Franssen et al. (2011)	*Pisum sativum*	454
Ibarra-Laclette et al. (2011)	*Utricularia gibba*	454
Su et al. (2011)	*Phalaenopsis aphrodite*	454
Pont et al. (2011)	*Triticum aestivum*	454
Bleeker et al. (2011)	*Solanum lycopersicum, S. habrochaites*	454
Blavet et al. (2011)	Eight *Silene* sp. and *Dianthus*	454
Villar et al. (2011)	*Eucalyptus*	454
Kaur et al. (2011)	*Lens culinaris*	454
Kalavacharla et al. (2011)	*Phaseolus vulgaris*	454
Lu et al. (2012)	*Capsicum annuum*	454
Meyer et al. (2012)	*Panicum hallii* var. filipes	454
Edwards et al. (2012)	*Ziziphus celata*	454
Desgagne-Penix et al. (2012)	*Papaver somniferum*	454
Angeloni et al. (2011)	*Scabiosa columbaria*	454 and Illumina
Garg et al. (2011)	*Cicer arietinum*	454 and Illumina
Krishnan et al. (2011)	*Azadirachta indica*	Illumina
Mutasa-Göttgens et al. (2012)	*Beta vulgaris*	Illumina
Gruenheit et al. (2012)	*Pachycladon fastigiatum, P. cheesemanii*	Illumina and Illumina paired end
Mizrachi et al. (2010)	*Eucalyptusgrandis* × *E. urophylla*	Illumina paired
Barrero et al. (2011)	*Euphorbia fischeriana*	Illumina paired
Xia et al. (2011)	*Hevea brasiliensis*	Illumina paired
Chibalina and Filatov (2011)	*Silene latifolia*	Illumina paired
Hao et al. (2011)	*Taxus marei*	Illumina paired
Tang et al. (2011)	*Siraitia grosvenorii*	Illumina paired
Wong et al. (2011)	*Acacia auriculiformis, A. mangium*	Illumina paired
Shi et al. (2011)	*Camellia sinensis*	Illumina paired
Hyun et al. (2012)	*Momordica cochinchensis*	Illumina paired
Hao et al. (2012)	*Polygonum cuspidatum*	Illumina paired
Huang et al. (2012)	*Millettia pinnata,*	Illumina paired
Gahlan et al. (2012)	*Picrorhiza kurrooa*	Illumina paired
Zhang et al. (2012)	*Arachis hypogaea*	Illumina paired
McKain et al. (2012)	Different Agavoideae	Illumina paired

Other major targets are primary (Dai et al., 2011; Franssen et al., 2011; King et al., 2011; Troncoso-Ponce et al., 2011) and secondary (Alagna et al., 2009; Wang et al., 2009; Bleeker et al., 2011; Desgagne-Penix et al., 2012) metabolism. Plants such as poppy for opium and other alkaloids, tomato for beneficial terpenoids, and Artemisia for artemisinin have been targeted by transcriptome

sequencing (**Table 1**). Adaptations to biotic (Barakat et al., 2009; Sun et al., 2011) and abiotic stress (Dassanayake et al., 2009; Villar et al., 2011) were studied in plants. Finally, transcriptomes of plants carrying a trait of interest such as C_4 photosynthesis (Bräutigam et al., 2011a; Gowik et al., 2011), weedy habitus (Riggins et al., 2010), being an orchid (Su et al., 2011), a carnivorous plant (Ibarra-Laclette et al., 2011), an ecological model (Blavet et al., 2011), a traditional biochemical model (Franssen et al., 2011), or an endangered species (Edwards et al., 2012), were analyzed. Since 454/Roche pyrosequencing was used, the number of sequenced reads is comparatively low, between 0.08 and 3.3 million reads (**Table 1**). The majority of the assemblies were realized with overlap consensus based assemblers such as CAP3 (Huang and Madan, 1999; four instances) or its implementation in the clustering pipeline TGICL (Pertea et al., 2003; five instances), which prefaces CAP3 with a megablast to reduce the number of sequences fed to CAP3 and hence RAM requirement. MIRA (Chevreux et al., 2004; one instance) and one of the multiple Newbler versions[5] (seven instances) were also frequently used. In four projects a combination of two assemblers was used. CLC[6], LEADS (Dai et al., 2011), Paracelsus Transcript Assembler (Novaes et al., 2008) and Seqman Ngen (Edwards et al., 2012) were each used in a single published assembly (**Table 1**). The different assemblies were quality controlled – if they were controlled at all – by different parameters. Hence it is difficult to compare the different assembly methods. All assemblies report the number of unigenes (the sum of assembled contigs and unassembled singletons) and either the N50 or the average length of the contigs. These two parameters can be compared with reference sequence numbers, average sizes and N50 from predicted transcriptomes of species with sequenced genomes. The parameters show that the assemblies are far from perfect and that none of the assemblers achieves a satisfactory reconstruction of an actual transcriptome. While the representation of the transcriptome was the expressed goal of these studies, none of them fully succeeded. Most of the assemblies were carried out either with Roche's Newbler or with a decades-old tool, CAP3. No marked improvements could be detected in the assembly parameters unigene number and average length over time (Table S1 in Supplementary Material).

Although one may be tempted to dismiss such error prone, incomplete assemblies, the majority of them have already proven themselves useful for downstream applications such as proteomics (Bräutigam et al., 2008; Franssen et al., 2011) or pathway reconstruction (Wang et al., 2009; Bräutigam et al., 2011a; Dai et al., 2011; Troncoso-Ponce et al., 2011; Desgagne-Penix et al., 2012). The databases were developed to provide a sequence resource for future experiments. The analysis of single genes involved in the C_4 photosynthetic pathway based on hypotheses derived from RNA-seq experiments has already been successful (Furumoto et al., 2011; Sommer et al., 2012). Hence even imperfect assemblies succeed in enabling future research. Downstream approaches that require perfect or near perfect unigenes such as the evolutionary analysis of gene family expansions will likely suffer more from the current shortcomings of these assemblies.

RNA-seq by Illumina sequencing was initially used for transcriptome sequencing in species with sequenced genomes (e.g., Vega-Arreguin et al., 2009; Li et al., 2011). It has been successfully applied to produce transcriptomes *de novo* (Table S1 in Supplementary Material). The technology appeals to researchers despite its comparatively short reads because it produces much larger coverage at the same or a lower price. However, it presents a new set of challenges for the assembly.

Similar to 454/Roche based sequencing projects, virtually all Illumina based RNA-seq experiments on non-model species have been conducted to produce a transcriptome database. RNA-seq using the Illumina technology was undertaken to analyze transcriptomes for plants of nutritional or medical value (Barrero et al., 2011; Hao et al., 2011, 2012; Krishnan et al., 2011; Tang et al., 2011; Gahlan et al., 2012; Hyun et al., 2012) or of commercial value (Mizrachi et al., 2010; Shi et al., 2011; Xia et al., 2011; Mutasa-Göttgens et al., 2012; Zhang et al., 2012). Two experiments addressed ecological and evolutionary questions, the evolution of sex chromosomes (Bergero and Charlesworth, 2011) and the phylogenetic positioning of species (McKain et al., 2012). The majority of sequences were produced with paired end technology. In this case, sequences from both ends of fragments of defined size are sequenced. The use of paired ends allows scaffolding: sequence reads are used to produce contigs. The information which reads belong together and their specific distance orders disconnected contigs on scaffolds. The unknown nucleotides in the gaps of scaffolds are caused by knowing the size of the gap but not the identity of the nucleotides and hence the nucleotides in the gap are denoted as Ns. One assembler that was originally developed for genome assemblies, SOAPdevono[7], has been used to assemble the majority of plant transcriptomes. Additional assemblers used include CLC, velvet (Zerbino and Birney, 2008)[8], AbySS (Simpson et al., 2009), and Trinity (Grabherr et al., 2011). In one of the projects a custom resolution algorithm for velvet was developed and used (Mizrachi et al., 2010; **Table 1**). This customized velvet version has produced the best assembly in terms of contig number and average contig length. Despite its success, the method has not been used for any of the other projects.

Finally, RNA-seq experiments have combined both 454/Roche and Illumina sequencing. Transcriptomes of chickpea and pincushion flower were produced using both technologies and hybrid assemblies (Angeloni et al., 2011; Garg et al., 2011). Although promising in prospect of complementary error correction, to date, true hybrid assembly approaches are limited to an assembly of one library (often 454) as a base transcriptome and subsequent correction of the consensus sequence by mapping the other read library (Illumina or SOLiD). Quality improvements of transcriptome hybrid assemblies have not yet been assessed in a comparative study. However, in the context of genome assembly it was shown that a stepwise (as explained above) hybrid assembly had a higher quality (according to the authors: comparable to Sanger-sequencing) than single library approaches (Aury et al., 2008). The use as well as the strategy of

[5]http://454.com/products/analysis-software/index.asp
[6]http://www.clcbio.com/
[7]http://soap.genomics.org.cn/soapdenovo.html
[8]http://www.ebi.ac.uk/∼zerbino/oases/

hybrid assemblies is currently vigorously discussed in the online community [9,10].

Overall, similar to the assemblies from 454/Roche RNA-seq experiments, those from Illumina technology suffer from limitations. It will be crucial to continue developing assemblers with enhanced capability while establishing standard quality controls to make assemblies from different species, technologies, and assembly strategies comparable.

ASSEMBLERS

Two principally different types of assemblers are available for RNA-seq data: overlap-layout-consensus (OLC) assemblers and Eulerian path assemblers which are based on de Bruijn graphs (summarized in Flicek and Birney, 2009).

Overlap-layout-consensus assemblers were developed for Sanger sequences. In principle, the assembler starts with a sequence read, looks at its sequence, and searches the read space for another read that contains an overlapping sequence. The overlap is specified by its length and the number or percentage of matching bases. The memory requirement for this operation depends on the number of reads to be searched. Thus, more reads require more computer power. Already during times of Sanger-sequencing, this method became inefficient with the available computers and a prefacing clustering step was added. This clustering step groups sequences deemed similar, for example by a megablast search (Pertea et al., 2003). The assembler then only searches the sequences in each cluster. The three most prominent examples for these OLC based assemblers are Newbler (Roche/454 Life Sciences, Branford, CT, USA), MIRA (Chevreux et al., 2004), and CAP3 (Huang and Madan, 1999; or TGICL which uses megablast and CAP3). While these assemblers are suitable for 454/Roche sequences, the number of reads generated with Illumina are simply too large to be processed. In an assessment of different assemblers with both simulated and real data, TGICL was superior to MIRA and CAP3 in its results (Bräutigam et al., 2011b). No new assemblers have been developed and used except for Newbler developed by the company 454/Roche itself.

To tackle Illumina-generated sequence reads, a new type of assembler was created. It is based on finding the Eulerian path through a de Bruijn graph (Pevzner et al., 2001). Essentially, this type of assembler breaks the whole sequence space in pieces of defined length, which are called k-mers. It then moves along the k-mers and creates a graph in the process. Identical overlaps of k-mers are merged and counted. If the assembler encounters differences, the graph will branch, if it subsequently encounters identity again, the graph will join the ends. That means that single nucleotide differences (SNDs) will produce bubbles (**Figure 1**; 2). Such SNDs can either represent a sequencing error or genetic variation in form of a single nucleotide polymorphism (SNP). Large bubbles and open ended branches can be caused by alternative splicing and alternative transcriptional starts and stops (**Figure 1**; 1). The presence of genomic DNA in the sample, improperly trimmed and filtered reads, sequencing errors, alternative splicing, and background transcription will lead to many more deviations

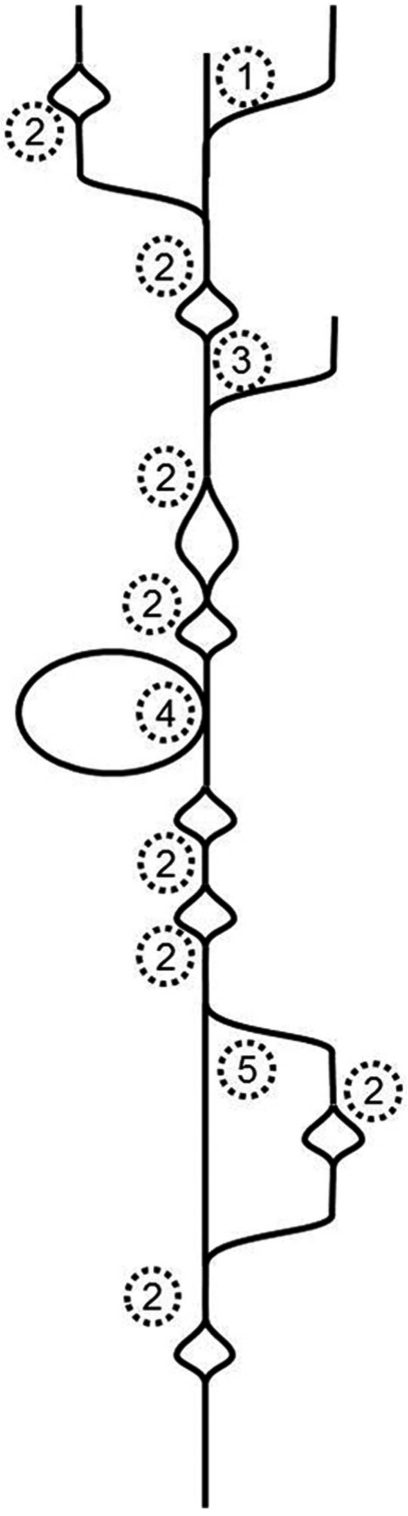


FIGURE 1 | Schematic de Bruijn graph of a single transcript; 1 alternative transcription start site *or* hybrid joining *or* DNA contamination; 2 SND caused by a sequencing error *or* a SNP *or* mutation after gene duplication; 3 alternative transcription start site *or* DNA contamination; 4 alternative exon use; 5 alternative exon use *or* mutations after recent gene duplication.

[9]http://www.seqanswers.com
[10]http://www.biostars.org

from the one transcript, which ideally should look like a straight line. In reality the graph has no straight lines but is full of bubbles and frayed ends (**Figure 1**). When such a graph is resolved, the researcher wants all "real differences" such as alternative splicing events, transcripts resulting from recently duplicated but still very similar genes, and genetic variation, for example from different alleles of a particular genetic locus, represented. However, all differences caused by technical errors should be removed. The only information available for the algorithm to resolve the graph is the number of instances observed for each k-mer. If such a graph is used for genome sequencing of organisms without complex genomes (i.e., not plants), the application for which it was developed, the graph can be resolved using the degree of coverage for each k-mer. In theory, the number of reads that cover each base in the graph should be equal for the whole graph. While this does not hold true for repetitive sequence elements, it can be used to resolve the remainder. Given 100-fold coverage in a genome homozygous at all loci, you would require that each k-mer is covered at least, say, 80 times to be called real. If the coverage is lower, it is likely a sequencing error.

The resolution of transcriptome graphs is very different from the resolution of genome graphs. The dynamic range of a leaf transcriptome spans at least five orders of magnitude (Bräutigam et al., 2011a; Gowik et al., 2011). Hence the coverage of a transcriptome is the polar opposite of even. SNPs and InDels present in natural populations cause uneven coverage. Transcripts with higher diversity in the population exhibit more changes (as represented by bubbles in **Figure 1**) than transcripts with lower diversity in the population. Alternative splicing and start and stop sites will cause differential coverage. If an exon is only used 10% of the time, it may not make it past the resolution cut-off.

To solve the problem of uneven coverage, the assemblers that were originally designed to produce genomic assemblies, such as ABySS, SOAPdenovo, or velvet, have been extended with add-ons for the assembly of transcriptomes, such as Trans-ABySS, SOAPdenovo-Trans, or velvet/Oases. Even given this amendment, assemblers do not succeed in assembly as evidenced by contig numbers that are much higher than the expected transcript number and average contig sizes much lower than that of an average transcriptome (Table S1 in Supplementary Material). Assemblers for short reads remain limited and both the development of new assemblers as well as post-assembly processing and parameter optimization is ongoing. The detection of genetic variation and transcript variants will likely require post-assembly read mapping and evaluation through the researcher.

CONSIDERATIONS FOR NGS TRANSCRIPTOME ASSEMBLY

The key differences between NGS and Sanger sequence reads are the number of reads and the length of the reads. Even using the long-read technology 454/Roche, the reads are only half to a third as long as compared to Sanger sequences. With a single NGS run, half a Gigabase to several Gigabases of sequence data is generated. In consequence, the challenge has shifted from efficiently generating sequence reads to efficiently assembling them. Given an error rate of ∼1% and 40,000 reads of 400 bases length for a gene of 1 kb, 160,000 incorrect base calls are expected. If these are randomly distributed, on average, each single base will be called incorrectly about 160 times. Even assuming error rates of only 0.1%, each base will still be called incorrectly 16 times. For this reason, there is a correlation between the number of contigs resulting from a transcript and the expression strength of the corresponding gene (Franssen et al., 2011). The large number of sequencing reads calls for intense sequence pruning. There are several software packages that include pruning pipelines, such as the fastx-toolkit[11], the fastQC software[12], and the RobiNA package (Lohse et al., 2012). Those are used to determine average quality per base in addition to other quality control parameters. Reads can be trimmed (pruned at the ends if bases are below a quality threshold), filtered (if internal bases are below a threshold), and purged from duplicates (merging multiple, identical reads into a single sequence). Unfortunately, the majority of assembly publications do not report their pruning pipeline and threshold values; they restrict themselves to stating the number of high quality bases that were fed into the assembly pipeline.

In theory, the error problem was solved if one were to assemble only reads with a high coverage cut-off during the graph resolution. In that case, sequencing errors were ignored because their k-mer numbers are too low. However, due to the large dynamic range of the transcriptome, low abundance genes, such as transcription factors and regulatory kinases, are underrepresented (Czechowski et al., 2004). These genes are discriminated against if the assembly is processed with high coverage cut-offs during resolution (Schliesky and Bräutigam, unpublished observations). They simply disappear. Similarly, rare transcript isoforms will also be discarded during the resolution step if high coverage is required.

Library normalization at least partially addresses the challenge of a high dynamic range. Normalization by digestion reduces the dynamic range by one order of magnitude (Christodoulou et al., 2001) but normalized libraries clearly retain some dynamic range (Franssen et al., 2011). While normalization likely improves the assembly, it comes at a cost: sequence information and quantitative information are no longer collected at the same time. If quantitative information is not required, normalization is highly recommended.

At least low coverage transcripts could be recovered if one knew before assembling how many reads are produced from each transcript and adjust the resolution algorithm accordingly for each piece of the graph. Possibly, a dynamic approach – assembly, read mapping on the preliminary assembly, re-assembly with sliding scale of resolution coverage cut-off – might be able to solve the problem. While none of the current transcriptome assemblers has implemented this strategy, its application for one Illumina plant transcriptome assembly may serve as the proof of concept for the approach (Mizrachi et al., 2010).

The key challenge in assembly is weeding out all variation caused by sequencing errors, library preparation, and other technical artifacts while keeping all variation caused by biological phenomena such as genetic variation, alternative splicing, and others.

[11] http://hannonlab.cshl.edu/fastx_toolkit/index.html
[12] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

ASSESSING THE ASSEMBLY

In principle, assessing an assembly is easy – it should accurately reflect the transcriptome of the sequenced tissue and species. In practice, the accurate transcriptome is unknown and not available for comparison. Two different approaches to overcome this problem can be envisioned. (i) Establishing assembly parameters with simulated reads from a reference species and transferring those to *de novo* sequencing and (ii) assembling *de novo* transcriptome and estimating reference parameters. While the first possibility has immediate appeal, there are a number of obstacles. The dynamic range of transcriptomes is different in different tissues and between species (Fluhr et al., 1986). A method optimized for a root transcriptome might not necessarily work well with a leaf transcriptome and *vice versa*. Different read length, paired end or single end sequencing, or different sequencing depth dictated by the available instrumentation and funding will likely change the parameters for the best possible assembly. Carrying out the optimization with a non-target dataset will also cause substantial time investment with little return in the beginning since not even a working assembly of the target transcriptome is created. For all these and possibly additional reasons, many researches will immediately start to work on the target transcriptome. If a common set of assessment parameters were developed, all possible transcriptomes could be measured against these parameters and thus compared with each other.

NUMBER OF UNIGENES

The number of unigenes expected from an assembly can be calculated with a Fermi estimate. The gene number for the majority of sequenced plant genomes is between 20,000 and 40,000. Using microarray data from *Arabidopsis*, one can estimate that about one half of the genes are expressed in leaves. Using these two numbers as approximations, the Fermi estimate for loci expected from a leaf transcriptome is about 15,000. While species with a very recently duplicated genome may have close to twice as many, none will have an order of magnitude more transcripts (compare to Table S1 in Supplementary Material). However, the number of unigenes can be easily manipulated while not gaining a better assembly. One strategy crops the unigenes by a minimal-length cut-off. While it facilitates subsequent read mappings it severely discriminates against "real," short transcripts. Another example, raising the coverage cut-off during graph resolution will reduce the number of unigenes. This strategy indeed removes unigenes constructed because of sequencing errors but it will also discriminate against low abundance transcripts as discussed above. It is thus important to combine these measures with the number of reference transcripts matching the unigenes.

NUMBER OF REFERENCE TRANSCRIPTS MATCHING THE UNIGENES

Once the assembly is complete, it needs to be compared to the most closely related reference species. The unigenes are matched to the reference sequence by Blast or Blat (Kent, 2002). While it is unknown how many reference transcripts should be tagged by the assembled unigenes, a higher number of tagged references indicate a more inclusive and thus better assembly. Genes that are not expressed will never be tagged but as long as the number of tagged genes increases during assembly optimization, the assembly is getting better in terms of inclusiveness.

NUMBER OF REFERENCE TRANSCRIPTS HIT BY READS COMPARED TO NUMBER OF REFERENCE TRANSCRIPTS HIT BY UNIGENES

It is possible to estimate the number of unigenes produced by the assembly. If the reads are at least 75 bases long after trimming and filtering, they can be mapped to a reference transcriptome provided that the reference species is reasonably closely related. However, in reality "reasonably close" will not be sufficient to produce a perfect mapping. Therefore (i) a traditional mapping program that allows for multiple mismatches (i.e., BLAST or BLAT) and (ii) mapping in protein-space (i.e., translated query against translated database; blatx or tblastx) improves the mapping success with respect to evolutionary distance. In theory, reference transcripts tagged by reads are expected to be tagged by unigenes. This assumption is only true if a loss-less assembler such as OLC assemblers are used. Reads that do not overlap with other reads are reported as singlets or singletons when using these assemblers. The resolution cut-off applied in graph-based assemblies will overlook unigenes if they are not covered by at least the coverage cut-off. Mapping reads to a reference results in estimated read numbers per locus. With these read numbers one can check how many reads are actually needed to produce a contig or a full length contig based on different assembly parameters such as k-mer size and coverage cut-off. Surprisingly, the assembly will also produce unigenes for which no read tagging was recorded. In that case, the setting of either Blat or Blast was too stringent too match the reads but the longer unigene produces a match. This quality control measure will overlook lineage specific transcripts that have no match in the reference transcriptome. While every genome sequencing approach does reveal lineage specific genes, the number of genes present in multiple plant lineages is vastly higher.

The ratio between reference sequences tagged by reads and those tagged by unigenes should ideally approach 1:1.

N50, AVERAGE LENGTH, MEDIAN LENGTH

These three parameters are always reported with genome assemblies. The N50 can be envisioned as follows: if you order the unigenes by their length and then start counting nucleotides at the largest unigene, the N50 will report the unigene length at which you have counted through half of the bases. While this is a sensible measure for genomes, it makes less sense for transcriptomes. After all, with genomes you expect as many contigs as you have chromosomes. In transcriptomes, you may have different N50s for different tissues of the same plant since different groups of genes are expressed. The same caveat is true for the average length and the median length.

While different (whole) transcriptomes indeed have slightly different parameters with regard to N50, average length and median length, the values are similar enough to yield an estimate for the expected values for an unknown transcriptome (compare to **Tables 1** and **2**).

LENGTH OF THE LONGEST UNIGENE

The length of the longest unigene might not represent a sensible measure. If the sequencing library was contaminated by genomic

Table 2 | Quality assessment parameters drawn from transcripts of publicly available genome databases.

Species	Genome size (Mbases)	Number of transcripts including isoforms	N50	GC%
Arabidopsis thaliana	120	41671	1912	42.27
Brassica rapa	485	41019	1482	46.28
Populus trichocarpa	481	45033	1845	42.29
Solanum lycopersicum	950	35802	1461	41.61
Oryza sativa	420	66338	2295	51.30
Setaria italica	515	40599	1811	52.75
Zea mays	2066	136770	1612	51.14

DNA, a large fraction of this DNA will come from the plastid genome. The plastome DNA is known to be AT-rich and thus survives the poly-A enrichment step during the Illumina mRNA enrichment protocol well (Schliesky, Mullick and Bräutigam et al., unpublished observations). Its presence leads to remarkably long contigs in the assembly albeit not quite to an assembly accurately representing the transcriptome. A second consequence of DNA contamination is the presence of many contigs matching transposon-like sequences which are also AT-rich. The complete or near complete presence of a unigene matching the longest nuclear transcript of a reference also only shows that the assembly parameters were ideal for that transcript but not for all transcripts in the sequenced library.

NUMBER OF ESTIMATED FULL LENGTH UNIGENES

While the length of the longest unigene may not be an ideal measure, the estimated number of full length unigenes reflects on the success of the assembly. The unigenes are matched to a transcriptome reference from a closely related species. While during evolution, genes will have extended or contracted, on average, their length will remain comparable. More unigenes that reach the length of the reference transcripts indicate a better assembly.

If no reference seems suitably close enough, it is still possible to compare the length distributions qualitatively. Comparing multiple publicly available plant transcriptome databases with respect to their length distributions demonstrates an overall pattern on what a transcriptome should possibly look like (e.g., ∼90% of the sequences between 200 and 3500 nt length). In practice that is not achieved because assembly software often produces a huge fraction of truncated transcripts between 0 and 200 nt length.

NUMBER OF HYBRID/READ-THROUGH UNIGENES

While full length unigenes are the goal of an assembly, no hybrid unigenes should be produced. These result from the joining of two target transcripts matching two different reference transcripts into one unigene. Two different kinds of hybrid unigenes can be produced. Illumina resequencing of *Arabidopsis* leaf transcriptomes identified unigenes that were assembled from adjacent transcripts (Schliesky, unpublished). Read mapping to the genome revealed that these hybrid unigenes resulted from read-through transcription. They thus likely reflect the true transcriptome. The second class of hybrid unigenes is undesirable. In this case, the similarity of sequences, sequencing errors, or incomplete read trimming and filtering cause the merging of two target transcripts into one reference unigene. A read mapping in this case identifies no evidence for this feature. Different assembly parameters favor or do not favor the creation of this second class of hybrids (Schliesky, unpublished) and thus hybrid detection should be included in the quality control. One strategy for hybrid detection by alignment to *Arabidopsis* could be designed as follows. Based on the outcome of an alignment, all unigenes that map to multiple genes get tagged as hybrid (also known as chimera or fusion genes), if the match takes place in distinct, i.e., non-repetitive, sections of the unigene sequence. Subsequently the chromosomal position is used to classify the type of hybrid to either read-through (matching neighboring genes) or second class hybrids (matching non-neighboring genes). A high proportion of second class hybrids points to a bad assembly algorithm, to bad assembly parameters (e.g., k-mer too large) or to a contamination of some sort (e.g., genomic DNA or low quality reads).

If no closely related reference is available, the hybrid detection strategy probably needs to be amended. With increasing evolutionary diversity mapping accuracy will decrease. Therefore mapping errors may lead to incorrectly detected hybrids. That may be solved by increasing the required matching length during mapping (increasing accuracy) at the cost of not mapping some unigenes at all (decreasing sensitivity). Alternatively, hybrid unigenes may be detected by mapping the reads back to the unigenes. At the position of error, read coverage is likely lower than in the adjacent regions. Detecting and cropping those bridging regions reliably will reduce the number of hybrid transcripts. This approach is based on same idea as an assembly algorithm with a sliding resolution window for per base coverage. If the quality assessment was completely independent of a reference sequence, lineage specific genes which have no match in reference database would also be included in the quality assessment.

EXAMPLE WORKFLOW

As a step toward comparable transcriptome assessments a collection of Perl and Unix scripts, which are automating parts of the assessment, is provided in this review. It resembles an example workflow (**Figure 2**, Supplementary Presentation 1) for assembling and assessing reads of *Arabidopsis* mRNA. This out-of-the-box pipeline consists of five blocks; (i) vigorous read pruning, (ii) assembling, (iii) mapping to a reference, (iv) collecting quality parameters, and (v) polishing the assembly for publication.

Carrying out transcriptome assembly in a standardized way has not been publicly pursued prior to this review. In order to keep the workflow repeatable and comparable we provide a step by step instruction set on how to use the supplemental scripts to assemble

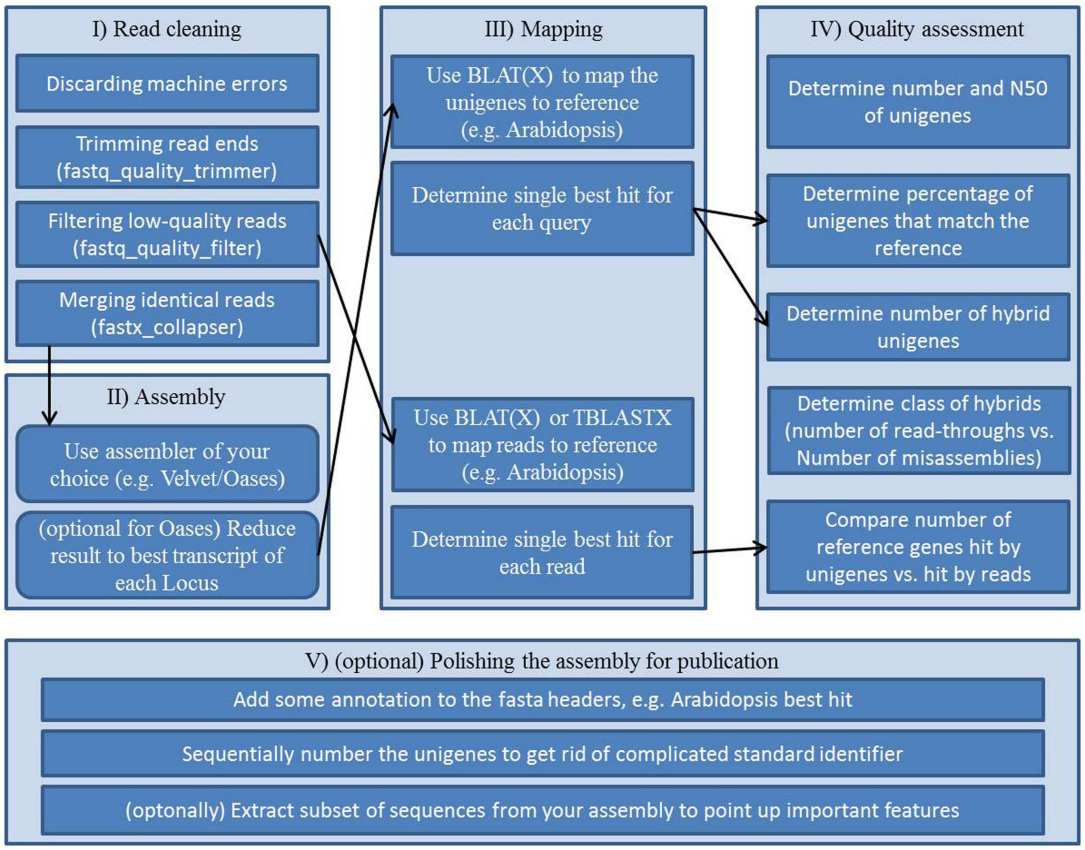


FIGURE 2 | Workflow scheme for a transcriptome assembly and quality assessment: (I) preprocessing of the raw reads, (II) assembly of processed reads, (III) mappings for annotation and for subsequent quality assessment, (IV) collecting quality information from assembly and mappings, (V) final polishing to create an easy to use, thus easy to share file from the assembly.

a sequencing run and conduct quality assessment on the assembly. Please be aware that the workflow including all scripts was designed with *Arabidopsis* as the target reference. Scripts might or might not be adaptable to other species. The workflow was established and tested on a Linux machine running 64 Bit Ubuntu 10.04 and having installed BioPerl, BioPython, the FASTX-toolkit, BLAT, and BLAST.

First, all scripts need to be extracted and copied into a folder (Supplemental Scripts 02–12), together with the raw reads (fastq.gz files) and the reference. Start a terminal and change to the directory containing the scripts. All commands needed are in Supplemental Script 1. Lines proceeded by a #-symbol present comment lines and are used for explanation. Illumina reads obtained from a sequencing facility are supplied as ∗.fastq.gz files. To unzip and concatenate them, the zcat command is used (Supplemental Script 1 line 4).

READ CLEANING (SUPPLEMENTAL SCRIPT 1 LINES 6–11)

While reads coming off the sequencer are not dirty in the traditional sense, they may contain low quality reads, adaptor sequences, and low quality bases. Reads are cleaned to remove as much non-biological variation as possible. As discussed previously read cleaning is crucial for a good assembly. The workflow starts by removing reads flagged as inappropriate by the sequencer (line 7). For quality trimming knowledge of the average overall base quality is needed. This is evaluated using the FASTX-Toolkit (line 8). Visual aids (e.g., fastq_quality_boxplot_graph.sh) may ease interpretation of the results. The stats file provides one line per base (i.e., in Illumina 101 bp reads 101 lines) and for each base a median quality score is calculated. Frequently, read quality will be low toward the end of the read. If at any point, say from base 86 to base 87, the median quality drops dramatically, the ideal quality cut-off will be in between this range. For sequencing runs with good library preparation and no problems during the sequencing we recommend a cut-off of 30.

The actual cleaning is conducted in three steps; (i) trimming (line 9), which prunes the ends off of the reads if they are below the defined quality cut-off and subsequently discards all reads that are shorter than a defined length cut-off (we suggest half the read length, i.e., 50) after trimming. (ii) Filtering (line 10), which discards all reads that do not meet the required quality cut-off with at least a defined length (in percent of the total read). For the majority of sequencing runs, the values suggested above are a good starting point. Trimming and filtering does not discard more than 15% of the reads if library preparation and sequencing went well. In other cases, values might have to be adjusted and

trimming and filtering values might have to be relaxed. (iii) Collapsing (line 11), since memory requirements are lower if fewer reads are assembled.

ASSEMBLY (SUPPLEMENTAL SCRIPT 1 LINES 13–27)

Lines 13–27 contain an out-of-the-box pipeline from cleaned reads to assembled best transcript isoforms using Velvet/Oases. The pipeline can be adapted for other assemblers. Velvet/Oases is called in three steps. In the first one, output directory, k-mer size and input files are declared (line 15). In the subsequent steps a de Bruijn Graph is built (line 16) and resolved with an algorithm optimized for transcriptomes, i.e., Oases (line 17). Oases outputs a huge amount of transcripts, which is due to the fact that Oases resolves bubbles and branches in the Graph into all possible transcript isoforms of a locus. The number of transcripts is, compared to the number of unique loci detected by Oases, frequently two (or more) times higher. Picking the best transcript for each locus is a challenge as there is no standard to what "best" means. The longest transcript is often the least supported (i.e., covered by k-mers), whereas the most supported often is the shortest one. To solve this problem a script (line 22, Supplemental Script 02) has recently been published on Google Code[13] (by Adrian Reich 2012) that essentially chooses the most supported transcript (i.e., highest k-mer coverage) that has at least XX% length of the longest transcript in this locus. In our hands a length cut-off of 20% showed the best results in subsequent quality assessment.

Many assembly papers include a length cut-off to reduce the number of transcripts. Although this curation is, in essence, cheating with the number of unigenes, the pipeline includes a Perl script for cropping the database (line 27, Supplemental Script 03).

MAPPING (SUPPLEMENTAL SCRIPT 1 LINES 29–46)

A major bottleneck – conceptually as well as computationally – if working with non-model species is the read mapping. When working on non-model species there is no sequenced genome available to use as a reference. Mapping to a close relative works if precautions are taken to account for the evolutionary distance. Modern mapping algorithms are designed for speed and allow only one mismatch. These algorithms will fail to map to a related reference. Therefore in cross-species mapping the use of traditional mapping algorithms like BLAST and BLAT in protein-space is recommended. While mapping unigenes to the reference (line 31) finishes in the order of minutes, mapping reads to the reference will take much longer (depending on the library size in the order of weeks). This limitation can be bypassed by parallelizing BLAT with a script (line 34–36, Supplemental Script 04) on the number of CPUs available. The script splits the read file, starts parallel single BLAT runs and merges the results. The number of CPUs can be changed within Supplemental Script 04 in line 3 (default is 2). Alternatively BLAST, which natively supports multiple CPUs, can be used for the mapping (line 39, 40).

Multiple mappings can be resolved to only one single best hit per query (i.e., per read) by using the best hit scripts for either BLAST (line 42, Supplemental Script 05) or BLAT (line 46, Supplemental Script 06).

[13] http://code.google.com/p/oases-to-csv/

QUALITY ASSESSMENT (SUPPLEMENTAL SCRIPT 1 LINES 48–87)

As discussed above the most frequently used measures to evaluate the quality of an assembly are number of unigenes and N50. A Perl script to calculate the read length histogram of a fasta file (line 50, Supplemental Script 07) was developed by Joseph Fass (modified from a script by Brad Sickler). The script produces a histogram that can be easily visualized, and calculates the number of unigenes, N25, N50, and N75.

The percentage of unigenes that match a reference are calculated using the total number of references and the number of matching unigenes. The total number of references is counted (line 54). The number of unigenes which map to a reference is produced by extracting the query identifiers from the mapping table and by counting unique occurrence (line 56). The mapping efficiency (ratio of mappable unigenes by total references) can be interpreted as a measure of completeness with the caveat that single tissue transcriptomes are not expected to represent a complete transcriptome.

Hybrid unigenes can be detected with the help of mapping. In hybrid unigenes, different sections of the unigene map to different loci in *Arabidopsis*. These hybrid unigenes can either be read-throughs of two adjacent genes or misassemblies. While it is desirable to have no hybrid unigenes that represent transcripts fused by the assembler, it might add to the understanding of cellular mechanisms to identify read-throughs. Therefore we provide two Perl scripts, which (i) detect any hybrid unigenes (line 60, Supplemental Script 08) and (ii) subsequently classify those as read-throughs or not (lines 63–67, Supplemental Script 09). While hybrid unigenes are undesirable in an assembly, they can be tolerated for single gene analysis. A read mapping provides visible cues whether coverage is even or whether parts of the unigene are only supported by few reads. Only with more and more transcriptomes being assembled and large scale comparisons enabled, hybrid unigenes will become an issue in comparison.

The quality of an assembly can also be measured by comparing the number of reference genes hit by unigenes with the number of reference genes hit by reads. This is based on the assumption that genes, which are expressed (i.e., hit by a read) will generate a transcript (i.e., unigene) during the assembly which maps to the same reference. Comparing the numbers of genes hit by reads (lines 70, 71) and by unigenes (lines 74, 75) provides a quick assessment whether those values are in the same range. Subsequently, it is assessed whether the reference genes hit by reads are also hit by unigenes. This question is answered using standard Unix commands and set theory. Given two files "genes hit by unigenes" and "genes hit by reads" with a unique set of identifiers in each, adding (i.e., concatenating) one file and twice the other file yields a new set which has each identifier either occurring once, twice, or three times. Extracting lines by count yields three groups, (i) genes only present in the file used once (line 84), (ii) genes only present in the file used twice (line 85) and (iii) genes that are present in both files and therefore commonly hit by unigenes and reads (line 86). A large percentage of the latter group indicates that the assembled transcripts reflect the expressed genes. An alternative way to determine the intersect between two files is based on the Unix "join" command (lines 90–92).

FINAL POLISH OF THE ASSEMBLY (SUPPLEMENTAL SCRIPTS 1 LINES 89–99)

Prior to publication, an annotated fasta database of the assembly needs to be generated. The scripts provided incorporate an annotation to the sequence headers, e.g., best hit in *Arabidopsis* (lines 96, 97, Supplemental Script 10) and number the identifiers of unigenes sequentially to get rid of awkward assembler headers (line 100, Supplemental Script 11). If only a subset of sequences are needed a Perl script (line 104, Supplemental Script 12) can extract it if given a one-per-line list of identifiers.

APPLYING THE WORKFLOW (QUICK AND DIRTY)

The complete workflow discussed in this review is attached as a script (Supplemental Presentation 1) and could be run unsupervised. This requires the fastq.gz files to be in the same folder as all the Supplemental Scripts along with an *Arabidopsis* reference that is named "TAIR10_cdna.fasta". Additionally Perl, Python, BioPerl, BioPython, BLAST, BLAT, Velvet, Oases, and the FASTX-Toolkit have to be installed on the system. The hardware requirements of the assembly in terms of memory are rather high. Assembly was limited to 50 M reads with 96 GB RAM available.

Due to these strict requirements, we strongly recommend reading and adjusting the workflow to your specific needs. All scripts have either a help output (if ran with - - help or - ? as parameter) or a Perldoc documentation (opened by running "perldoc *script_name*") or both.

CONCLUSION

Next generation sequencing and transcriptome assembly have already proven beneficial for research. However, current assemblies are still far away from an accurate representation of a transcriptome. Detailed description of the assembly method including read treatment prior to assembly, assembly parameters, and stringent quality control will make different assemblies more comparable and will make it easier to reproduce successful assemblies. This first attempt to bring the quality assessment in line helps to make transcriptomic resources much more comparable and reusable for the community. At the very least, each assembly publication should include a fasta file with all unigenes. Until full length single molecule sequencing for transcriptome sequences becomes technically feasible, transcriptome assembly will remain the major bottle neck during transcriptome sequencing. We are not there yet!

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Plant_Systems_Biology/10.3389/fpls.2012.00220/abstract.

REFERENCES

Alagna, F., D'Agostino, N., Torchia, L., Servili, M., Rao, R., Pietrella, M., Giuliano, G., Chiusano, M. L., Baldoni, L., and Perrotta, G. (2009). Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* 10, 399. doi:10.1186/1471-2164-10-399

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

Aury, J.-M., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., Poulain, J., Anthouard, V., Scarpelli, C., Artiguenave, F., and Wincker, P. (2008). High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* 9, 603. doi:10.1186/1471-2164-9-603

Barakat, A., Diloreto, D. S., Zhang, Y., Smith, C., Baier, K., Powell, W. A., Wheeler, N., Sederoff, R., and Carlson, J. E. (2009). Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (Castanea mollissima) in response to the chestnut blight infection. *BMC Plant Biol.* 9, 51. doi:10.1186/1471-2229-9-51

Barrero, R. A., Chapman, B., Yang, Y. F., Moolhuijzen, P., Keeble-Gagnere, G., Zhang, N., Tang, Q., Bellgard, M. I., and Qiu, D. Y. (2011). De novo assembly of Euphorbia fischeriana root transcriptome identifies prostratin pathway related genes. *BMC Genomics* 12, 600. doi:10.1186/1471-2164-12-600

Bergero, R., and Charlesworth, D. (2011). Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. *Curr. Biol.* 21, 1470–1474.

Blavet, N., Charif, D., Oger-Desfeux, C., Marais, G. A. B., and Widmer, A. (2011). Comparative high-throughput transcriptome sequencing and development of SiESTa, the Silene EST annotation database. *BMC Genomics* 12, 376. doi:10.1186/1471-2164-12-376

Bleeker, P. M., Spyropoulou, E. A., Diergaarde, P. J., Volpin, H., De Both, M. T. J., Zerbe, P., Bohlmann, J., Falara, V., Matsuba, Y., Pichersky, E., Haring, M. A., and Schuurink, R. C. (2011). RNA-seq discovery, functional characterization, and comparison of sesquiterpene synthases from Solanum lycopersicum and Solanum habrochaites trichomes. *Plant Mol. Biol.* 77, 323–336.

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

Bräutigam, A., Mullick, T., Schliesky, S., and Weber, A. P. M. (2011b). Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. *J. Exp. Bot.* 62, 3093–3102.

Bräutigam, A., Shrestha, R. P., Whitten, D., Wilkerson, C. G., Carr, K. M., Froehlich, J. E., and Weber, A. P. M. (2008). 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.

Cantu, D., Pearce, S., Distelfeld, A., Christiansen, M., Uauy, C., Akhunov, E., Fahima, T., and Dubcovsky, J. (2011). Effect of the down-regulation of the high Grain Protein Content (GPC) genes on the wheat transcriptome during monocarpic senescence. *BMC Genomics* 12, 492. doi: 10.1186/1471-2164-12-492

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Muller, W. E. G., Wetter, T., and Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14, 1147–1159.

Chibalina, M. V., and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr. Biol.* 21, 1475–1479.

Christodoulou, D. C., Gorham, J. M., Herman, D. S., and Seidman, J. G. (2001). Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Curr. Protoc. Mol. Biol.* 94, 4.12.1–4.12.11.

Czechowski, T., Bari, R. P., Stitt, M., Scheible, W. R., and Udvardi, M. K. (2004). Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* 38, 366–379.

Dai, N., Cohen, S., Portnoy, V., Tzuri, G., Harel-Beja, R., Pompan-Lotan, M., Carmi, N., Zhang, G. F., Diber, A., Pollock, S., Karchi, H., Yeselson, Y., Petreikov, M., Shen, S., Sahar, U., Hovav, R., Lewinsohn, E., Tadmor, Y., Granot, D., Ophir, R., Sherman, A., Fei, Z. J., Giovannoni, J., Burger, Y., Katzir, N., and Schaffer, A. A. (2011). Metabolism of soluble sugars in developing melon fruit: a global transcriptional view of the metabolic transition to sucrose accumulation. *Plant Mol. Biol.* 76, 1–18.

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

Der, J., Barker, M., Wickett, N., Depamphilis, C., and Wolf, P. (2011). De novo characterization of the gametophyte transcriptome in bracken fern, Pteridium aquilinum. *BMC Genomics* 12, 99. doi: 10.1186/1471-2164-12-99

Desgagne-Penix, I., Farrow, S. C., Cram, D., Nowak, J., and Facchini, P. J. (2012). Integration of deep transcript and targeted metabolite profiles for eight cultivars of opium poppy. *Plant Mol. Biol.* 79, 295–313.

Edwards, C. E., Parchman, T. L., and Weekley, C. W. (2012). Assembly, gene annotation and marker development using 454 floral transcriptome sequences in Ziziphus celata (Rhamnaceae), a highly endangered, florida endemic plant. *DNA Res.* 19, 1–9.

Emrich, S. J., Barbazuk, W. B., Li, L., and Schnable, P. S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* 17, 69–73.

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

Fluhr, R., Moses, P., Morelli, G., Coruzzi, G., and Chua, N. H. (1986). Expression dynamics of the pea rbcS multigene family and organ distribution of the transcripts. *EMBO J.* 5, 2063–2071.

Franssen, S.U., Shrestha, R.P., Brautigam, A., Bornberg-Bauer, E., and Weber, A.P.M. (2011). Comprehensive transcriptome analysis of the highly complex Pisum sativum genome using next generation sequencing. *BMC Genomics* 12, 227. doi:10.1186/1471-2164-12-227

Furumoto, T., Yamaguchi, T., Ohshima-Ichie, Y., Nakamura, M., Tsuchida-Iwata, Y., Shimamura, M., Ohnishi, J., Hata, S., Gowik, U., Westhoff, P., Brautigam, A., Weber, A. P. M., and Izui, K. (2011). A plastidial sodium-dependent pyruvate transporter. *Nature* 476, 472–475.

Gahlan, P., Singh, H. R., Shankar, R., Sharma, N., Kumari, A., Chawla, V., Ahuja, P. S., and Kumar, S. (2012). De novo sequencing and characterization of Picrorhiza kurrooa transcriptome at two temperatures showed major transcriptome adjustments. *BMC Genomics* 13, 126. doi:10.1186/1471-2164-13-126

Garg, R., Patel, R. K., Jhanwar, S., Priya, P., Bhattacharjee, A., Yadav, G., Bhatia, S., Chattopadhyay, D., Tyagi, A. K., and Jain, M. (2011). Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. *Plant Physiol.* 156, 1661–1678.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11, 759–769.

Gowik, U., Brautigam, 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.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. D., Chen, Z. H., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652.

Gruenheit, N., Deusch, O., Esser, C., Becker, M., Voelckel, C., and Lockhart, P. (2012). Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. *BMC Genomics* 13, 92. doi: 10.1186/1471-2164-13-92

Guo, S. G., Zheng, Y., Joung, J. G., Liu, S. Q., Zhang, Z. H., Crasta, O. R., Sobral, B. W., Xu, Y., Huang, S. W., and Fei, Z. J. (2010). Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genomics* 11, 384. doi:10.1186/1471-2164-11-384

Hao, D. C., Ge, G. B., Xiao, P. G., Zhang, Y. Y., and Yang, L. (2011). The first insight into the tissue specific taxus transcriptome via illumina second generation sequencing. *PLoS ONE* 6, e21220. doi:10.1371/journal.pone.0021220

Hao, D. C., Ma, P., Mu, J., Chen, S. L., Xiao, P. G., Peng, Y., Huo, L., Xu, L. J., and Sun, C. (2012). De novo characterization of the root transcriptome of a traditional Chinese medicinal plant Polygonum cuspidatum. *Sci. China Life Sci.* 55, 452–466.

Hiremath, P. J., Farmer, A., Cannon, S. B., Woodward, J., Kudapa, H., Tuteja, R., Kumar, A., Bhanuprakash, A., Mulaosmanovic, B., Gujaria, N., Krishnamurthy, L., Gaur, P. M., Kavikishor, P. B., Shah, T., Srinivasan, R., Lohse, M., Xiao, Y. L., Town, C. D., Cook, D. R., May, G. D., and Varshney, R. K. (2011). Large-scale transcriptome analysis in chickpea (Cicer arietinum L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. *Plant Biotechnol. J.* 9, 922–931.

Huang, J., Lu, X., Yan, H., Chen, S., Zhang, W., Huang, R., and Zheng, Y. (2012). Transcriptome characterization and sequencing-based identification of salt-responsive genes in Millettia pinnata, a semi-mangrove plant. *DNA Res.* 19, 195–207.

Huang, X. Q., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877.

Hyun, T. K., Rim, Y., Jang, H. J., Kim, C. H., Park, J., Kumar, R., Lee, S., Kim, B. C., Bhak, J., Binh, Q., Kim, S. W., Lee, S. Y., and Kim, J. Y. (2012). De novo transcriptome sequencing of Momordica cochinchinensis to identify genes involved in the carotenoid biosynthesis. *Plant Mol. Biol.* 79, 413–427.

Ibarra-Laclette, E., Albert, V. A., Perez-Torres, C. A., Zamudio-Hernandez, F., Ortega-Estrada, M. D., Herrera-Estrella, A., and Herrera-Estrella, L. (2011). Transcriptomics and molecular evolutionary rate analysis of the bladderwort (Utricularia), a carnivorous plant with a minimal genome. *BMC Plant Biol.* 11, 101. doi:10.1186/1471-2229-11-101

Kalavacharla, V., Liu, Z., Meyers, B., Thimmapuram, J., and Melmaiee, K. (2011). Identification and analysis of common bean (Phaseolus vulgaris L.) transcriptomes by massively parallel pyrosequencing. *BMC Plant Biol.* 11, 135. doi: 10.1186/1471-2229-11-135

Kaur, S., Cogan, N. O. I., Pembleton, L. W., Shinozuka, M., Savin, K. W., Materne, M., and Forster, J. W. (2011). Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genomics* 12, 265. doi:10.1186/1471-2164-12-265

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

King, A. J., Li, Y., and Graham, I. A. (2011). Profiling the developing Jatropha curcas L. Seed transcriptome by pyrosequencing. *Bioenergy Res.* 4, 211–221. doi:10.1007/s12155-011-9114-x

Krishnan, N. M., Pattnaik, S., Deepak, S. A., Hariharan, A. K., Gaur, P., Chaudhary, R., Jain, P., Vaidyanathan, S., Krishna, P. G. B., and Panda, B. (2011). De novo sequencing and assembly of Azadirachta indica fruit transcriptome. *Curr. Sci.* 101, 1553–1561.

Li, P. H., Ponnala, L., Gandotra, N., Wang, L., Si, Y. Q., Tausta, S. L., Kebrom, T. H., Provart, N., Patel, R., Myers, C. R., Reidel, E. J., Turgeon, R., Liu, P., Sun, Q., Nelson, T., and Brutnell, T. P. (2011). The developmental dynamics of the maize leaf transcriptome. *Nat. Genet.* 42, 1060–1067.

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., and Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* 40, W622–W627.

Lu, F.-H., Cho, M.-C., and Park, Y.-J. (2012). Transcriptome profiling and molecular marker discovery in red pepper, Capsicum annuum L. TF68. *Mol. Biol. Rep.* 39, 3327–3335.

McKain, M. R., Wickett, N., Zhang, Y., Ayyampalayam, S., McCombie, W. R., Chase, M. W., Pires, J. C., Depamphilis, C. W., and Leebens-Mack, J. (2012). Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in agavoideae (Asparagaceae). *Am. J. Bot.* 99, 397–406.

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

Meyer, E., Logan, T. L., and Juenger, T. E. (2012). Transcriptome analysis and gene expression atlas for Panicum hallii var. filipes, a diploid model for biofuel research. *Plant J.* 70, 879–890.

Mizrachi, E., Hefer, C. A., Ranik, M., Joubert, F., and Myburg, A. A. (2010). De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by illumina mRNA-seq. *BMC Genomics* 11, 681. doi:10.1186/1471-2164-11-681

Mutasa-Göttgens, E. S., Joshi, A., Holmes, H. F., Hedden, P., and Gottgens, B. (2012). A new RNASeq-based reference transcriptome for sugar beet and its application in transcriptome-scale analysis of vernalization and gibberellin responses. *BMC Genomics* 13, 99. doi:10.1186/1471-2164-13-99

Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J., Grattapaglia, D., Sederoff, R. R., and Kirst, M. (2008). High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *BMC Genomics* 9, 312. doi:10.1186/1471-2164-9-312

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y. H., White, J., Cheung, F., Parvizi, B., Tsai, J., and Quackenbush, J. (2003). TIGR gene indices clustering tools (TGICL): a

software system for fast clustering of large EST datasets. *Bioinformatics* 19, 651–652.

Pevzner, P. A., Tang, H. X., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U.S.A.* 98, 9748–9753.

Pont, C., Murat, F., Confolent, C., Balzergue, S., and Salse, J. (2011). RNA-seq in grain unveils fate of neo- and paleopolyploidization events in bread wheat (Triticum aestivum L.). *Genome Biol.* 12, R119.

Riggins, C. W., Peng, Y. H., Stewart, C. N., and Tranel, P. J. (2010). Characterization of de novo transcriptome for waterhemp (Amaranthus tuberculatus) using GS-FLX 454 pyrosequencing and its application for studies of herbicide target-site genes. *Pest Manag. Sci.* 66, 1042–1052.

Shi, C. Y., Yang, H., Wei, C. L., Yu, O., Zhang, Z. Z., Jiang, C. J., Sun, J., Li, Y. Y., Chen, Q., Xia, T., and Wan, X. C. (2011). Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics* 12, 131. doi:10.1186/1471-2164-12-131

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123.

Sommer, M., Bräutigam, A., and Weber, A. P. M. (2012). The dicotyledonous NAD malic enzyme C4 plant Cleome gynandra displays age-dependent plasticity of C4 decarboxylation biochemistry. *Plant Biol.* 14, 621–629.

Su, C. L., Chao, Y. T., Chang, Y. C. A., Chen, W. C., Chen, C. Y., Lee, A. Y., Hwa, K. T., and Shih, M. C. (2011). De novo assembly of expressed transcripts and global analysis of the Phalaenopsis aphrodite transcriptome. *Plant Cell Physiol.* 52, 1501–1514.

Sun, H., Paulin, L., Alatalo, E., and Asiegbu, F. O. (2011). Response of living tissues of Pinus sylvestris to the saprotrophic biocontrol fungus Phlebiopsis gigantea. *Tree Physiol.* 31, 438–451.

Swarbreck, S. M., Sudderth, E. A., St. Clair, S. B., Salve, R., Castanha, C., Torn, M. S., Ackerly, D. D., and Andersen, G. L. (2011). Linking leaf transcript levels to whole plant analyses provides mechanistic insights to the impact of warming and altered water availability in an annual grass. *Glob. Change Biol.* 17, 1577–1594.

Tang, Q., Ma, X. J., Mo, C. M., Wilson, I. W., Song, C., Zhao, H., Yang, Y. F., Fu, W., and Qiu, D. Y. (2011). An efficient approach to finding Siraitia grosvenorii triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. *BMC Genomics* 12, 343. doi:10.1186/1471-2164-12-343

Troncoso-Ponce, M. A., Kilaru, A., Cao, X., Durrett, T. P., Fan, J. L., Jensen, J. K., Thrower, N. A., Pauly, M., Wilkerson, C., and Ohlrogge, J. B. (2011). Comparative deep transcriptional profiling of four developing oilseeds. *Plant J.* 68, 1014–1027.

Vega-Arreguin, J. C., Ibarra-Laclette, E., Jimenez-Moraila, B., Martinez, O., Vielle-Calzada, J. P., Herrera-Estrella, L., and Herrera-Estrella, A. (2009). Deep sampling of the Palomero maize transcriptome by a high throughput strategy of pyrosequencing. *BMC Genomics* 10, 299. doi:10.1186/1471-2164-10-299

Villar, E., Klopp, C., Noirot, C., Novaes, E., Kirst, M., Plomion, C., and Gion, J. M. (2011). RNA-Seq reveals genotype-specific molecular responses to water deficit in eucalyptus. *BMC Genomics* 12, 538. doi:10.1186/1471-2164-12-538

Wang, W., Wang, Y. J., Zhang, Q., Qi, Y., and Guo, D. J. (2009). Global characterization of Artemisia annua glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* 10, 465. doi:10.1186/1471-2164-10-465

Weber, A. P. M., Weber, K. L., Carr, K., Wilkerson, C., and Ohlrogge, J. B. (2007). Sampling the arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol.* 144, 32–42.

Wong, M., Cannon, C., and Wickneswari, R. (2011). Identification of lignin genes and regulatory sequences involved in secondary cell wall formation in Acacia auriculiformis and Acacia mangium via de novo transcriptome sequencing. *BMC Genomics* 12, 342. doi: 10.1186/1471-2164-12-342

Xia, Z. H., Xu, H. M., Zhai, J. L., Li, D. J., Luo, H. L., He, C. Z., and Huang, X. (2011). RNA-Seq analysis and de novo transcriptome assembly of Hevea brasiliensis. *Plant Mol. Biol.* 77, 299–308.

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829.

Zhang, J. A., Liang, S., Duan, J. L., Wang, J., Chen, S. L., Cheng, Z. S., Zhang, Q., Liang, X. Q., and Li, Y. R. (2012). De novo assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in Peanut (Arachis hypogaea L.). *BMC Genomics* 13, 90. doi:10.1186/1471-2164-13-90

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

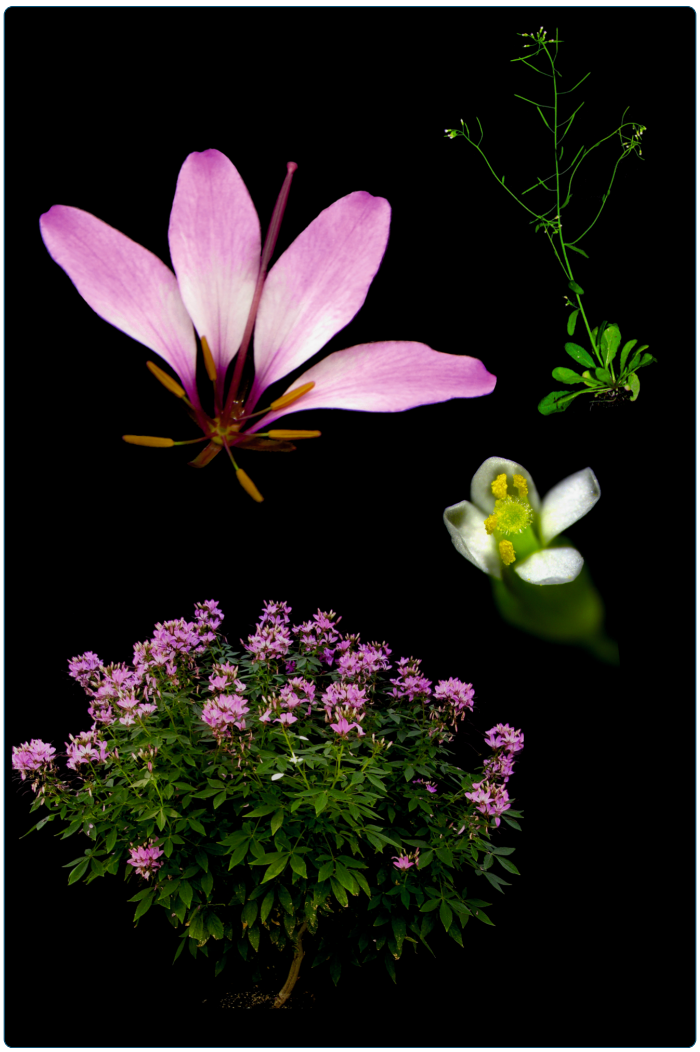
Received: 06 August 2012; accepted: 05 September 2012; published online: 25 September 2012.

*Citation: Schliesky S, Gowik U, Weber APM and Bräutigam A (2012) RNA-seq assembly – are we there yet? Front. Plant Sci. **3**:220. doi: 10.3389/fpls.2012.00220*

This article was submitted to Frontiers in Plant Systems Biology, a specialty of Frontiers in Plant Science.

4.1.6 Analysis of the floral transcriptome of Tarenaya hassleriana (Cleomaceae), a member of the sister group to the Brassicaceae: towards understanding the base of morphological diversity in Brassicales

BMC Genomics



Analysis of the floral transcriptome of *Tarenaya hassleriana* (Cleomaceae), a member of the sister group to the Brassicaceae: towards understanding the base of morphological diversity in Brassicales

Bhide *et al.*

BioMed Central

Bhide *et al. BMC Genomics* 2014, **15**:140
http://www.biomedcentral.com/1471-2164/15/140

Bhide et al. BMC Genomics 2014, 15:140
http://www.biomedcentral.com/1471-2164/15/140

BMC Genomics

RESEARCH ARTICLE Open Access

Analysis of the floral transcriptome of *Tarenaya hassleriana* (Cleomaceae), a member of the sister group to the Brassicaceae: towards understanding the base of morphological diversity in Brassicales

Amey Bhide[1], Simon Schliesky[2], Marlis Reich[3], Andreas PM Weber[2] and Annette Becker[1*]

Abstract

Background: *Arabidopsis thaliana*, a member of the Brassicaceae family is the dominant genetic model plant. However, while the flowers within the Brassicaceae members are rather uniform, mainly radially symmetrical, mostly white with fixed organ numbers, species within the Cleomaceae, the sister family to the Brassicaceae show a more variable floral morphology. We were interested in understanding the molecular basis for these morphological differences. To this end, the floral transcriptome of a hybrid *Tarenaya hassleriana*, a Cleomaceae with monosymmetric, bright purple flowers was sequenced, annotated and analyzed in respect to floral regulators.

Results: We obtained a comprehensive floral transcriptome with high depth and coverage close to saturation analyzed using rarefaction analysis a method well known in biodiversity studies. Gene expression was analyzed by calculating reads per kilobase gene model per million reads (RPKM) and for selected genes in silico expression data was corroborated by qRT-PCR analysis. Candidate transcription factors were identified based on differences in expression pattern between *A. thaliana* and *T. hassleriana*, which are likely key regulators of the *T. hassleriana* specific floral characters such as coloration and male sterility in the hybrid plant used. Analysis of lineage specific genes was carried out with members of the fabids and malvids.

Conclusions: The floral transcriptome of *T. hassleriana* provides insights into key pathways involved in the regulation of late anthocyanin biosynthesis, male fertility, flowering time and organ growth regulation which are unique traits compared the model organism *A. thaliana*. Analysis of lineage specific genes carried out with members of the fabids and malvids suggests an extensive gene birth rate in the lineage leading to core Brassicales while only few genes were potentially lost during core Brassicales evolution, which possibly reflects the result of the At-β whole genome duplication. Our analysis should facilitate further analyses into the molecular mechanisms of floral morphogenesis and pigmentation and the mechanisms underlying the rather diverse floral morphologies in the Cleomaceae.

Keywords: *Tarenaya hassleriana*, *Arabidopsis thaliana*, Floral transcriptome, Cleomaceae, Brassicaceae, Brassicales, 454 sequencing, Anthocyanins, Flower development

* Correspondence: annette.becker@bot1.bio.uni-giessen.de
[1]Justus-Liebig-Universität Gießen, Institute of Botany, Plant Development Group, Heinrich-Buff-Ring 38, 35392 Gießen, Germany
Full list of author information is available at the end of the article

Background

Tarenaya hassleriana, formerly known as *Cleome hassleriana* and sometimes erroneously referred to as *Cleome spinosa* [1] is a quick growing herbaceous perennial, native to Brazil and adjoining South American countries. The species belongs to the section Tarenaya and the subgenus Neocleome within the Cleomaceae [2] which includes roughly 300 species distributed throughout the tropical and subtropical regions of the world [3,4]. The family Cleomaceae belongs to the order Brassicales and previously Cleomaceae were thought to be more closely related to Capparaeceae but recent phylogenetic studies indicate that Cleomaceae are more closely related to and a sister family to Brassicaceae [3,5]. Molecular clock analyses suggests that Cleomaceae and Brassicaceae diverged from each other around 24.2 – 49.4 Million Years Ago (MYA) [6,7].

Analysis of normalized expressed sequence tag (EST) sequences in *T. hassleriana* and comparative genome analysis in *Carica papaya*, both members of the Brassicales, and in *Arabidopsis thaliana* belonging to Brassicaceae revealed that Cleomaceae share the most ancient gamma (γ) whole genome duplication (WGD) with both *C. papaya* and *A. thaliana*. The sister families Brassicaceae and Cleomaceae also share the more recent beta (β) WGD which is lacking in *C. papaya*. However, the third and most recent alpha (α) WGD has occurred independently in Brassicaceae and Cleomaceae. The *T. hassleriana* α WGD (Th- α) is a genome triplication and occurred approximately 13.7 MYA, while the *Arabidopsis thaliana* α WGD (At- α) happened around 23.3 MYA [8]. In spite of the recent Th- α triplication event the genome of *T. hassleriana* is only 1.9 times the size of that of *A. thaliana* [6] and around half the size of the *C. papaya* genome. The small genome size of *T. hassleriana* indicates rapid diploidization, and a faster subsequent gene loss when compared to *A. thaliana* [6].

Cleomaceae are being intensively studied as C4 type photosynthesis evolved de novo in this group of plants. While *A. thaliana* and other Brassicaceae are C3 plants, C4 photosynthesis evolved in Cleomaceae at least three times independently in *Gynandropsis gynandra*, *Cleome oxalidea*, and *Cleome angustifolia*. *Cleome paradoxa* shows a C3 – C4 intermediate anatomy and physiology thus making Cleomaceae a model system to study C3 – C4 evolution [9]. Comparative leaf transcriptome studies by RNA-Seq have been carried out in *G. gynandra* (C4) and *T. hassleriana* (C3) to elucidate and identify novel genes and gene networks responsible for the C4 anatomy [10].

T. hassleriana (Figure 1) is also called the spider flower plant due to the long stamens which appear like appendages of spiders and is a popular ornamental plant owing to its colorful and abundant flowers. Adult plants can grow about five feet tall and several feet in diameter

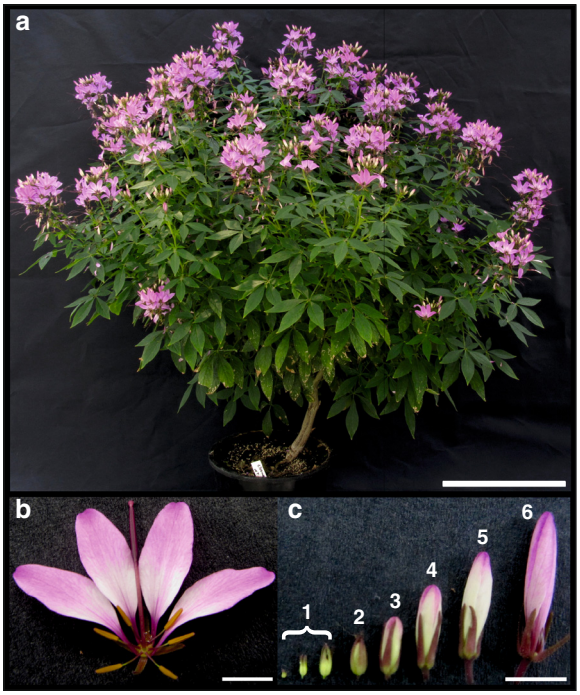


Figure 1 *Tarenaya hassleriana* plant. a) Morphology of a flowering *T. hassleriana* plant. **b)** Flower at anthesis showing four small purple sepals, four showy pink petals, six yellow stamens and a central gynoecium. **c)** Bud stages 1–6 characterized by bud length, bud stage 1 (<2.5 mm), bud stage 2 (3–4 mm), bud stage 3 (4–5 mm), bud stage 4 (6–8 mm), bud stage 5 (8–10 mm), bud stage 6 (12–15 mm). Scale bar 30 cm for a and 5 mm for b and c.

with several lateral branches. The stem and the lateral branches are soft and succulent but the main stem and older branches become woody with age. The leaves are palmate with 3–5 folioles per leaf (Figure 1a). Plants start flowering while they are in the juvenile stages and most of vegetative growth overlaps with the flowering period [11].

A typical *T. hassleriana* flower is zygomorphic unlike the disymmetric *A. thaliana* flower. Each flower has four sepals, four petals, six stamens, and a single gynoecium composed of two fused carpels (Figure 1b). The flower buds are laid out in a disymmetric bauplan during the early developmental stages which changes near anthesis and the mature flowers become zygomorphic. Conversely, in *A. thaliana* early developmental bud stages are monosymmetric and the flowers become disymmetric near anthesis [12]. *C. papaya* flowers on the other hand are actinomorphic at anthesis but early development has not been characterized yet. *T. hassleriana* inflorescences produce hermaphroditic, female, or male only flowers such that fruits are only periodically formed. The synchronous and alternate appearance of male, female, and hermaphroditic flowers in a raceme favors out-crossing, and prevents selfing except in the case of the hermaphroditic flowers [11]. This feature distinguishes *T. hassleriana* from most

plants which are either dioecious (like *C. papaya*) with separate male and female plants which rarely produce hermaphroditic flowers, or monoecious like *A. thaliana* which is an obligate self-pollinated plant with hermaphroditic flowers. *T. hassleriana* plants are very prolific, they reseed and establish in suitable environments very easily and escape from cultivation, often becoming invasive in subtropical countries like Japan, New Zealand, parts of Australia, and the United States of America [13]. Hence many horticultural varieties, possibly like the one used in this study were developed to be sterile so that they cannot establish in non-native environments.

Also unlike most Brassicales *T. hassleriana* flowers are very colorful due to the presence of various anthocyanins and show 'petal fading' i.e. loss of pigmentation and dry matter associated with anthesis. This phenomenon coupled with favored cross pollination may suggest a specific role in flower - pollinator interactions or simply an age related phenomenon [14].

A close relationship to *A. thaliana* facilitates the analysis of *T. hassleriana* specific traits, such as flower coloration, alternating development of three types of flowers and flexible shifts from vegetative to reproductive growth, which are all not found in *A. thaliana*. Here, we describe the floral transcriptome sequence along with Transcriptome Sequencing Expression (TSE) of a horticultural *T. hassleriana* hybrid as a starting point for further analysis of Tarenaya flower development. Expression analysis by qRT-PCR documents the robustness of the TSE and rarefaction analysis shows that the transcriptome sequencing covers even rare transcripts. Candidate genes that may be involved in the *T. hassleriana*-specific flower developmental processes have been identified and are presented here.

Methods

Plant material and growth parameters

A *T. hassleriana* hybrid plant was obtained from a local garden center. It was grown in 3:1 mixture of a peat and sand based potting soil with perlite supplemented with 2 g/l Osmocote® slow release fertilizer (Scotts Deutschland GmbH, Nordhorn, Germany). The plant was grown in a greenhouse under long day growth conditions (17 hours light and 7 hours dark) with light varying between 80 and 700 $\mu mol.m^{-2}.s^{-1}$ photons. Supplemental lighting of 70 $\mu mol.m^{-2}.s^{-1}$ was provided throughout the photoperiod. The temperatures in the green house varied between 20°C (day) and 16°C (night).

Tissue collection, nucleic acid extraction, and cDNA synthesis

T. hassleriana floral tissue was collected for RNA and DNA extraction. The collected flower tissue was composed of equal quantity (by mass) of flowers at anthesis and each of the 6 floral bud stages defined by bud length as shown in Figure 1c).

For the floral transcriptome sequencing total RNA was extracted from the *T. hassleriana* floral tissue using guanidium thiocyanate-phenol-chloroform extraction protocol [15]. The $polyA^+$ mRNA was isolated using the Oligotex mRNA Minikit (Qiagen, Hilden, Germany) according to the manufacturer's instruction. The purified mRNA was analyzed for quality and quantity using the Eukaryote Total RNA Pico assay of the Agilent 2100 Bioanalyzer (Agilent Technologies, Böblingen, Germany). For the qRT-PCRs total RNA was isolated from floral tissues with the plantRNA Kit-OLS® (Omni Life Science, Bremen, Germany) following the manufacturer's protocol. Genomic DNA was extracted from the floral tissue using the DNeasy® Plant Mini kit (Qiagen, Hilden, Germany) according to the manual.

Library preparation and 454 pyrosequencing

200 ng of purified $polyA^+$ mRNA was used to synthesize the cDNA for sequencing with Roche Rapid Library kit (Roche) following the manual. A massively parallel pyrosequencing run was performed on a GS FLX using Titanium chemicals (Roche) with a split picotiterplate allowing two replicates to run at the same time.

Assembly, annotation and gene expression

All reads together were de novo assembled using CLC Genomics Workbench 4.9 (clcBio, Aarhus Denmark). Default parameters were chosen for the assembly. The resulting 49321 contigs were annotated against TAIR10 coding sequences (representative gene model 20110103). A reciprocal BLATX mapping was performed [16] and the best bi-directional hit per contig was kept as annotation. Chimeric contigs were determined with the pipeline provided in [17].

Gene expression was determined by mapping the reads to TAIR10 coding sequences using BLATx. The single best hit for each read was counted. Expression values were normalized to reads per Kilobase gene model per mappable million (RPKM). All reads were additionally mapped to the *T. hassleriana* floral transcriptome contigs with CLC Genomics Workbench for subsequent rarefaction analysis. The expression data for the *T. hassleriana* leaf transcriptome was obtained from Bräutigam et al. [10].

Lineage specific gene detection

Based on the *T. hassleriana* floral contigs, mappings to the transcriptomes of *A. thaliana*, *B. rapa*, *C. papaya*, and *P. trichocarpa* with an e-value cutoff of 10^{-10} were created in proteinspace. From those all against all mappings the 15 overlapping sets and the residual *T. hassleriana* specific set were determined using R's set methods [18].

Quantitative reverse transcription PCR (qRT-PCR)

For the qRT-PCRs the first strand cDNA was synthesized with the RevertAid™ H Minus First Strand cDNA Synthesis Kit (Fermentas, St.Leon-Rot, Germany) according to the manufacturer's protocol using an universal oligo(dT) (T_{18}) primer. qRT-PCR experiments were performed according to the MIQE guidelines [19]. Exon spanning primers were then generated using PerlPrimer 1.1.21. [20]. A primer efficiency test was carried out and all the primers were tested with genomic DNA to ensure cDNA specificity. (Primer sequences are provided in Additional file 1: Table S2).

The qRT-PCR assay was performed in 96 well plates using the LightCyler®480 II (Roche, Mannheim, Germany) and analyzed with the LCS480 1.5.0.39 software. Each reaction was composed of 10 μl of 2x DyNAmo™ Flash SYBR® Green qPCR Mastermix (Biozym Scientific GmbH, Oldendorf Germany), 2 μl each of 10 μM forward and reverse primers, 1 μl H_2O and 5 μl of diluted cDNA template. Standard dose response (SDR) curves were constructed for all the genes by using serial dilutions (1:50 to 1:50,000) of the cDNA template. Each reaction was performed in biological duplicates and technical triplicates along with water and RNA controls for each primer pair. The *T. hassleriana ACTIN7 (ACT7)* gene served as an internal control. The following PCR program was used: 7 min at 95°C; 45 cycles of 10 s at 95°C, 15 s at 60°C, 15 s at 72°C, followed by a melting curve of 5 s at 95°C, 1 min at 65°C and 30 s at 97°C. The Absolute Quantification analysis and the quantification cycle (Cq) were calculated according to the Fit Points method using the LCS480 1.5.0.39 software. The amplification efficiency was calculated using the SDR for each gene. The raw data were analyzed according to the relative standard curve method and the fold difference between the expression of *ACT7* and the genes of interest was calculated using the comparative Cq method (ΔΔCq) [21]. A one way ANOVA was performed to calculate the statistical significance of the difference between the three expression values.

Comparison of *A. thaliana* and *T. hassleriana* floral gene expression and GO annotations

In order to identify genes that may play a role in the *T. hassleriana* specific floral traits, transcripts specific for the *T. hassleriana* flower, not expressed in the *A. thaliana* flower and vice versa were identified. Microarray expression data [22] for *A. thaliana* flower stages 1–6, 9, 10–11, 12, 15 (ATGE_29_A2, B2, C2; ATGE_31_A2, B2, C2; ATGE_32_A2, B2, C2; ATGE_33_A2, B2, C2; ATGE_39_A2, B2, C2) were downloaded from

The Arabidopsis Information Resource (TAIR), http://arabidopsis.org/servlets/TairObject?type=hyb_descr_collection&id=1006710873#497, on, Nov 20 2012.

Of the 22,746 microarray probes hybridizing to 23,570 genes, only 21,107 probes hybridizing to unique transcripts were considered for the analysis. A dataset corresponding to the expression of these 21,107 transcripts in the *A. thaliana* floral transcriptome was compiled. Expression of a gene in at least one floral stage and sample subset was considered as presence of the transcript in the *A. thaliana* floral transcriptome. The presence or absence of homologous transcripts in the *T. hassleriana* floral transcriptome was analyzed. A list of putative *T. hassleriana* orthologs of *A. thaliana* genes expressed in *T. hassleriana* floral transcriptome but not in the *A. thaliana* floral transcriptome was constructed. Also, transcripts present in the *A. thaliana* flower transcriptome but homologs absent in the *T. hassleriana* transcriptome were identified. Gene Ontology (GO) annotations were assigned to genes expressed exclusively in the *A. thaliana* or *T. hassleriana* transcriptome using the online tool for functional annotation Blast2GO® [23] by performing a BLASTX with a cutoff value of $1e^{-100}$ as this value showed robust matches of GO annotations to TAIR annotations.

GO annotations were assigned to *T. hassleriana* lineage specific sequences, and other sequences shared by Cleomaceae with the Brassicaceae, Brassicales or lost in the Brassicaceae using Blast2GO® [23] by performing BLASTX and BLASTN with cutoff values of $1e^{-10}$.

Rarefaction analysis

Rarefaction analysis is commonly used in ecological research defining species richness as a function of sequencing effort. Such an analysis can be broadened to genomics as long as the data are distributed as described in the original paper defining the underlying equation [24]. Hale et al. [25] already calculated rarefaction curves for transcriptome analysis of a polyploid lake sturgeon. Here, we applied rarefaction analysis to ascertain whether sequencing depth and coverage was sufficient to draw a comprehensive picture of the transcriptome of Cleome. Thus, three different libraries were created: one data set for each biological replica as well as a merged one. Data sets were constructed by listing each gene (defined by a contig) with its read support. Rarefaction curves were calculated using the program aRarefactWin (https://www.uga.edu/strata/software/). Hereby, genes were randomly resampled and it was recorded which gene of the library was identified with which frequency. This procedure was repeated 1,000 times. Then, the average number of each gene found was plotted for different read numbers drawing a curve whose slope indicated if sequencing effort was deep enough. This was the case when the curve flattened and ran into a plateau.

Results

Sequencing

Massively parallel pyrosequencing of two samples of *T. hassleriana* (Additional file 1: Table S1) yielded 1,254,286 sequencing reads in total. The sequencing raw data are

deposited in the DDBJ (DNA Data Bank Japan, http://trace.ddbj.nig.ac.jp/index_e.html) under the experiments SRR1051360 and SRX393170 https://trace.ddbj.nig.ac.jp/DRASearch/run?acc=SRR1051360 and (https://trace.ddbj.nig.ac.jp/DRASearch/experiment?acc=SRX393170) The histogram of reads by length (Additional file 2: Figure S1a) showed an average read length of ~316 nucleotides. Roughly 45% of the reads could be mapped against *A. thaliana* TAIR10 coding sequences for counting gene expression.

Assembling the reads de novo resulted in 49,237 contigs with an N50 of 690 bases (Additional file 2: Figure S1b). Of these, 41,320 could be annotated by mapping against Arabidopsis. 1.1% (537) chimeric contigs could be detected in the assembly.

Rarefaction analysis

Rarefied libraries were constructed separately for the two biological replicates 1 and 2 and a merged sample to illustrate possible differences in gene discovery rates. Although the gene discovery rate of replicate 1 was less than the one of replicate 2 (Figure 2), the curves for both replicates indicated that a larger part of the *T. hassleriana* floral transcriptome was detected as the curves already flattened. However, the merging of the information of both libraries affected the overall output as the rarefaction curve reached nearly a plateau (Figure 2). This shows that each library comprised genes not detected with the other one. Thus, the merged data set allows drawing a detailed view of the transcriptome of *T. hassleriana*. Increasing the sequencing depth would only result in the detection of extremely rare genes.

qRT-PCR expression analysis validates transcriptome sequencing expression (TSE)

The robustness of expression data generated by the transcriptome sequencing was analyzed independently using a qRT-PCR assay (Figure 3). A normalized expression profile for *T. hassleriana* reads mapped to *A. thaliana* CDS sequences was created by calculating the ratio of reads mapped to an individual gene against the reads mapped to *A. thaliana ACT7*. A subset of 14 genes was randomly chosen to represent genes with high (normalized expression ratio 1.0 – 10.0, Figure 4a), moderate (normalized expression ratio 0.3 – 1.0, Figure 4b) and low (normalized expression ratio 0.05 – 0.3, Figure 4c) expression levels. The expression of the putative *T. hassleriana* orthologs of the *A. thaliana genes RBCS1A, MVP1, GAPC1, TT4, BGLUC19, GAMMAVPE, ATP3, SCE1A, SFGH, ARF6, PGLUHYD, GI, OMR1*, and *SPL7* was analyzed in *T. hassleriana* floral tissue (*A. thaliana* gene identifier, full gene names are shown in Additional file 1: Table S3). The qRT-PCR expression data were also normalized to the expression of the *T. hassleriana ACT7*.

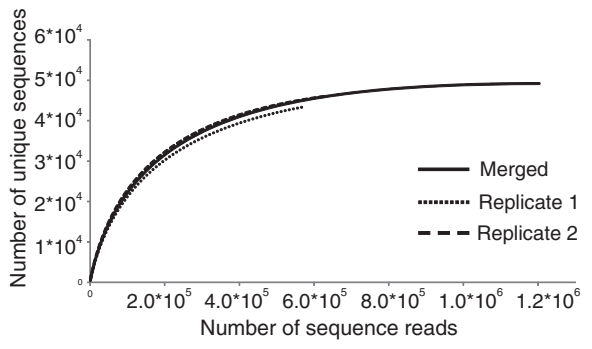


Figure 2 Rarefaction for transcriptome depth and coverage analysis. Rarefaction curves for both libraries representing expected gene coverage rates. Curves were constructed for the two biological replicates 1 and 2 and a merged sample.

Generally we found a better match of transcript abundance detected by qRT-PCR in *T. hassleriana* as compared to reads mapped to the *A. thaliana* orthologs (TSE1) than to the *T. hassleriana* contigs (TSE2). A correlation plot for the comparison of expression measured by qRT-PCR and TSE was generated (Additional file 2: Figure S3). When all the 14 gene expressions by the two methods were plotted a positive linear correlation was observed (Additional file 2: Figure S3a) as indicated by a R^2 value 0.55. The expression of *MVP1* and *BGLUC19* gene homologs which belong to big gene families with 41 and 66 homologs in *A. thaliana* respectively was the most significant outlier in this plot. When the expression data for the *MVP1* and *BGLUC19* gene homologs were removed and the data plotted again a very strong positive linear correlation between TSE1 and qRT-PCR expression values was obtained with an R^2 value 0.91 (Additional file 2: Figure S3b). This indicated that TSE1 approach for measuring gene expression was very robust except for genes belonging to large gene families with highly similar homologs in which case the read mapping may be incorrect. Nonetheless a positive linear expression correlation for all genes corroborates the TSE1 expression data. In particular, similar normalized fold expression between qRT-PCR data and reads mapped to the *A. thaliana* orthologs were observed in the genes *RBCS1A* (high expression), *ATP3, SCE1A, SFGH* (moderate expression), and *ARF6, PGLUHYD, OMR1*, and *SPL7* (low expression) P >0,01 (Additional file 1: Table S4 shows the comparative P values for the ANOVA tests). In case of *T. hassleriana* homologs of genes *GAMMAVPE, MVP1* and *TT4* the transcript abundance detected by qRT-PCR was more similar when reads were mapped to the *T. hassleriana* contigs (TSE2) P > 0,01. In case of the *GAPC1* and *BGLUC19* homologs the difference between qRT-PCR expression and TSE1 and TSE2 was statistically significant P < 0,01. It was further observed that the number of reads mapped to the *T. hassleriana* contigs was in all

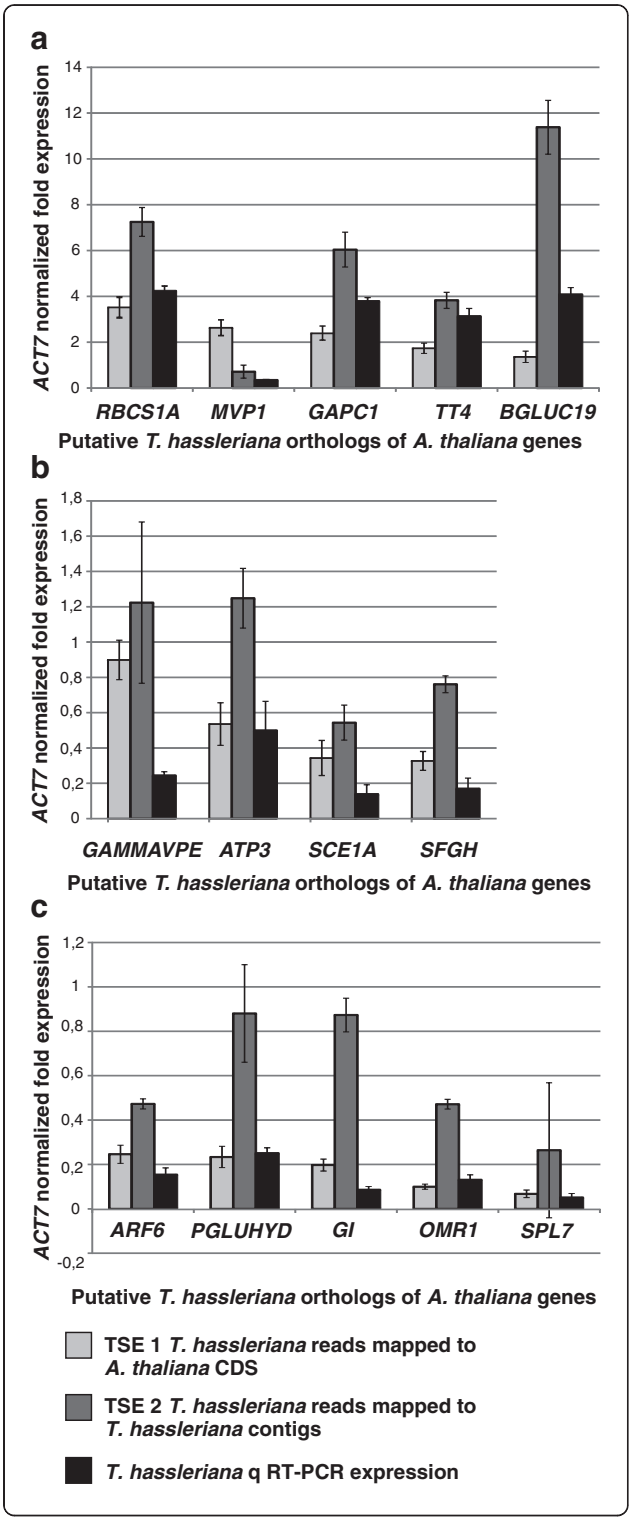


Figure 3 Comparative analysis of Transcriptome Sequencing based Expression data (TSE) with qRT-PCR expression data. The TSE and the qRT-PCR expression were normalized with the putative *T. hassleriana ACTIN7* homolog. The first column represents the TSE 1 where the *T. hassleriana* reads are mapped on to the *A. thaliana* CDS sequences; the second column represents the TSE 2 where the *T. hassleriana* reads are mapped on the *T. hassleriana* contigs and the third column represents the qRT-PCR expression data. **a**: Comparison of TSE and qRT-PCR in genes with high expression (750–2000 RPKM, **b**: Comparison of TSE and qRT-PCR in genes with moderate expression (150–300 RPKM) and **c**: Comparison of TSE and qRT-PCR in genes with low gene expression (25–150 RPKM). The error bars represent the standard deviation and the P-values for statistical significance between expression values are presented in Additional file 1: Table S4.

cases, with the exception of *TT4*, grossly overestimating gene expression.

Expression of genes controlling floral traits in the flower and leaf transcriptome

Genes controlling various floral traits and flower development in *A. thaliana*, *Antirrhinum majus*, *Fagopyrum esculentum* etc. were identified based on literature [26-31]. The expression pattern of their putative *T. hassleriana* orthologs identified by a bidirectional BLATX search with the *A. thaliana* CDS sequences was analyzed in the flower and leaf transcriptomes to learn more about the regulation of the special floral traits of *T. hassleriana* (Figure 4). The selected genes were first grouped into different classes such as homeotic transcription factors, regulators of homeotic genes etc. and ordered within their groups according to transcript abundance. Of the genes analyzed, 49 (41.9%) were specific to the flower transcriptome and not found in the leaf transcriptome. (*A. thaliana* gene identifier, full gene names are shown in Additional file 1: Table S3).

Amongst the putative class ABCDE homeotic transcription factor orthologs, the highest expression was observed among the class B gene homologs *AP3* and *PI* and the class E gene homologs *SEP1* and *SEP3*. The putative ortholog of the C class gene *AG* was expressed at a 10 fold lower magnitude compared to the class B and E genes. The expression of the putative orthologs of the D class genes *SHP1*, *SHP2* and *STK* the expression of which regulates the ovule and fruit development in *A. thaliana* was found to be considerably lower, when compared to the class ABCE genes. *AP3, SEP3, SEP1*, and *STK* transcripts were not present in the leaf transcriptome while *PI, AP1, AP2*, and *SEP4* are expressed at a very low level in leaves. In addition to these, 25 other putative MADS box transcription factors without floral homeotic function that are members of the MIKC, Mα, Mβ, Mγ, Mδ subfamilies were also found to be expressed in the floral transcriptome.

Amongst the genes putatively regulating the class ABCDE homeotic transcription factors, the *LUG*, *LUH*

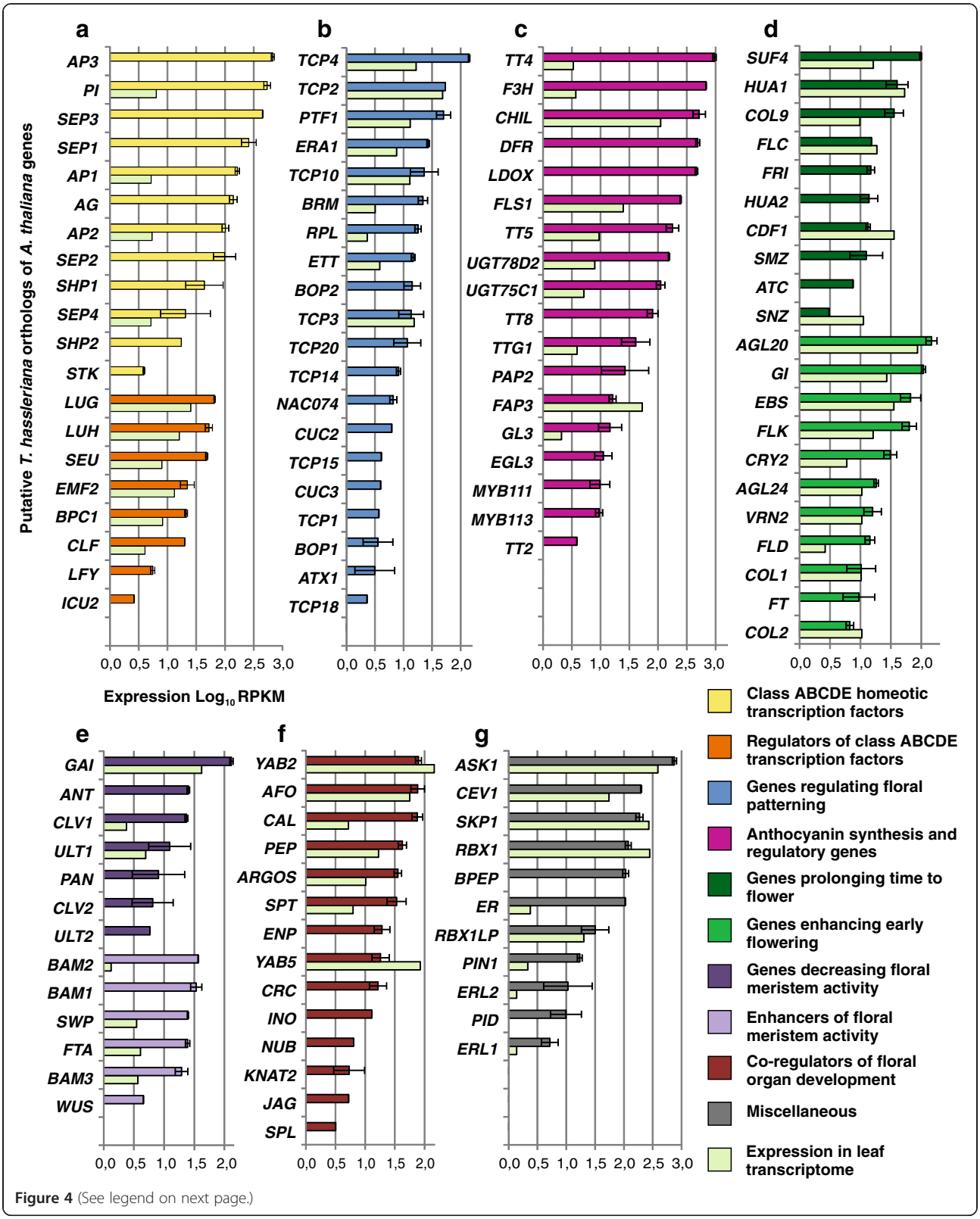


Figure 4 (See legend on next page.)

(See figure on previous page.)
Figure 4 Expression of *T. hassleriana* orthologs of *A. thaliana* genes regulating various floral characteristics in the *T. hassleriana* flower and leaf transcriptomes. The putative orthologs are plotted on the Y axis and the Transcriptome Sequencing Expression (TSE 1) which is the $\log_{(10)}$ of RPKM is plotted on the X axis. The error bars show standard deviation; **a**: Expression of class ABCDE homeotic transcription factors and their regulators in the floral and leaf transcriptomes of *T. hassleriana*, **b**: Genes regulating patterning and symmetry, **c**: Genes involved in synthesis and regulation of anthocyanins, **d**: Genes regulating time to flower, **e**: Positive and negative regulators of floral meristem activity, **f**: Co-regulators of floral organ development, **g**: Miscellaneous group genes involved in flower development.

and *SEU* orthologs showed highest expression in the flower transcriptome, while a 10 fold lower expression of these genes was observed in the leaf transcriptome. The expression of the *LFY* homolog was also observed in the floral transcriptome albeit at very low levels. Interestingly, putative homologs of genes regulating class B gene activities like *UFO* and *SUP* and class A gene activity like *SAP* were not identified in the floral transcriptome library. The orthologs for genes regulating patterning and symmetry also showed expression in the floral transcriptome. The putative orthologs of *TCP4, TCP2*, and *PTF1* showed the highest expression. The expression of these genes was also observed in the leaf transcriptome; in case of the *TCP4* ortholog a 100 fold higher expression was observed in the floral transcriptome when compared to the leaf transcriptome, while the *TCP2* ortholog expression was almost equal in both the transcriptomes. Comparatively low expression of other putative patterning gene orthologs like *TCP14, TCP15, TCP18, CUC2* and *CUC3* was also observed specific to the floral transcriptome.

While the *A. thaliana* flowers are mostly free of pigments, the petals and reproductive organs of *T. hassleriana* are pink and dark magenta and hence the expression of putative orthologs of genes regulating anthocyanin production, regulation, and deposition was analyzed. Very high expression was observed for the putative orthologs of *TT4, F3H, CHIL, DFR* and *LDOX*. Most genes show a higher expression in flowers than in leaves and for several, such as *DFR, LDOX*, and *TT8*, expression is specific to the flower suggesting key roles in flower pigmentation. Very low expression of *TT4* and *F3H* orthologs (about 300 and 200 fold lower respectively) was observed also in the leaf transcriptome, whereas the *CHIL* ortholog expression was only about 5 fold lower in the leaf transcriptome. The spatiotemporal expression pattern of *A. thaliana* orthologs of these genes was investigated in *A. thaliana* using the Arabidopsis eFP Browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi) [32]. The expression patterns for the homologs of *TT4, F3H, CHIL*, and *FLS1* was very similar in *T. hassleriana* and *A. thaliana*. The enzymes encoded by these genes are required for the synthesis of flavonoids like quercitin, dihydroquercitin, myricetin etc., which are intermediates of anthocyanin biosynthesis. The products of the genes *DFR, LDOX, UGTD2* which were found to be expressed in in the *T. hassleriana* floral transcriptome but only in senescing leaves in *A. thaliana* (Table 1) are involved in downstream processes that catalyze the conversion of the flavonoids into anthocyanins like Pelargonidin and Cyanidin which determine the characteristic pink-magenta flower color. Genes like *PAP2, MYB111, MYB113*, and *EGL3* are regulators of flavonoid and anthocyanin biosynthesis and were also expressed in *T. hassleriana* floral tissue whereas in *A. thaliana* their expression was restricted to senescing leaves and seeds during early stages of embryo development.

Expression of gene orthologs governing time to flower was also analyzed. Expression of both antagonistic groups of genes that prolong time to flower or enhance the transition into flowering was observed. Among the orthologs inducing flowering *AGL20, GI, EBS*, and *FLK* had the highest expression; expression of these genes was also observed in the leaf transcriptome at very comparable levels. Amongst the orthologs of genes delaying flowering *SUF4, HUA1, COL9*, and *FLC* had high levels of expression which was also observed at comparable levels in the leaf transcriptome. The orthologs of *FRI, HUA2, SMZ*, and *ATC* showed moderate to low floral transcriptome specific expression.

T. hassleriana homologs of meristem activity regulators, such as *GAI, ANT* and *CLV1* which are involved in decreasing meristem proliferation was observed at high levels in the flower and varying levels in the leaf transcriptome while *ANT* expression was not detected in the leaf transcriptome. Putative homologs of genes *BAM1, BAM2, BAM3* and *WUS* which enhance meristem proliferation were also found to have moderate expression levels in the floral transcriptome. Interestingly, putative homologs for *FTA*, *ERA1*, and *STM*, were found to be expressed in the floral transcriptome as their *A. thaliana* counterparts show very low expression the flower.

Another important category of gene orthologs analyzed for expression are the genes that co-regulate floral organ development alongside the ABCDE floral homeotic transcription factors. High expression was observed in case of orthologs of *YAB2, AFO* and *PEP* in both the floral and leaf transcriptomes whereas the expression of the *CAL* ortholog was about 100 fold higher in the flower transcriptome. Other floral organ developmental regulators, such as *ENP, CRC, INO, NUB, JAG*, and *SPL* were not identified in the *T. hassleriana* leaf transcriptome, but only in floral transcriptome whereas they are also expressed in *A. thaliana* leaves at very low levels.

Table 1 Genes putatively involved in anthocyanin synthesis, regulation, and deposition found in the floral transcriptome of *T. hassleriana* and the expression of their putative orthologs in *A. thaliana* tissues and developmental stages

Gene homologs expressed in *T. hassleriana* floral transcriptome	Expression in *A. thaliana*
TT4	Buds, senescent leaf, seed (globular embryo stage)
F3H	Buds, petal, seed (globular and torpedo stage embryo)
CHIL	Buds, petal, young silique, seeds (globular and torpedo stage embryo)
DFR	Senescent leaf, young silique, seed (heart stage embryo)
LDOX	Senescent leaf, young silique, seed (heart stage embryo)
FLS1	Buds, petal, seeds (torpedo and walking stick stage embryo)
TT5	Buds, petal, carpel, seed (globular and heart stage embryo)
UGTD2	Senescent leaf, seed (curled cotyledon, green cotyledon stage embryo)
UGTC1	Senescent leaf
TT8	Young siliques, seeds (heart, walking stick stage embryo)
TTG1	All plant organs, high expression in cauline and senescent leaves, young siliques, seeds (heart and torpedo stage embryo)
PAP2	Senescent leaf
FAP3	Cauline leaf, young siliques, seeds (Heart, torpedo, walking stage embryo)
GL3	Expression data not available
EGL3	Shoot apex (vegetative, floral transition, inflorescence), young silique, seeds (globular, torpedo, walking stick stage embryo)
MYB111	Petals, shoot apex (inflorescence)
MYB113	All plant organs, high expression in pollen, seeds (curled cotyledon and green cotyledon embryo stage)
TT2	Young siliques, seeds (globular and heart stage embryo)

No expression was observed for *ROXY* gene homologs which are responsible for anther and male gametophyte development downstream of *SPL*.

Other putative homologs of *A. thaliana* floral regulators were identified amongst them were the highly expressed homologs of genes *ASK1, CEV1, SKP1, RBX1*, which are part of SCF ubiquitin protein ligase complexes which regulate multiple aspects of flower development together with *UFO* in *A. thaliana* [33]. Homologs of genes *ER, ERL1* and *ERL2* which are protein kinases that influence meristem cell fate and patterning in the inflorescence meristem were also highly expressed. Interestingly the homolog of *BPEP* was found to be expressed only in the floral transcriptome, while the two distinct *BPEP* transcripts in *A. thaliana* are expressed in the floral as well as in vegetative organs respectively. Homologs of genes *PIN1* and *PID* were also expressed which are known to affect size, floral organ number and total number of flowers in *A. thaliana*.

This in silico expression analysis of genes related to flower development demonstrates that with the chosen RNAseq method we are able to monitor gene expression in logarithmic scales covering more than two magnitudes. In addition, the two library preparations for this sequencing experiment show only rarely any difference in RPKM. Detailed expression analysis of putative *T. hassleriana* homologs of *A. thaliana* genes in the *T. hassleriana* floral transcriptome is provided in Additional file 3 along with the AGI identifiers.

Characterization of genes putatively governing sterility in *T. hassleriana*

The particular *T. hassleriana* hybrid used in this study was sterile. While orthologs of *A. thaliana* regulators of anther development were expressed in the *T. hassleriana* flower, no expression of *ROXY1* and *ROXY2* was detected. These two genes redundantly control the anther lobe and pollen mother cell differentiation downstream of *SPL* [34]. The genome of one of the parents of this hybrid, *T. hassleriana* Purple Queen (ES1100) was recently published [35] and this plant, unlike its hybrid offspring is fertile. Only *ROXY1* ortholog was found in the *T. hassleriana* genome To learn more about the possible causes for the sterility we compared the expression pattern of homologs of *SPL, ROXY1* and their *A. thaliana* downstream targets *DYT1* and *MYB35* affecting stamen development and microsporogenesis in these two plants by qRT-PCR at small, medium and large buds (Figure 5).

Expression analysis by qRT-PCR indeed revealed that the expression of the *ROXY1* homolog was very low (10^3 fold lower compared to *ACT7*) and well beyond the scope of detection by RNA seq. *ROXY1* expression was down regulated in the sterile hybrid only at bud stage M when compared to the fertile parent the (Figure 5b) whereas it was similar to the parent at the younger and later developmental stages. Along with the down regulation of *ROXY1*, expression for the *DYT1* and *MYB35* homologs which most likely act downstream of *ROXY1* was also down regulated in stage M buds. In stages other than M, the expression of *DYT1* and *MYB35* homologs in the sterile *T. hassleriana* hybrid was several fold higher than the respective expression in fertile parent buds in both the early and late developmental stages. Expression of the *SPL* homolog in the sterile hybrid buds was 3–4 fold higher than the fertile plant buds in stages S and M whereas in stage latter L the expression was 8 fold. Thus our expression data suggest that the complex network governing

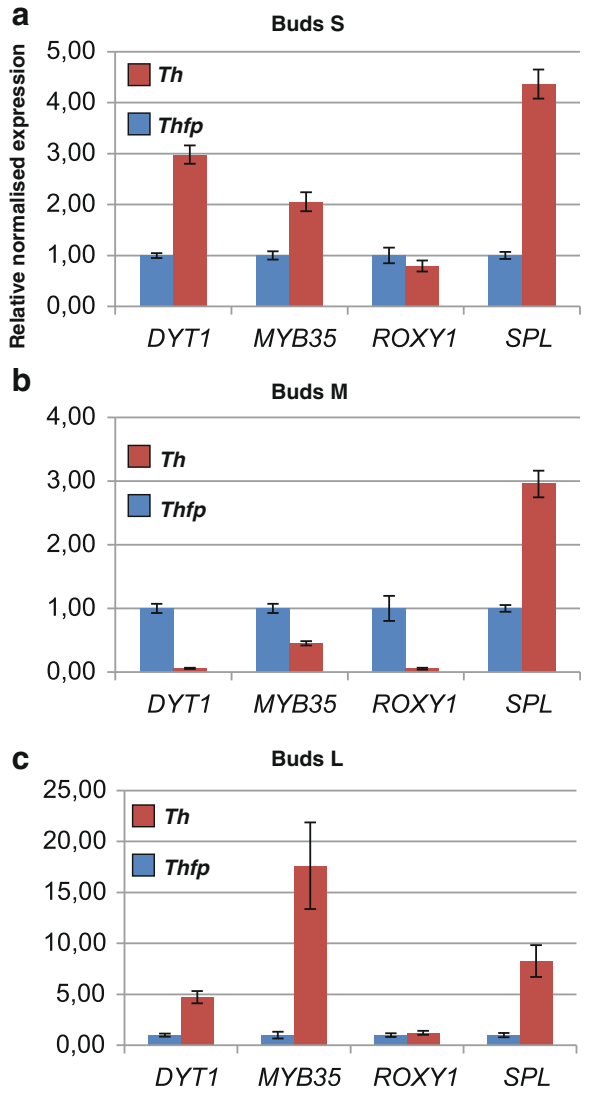


Figure 5 Comparative expression analysis of genes governing stamen development and microsporogenesis in the sterile *T. hassleriana hybrid* (Th) and the fertile *T. hassleriana* parent (Thfp) in a: Small buds, stage 1 (S), b: Medium sized buds, stage 2 and 3 (M), and c: Large buds (stage 4 to 6) and open flowers, (L).

stamen development and microsporogenesis is disrupted in the *T. hassleriana* hybrid which could provide a causal link to its sterility.

Characterization of *T. hassleriana* floral transcriptome specific genes in comparison to *A. thaliana*

We described above that the flower of *T. hassleriana* is morphologically distinct from the *A. thaliana* flower and our aim was to identify genes that may contribute to the differences by comparing the *A. thaliana* floral transcriptome with that of *T. hassleriana*. However, as our data are based on RPKM and the *A. thaliana* are microarray data the two datasets may be compared only qualitatively but not quantitatively. We thus chose the more careful approach to score only for presence/absence of transcripts of *A. thaliana/T. hassleriana* putatively orthologous gene pairs. Of the 21,107 genes in *A. thaliana* for which microarray expression data for the floral transcriptome could be compiled, ~1200 genes were not expressed in the *A. thaliana*. The expression analysis of these gene homologs in the *T. hassleriana* revealed that a majority of these genes (~750) were also not expressed in the *T. hassleriana* floral transcriptome. But 351 gene homologs were identified that were expressed differentially amongst the floral transcriptomes of the two species. These differentially expressed Tarenaya transcripts were assigned GO annotations using Blast2GO® by performing a BLASTX search with a cut off value of e^{-100} to identify the molecular processes that are distinct between *T. hassleriana* and *A. thaliana*. 81 genes were annotated as genes with unknown function. The remaining 270 genes were assigned multiple GO annotations based on the biological processes associated with the function of these genes (Additional file 1: Table S5). Of special interest were genes annotated to be involved in anthocyanin accumulation, cell growth, flower development and other developmental processes. Candidate genes were selected for further analysis (Table 2). High expression of *PGP10* homolog, a gene involved in anthocyanin accumulation in response to UV light was observed in the *T. hassleriana* floral transcriptome whereas its expression is limited to pollen in *A. thaliana*. The homolog of *TTFP* which codes for a tyrosine transaminase family protein was also expressed at high levels in the *T. hassleriana* floral transcriptome; this gene is involved in regulation of cell growth in response to external stimulus and is primarily expressed in the roots of *A. thaliana*. Other notable gene homologs involved in various aspects of cell growth were *LRX2, HAT4* and *PIP5K3*. Of the gene homologs involved in various aspects of floral development, prominent were *ICMTA* and *TEM2*. *ICMTA* is an enzyme belonging to the methyltransferase family, which is inducted during floral morphogenesis. *TEM2* is a transcription factor known for its role in flowering time regulation by controlling *FT* expression. Amongst the genes annotated as genes governing various aspects of development were *JAL33, MTSP1*, *EMB2217* and *GLUDOXRP* which are involved in embryo and root development.

Identification and characterization of Cleome lineage specific genes

To identify genes shared between Cleome and other closely related rosids and genes that are specific to the Cleome lineage a BLASTX search with a cut off value of e^{-10} was performed with the 49,237 Tarenaya floral transcriptome contigs against the *A. thaliana, Brassica rapa, C. papaya*

Table 2 Selection of homologous gene pairs in which the homologs of *T. hassleriana* are expressed in the flower and the *A. thaliana* homolog expression is absent from the flower

Gene abbreviation	Process/protein family	GO ID	*T. hassleriana* TSE (RPKM)	Expression in *A. thaliana*
Anthocyanin accumulation				
PGP10	multidrug pheromone mdr abc transporter family	GO:0043481	63.79	Mature pollen
Cell growth				
TTFP	tyrosine transaminase family protein	GO:0001560	194.20	Root
ARP2	actin-related protein 2-like	GO:0009825	28.21	Senescent leaf, cauline leaf, buds, flower, inflorescence shoot apex
HB-2	homeodomain-leucine zipper protein	GO:0009826	28.07	Young leaf, mature leaf, cauline leaf, senescent leaf, pedicel, seed (torpedo stage embryo)
LRX2	leucine-rich repeat extensin-like protein 1	GO:0009826	22.97	Young leaf, pollen, seed (cotyledon stage embryo)
PLLSP	pectate lyase family protein	GO:0042547	12.85	Young leaf petiole, mature leaf (distal end), seed (curled cotyledon stage embryo)
Flower development				
ICMTA	protein-s-isoprenylcysteine o-methyltransferase a	GO:0009908	37,04	Young leaf, cauline leaf, senescent leaf, young silique, seed (heart and torpedo stage embryo)
SBP3	selenium-binding protein	GO:0048573	24,26	Imbibed seed
BTB/POZ P	BTB/POZ domain-containing protein	GO:0048439	12,55	Petals stamens
TEM2	ap2 erf and b3 domain-containing transcription factor rav2	GO:0009910	12,28	Cotyledon, young leaf, senescent leaf
Development				
JAL33	jacalin-like lectin domain-containing protein	GO:0009793	302,33	Root, hypocotyl
MTSP2	caffeoyl- o-methyltransferase	GO:0048316	47,03	Seed (curled cotyledon, green cotyledon embryo stage), dry seed
MTSP1	s-adenosyl-l-methionine-dependent methyltransferase-like protein	GO:0010089	40,59	Seed (walking stick, curled cotyledon, green cotyledon embryo stage)
LRRTPKP	lrr receptor-like serine threonine-protein kinase rch1-like	GO:0048443	15,26	Root, seed (torpedo stage embryo), imbibed seed
CYP705A27	cytochrome p450	GO:0048589	11,86	Root, seed (cotyledon embryo stage), dry seed
EMB2271	u3 small nucleolar rna-interacting protein 2-like	GO:0009553	11,81	Stamen
CYP705A	cytochrome p450	GO:0048589	11,54	Root
GLUDOXRP	glutaredoxin-related protein	GO:0048653	7,40	Pollen, seed (walking stick, curled cotyledon, green cotyledon stage embryo)
LRRRPK	receptor-like protein kinase 2-like	GO:0048443	3,50	Imbibed seed, root

(all malvids, order Brassicales) and *Populus trichocarpa* (fabid, order Malpighiales) protein databases in a systematic manner (Figure 6a). This allows the assessment of gene births and gene losses in the rosid lineage. Figure 6b shows the result of the comparative analysis: A large number of the contigs 37,989 (subset I) represent the sequences shared between malvids and fabids. According to our analysis, only 684 genes are shared between all Brassicales, but 1375 genes (subset B) are shared between the core Brassicales, which include *T. hassleriana, A. thaliana,* and *B. rapa* [36]. This suggests a high rate of gene births in the lineage leading to core Brassicales after their split from the lineage leading to *C. papaya*. Conversely, 148 genes (subset K) are shared between *T. hassleriana, C. papaya* and *P. trichocarpa* and not found in the Brassicaceae suggesting that these genes were lost in the lineage leading to *A. thaliana* and *B. rapa* after its separation from the lineage leading to *T. hassleriana*. Another 132 (subset G) genes are found only in *C. papaya* and *T. hassleriana* indicating that these are Brassicales-specific genes that were lost in the Brassicaceae. 453 genes are shared between *T. hassleriana* and *A. thaliana* but not found in *B. rapa* suggesting that they were lost in the lineage leading to *B. rapa*. Conversely, only 246 genes were lost in the lineage leading to *A. thaliana* and are shared between *B. rapa* and *T. hassleriana* (subset C).

An astonishing number of 5600 contigs (subset Z) could not be matched with high confidence to any other sequence

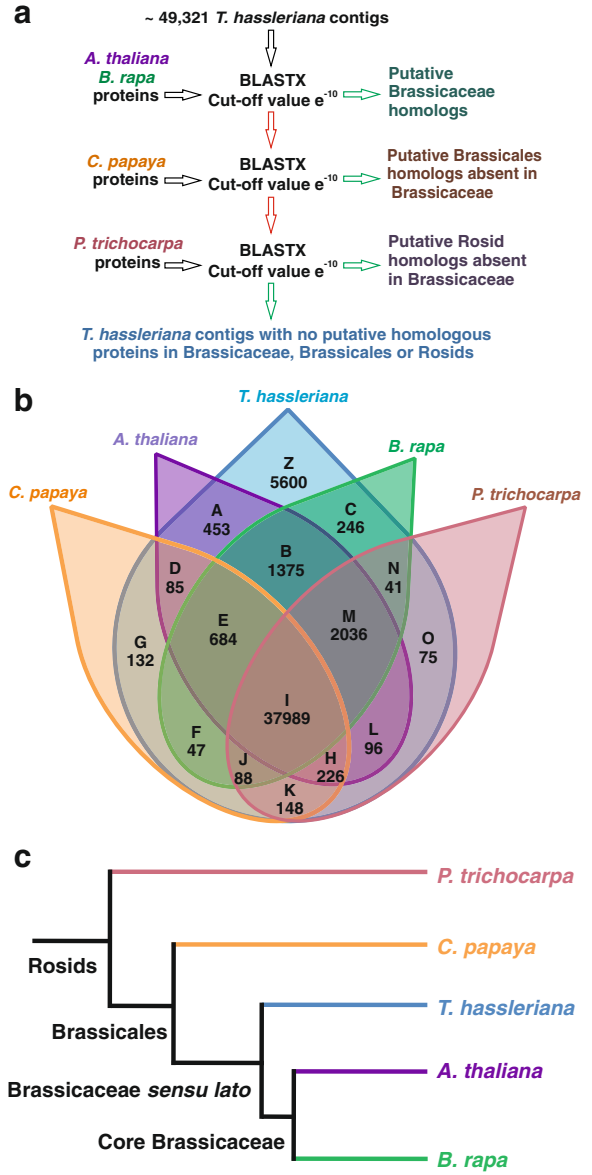


Figure 6 Identification and representation of lineage specific genes in selected rosid taxa. 6**a**: Strategy for the identification of lineage specific genes by BLASTX based sequence comparisons; 6**b**: Venn diagram representing lineage specific genes and genes shared between the different Rosid lineages, the letters in bold indicate the subset of sequences and the number indicates the number of genes/contigs included in the subset.; B: Genes shared between *A. thaliana, B. rapa* (core Brassicaceae) and *T. hassleriana*; E: Genes shared between *C. papaya, A. thaliana, B. rapa* and *T. hassleriana* (Brassicales); I: Genes shared by *A. thaliana, B. rapa, C. papaya, P. trichocarpa* and *T. hassleriana* (all Rosids); K: Genes shared by *C. papaya, P. trichocarpa* and *T. hassleriana* (potentially lost in Brassicaceae); O: Genes shared by *P. trichocarpa* and *T. hassleriana* (Lost in Brassicales); Z: contigs representing genes present only in *T. hassleriana* not shared with the other lineages. 6**c**: A schematic representation of the relationship of rosids used for the identification of Tarenaya specific genes.

from *P. trichocarpa, C. papaya, A. thaliana* and *B. rapa*. Of these contigs only 82 could be assigned to 353 GO terms, but a vast majority of the contigs could not be annotated attributing to no significant BLAST hits. A sequence length histogram for these contigs (Additional file 2: Figure S2) shows a bias towards shorter sequences when compared to the sequence length histogram of all contigs (Additional file 2: Figure S1b) suggesting that these were too short for proper annotation and/or may represent 5' and 3' UTR regions of transcripts. Another reason for the small number of annotated genes is because most of the current annotations are based on *A. thaliana*, *B. rapa* and *P. trichocarpa* and we already subtracted the sequences orthologous to them. The GO annotations for the *T. hassleriana* specific genes are the following: cellular process (26.98%), metabolic process (29.36%), response to stimulus (4.76%), biological regulation (4.7%), development (1.5%), cell proliferation (1.5%), reproduction (6.34%) and signaling processes (4.76%) (Additional file 1: Table S6).

Discussion

In this work we present the floral transcriptome sequence of *T. hassleriana*, which is a member of the Cleomaceae and thus a sister taxon to the Brassicaceae. The transcriptome was analyzed by rarefaction analysis and shown to be of sufficient depth to also identify rare transcripts. As normalization was not carried out, abundance of transcripts could be assessed in silico and compared to qRT-PCR data. The leaf transcriptome of *T. hassleriana* has been published earlier [10] allowing for comparison of transcript abundance between the leaf and the floral transcriptome. We also attempted to compare the expression of genes represented in the floral transcriptome with the expression of their respective *A. thaliana* orthologs based on presence/absence of expression in the microarray dataset [22] including all flower developmental stages. In addition, we are able to identify 5600 putative transcripts that are specific to the Cleomaceae and 684, which are shared only among the Brassicales *C. papaya*, *T. hassleriana, B. rapa*, and *A. thaliana*.

During assembly, annotation and analysis of the reads obtained by 454 sequencing we observed several challenges. In our study we correlated the in silico floral transcriptome expression in *T. hassleriana* with the conventional qRT-PCR expression of arbitrarily chosen genes with low, moderate, and high expression levels for validation of the transcriptome sequencing expression data. For the in silico expression data we applied two approaches, one was to map the individual reads to the annotated A. thaliana CDS sequences (TSE1), with the advantage that expression data for putative *T. hassleriana* orthologs of *A. thaliana* genes can be generated without a prior genome sequence information of *T. hassleriana*, thus individual reads are not assembled into contigs. The second

approach was the de novo assembly of the reads into contigs that are then annotated and the reads are mapped on to these contigs (TSE2). Comparing both methods with qRT-PCR data, the TSE1 approach clearly matches better than the TSE2 approach. One reason for this finding is the presence of chimeric contigs composed of more than one gene, in such cases reads to multiple genes are mapped onto the same contig flaring up the expression. This problem is avoided in TSE1 when the reads are mapped onto orthologous sequences in *A. thaliana*. Another reason for the disparity between the qRT-PCR expression and TSE2 is due to assembly of contigs with additional non coding nucleotide sequences. This phenomenon was observed in the case of *RBCS1A* amongst the genes analyzed. The assembled contig was 838 nucleotides long whereas the coding sequence of this gene is ~540 nucleotides across many plant lineages. The additional 295 nucleotides at the 3' end could represent the 3' UTR nonetheless reads would be mapped to such sequences leading to an overestimation of expression. The third reason for the differences in the expression between qRT-PCR and TSE2 may be the length of the assembled contig versus transcript size, as larger transcripts are fragmented prior cDNA library preparation. For TSE1, qRT-PCR expression data were normalized to *A. thaliana* CDS lengths and in case of TSE2 to *T. hassleriana* contig length. The necessity of normalization was seen in case of gene *SPL7*. The *SPL7* contig length was 629 nucleotides whereas the coding sequence of *SPL7* in *A. thaliana* is 2406 nucleotides, thus when the expression is normalized for length of the contig it leads to a much higher expression than when normalized to the length of the *A. thaliana* ortholog. In case where the contig length matched to the coding sequence length and when the contig had very low or almost no unknown sequences incorporated the qRT-PCR expression matched very well to the TSE2 as was observed in case of *TT4*.

However, since *T. hassleriana* and *A. thaliana* have, independent α-WGDs, the retention and loss of gene copies following the duplication will be different. By mapping the *T. hassleriana* reads onto the *A. thaliana* orthologs identical sets of orthologous gene copies are assumed for both species leading to over or underestimation of transcript abundance. These illustrated pitfalls for calculating gene expression from RNAseq experiments without the availability of a high-quality reference transcriptome or genome require thorough independent validation of gene expression data.

This work was initiated as a primer to identify genes that may contribute to the morphological differences between the *T. hassleriana* and the *A. thaliana* flower. Our focus was mainly on coloration, flowering time, and floral organ size as these are traits that show obvious differences between the two species.

T. hassleriana petals show a deep pink coloration which, when the flower opens, fades into light pink after a few days of exposure to the sun. While in most species only the epidermal petal cell layer is pigmented, *T. hassleriana* also has pigmented mesophyll cells [14]; suggesting an expansion of the anthocyanin regulation and biosynthesis pathway from petal to mesophyll cells. The pink pigments found in *T. hassleriana* flowers are acetylated cyanidin diglucoside (sophorosyl)-5-glucosides and acetylated pelargonidin sophorosyl-5-glucosides [14]. All genes required for the synthesis of pelargonidin-3-glucoside and cyanidin-3-glucoside are present in the flower transcriptome. Genes encoding proteins required for the early steps of anthocyanin up to the flavonoid myricetin are also found expressed in the leaf transcriptome, while the genes participating in later steps such as *DFR* and *LDOX* are restricted to the flower. These two enzymes are also not expressed in *A. thaliana* flowers but during seed development and late stages of leaf senescence [32].

Transcription factors of the MYB, bHLH, and WD40 families regulate the expression of anthocyanin biosynthesis genes in *A. thaliana* and Zea mays. While early biosynthesis genes, and their regulators such as *AtMYB11*, *AtMYB12*, and *AtMYB111* are involved in the production of flavonols, late biosynthesis genes and their regulators are required for the synthesis of anthocyanins from flavonols [37] and references therein. While the putative T. hassleriana orthologs of *AtMYB11* and *AtMYB12* are hardly expressed in the flower transcriptome, the putative *AtMYB111* ortholog shows very strong and flower specific expression suggesting a more prominent role for this gene in the regulation of early biosynthesis genes than for the putative orthologs of *AtMYB11* and *AtMYB12*. Orthologs of the regulators of late anthocyanin biosynthesis in *A. thaliana AtTTG1* (WD40 family member), *AtTT8, AtGL3, AtEGL3* (all bHLH family members) and *AtPAP2* (MYB family member) are also found expressed in the *T. hassleriana* flowers. The *T. hassleriana* orthologs of *A. thaliana* genes *AtTTG1, AtTT8, AtGL3, AtEGL3*, and *AtPAP2* forming the late anthocyanin biosynthesis regulatory complex show an approximately similar transcript abundance suggesting that they may function in a complex similar to the one in *A. thaliana*, only with an expression domain expanded to the floral organs.

A. thaliana late regulators are mainly expressed in senescing leaves and during seed development (Table 2), but most likely, their expression domain in *T. hassleriana* has expanded into the flower leading to the pink coloration of the floral organs. A similar situation is found in petunia, where, among others genes *AN1, AN11, AN2* and *AN4* form complexes similar to that in *A. thaliana* to regulate anthocyanin biosynthesis in the flower [38].

T. hassleriana has, unlike *A. thaliana*, large oval shaped petals, and indeed orthologs of genes involved in limiting

growth of floral organs were found to be hardly expressed in the *T. hassleriana* floral transcriptome. *BIG BROTHER*, encoding for a E3 ubiquitin-ligase represses cell proliferation in all *A. thaliana* proliferating tissues and is expressed strongly and uniformly in all developmental stages of the flower independently of other pathways while being a direct target of the petal organ identity gene *AP3* [39-41]. In the *T. hassleriana* floral transcriptome it has a very low expression of 9 RPKM, suggesting that this may be a reasonable candidate to account for the differences in petal size between the two species.

The particular *T. hassleriana* hybrid used in this study is sterile even though it produces all the floral organ whorls in the right number and position. However, even though the anthers developed, they did not produce any pollen and also did not dehisce rendering the plants male sterile. While orthologs of *A. thaliana* regulators of anther development were expressed in the the *T. hassleriana* flower, no expression of *ROXY1* and *ROXY2* was detected. These two genes redundantly control the anther lobe and pollen mother cell differentiation downstream of *SPL* in *A. thaliana* [34]. Moreover, only very low expression (6 RPKM) was observed for the *T. hassleriana* ortholog of *DYT1* which acts directly downstream of the *ROXY* genes. The phenotype of the *T. hassleriana* anthers also resembles the *roxy1 roxy2* double mutant anther phenotype in *A. thaliana*, suggesting that our *T. hassleriana* hybrid may lack functional *ROXY* genes leading to male sterility. We corroborated this observation by qRT-PCR expression data which not only detected very low *ROXY1* expression (the only *ROXY* ortholog in *T. hassleriana* genome [35]) in the mid developmental stage but also showed the de-regulation of expression of the upstream and downstream genes throughout bud development which may provide a cause for the male sterility.

T. hassleriana is perpetually flowering and a sharp transition to flowering as in *A. thaliana* cannot be observed. Several genes involved in flowering time regulation in *A. thaliana* are differently regulated in leaves and flowers and we compared the expression of their orthologs in the flower and leaf transcriptomes. FRI is a protein involved in activating transcription via chromatin remodeling of the central floral repressor *FLC* in *A. thaliana* [42] and is expressed rather uniformly throughout the plant. However, in *T. hassleriana*, *FRI* ortholog expression is not found in leaves. This may suggest a different mechanism for *FLC* ortholog activation in *T. hassleriana* leaves, as *FLC* is expressed there without the presence of *FRI*.

Interestingly; the expression of two more genes most likely involved in the change from vegetative to reproductive phase in *T. hassleriana* is different from *A. thaliana*. The *A. thaliana* gene *SMZ* is expressed in young seedlings, during floral transition and seed maturation [43] unlike its *T. hassleriana* ortholog which is expressed in flowers and developing buds. Possibly, the *T. hassleriana SMZ* has function different from its *A. thaliana* ortholog, which is a rather strong repressor of flowering.

Another candidate gene in the group of flowering time regulators that are differentially regulated in *A. thaliana* and *T. hassleriana* is *ATC*. In *A. thaliana*, *ATC* is strongly expressed in the root and a small fraction (1-5%) of its mRNAs moves a long distance to the plant's apex [44]. Notably, we find a significant amount of reads in flower tissue, too many to attribute them to long distance RNA transport. More likely, the *ATC* homolog is expressed in *T. hassleriana* floral tissue and may be transported throughout the plant to enable the vegetative shoots to first reach sufficient size to start flowering.

Conclusions

Taken together we conclude from our expression data that a number of floral regulators show expression distinct from that in *A. thaliana* suggesting that differences in life history traits such as perpetual flowering and pigmentation may be regulated by similar components of regulatory networks in *A. thaliana* and *T. hassleriana* that are highly conserved in coding sequence but expressed in a different way in the two species, suggesting that modifications in expression pattern account for a large part of the diversity in flowers and plant life history traits.

Additional files

Additional file 1: Table S1. 454 sequencing statistics. **Table S2.** Sequences of the oligonucleotides used for the qRT-PCR. **Table S3.** List of gene names along with their abbreviations and AGI identifiers. **Table S4.** P- value calculations using one way ANOVA for analyzing the statistical significance of difference between expression values by qRT-PCR and Transcriptome Sequencing Expression (TSE). **Table S5.** GO annotation of putatively homologous gene pairs expressed in the *T. hassleriana* floral transcriptome but not expressed in the *A. thaliana* floral transcriptome. **Table S6.** GO annotation of *T. hassleriana* specific sequences not found in *A. thaliana*, *B. rapa*, *C. papaya*, and *P. trichopoda* using Blast2GO® with BLASTX searches.

Additional file 2: Figure S1. 454 sequencing statistics. **Figure S2.** Read length distribution of *T. hassleriana* lineage specific contigs without any GO annotation. **Figure S3.** Correlation plot of TSE1 expression by RNA seq and qRT-PCR gene expression.

Additional file 3: *T. hassleriana* floral transcriptome gene expression.

Abbreviations

WGD: Whole genome duplication; At-α: *A. thaliana* alpha WGD; At-β: *A. thaliana* beta WGD; Cq: Quantification cycle; EST: Expressed sequence tag; MYA: Million years ago; RPKM: Reads per kilobase gene model per mappable million; SDR: Standard dose response; Th-α: *T. hassleriana* alpha WGD; TSE: Transcriptome sequencing expression; TSE1: TSE mapped to *A. thaliana* CDS sequences; TSE2: TSE mapped to *T. hassleriana* contigs.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

A. Bhide performed the molecular biology experiments, comparative and in silico expression analysis and drafted the manuscript. SS carried out RNAseq

assembly, annotation, in silico expression analysis, lineage specific gene discovery. MR performed the rarefaction analysis, A. Becker coordinated and designed the study, and helped to draft the manuscript. A. Becker and APMW conceived the study. All authors helped to improve the manuscript, read, and approved of the final manuscript.

Acknowledgments
We thank the University of Bremen and the Justus-Liebig University for funding of A. Bhide's position, work in A. Becker's lab is largely funded by the DFG (German Research Foundation). A.P.M.W. acknowledges funding by DFG Priority Program 1529 (Adaptomics). Library preparation and subsequent 454 sequencing was performed by René Deenen at the Biomedical Research Center (BMFZ) of the Heinrich-Heine-University Düsseldorf.

Author details
[1]Justus-Liebig-Universität Gießen, Institute of Botany, Plant Development Group, Heinrich-Buff-Ring 38, 35392 Gießen, Germany. [2]Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS) Heinrich-Heine-University, Universitätsstr. 1, D-40225 Düsseldorf, Germany. [3]Department of Biology and Chemistry, University of Bremen, Leobener Str. NW2, D- 28359 Bremen, Germany.

Received: 8 August 2013 Accepted: 6 February 2014
Published: 19 February 2014

References

1. Iltis HH, Cochrane TS: **Studies in the Cleomaceae V: a new genus and ten new combinations for the flora of North America.** *J Bot Nomencl* 2007, **17:**447–451.
2. Inda LA, Torrecilla P, Catalán P, Ruiz-Zapata T: **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 Syst Evol* 2008, **274:**111–126.
3. Hall JC, Sytsma KJ, Iltis HH: **Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data.** *Am J Bot* 2002, **89:**1826–1842.
4. Kers LE: **Capparaceae**. In *Flowering Plants Dicotyledons.* Edited by Kubitzki K, Bayer C. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003:36–56.
5. Rodman J, Soltis P, Soltis D, Sytsma K, Karol K: **Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies.** *Am J Bot* 1998, **85:**997.
6. Schranz ME: **Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae.** *Plant Cell Online* 2006, **18:**1152–1165.
7. Couvreur TLP, Franzke A, Al-Shehbaz IA, Bakker FT, Koch MA, Mummenhoff K: **Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae).** *Mol Biol Evol* 2009, **27:**55–71.
8. Barker MS, Vogel H, Schranz ME: **Paleopolyploidy in the Brassicales: analyses of the cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales.** *Genome Biol Evol* 2009, **1:**391–399.
9. Koteyeva NK, Voznesenskaya EV, Roalson EH, Edwards GE: **Diversity in forms of C4 in the genus Cleome (Cleomaceae).** *Ann Bot* 2011, **107:**269–283.
10. Brautigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Mass J, Lercher MJ, Westhoff P, Hibberd JM, Weber APM: **An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species.** *Plant Physiol* 2011, **155:**142–156.
11. Stout AB: *Alternation of Sexes and Intermittent Production of Fruit in the Spider Flower (cleome Spinosa).* New York: New York Botanical Garden; 1923 [*Contributions from the New York Botanical Garden*].
12. Patchell MJ, Bolton MC, Mankowski P, Hall JC: **Comparative floral development in Cleomaceae reveals two distinct pathways leading to monosymmetry.** *Int J Plant Sci* 2011, **172:**352–365.
13. Randall RP: *A Global Compendium of Weeds.* 2nd edition. 2012.
14. Nozzolillo C, Amiguet VT, Bily AC, Harris CS, Saleem A, Andersen, Oyvind M, Jordheim M: **Novel aspects of the flowers and floral pigmentation of two Cleome species (Cleomaceae), C. hassleriana and C. serrulata.** *Biochem Syst Ecol* 2010, **38:**361–369.
15. Chomczynski P, Sacchi N: **Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.** *Anal Biochem* 1987, **162:**156–159.
16. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656–664.
17. Schliesky S, Gowik U, Weber A, Andreas PM, Bräutigam A: **RNA-Seq assembly - are we there yet?** *Front Plant Sci* 2012, **3**.
18. R Core Team: *R: A Language and Environment for Statistical Computing.* Vienna, Austria; 2013.
19. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, Wittwer CT: **The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.** *Clin Chem* 2009, **55:**611–622.
20. Marshall OJ: **PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR.** *Bioinformatics* 2004, **20:**2471–2472.
21. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative 5PCR6 and the 2-\textgreek5DD6CT method.** *Methods* 2001, **25:**402–408.
22. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU: **A gene expression map of Arabidopsis thaliana development.** *Nat Genet* 2005, **37:**501–506.
23. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674–3676.
24. Tipper JC: **Rarefaction and rarefiction; the use and abuse of a method in paleoecology.** *Paleobiology* 1979, **5:**423–434.
25. Hale MC, McCormick CR, Jackson JR, DeWoody JA: **Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (Acipenser fulvescens): the relative merits of normalization and rarefaction in gene discovery.** *BMC Genomics* 2009, **10:**203.
26. Lohmann JU, Weigel D: **Building beauty: the genetic control of floral patterning.** *Dev Cell* 2002, **2:**135–142.
27. Broun P: **Transcriptional control of flavonoid biosynthesis: a complex network of conserved regulators involved in multiple aspects of differentiation in Arabidopsis.** *Curr Opin Plant Biol* 2005, **8:**272–279.
28. Weiss J, Delgado-Benarroch L, Egea-Cortines M: **Genetic control of floral size and proportions.** *Int J Dev Biol* 2005, **49:**513–525.
29. Zhang X, Feng B, Zhang Q, Zhang D, Altman N, Ma H: **Genome-wide expression profiling and identification of gene activities during early flower development in Arabidopsis.** *Plant Mol Biol* 2005, **58:**401–419.
30. Gou J, Felippes FF, Liu C, Weigel D, Wang J: **Negative regulation of anthocyanin biosynthesis in Arabidopsis by a miR156-targeted SPL transcription factor.** *Plant Cell Online* 2011, **23:**1512–1522.
31. Logacheva MD, Kasianov AS, Vinogradov DV, Samigullin TH, Gelfand MS, Makeev VJ, Penin AA: **De novo sequencing and characterization of floral transcriptome in two species of buckwheat (Fagopyrum).** *BMC Genomics* 2011, **12:**30.
32. Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ, Baxter I: **An "Electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets.** *PloS One* 2007, **2:**e718.
33. Ni W: **Regulation of flower development in Arabidopsis by SCF complexes.** *Plant Physiol* 2004, **134:**1574–1585.
34. Xing S, Zachgo S: **ROXY1 and ROXY2, two Arabidopsis glutaredoxin genes, are required for anther development.** *Plant J* 2008, **53:**790–801.
35. Cheng S, van den Bergh E, Zeng P, Zhong X, Xu J, Liu X, Hofberger J, de Bruijn S, Bhide AS, Kuelahoglu C, Bian C, Chen J, Fan G, Kaufmann K, Hall JC, Becker A, Brautigam A, Weber APM, Shi C, Zheng Z, Li W, Lv M, Tao Y, Wang J, Zou H, Quan Z, Hibberd JM, Zhang G, Zhu X, Xu X, *et al*: **The tarenaya hassleriana genome provides insight into reproductive trait and genome evolution of crucifers.** *Plant Cell* 2013, **25:**2813–2830.
36. Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S: **Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana.** *Proc Natl Acad Sci* 2010, **107:**18724–18728.
37. Petroni K, Tonelli C: **Recent advances on the regulation of anthocyanin synthesis in reproductive organs.** *Plant Sci* 2011, **181:**219–229.
38. Quattrocchio F: **PH4 of petunia is an R2R3 MYB protein that activates vacuolar acidification through interactions with basic-helix-loop-helix transcription factors of the anthocyanin pathway.** *Plant Cell Online* 2006, **18:**1274–1291.
39. Disch S, Anastasiou E, Sharma VK, Laux T, Fletcher JC, Lenhard M: **The E3 ubiquitin ligase BIG BROTHER controls Arabidopsis organ size in a dosage-dependent manner.** *Curr Biol* 2006, **16:**272–279.
40. Wuest SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, Hanczaryk K, Lohan AJ, Loftus B, Graciet E, Wellmer F: **Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA.** *Proc Natl Acad Sci* 2012, **109:**13452–13457.

41. Krizek BA, Anderson JT: **Control of flower size.** *J Exp Bot* 2013, **64:**1427–1437.
42. Choi K, Kim J, Hwang H, Kim S, Park C, Kim SY, Lee I: **The FRIGIDA complex activates transcription of FLC, a strong flowering repressor in Arabidopsis, by recruiting chromatin modification factors.** *Plant Cell Online* 2011, **23:**289–303.
43. Mathieu J, Yant LJ, Mürdter F, Küttner F, Schmid M, Dean C: **Repression of flowering by the miR172 target SMZ.** *PLoS Biol* 2009, **7:**e1000148.
44. Huang N, Jane W, Chen J, Yu T: **Arabidopsis thaliana CENTRORADIALIS homologue (ATC) acts systemically to inhibit floral initiation in Arabidopsis.** *Plant J* 2012, **72:**175–184.

doi:10.1186/1471-2164-15-140
Cite this article as: Bhide *et al.*: Analysis of the floral transcriptome of *Tarenaya hassleriana* (Cleomaceae), a member of the sister group to the Brassicaceae: towards understanding the base of morphological diversity in Brassicales. *BMC Genomics* 2014 **15**:140.

4.1.7 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 Advance Access published March 18, 2014

Journal of Experimental Botany
doi:10.1093/jxb/eru100
This paper is available online free of all access charges (see http://jxb.oxfordjournals.org/open_access.html for further details)

RESEARCH PAPER

Towards an integrative model of C_4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C_4 species

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

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

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

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

Abstract

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

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

Introduction

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

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

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

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

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

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

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

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

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

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

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

Materials and methods

Plant growth and harvesting

Megathyrsus maximus (Collection of the Botanical Garden Düsseldorf) and *D. clandestinum* (grown from seed obtained from B&T World Seeds, Perpignan, France) plants were grown with 16 h of light at 24 °C. *Dichanthelium clandestinum* was maintained vegetatively. Harvesting was scheduled to the eight-leaf stage, which was 3–5 weeks after germination or tiller initiation. In the middle of the light period, the third leaf from the top—the third youngest—was sampled in three replicates for sequencing (one for 454 and two for Illumina sequencing) and five replicates for enzyme activities, and quenched in liquid nitrogen immediately after cutting. Pools of 20 plants per sample were harvested.

Enzyme activities

C_4 decarboxylation enzymes were extracted from frozen, ground leaves using 1 ml of buffer [25 mM TRIS-HCl (pH 7.5), 1 mM $MgSO_4$, 1 mM EDTA, 5 mM dithiothreitol (DTT), 0.2 mM phenylmethylsulphonyl fluoride (PMSF), and 10% (v/v) glycerol] per 10 mg of leaf powder. After desalting using NAP-5 size exclusion columns, enzyme activities of PEP-CK (Walker *et al.*, 1995), NAD-ME, and NADP-ME (Hatch and Mau, 1977) were measured photometrically based on the absorption change of NAD(P)H at 340 nm.

CO_2 assimilation rates and isotope discrimination

For three replicates of both species, the net leaf photosynthetic assimilation rate (A) was measured using a Li-Cor LI-6400XT infrared gas exchange analyser (LI-COR Inc., Lincoln, NE, USA). CO_2-dependent assimilation curves (A–C_i) were measured at 1500 µmol m^{-2} s^{-1} constant light. Light-dependent assimilation curves were measured at a constant external CO_2 concentration of 400 ppm.

For ^{13}C isotope discrimination, leaf powder was dried and analysed using the isotope ratio mass spectrometer IsoPrime 100 (IsoPrime Ltd, Cheadle, Manchester, UK). Results were expressed as relative values compared with the international standard (Vienna-PeeDee Belemnite).

RNA extraction and sequencing

Isolation of total RNA from ground tissue of *M. maximus* was performed using a guanidium thiocyanate extraction followed by an ethanol and a lithium chloride precipitation, as described by Chomczynski and Sacchi (1987). Extraction of total RNA from *D. clandestinum* was performed using a TRIS-borate buffer to cope with large amounts of polysaccharides, as described by Westhoff and Herrmann (1988). mRNA for 454 library preparation was enriched by using Qiagen Oligotex poly(A)-binding silicone beads and further prepared for sequencing as described in Weber *et al.* (2007). For Illumina sequencing two replicates of total RNA were used per sample. Library preparation and sequencing were carried out according to the manufacturer's suggestions by the local NGS facility (BMFZ, Biologisch-Medizinisches Forschungszentrum, Düsseldorf), using Roche Titanium chemicals for 454 and the TruSeq library kit for Illumina HiSeq 2000. Long and short read raw data were submitted to the short read archive (SUB440021, *D. clandestinum*; SUB439950, *M. maximus*).

Sequence assembly and expression statistics

De novo assembly was done using CAP3 (Huang and Madan, 1999) using default parameters on cleaned 454 reads. Reads were cleaned by trimming low quality ends, discarding reads of overall minor quality, and removal of exact duplicates using scripts of the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) as described in Schliesky *et al.* (2012). Contigs were annotated by BLAST best hit mapping to *S. italica* (v164) representative coding sequences. Quantitative expression was determined by mapping of all Illumina reads against *S. italica* representative coding sequences (v164) using BLAT (Kent, 2002) and counting the best hit for each read. Zero counts were treated as true 0. Expression was normalized to reads per mappable million and per kilobase (rpkm) *Setaria* CDS. Eight rpkm were chosen as the threshold of expression to discriminate background transcription. Differential expression was determined by DESeq (Anders and Huber, 2010), a negative binomial test, in R (R Development Core Team, 2012). A significance threshold of 0.05 was applied after Bonferroni correction for multiple hypothesis testing and is reported in Supplementary Table S3 available at *JXB* online. For all single genes mentioned in the text, changes in expression were confirmed using the 454 data set which was also mapped across species to *S. italica* as described in Bräutigam *et al.* (2011) (Supplementary Table S3). Pathway enrichment was determined by Benjamini–Hochberg correction (Benjamini and Hochberg, 1995). Fisher's exact test was used to test for over-/under-representation of MapMan categories.

Meta comparison of functional categories

Expression data for *B. dystachyon*, *S. bicolor*, and *O. sativa* were previously published by Davidson *et al.* (2012). Transcript sequences for mature *Z. mays* leaves (+4 cm sample) were obtained from the short read archive SRA012297 (Li *et al.*, 2010) and mapped to *S. italica* representative coding sequences. Expression data for five Flaveria species were taken from Gowik *et al.* (2011). Expression data for *Cleome gynandra* (C_4) and *Tarenaya hassleriana* (C_3) were taken from Bräutigam *et al.* (2011). The samples were produced in

different laboratories and with different sequencing technologies. Only the presence of C_4-related traits was interpreted, as absence calls may be due to inconsistent sampling with regard to leaf developmental state, time of day, and other variables.

EC (enzyme classifiers; Schomburg *et al.*, 2013) and Pfam (protein family; Sonnhammer *et al.*, 1997) annotations were added to the two reference transcriptomes, *S. italica* CDS (v164) and *Arabidopsis thaliana* CDS (TAIR10). Reduction of data complexity to functional classifiers was achieved by summing up all expression values mapping to the same EC or Pfam. Venn diagram sets were built through logical operators; that is, expression is higher/lower in all C_4 versus C_3 comparisons (see also Supplementary Table S2 at *JXB* online). Comparison pairs were chosen according to the sequencing method and experimenter: *M. maximus* versus *D. clandestinum* (this study), *S. bicolor* versus *O. sativa* and versus *B. dystachyon* (all from Davidson *et al.* 2012); *Z. mays* (Li *et al.* 2011) and *S. italica* (Bennetzen *et al.* 2012) were orphan data sets as no comparison partner was sequenced with the same technology and both were compared against *B. dystachion* as the C_3 reference. The dicots were compared as previously published (Bräutigam *et al.* 2011; Gowik *et al.*, 2011).

Leaf cross-sections for confocal microscopy

Fresh mature leaves (upper third of the leaf) of *M. maximus* and *D. clandestinum* were cut transversally and fixed in PBST [1× PBS buffer (137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.4 mM KH_2PO_4); 1% (v/v) Tween-20; 3% (v/v) glutaraldehyde] overnight at room temperature. Leaf cross-sections were stained with 0.1% 4',6-diamidino-2-phenylindole (DAPI) solution in phosphate-buffered saline (PBS) for 30 min. Subsequently, cross-sections were analysed with an LSM 780 (Zeiss) confocal microscope with a ×40 objective. Z-stack images were processed with LSM Zeiss software to produce maximum intensity overlay images.

Results

D. clandestinum *is well suited for a C_3 comparison with* M. maximus

The PACMAD clade of the grasses is exceptionally rich in C_4 plants (Christin *et al.*, 2013) to the point that it is difficult to identify and cultivate closely related C_3 species for comparative analyses. To confirm that *D. clandestinum* is a *bona fide* C_3 plant and to confirm the biochemical subtype of the C_4 plant *M. maximus*, different parameters were tested. The measured enzyme activities, stable isotopic carbon discrimination, A–C_i curves, and light curves indicated that *D. clandestinum* indeed represents a C_3 plant (Fig. 1). *Megathyrsus maximus* has high NAD-ME and PEP-CK enzyme activities as compared with *D. clandestinum*, but comparable activities of the NADP-ME decarboxylation enzyme (Fig. 1A). *Dichanthelium clandestinum* discriminates against ^{13}C at a $\delta^{13}C$ ratio of –30‰, while *M. maximus* shows C_4 typical

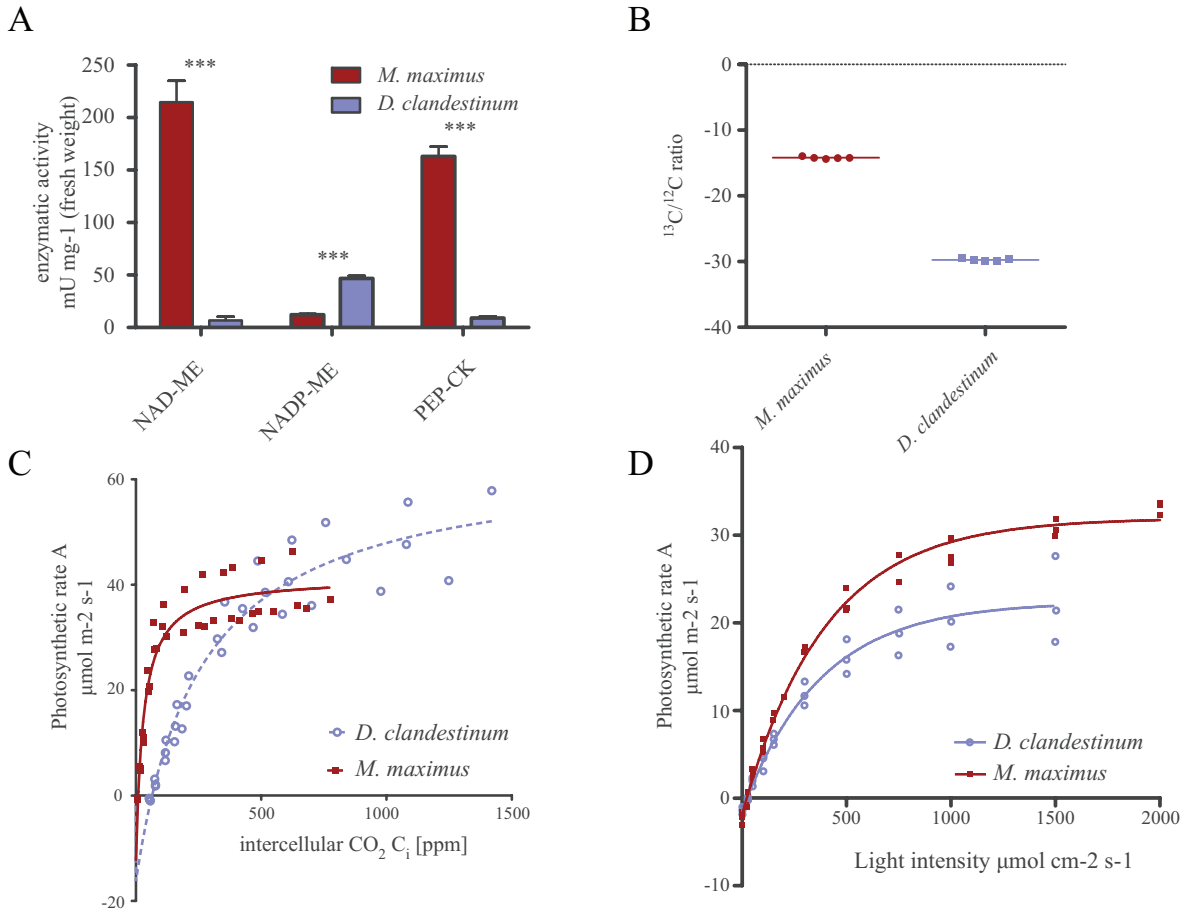


Fig. 1. Physiological characterization of *Megathyrsus maximus* and *Dicanthelium clandestinum*. Activity of the decarboxylation enzymes in *M maximus* and *D. clandestinum* (A); $^{13}C/^{12}C$ stable isotope ratio (B); A–C_i curves at 1500 µE (C); and light curves at 400 ppm CO_2 (D). ***P<0.001. (This figure is available in colour at *JXB* online.)

relaxation of carbon isotope discrimination with a $\delta^{13}C$ ratio of –13‰ (Fig. 1B). The A–C_i curve of *M. maximus* shows a low CO_2 compensation point of 9 ppm and saturation of the net carbon fixation rate at 41 µmol m^{-2} s^{-1}. The A–C_i curve of *D. clandestinum* plants grown alongside *M. maximus* indicates a CO_2 compensation point of 65 ppm and does not saturate even with high CO_2 concentrations, as is typical for a C_3 plant (Fig. 1C). The light response curves of CO_2 assimilation show similar rates for both types of plants at very low light intensities, with *M. maximus* continuously outgaining *D. clandestinum* as light increases. Thus, *M. maximus* has slightly higher quantum efficiency and saturates at a higher light intensity compared with *D. clandestinum* (Fig. 1D). In summary, the physiological data indicate that *D. clandestinum* is a suitable comparison partner for *M. maximus* due to its phylogenetic proximity and physiological characteristics typical of C_3 plants.

Quantitative and qualitative transcriptome information

The transcriptomes of both grass species were determined by RNA-Seq using two complementary technologies to gain quantitative gene expression information and provide a sequence resource optimized for C_4 unigene assembly. RNA-Seq libraries from two biological replicates of *M. maximus* and two biological replicates of *D. clandestinum* were sequenced with Illumina HiSeq2000 technology and yielded upwards of 53 million reads per replicate, of which >48 million reads were of high quality (Table 1). Reads were mapped cross-species to a closely related reference sequence database derived from the *S. italica* genome (Bennetzen *et al.*, 2012) and between 66% and 74% of reads matched the reference sequence database (Table 1). In the reference sequence database, 13 043 genes were matched with >8 rpkm, of which 792 were detected as differentially up-regulated in C_4 and 376 were detected as differentially down-regulated in C_4 (Table 1). In addition, 1.1 million and 0.9 million 454/Roche Titanium reads were generated and assembled for *M. maximus* and *D. clandestinum*, respectively, and mapped onto *S. italica* as a quality control for the Illumina mapping. The majority of gene expression differences followed similar trends in the 454 mapping or were not detected among the 454 reads; only 12 genes displayed inversely regulated patterns with the different sequencing technologies. Reads were filtered and trimmed based on a Phred score of 30 and assembled with CAP3 (Huang and Madan, 1999) to provide a reliable database of unigenes. C_4 cycle genes were covered by unigenes with full length (Supplementary Table S1 at *JXB* online). About 40 000 unigenes were generated for each species (Table 1).

Genes commonly up- or down-regulated in all C_4 decarboxylation types

Comparative RNA-Seq data for NADP-ME species versus C_3 sister species (Gowik *et al.*, 2011) and for NAD-ME species versus C_3 sister species (Bräutigam *et al.*, 2011), three RNA-Seq data sets for *S. bicolor*, *O. sativa*, and *B. dystachyon* from one comparative experiment (Davidson *et al.*, 2012), as well as orphan RNA-Seq data sets for two PACMAD NADP-ME grasses, *Z. mays* (Li *et al.*, 2011) and *S. italica* (Li *et al.*, 2011), are publicly available. By combining the public data with data from this study, the up- and down-regulated core C_4 genes altered in all C_4 species were identified.

Gene by gene comparisons may be limited between different C_3–C_4 species comparison pairs since for known C_4 genes, most notably PEPC, recruitment of paralogous genes has already been demonstrated (Westhoff and Gowik, 2004; Besnard *et al.*, 2009; Christin and Besnard, 2009). In addition, a function may be distributed among multiple genes, each of which singly does not appear changed. To overcome the inherent limitations of orthologous gene pair comparisons when analysing multiple species pairs, reads were summed to categories which represent a function rather than a particular gene. Enzymes were identified in the reference species *A. thaliana* and *S. italica* on the basis of EC numbers which cover ~5000 different enzymes (Schomburg *et al.*, 2013), of which 1073 are present in the references, and reads for each gene were summed based on the EC number. For example, reads mapping to different isogenes encoding PEPC are no longer represented by the gene identifier but they have been collapsed onto the EC number representing PEPC function (4.1.1.31). All proteins in both reference species were also assigned to their protein family on the basis of Pfam domains (Sonnhammer *et al.*, 1997), of which 4073 unique combinations are present in the references, and reads for each gene were summed based on the Pfam domain combination. Consequently, PEPC is no longer represented by a gene identifier but its function is represented by its Pfam domain combination pf00311. The functions up-regulated or down-regulated in all C_4 species compared with their related

Table 1. Sequencing, mapping, and assembly statistics for *Megathyrus maximus* and *Dicanthelium clandestinum*

Read mapping	*Megathyrus maximus* 1	*Megathyrus maximus* 2	*Dicanthelium clandestinum* 1	*Dicanthelium clandestinum* 2
No. of Illumina reads	61 703 536	56 780 148	53 079 709	56 765 538
No. of cleaned reads	56 470 008	52 282 627	48 160 148	50 328 269
Mappable reads (%)	41 570 126 (73.6%)	38 848 638 (74.3%)	34 151 633 (70.9%)	33 311 704 (66.2%)
No. of 454 reads	1 152 766		971 065	
No. of contigs in assembly	39 565		40 320	
Setaria CDS with >8 rpkm	13 043			
Differentially up-regulated	792			
Differentially down-regulated	376			

C_3 species and those limited to the two NAD-ME type based species were then analysed (Fig. 2A–D; Supplementary Table S2 at *JXB* online).

The functional analysis based on EC numbers indicated a consistent up-regulation of 16 functions in all C_4 comparisons. The C_4 enzymes with PPDK, PPase, AMK, PEPC, aspartate aminotransferase (AspAT), NADP-dependent malate dehydrogenase (NADP-MDH), and ME are up-regulated in all comparisons (Fig. 2A). In addition, one function related to starch synthesis, two functions related to sucrose synthesis, and six functions currently unlinked to C_4 were identified (Fig. 2A; Supplementary Table S2 at *JXB* online). Both NAD-ME species have 135 up-regulated functions in common, including PEP-CK, alanine aminotransferase (AlaAT), pyruvate dehydrogenase (PDH) kinase, and nine enzymes involved in purine synthesis and turnover (Fig. 2A). The 37 functions down-regulated in all C_4 comparisons include four of the Calvin–Benson (CBB) cycle and eight related to photorespiration (Fig. 2B). The down-regulated functions in both NAD-ME-type comparisons included aspartate kinase and aspartate oxidase, eight functions of pyrimidine synthesis, four of the CBB cycle, 11 of chlorophyll synthesis, and 16 of translation (Fig. 2B).

The functional analysis based on Pfam domain combinations showed 34 up-regulated functions in all C_4 species including PEPC, PPDK, phospho*enol*pyruvate phosphate translocator (PPT), and ME. Four photosystem-related functions, two functions related to starch synthesis, and one related to sucrose synthesis are also among those up-regulated (Fig. 2C). The 413 NAD-ME-type related up-regulated functions include PEP-CK, the pyruvate transporter (BASS2, Furumoto *et al.*, 2011), and the sodium:hydrogen antiporter (NHD; Furumoto *et al.*, 2011), all detected with high fold changes (Fig. 2C; Supplementary Table S2 at *JXB* online). Among the 38 down-regulated functions are the CBB cycle, photorespiration, and translation (Fig. 2D).

The analyses of C_4-related functions extend the known C_4 up-regulated traits to sucrose and starch synthesis and the C_4 down-regulated traits to the CBB cycle, photorespiratory functions, and translation. They also provide candidates for as yet unknown functions which may be C_4 related. The NAD-ME-type related functions include those that prevent the leakage of C_4 cycle metabolites into general metabolism.

The PEP-CK decarboxylation subtype is qualitatively similar to but quantitatively distinct from the NAD-ME

Given the blueprint of NAD-ME C_4 photosynthesis (Bräutigam *et al.*, 2011), it was tested whether the differentially regulated functions in the PEP-CK species are those already identified for the NAD-ME species. The occurrence of PEP-CK activity in species previously classified as NADP-ME and NAD-ME species and recent modelling efforts raised the question of whether the classification of PEP-CK as its own C_4 type is warranted (Wang *et al.*, 2014).

The C_4 genes were extracted from the complete data set (Supplementary Table S3 at *JXB* online) and compared with those of *C. gynandra* (Bräutigam *et al.*, 2011). *Megathyrsus maximus* and *C. gynandra* show elevated expression of enzymes and transporters known to be required for C_4 photosynthesis (Table 2). *Megathyrsus maximus* showed significantly increased transcripts encoding BASS2, NHD, PPDK, and PPT, which is similar to the dicotyledonous NAD-ME C_4 species *C. gynandra*. In comparison with a C_3 reference, the up-regulation of these transcripts was between 27-fold and 67.5-fold in *M. maximus* and between 15-fold and 226-fold

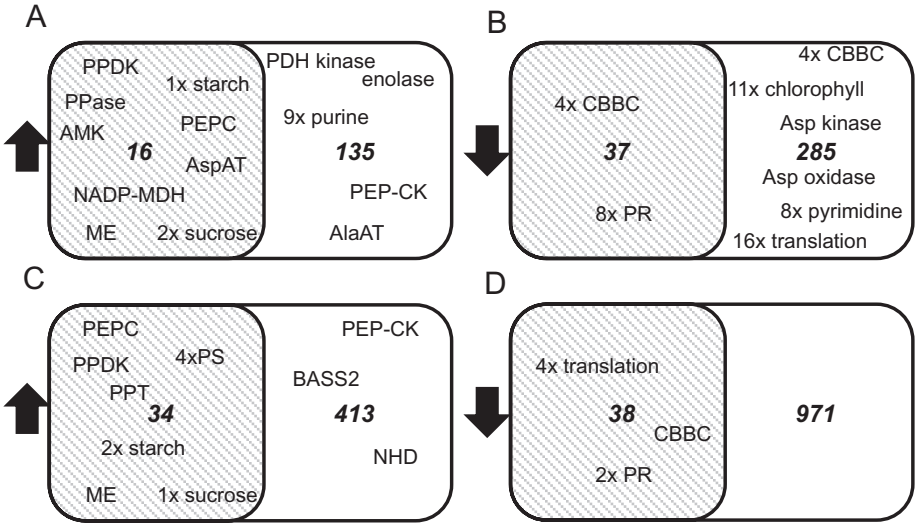


Fig. 2. Shared expression based on function in NAD-ME (white set) versus all C_4 species (grey set). Up- and down-regulated functions are based on expression of functions represented by enzyme classifiers (EC) (A, B) and by Pfam domain combinations (PDC) (C, D). PPDK, pyruvate phosphate dikinase; PPase, inorganic pyrosphate phosphorylase; AMK, adenosine monophosphate kinase; PEPC, phospho*enol*pyruvate carboxylase; AspAT, aspartate aminotransferase; MDH, malate dehydrogenase; ME, malic enzyme; PDH, pyruvate dehydrogenase; PEP-CK, phospho*enol*pyruvate carboxykinase; AlaAT, alanine aminotransferase; CBBC, Calvin–Benson–Bassham cycle; PR, photorespiration; Asp, aspartate; PPT, phospho*enol*pyruvate phosphate translocator; PS, photosynthesis; BASS2, pyruvate transporter; NHD sodium proton antiporter; all functions are listed in Supplementary Table S2 at *JXB* online. (This figure is available in colour at *JXB* online.)

Table 2. The expression of C_4 cycle genes of *Megathyrsus maximus* in comparison with *Dicanthelium clandestinum* and *Cleome gynandra*, and their location in *M. maximus*

Module	Gene name	*Setaria* ID	Function	Predicted location of translated protein	*M. maximus* expression (rpkm)	*D. clandestinum* expression (rpkm)	Fold- change	Significantly changed (DESeq, Bonferroni)	Fold change of function in *C. gynandra*
Regeneration	BASS2	Si001591m	Pyruvate sodium symport	Chloroplast	2797	69	40.5	Yes	87.3
	NHD	Si029362m	Sodium proton antiport	Chloroplast	838	31	27.0	Yes	**15.9**
	PPDK	Si021174m	Pyruvate→PEP	Chloroplast	13380	283	47.3	Yes	226.4
	PPa	Si017993m	Pyrophosphate→phosphate	Chloroplast	450.5	158.5	2.8	NS	3.2
	AMK	Si017707m	AMP→ADP	Chloroplast	985.5	114.5	8.6	NS	**8.9**
	PPT	Si013874m	PEP phosphate antiport	Chloroplast	405	6	67.5	Yes	15.0
Carboxylation	PEPC	Si005789m	PEP→OAA	Cytosol	18393	303.5	60.6	Yes	**77.6**
C_4 transfer acid	AspAT	Si001361m	Asp↔OAA	Cytosol	1273	79	16.1	Yes	2[a]
	GAP-DH	Si014034m	3-GPA→TP	Cytosol	4544	1538	3.0	NS	0.2
	MDH	Si036550m	Malate↔OAA	Cytosol	735	452	1.6	NS	0.44[a]
Decarboxylation	DIC	Si014081m	Malate phosphate antiport	mitochondrion	455	114	4.0	NS	519.0
NAD-ME	PIC	Si017569m	Phosphate proton symport	Mitochondrion	225	96	2.3	NS	2.5
	ME	Si000645m& Si034747m[b]	Malate→pyruvate	Mitochondrion	1299	230	5.6	NS	20.3
	Unknown/diffusion?		Pyruvate export						
Decarboxylation	PEP-CK	Si034404m	OAA→PEP	Cytosol	8819	99	89.5	Yes	8.6
PEP-CK	AAC	Si017474m	ATP ADP/P antiport	Mitochondrion	461	150	3.1	NS	0.4

Bold indicates use of a paralogous gene.
NS, non significant.
[a] A paralogue in a different compartment is up-regulated.
[b] Reads map to both malic enzymes

in *C. gynandra*. PPDK induction, however, was lower in *M. maximus* compared with *C. gynandra*, which might indicate increased regeneration of PEP by PEP-CK rather than PPDK. Both species also showed changes in AMK and PPase expression, but these were not expressed to a significantly higher extent in *M. maximus*. The NHD and AMK expressed at high levels are paralogous to the same proteins required for the C_4 cycle in the dicotyledonous plant (Table 2). The carboxylation enzyme PEPC was significantly up-regulated in both the dicot and the monocot, again using paralogues (Table 2). For the generation of the C_4 transfer acids malate and aspartate, only cytosolic AspAT was significantly up-regulated in *M. maximus*, while no up-regulation of the cytosolic isozyme was observed in *C. gynandra*. Cytosolic targeting was determined by localization prediction of the full-length protein of *M. maximus* (Supplementary Table S4 at *JXB* online). The most abundant transcript encoding MDH also encoded a cytosolic isozyme, suggesting use of the NAD-MDH form (Supplementary Table S4).

Two different decarboxylation modules using NAD-ME and PEP-CK, respectively, are active in the plants (Fig. 1A). In *M. maximus*, neither the transport protein DIC, responsible for antiport of malate into mitochondria against phosphate (Palmieri *et al.*, 2008), and PIC, responsible for symport of phosphate and protons (Pratt *et al.*, 1991; Hamel *et al.*, 2004), nor the decarboxylation enzyme NAD-ME were significantly changed, although all were up-regulated between 2.3- and 5.6-fold (Table 2). This is in stark contrast to the up-regulation detected for DIC and NAD-ME in *C. gynandra* which was between 20-and 519-fold. No candidate for pyruvate export from the mitochondria could be identified. The situation is reversed for the PEP-CK module where PEP-CK was significantly up-regulated 90-fold in *M. maximus* but only 8.6-fold in *C. gynandra*. The mitochondrial ATP-ADP translocase, AAC (Haferkamp *et al.*, 2002), is up-regulated in *M. maximus*, but not to a significant degree (Table 2). Orthologous AlaATs are significantly up-regulated by 37-fold in both species. Unlike the *C. gynandra* protein, which is predicted to be targeted to mitochondria, the *M. maximus* protein is predicted to be cytosolic (Supplementary Table S4 at *JXB* online). The *M. maximus* AlaAT protein showed a shortened N-terminus when aligned to the *S. italica* gene (Supplementary Table S4), hence *in silico* targeting predicted cytosolic localization. Finally, non-significant up-regulation of TPT and plastidic GAP-DH was detected in *M. maximus* to comparable levels as in *C. gynandra* (Table 2).

In addition to single gene analysis, differentially regulated genes were subjected to pathway enrichment analysis to detect changes in gene expression for whole pathways such as the CBB cycle, photorespiration, and photosynthesis. None of the pathways was significantly enriched among the differentially regulated genes (Supplementary Table S5 at *JXB* online).

The gene-by-gene and enrichment analyses revealed a similar but not identical blueprint for the PEP-CK species compared with the NAD-ME species. The core cycle blueprint was amended to include a companion transporter for the malate phosphate antiporter DIC, which couples it to the proton gradient with phosphate proton symport through PIC.

Energy requirements derived from the PEP-CK blueprint

The energy requirements of intracellular transport reactions were not considered when the energy balance of C_4 photosynthesis was originally calculated (i.e. Kanai and Edwards, 1999), although pyruvate transport was hypothesized to be active based on measurements of the metabolite concentration gradients in maize (Stitt and Heldt, 1985). To assess the energy requirements of the PEP-CK-based C_4 cycle, the

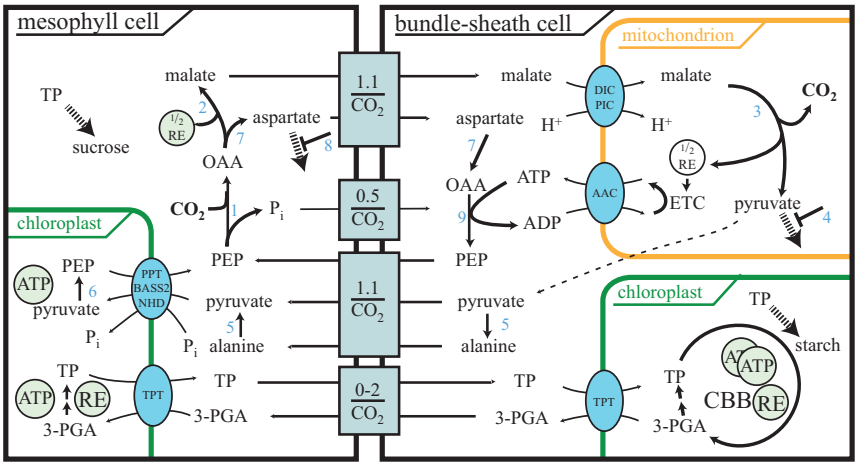


Fig. 3. Extended model for NAD-ME with high PEP-CK activity. Transport modules, consisting of one or more transporters, are shown together with the net transport through the module. Abbreviations: (1) Phospho*enol*pyruvate carboxylase; (2) malate dehydrogenase; (3) NAD-dependent malic enzyme (NAD-ME); (4) pyruvate dehydrogenase kinase; (5) alanine aminotransferase; (6) pyruvate, phosphate dikinase; (7) aspartate aminotransferase; (8) aspartate oxidase and aspartate kinase; (9) phospho*enol*pyruvate carboxykinase (PEP-CK); 3-PGA, 3-phosphoglyceric acid; TP, triose-phosphate; CBB, Calvin–Benson–Bassham cycle; OAA, oxaloacetic acid; RE, reducing equivalent; BASS2, pyruvate transporter; NHD, sodium proton antiporter; PPT, phospho*enol*pyruvate phosphate translocator; TPT, triose-phosphate phosphate translocator; ETC, electron transfer chain. Dashed arrows represent leakage to general metabolism. (This figure is available in colour at *JXB* online.)

amended blueprint was translated into a model of PEP-CK C_4 photosynthesis (Fig. 3).

Energy requirements are calculated following one turn of the cycle (Fig. 3): after PEP is carboxylated to OAA, half of the OAA is reduced to malate (Hatch *et al.*, 1988), requiring on average 0.5 REs derived from photosynthesis for each CO_2 (Fig. 3). The remaining OAA is transported as aspartate (Fig. 3). At the bundle sheath mitochondria, malate exchange for phosphate via DIC is coupled to the proton gradient via phosphate proton symport by PIC (Fig. 3). This process consumes the proton gradient of mitochondria. The proton gradient is also used to drive mitochondrial ATP synthesis for the PEP-CK reaction which decarboxylates OAA to PEP (Fig. 3) and is regenerated by oxidizing the NADH produced by malate decarboxylation (Fig. 3). The carboxylation, transfer, and decarboxylation thus consume on average 0.5 NADPH per CO_2 generated in photosynthetic electron transfer. During regeneration, the PPDK reactions require 2 ATP for the regeneration of each pyruvate but, since only half of the flux runs through malate decarboxylation and therefore pyruvate, only 1 ATP is required for each CO_2. The PPDK reaction is driven towards PEP regeneration by the PPase, which splits the energy-rich bond of pyrophosphate and makes the PPDK reaction irreversible *in vivo*. The production of PEP and its export through PPT creates the proton gradient required to import pyruvate and cycle sodium through the transport system (Fig. 3). Although the active transport of pyruvate is driven by the proton gradient, it requires no additional input of energy beyond that expended for the PPDK reaction (Furumoto *et al.*, 2011). The regeneration phase thus requires 1 ATP in total. The CBB cycle requires 3 ATP and 2 REs from photosynthesis, which may be consumed in the bundle sheath or mesophyll.

The total PEP-CK-based C_4 cycle, assuming no overcycling, thus requires 4 ATP and 2.5 NADPH from the photosynthetic electron transfer chain while solely NADP-ME-based C_4 photosynthesis requires 5 ATP and 2 NADPH and C_3 photosynthesis requires 3 ATP and 2 NADPH for each CO_2 (Kanai and Edwards, 1999). Engineering a PEP-CK-type C_4 cycle will thus avoid the adjustments required for the photosynthetic electron transfer chain since the demands in terms of the ATP and NADPH ratio are almost the same as in C_3 plants.

Intercellular transport derived from the PEP-CK blueprint

Engineering a C_4 cycle may require modifications to the symplastic transport interface (Weber and Bräutigam, 2013). To estimate the difference in intercellular transport for each MC, intercellular transport events between C_4 and C_3 were compared. Data from the scheme depicted in Fig. 3 were combined with anatomical data (Supplementary Fig. S1 at *JXB* online) and photosynthetic rates (Fig. 1C).

Since transport events are assessed per MC and not per leaf area, the number of MCs per leaf area was determined. In the C_4 plants, photosynthesis requires the MC and its adjacent BSC; in the C_3 plant, each MC is a self-contained unit. Microscopic imaging of leaf cross-sections revealed typical Kranz anatomy in *M. maximus* with large BSCs, each of which was connected to multiple MCs (Supplementary Fig. S1 at *JXB* online). The density of MCs was almost twice as high in the C_3 leaf compared with the C_4 leaf (Table 3). Since the photosynthetic rate per leaf area is also higher in *M. maximus* (Fig. 1C), almost twice as much CO_2 is fixed in each MC–BSC pair in *M. maximus* compared with an MC of *D. clandestinum* (5.4 versus 2.6 pmol CO_2 per unit and second). In *D. clandestinum*, only sucrose transport is required across the MC wall. Since each sucrose molecule carries 12 carbons, and since only half of the carbon is exported at any given time, with the remainder stored as starch, the assimilation of one molecule of CO_2 requires $1/12 \times 1/2 = 0.042$ transport events in the C_3 plant (Table 3). In contrast, the PEP-CK-based C_4 cycle requires between 2.75 and 4.75 transport events depending on the extent of RE shuttling because the C_4 acids, the C_3 acids, balancing phosphates, and REs are transported (Table 3). The total number of transport events is estimated by multiplying the number of CO_2 molecules assimilated with the number of transport events required for each CO_2 as 11.6–20.1 pmol s^{-1} in the C_4 species while for C_3 it is 0.1 pmol s^{-1}. C_4 photosynthesis requires between 100- and

Table 3. *Parameters for the calculation of transport requirements for the PEP-CK/NAD-ME C_4 cycle show that C4 photosynthesis requires 100–200 times more transport events*

Cell density was estimated from Supplementary Fig. S1 at JXB online and divided by photosynthetic parameters derived from Fig. 1 to yield the photosynthetic rate per cell (A). C_4 cycle transport requirements were derived from Fig. 3 and summed to calculate total transport events (B). Total transport events through plasmodesmata are calculated as A×B.

		M. maximus	***D. clandestinum***
Photosynthetic parameter	Photosynthetic cell density (Giga photosynthetic units m^{-2})	6.987	12.5
	Photosynthetic rate at 400 ppm (μmol m^{-2} s^{-1})	29.6	20.8
A	Photosynthetic rate CO_2 per cell (pmol CO_2 pu^{-1} s^{-1})	4.2	1.7
Metabolic parameter	C_4 acid (malate, aspartate)	1.1	
(transport events per CO_2)	C_3 acid (PEP, pyruvate, alanine)	1.1	
	Phosphate balance (P_i; 50% PEP assumed)	0.55	
	RE shuttle (triose-phosphate, 3-PGA)	0–2	
	Sucrose export		0.042
B	Total no. of transport events (transport events CO_2^{-1} pu^{-1})	2.75–4.75	0.042
A×B	No. of transport events per cell (pmol transport events s^{-1})	11.6–20.0	0.1

200-fold more transport events than C_3 photosynthesis, such that the intercellular transport capacity needs to be increased by approximately two orders of magnitude in C_4 (Table 3).

Engineering of the C_4 cycle will thus almost certainly require engineering of the BSC–MC interface, as it is highly unlikely that an existing C_3 MC could support the >100-fold increased symplastic flux.

Discussion

Assembly and mapping characteristics

This study was designed to compare two closely related C_3 and C_4 species to increase the probability of detecting C_4-related rather than species-related differences. While for several C_3 grass species, such as rice and *Brachypodium*, the genomes have already been sequenced and thus could serve as C_3 reference for comparative transcriptome sequencing, all of these belong to the BEP clade and have thus diverged 45–55 Myr ago from *M. maximus* (Grass Phylogeny Working Group II, 2012), which belongs to the PACMAD clade. *Dichanthelium clandestinum* was chosen as a C_3 species from within the PACMAD clade for the transcriptomic comparison presented here. Although the precise phylogenetic position of the *Dichanthelium* clade of Paniceae, which includes *D. clandestinum*, has not been determined, it was recently placed as sister to the group, which contains *M. maximus* (Grass Phylogeny Working Group II, 2012), with a divergence time of 14–22 Myr ago (Vicentini *et al.*, 2008). For quantification of steady-state transcript amounts, the RNA-Seq reads were mapped onto the coding sequences predicted from the *Setaria* genome. The closer phylogenetic proximity of *M. maximus* to *Setaria* is represented in the slightly higher mapping efficiency of its reads (Table 1). Overall, the mapping efficiency is above that of the *Flaveria* species on *Arabidopsis* (Gowik *et al.*, 2011*a*) but below that of the *Cleomaceae* on *Arabidopsis* (Bräutigam *et al.*, 2011). The disadvantage of a slightly uneven mapping efficiency was, however, outweighed by mapping reads from both species onto a common genome-based reference sequence, which enabled normalization to reads per kilobase per million reads. In addition, low abundance transcripts are frequently under-represented in contig assemblies, while high abundance transcripts were fragmented into multiple contigs per transcript. Establishing orthology, while possible with tools such as OrthoMCL, requires assumptions about similarities. Mapping onto a reference database as previously successfully established (Bräutigam *et al.*, 2011, Gowik *et al.*, 2011) was chosen to overcome this problem.

Contig assembly from Illumina reads results in fragmented contigs, especially for the high abundance contigs, as observed previously in other RNA-Seq projects (Bräutigam and Gowik, 2010; Franssen *et al.*, 2011; Schliesky *et al.*, 2012). The C_4 transcripts are among the most highly expressed transcripts in leaves of C_4 plants (Bräutigam *et al.*, 2011). To produce high confidence contigs, the transcriptome was sequenced by a long read technology, the reads cleaned with a high base quality threshold of Phred=30, and assembled with CAP3. Within the database, full-length contigs for all candidate C_4 genes were identified (Supplementary Tables S1, S4 at *JXB* online), validating a hybrid approach to quantification and database generation (Bräutigam and Gowik, 2010).

Are NAD-ME and the PEP-CK distinct subtypes of C_4 photosynthesis?

The three classical subtypes of C_4 photosynthesis, NADP-ME, NAD-ME, and PEP-CK, have been analysed by comparative transcriptome sequencing (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011; this study). If the two C_4 types NAD-ME and PEP-CK which both rely wholly or partially on NAD-ME-based decarboxylation were fundamentally different, major differences in the transcriptional profile would be expected. However, quantification of transcript abundance showed that the functions up-regulated in the NAD-ME plant *C. gynandra*, which shows some PEP-CK activity (Sommer *et al.*, 2012), and the PEP-CK plant *M. maximus*, which displays high PEP-CK activity, are quite similar.

The bicarbonate acceptor regeneration module is essentially identical. Both plant species belong to the sodium pyruvate transport group, as defined by Aoki *et al.* (1992), and show joint up-regulation of not only the sodium pyruvate symporter BASS2 (Furumoto *et al.*, 2011), but also the companion sodium:hydrogen antiporter NHD, and the PEP phosphate antiporter PPT (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011; Table 2). The generation of the transfer acids appears to be cytosolic as neither of the two plastidial dicarboxylate transporters, DiT1 (OAA/malate antiporter) (Weber *et al.*, 1995; Kinoshita *et al.*, 2011) and DiT2 (OAA/aspartate antiporter) (Renne *et al.*, 2003), was up-regulated (Supplementary Table S3 at *JXB* online) and the most abundant contigs encoding AspAT and MDH were predicted to be cytosolic (Table 2; Supplementary Table S4). The cytosolic localization relaxes the need to up-regulate organellar transporters, which are required to import substrates and export products. The two species use differentially localized AspATs, a mitochondrial isozyme in the case of *C. gynandra* (Sommer *et al.*, 2012) and a cytosolic one in the case of *M. maximus* (Table 2; Toledo-Silva *et al.*, 2013). For the decarboxylation process, both species use a combination of PEP-CK and NAD-ME and consequently have the same functions up-regulated. The degree of up-regulation, however, mirrors the enzyme activity differences, with PEP-CK transcripts being much more induced in *M. maximus* and NAD-ME and associated transporters much more induced in *C. gynandra* (Table 2). Hence the difference in decarboxylation biochemistry between both species rests in an altered balance between NAD-ME and PEP-CK activities, while the overall pathway is very similar.

At least part of the C_3 acid transport is accomplished through alanine to balance the amino groups between MCs and BSCs. The up-regulated AlaAT for both plants is an orthologous pair, which is targeted to organelles in *C. gynandra* (Bräutigam *et al.*, 2011; Sommer *et al.*, 2012) and *S. italica* (Supplementary Table S4 at *JXB* online). However, enzyme activity measurements placed high AlaAT activity in the

cytosol of *M. maximus* (Chapman and Hatch, 1983). The *in silico* translation of the *M. maximus* transcript revealed that it encodes a truncated version of AlaAT in comparison with the *Setaria* gene, in which a potential start ATG in-frame with the coding sequence is prefaced by a stop codon. The shortened protein is predicted to be cytosolic (Supplementary Table S4). Hence, the cytosolic AlaAT activity in *M. maximus* appears to have evolved by loss of the target peptide of an originally organellar-targeted protein. The simpler cycle suggests that the *M. maximus* blueprint is easier to engineer compared with the blueprints of NAD-ME (Bräutigam *et al.*, 2011; Sommer *et al.*, 2012) and NADP-ME species (Gowik *et al.*, 2011; Pick *et al.*, 2011; Denton *et al.*, 2013; Weber and Bräutigam, 2013).

Multiple species which had previously been grouped as NADP-ME or NAD-ME plants have different degrees of PEP-CK activity (Walker *et al.*, 1997; Pick *et al.*, 2011; Sommer *et al.*, 2012; Muhaidat and McKown, 2013) and modelling shows the advantages of supplemental PEP-CK activity in conferring environmental robustness to the pathway (Wang *et al.*, 2014), raising the question as to whether PEP-CK-type plants deserve their own group. While the functions up-regulated in *C. gynandra* and *M. maximus* are similar, there are differences with regard to localization of the enzymes generating the transfer acids. Whether the different enzyme localizations are tightly associated with the type and degree of use of the decarboxylation enzymes remains to be determined once additional transcriptomes are sequenced and a global view is enabled on more than just one prototypical species for each historical C_4 type. For engineering, it is probably advisable to follow the blueprint of a particular species since it is currently not clear whether differences in transfer acid generation are only species specific or are tied to other processes such as decarboxylation enzymes and therefore functionally relevant.

An extended model of C_4 photosynthesis with high PEP-CK activity

Understanding the evolution of C_4 metabolism and re-engineering a C_4 cycle in a C_3 plant requires a mechanistic understanding of the parts making up the system (Denton *et al.*, 2013). The global transcriptomics analysis of *M. maximus* compared with *D. clandestinum* enabled the extension of the C_4 metabolism model presented earlier for *M. maximus* (Hatch, 1987) and *C. gynandra* (Bräutigam *et al.*, 2011; Sommer *et al.*, 2012).

Transport processes and core cycle

The *M. maximus* analysis confirmed DIC as the mitochondrial malate importer (Table 2; Bräutigam *et al.*, 2011). The companion transporter, which couples malate transport to the proton gradient of the mitochondria and supplies mitochondria with inorganic phosphate for ATP production, is probably PIC (Hamel *et al.*, 2004; Table 2; Fig. 3). The only transporter which remains unknown at the molecular level is the mitochondrial pyruvate exporter. The candidate pyruvate transport protein, the human mitochondrial pyruvate carrier (MPC) (Bricker *et al.*, 2012; Herzig *et al.*, 2012), is not differentially expressed in *C. gynandra* and *M. maximus*. Potentially, pyruvate can traverse biomembranes in its protonated form by simple diffusion (Benning, 1986), although this is unlikely in a cellular context given that only one out of 10^5 molecules of pyruvic acid occurs in the protonated form at physiological pH values. Although early models did not take a reducing equivalent shuttle across both chloroplast envelopes into account for PEP-CK species (Hatch, 1987), possibly because *M. maximus* lacks chloroplast dimorphism (Yoshimura *et al.*, 2004), measurements of enzyme activity confirmed glyceraldehyde dehydrogenase in both MCs and BSCs of *U. panicoides* (Ku and Edwards, 1975), and RNA-Seq indicated modest up-regulation of the necessary transporters and enzymes (Table 2). Engineering a C_4 cycle will critically depend on correctly enabling the transport of substrates through transporters and companion transporters (Weber and von Caemmerer, 2010; Fig. 3). Balancing reducing power between MCs and BSCs via triose-phosphate/phosphate translocators in chloroplasts in both MCs and BSCs appears also to be required in species which lack chloroplast dimorphism (Table 2; Yoshimura *et al.*, 2004).

Knowledge about the intracellular transport proteins involved in C_4 photosynthesis has recently improved significantly (compare with Weber and von Caemmerer, 2010; Bräutigam and Weber, 2011; Denton *et al.*, 2013; Weber and Bräutigam, 2013), largely due to RNA-Seq-enabled identification and characterization of the chloroplast pyruvate transporter (Furumoto *et al.*, 2011), and the placement of several known transport proteins in the C_4 pathway (Taniguchi *et al.*, 2003; Bräutigam *et al.*, 2011; Gowik *et al.*, 2011*a*; Kinoshita *et al.*, 2011). However, information about the intercellular transport has not progressed since the discovery of sieve element-like plasmodesmata plates in the MC–BSC interface (Evert *et al.*, 1977; Botha, 1992).

The difference in total transport events between the C_3 and the C_4 species was estimated using the data provided by the model shown in Fig. 3, by images of the cellular architecture (Supplementary Fig. S1 at *JXB* online), and by photosynthetic rate measurements (Fig. 1C). The large difference in the requirement for intracellular transport between C_4 and C_3 pathways is not predominantly driven by the rather small differences in photosynthetic rates (Fig. 1C), but by two other factors: the number of MCs per leaf area and the number of transport events required for each CO_2 assimilated. The large BSCs, each of which borders several MCs, and the fact that *M. maximus* requires two cells in each photosynthetic unit means that the C_3 grass has almost twice as many photosynthetic units in the same leaf area. The net CO_2 assimilation capacity is thus not only higher by the ~20% higher photosynthetic rate per leaf area but—if normalized to the number of MCs—is almost twice as high for each unit. The second factor is the number of transport processes occurring over each interface. Intercellular transport for each C_3 cell is very low, 0.042 events per CO_2 assimilated for an MC. The transport events for the C_4 cycle are more difficult to estimate since, in addition to the comparatively fixed flux of C_4 and C_3 acids in the cycle, the PEP-balancing phosphate flux and the RE shuttle yield variable fluxes. However, even using

the lowest possible estimates, a >100-fold difference in transport events is predicted between the C_4 and C_3 plant interfaces. The interface itself is probably optimized for a balance of openness to enable the flux and closed-ness to enrich the CO_2 at the site of Rubisco, since different light intensities correlate with photosynthetic rates and plasmodesmatal density in *M. maximus* (Sowinski *et al.*, 2007). The fold change in transport events across the interface is in the range of the fold change expression changes for the C_4 genes (Tables 2, 23). The evolution and hence also re-engineering of the C_4 cycle must adapt the intercellular interface.

Accessory pathways to the core cycle

It is tempting to limit engineering efforts to the major transcriptional changes and therefore to the core cycle. However, accessory pathways to the core C_4 cycle may play a major role in adapting the underlying metabolism to the presence of the carbon-concentrating pump.

The comparison of multiple different C_3–C_4 pairs and therefore C_4 origins with each other provides a method to identify differentially regulated functions with biological significance, once the problem of paralogous genes carrying out the functions is overcome. By mapping RNA-Seq data to EC numbers and Pfam domains rather than individual genes, it has been possible to identify core C_4 genes (Fig. 2; Supplementary Table S2 at *JXB* online), which indicates that these methods are suitable to pick up additional C_4-related functions.

Both methods picked up functions involved in starch metabolism and sucrose synthesis (Fig. 2A, C). In the EC-based mapping, the sucrose synthesis pathway was present with two functions, the UDP-glucose pyrophosphorylase and the sucrose-phosphate synthase. Sucrose-phosphate synthase is the rate-limiting enzyme for sucrose synthesis in the C_3 plant *A. thaliana* (Häusler *et al.*, 2000; Strand *et al.*, 2000; Koch, 2004) and NDP sugar pyrophosphorylases are comparatively slow enzymes. The surplus of fixed carbon (Fig. 1C, D) leads to a surplus of triose-phosphates. In *Z. mays*, *Panicum miliaceum*, and *Brachiaria erucaeformis*, sucrose synthesis is localized to the mesophyll (Usuda and Edwards, 1980), which may also be the case in *M. maximus*. Both the localization of sucrose synthesis and the higher carbon assimilation rate contribute to more triose-phosphate at the site of sucrose synthesis and hence the need for greater sequestration (Fig. 3). Similarly, the higher rate of CO_2 assimilation (Fig. 2) and the localization of starch storage in the BSCs (Majeran and van Wijk, 2009; Majeran *et al.*, 2010) probably also require higher rates of starch synthesis to sequester the triose-phosphates efficiently (Figs 2, 3). When considering the engineering of C_4 photosynthesis, the sequestration of triose-phosphates is probably of low priority compared with the engineering of the enzymes and transport proteins, yet not adding these functions for triose-phosphate sequestration will probably limit the system to the capacity of C_3 photosynthetic plants, a 20% loss of potential productivity.

Insulating the C_4 cycle from other metabolic networks is also probably critical to avoid loss of cycle intermediates. No obvious proteins with functions in this context were identified in comparisons across all C_4 data sets (Fig. 2; Supplementary Table S2 at *JXB* online), although the uncharacterized functions may include such insulators (Supplementary Table S2). The analysis of only NAD-ME-based C_4 photosynthesis registered changes, which represent the overlap between the dicot *C. gynandra* and the grass *M. maximus*. Both species produce pyruvate in their mitochondria (Table 2; Bräutigam *et al.*, 2011) and use aspartate as a dominant transfer acid. Both NAD-ME species show higher PDH kinase and reduced aspartate kinase and aspartate oxidase transcript amounts (Fig. 2). These three enzymes control metabolite exit from the C_4 cycle as PDH kinase gates pyruvate decarboxylation for entry into the tricarboxylic acid (TCA) cycle, aspartate kinase controls entry into aspartate-derived amino acid metabolism, and aspartate oxidase controls entry into NAD synthesis. The leaking of cycle intermediates into other metabolism despite the insulation can be indirectly seen in the labelling pattern obtained by $^{14}CO_2$ feeding. If metabolites from the cycle are consumed, they need to be replaced from the CBB cycle and will thus carry label in C_2–C_4 of the four-carbon compounds and lead to label in the three-carbon compounds, which—if only cycling—should show no label at all. Indeed, labelling studies identified delayed labelling in both groups (e.g. Hatch, 1979), indicating that leaking of intermediates does occur. When engineering a C_4 cycle into a C_3 plant, limiting the leakage of cycle intermediates is probably required for all cycle metabolites to keep the cycle running robustly.

Two to three pathways are commonly down-regulated: the CBB cycle, photorespiration, and protein synthesis (Fig. 2). Reduced expression of these functions in C_4 species may not be required to engineer efficient CO_2 capture. However, reduced expression of CBB, photorespiration, and protein translation (Fig. 2) may be necessary to realize the nitrogen-saving benefits of C_4 photosynthesis which are common to C_4 plants (Sage, 2004).

NAD-ME species show an unusual pattern with regard to nucleotide metabolism; several functions of purine metabolism are up-regulated while several functions of pyrimidine synthesis are down-regulated. While one may speculate that the changes in purine metabolism are due to the altered ATP usage in these plants, the functional reason for these changes remains unknown.

Previous global transcriptome analyses found that genes encoding components of photosynthetic cyclic electron flow (CEF) were significantly up-regulated (Bräutigam *et al.*, 2011; Gowik *et al.*, 2011), raising the question of whether such alterations to photosynthesis are required in all C_4 subtypes. The present analysis did not indicate differences in CEF in *M. maximus* compared with *D. clandestinum* (Supplementary Table S5 at *JXB* online). The reason lies in the high PEP-CK activity, which is fuelled by malate oxidation in mitochondria (Fig. 3). Malate is generated using photosynthetic REs leading to a 4:2.5 ATP:NADPH production ratio in photosynthesis which is very similar to that of C_3 photosynthesis at a 3:2 ratio and in contrast to the classical C_4 calculation of 5:2 (Kanai and Edwards, 1999). If considering engineering, a C_4 cycle with high PEP-CK activity together with malate decarboxylation in mitochondria removes the requirement

for dimorphic chloroplasts, which results in one less feature to be engineered.

It is tempting to think that the type of C_4 photosynthesis realized in *M. maximus* is less efficient because of higher energy input for the C_4 cycle (Fig. 3) and because of oxygen production in the bundle sheath, which increases the potential for photorespiration. Elevated photorespiration is indeed a feature of *M. maximus* (Furbank and Badger, 1982; Ohnishi and Kanai, 1983; Farineau *et al.*, 1984). However, the quantum efficiency of *M. maximus* is indistinguishable from that of *Z. mays* or *S. bicolor* (Ehleringer and Pearcy, 1983). It is surprising that the energy requirements derived from the model (Fig. 3) and the photorespiratory rate (Furbank and Badger, 1982; Ohnishi and Kanai, 1983; Farineau *et al.*, 1984) do not predict quantum efficiency.

The blueprint of C_4 metabolism in *M. maximus* is simpler compared with that of NAD-ME and NADP-ME plants, because the generation of transfer acids requires fewer adjustments in intracellular transport capacity and photosynthetic electron transfer, and at least some part of the insulators that prevent leakage of C_4 cycle intermediates into general metabolism are known. Thus, it represents an attractive target for engineering the C_4 cycle into a C_3 crop plant.

Supplementara data

Supplementary data are available at *JXB* online.

Figure S1. Cross-sections of *D. clandestinum* and *M. maximus.*

Table S1. *D. clandestinum* and *M. maximus* unigene fasta files.

Table S2. Excel table of Pfam and EC function analysis for all genes.

Table S3. Excel table of quantitative gene expression information including statistical analysis.

Table S4. Text document of selected full-length unigenes including alignment to *S. italica* genes and targeting prediction.

Table S5. Excel table of enrichment analysis for pathways.

Acknowledgements

The authors acknowledge excellent technical support for metabolite analysis by Katrin L. Weber and Elisabeth Klemp, and for RNA sequencing by the BMFZ, HHU Düsseldorf. The authors thank Alisandra Denton and especially Richard Leegood and Urte Schlüter for helpful comments on the manuscript. This work was supported by grants of the Deutsche Forschungsgemeinschaft to APMW (IRTG 1525 and EXC 1028 to APMW) and of the European Union Framework 7 Program (3to4 to APMW and CPO).

References

Agostino A, Heldt HW, Hatch MD. 1996. Mitochondrial respiration in relation to photosynthetic C_4 acid decarboxylation in C_4 species. *Australian Journal of Plant Physiology* **23,** 1–7.

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

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106.

Aoki N, Ohnishi J, Kanai R. 1992. 2 Different mechanisms for transport of pyruvate into mesophyll chloroplasts of C_4 plants—a comparative-study. *Plant and Cell Physiology* **33,** 805–809.

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

Bennetzen JL, Schmutz J, Wang H, *et al*. 2012. Reference genome sequence of the model plant Setaria. *Nature Biotechnology* **30,** 555–559.

Benning C. 1986. Evidence supporting a model of voltage-dependent uptake of auxin into cucurbita vesicles. *Planta* **169,** 228–237.

Besnard G, Muasya AM, Russier F, Roalson EH, Salamin N, Christin PA. 2009. Phylogenomics of C_4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. *Molecular Biology and Evolution* **26,** 1909–1919.

Botha CEJ. 1992. Plasmodesmatal distribution, structure and frequency in relation to assimilation in C_3 and C_4 grasses in Southern Africa. *Planta* **187,** 348–358.

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

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

Bräutigam A, Weber APM. 2011. Transport processes—connecting the reactions of C_4 photosynthesis: In: Raghavendra AS, Sage RF, eds. *C_4 photosynthesis and related CO_2 concentrating mechanism. Advances in photosynthesis and respiration* , Vol. **32**. Dordrecht: Springer, 199–219

Bricker DK, Taylor EB, Schell JC, *et al*. 2012. A mitochondrial pyruvate carrier required for pyruvate uptake in yeast, *Drosophila*, and humans. *Science* **337,** 96–100.

Burnell JN, Hatch MD. 1988a. Photosynthesis in phospho*enol*pyruvate carboxykinase-type-C_4 plants—photosynthetic activities of isolated bundle sheath-cells from *Urochloa panicoides*. *Archives of Biochemistry and Biophysics* **260,** 177–186.

Burnell JN, Hatch MD. 1988b. Photosynthesis in phospho*enol*pyruvate carboxykinase-type-C_4 plants—pathways of C_4 acid decarboxylation in bundle sheath-cells of *Urochloa panicoides*. *Archives of Biochemistry and Biophysics* **260,** 187–199.

Chapman KSR, Hatch MD. 1983. Intracellular location of phospho*enol*pyruvate carboxykinase and other C_4 photosynthetic enzymes in mesophyll and bundle sheath protoplasts of *Panicum maximum*. *Plant Science Letters* **29,** 145–154.

Chomczynski P, Sacchi N. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Analytical Biochemistry* **162,** 156–159.

Christin PA, Besnard G. 2009. Two independent C_4 origins in Aristidoideae (Poaceae) revealed by the recruitment of distinct phospho*enol*pyruvate carboxylase genes. *American Journal of Botany* **96,** 2234–2239.

Christin PA, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ. 2013. Anatomical enablers and the evolution of C_4 photosynthesis in grasses. *Proceedings of the National Academy of Sciences, USA* **110,** 1381–1386.

Davidson RM, Gowda M, Moghe G, Lin HN, Vaillancourt B, Shiu SH, Jiang N, Buell CR. 2012. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *The Plant Journal* **71,** 492–502.

Denton AK, Simon R, Weber APM. 2013. C_4 photosynthesis: from evolutionary analyses to synthetic reconstruction of the trait. *Current Opinion in Plant Biology* **16,** 315–321.

Ehleringer J, Pearcy RW. 1983. Variation in quantum yield for CO_2 uptake among C_3 and C_4 plants. *Plant Physiology* **73,** 555–559.

Evert RF, Eschrich W, Heyser W. 1977. Distribution and structure of plasmodesmata in mesophyll and bundle-sheath cells of *Zea mays* L. *Planta* **136,** 77–89.

Farineau J, Lelandais M, Morot-Gaudry JF. 1984. Operation of the glycolate pathway in isolated bundle sheath strands of maize and *Panicum maximum*. *Physiologia Plantarum* **60,** 208–214.

Franssen SU, Shrestha RP, Brautigam A, Bornberg-Bauer E, Weber APM. 2011. Comprehensive transcriptome analysis of the highly

complex *Pisum sativum* genome using next generation sequencing. *BMC Bioinformatics* **12,** 227

Furbank RT, Badger MR. 1982. Photosynthetic oxygen exchange in attached leaves of C_4 monocotyledons. *Australian Journal of Plant Physiology* **9,** 553–558.

Furumoto T, Yamaguchi T, Ohshima-Ichie Y, *et al*. 2011. A plastidial sodium-dependent pyruvate transporter. *Nature* **476,** 472–473.

Gowik U, Brautigam A, Weber KL, Weber APM, Westhoff P. 2011. Evolution of C_4 photosynthesis in the genus Flaveria: how many and which genes does it take to make C_4? *The Plant Cell* **23,** 2087–2105.

Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins. *New Phytologist* **193,** 304–312.

Haferkamp I, Hackstein JHP, Voncken FGJ, Schmit G, Tjaden J. 2002. Functional integration of mitochondrial and hydrogenosomal ADP/ATP carriers in the *Escherichia coli* membrane reveals different biochemical characteristics for plants, mammals and anaerobic chytrids. *European Journal of Biochemistry* **269,** 3172–3181.

Hamel P, Saint-Georges Y, de Pinto B, Lachacinski N, Altamura N, Dujardin G. 2004. Redundancy in the function of mitochondrial phosphate transport in *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. *Molecular Microbiology* **51,** 307–317.

Hatch MD. 1979. Mechanism of C_4 photosynthesis in *Chloris gayana*—pool sizes and kinetics of CO_2-C14 incorporation into 4-carbon and 3-carbon intermediates. *Archives of Biochemistry and Biophysics* **194,** 117–127.

Hatch MD. 1987. C_4 photosynthesis—a unique blend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta* **895,** 81–106.

Hatch MD, Agostino A, Burnell JN. 1988. Photosynthesis in phospho*enol*pyruvate carboxykinase-type C_4 plants—activity and role of mitochondria in bundle sheath-cells. *Archives of Biochemistry and Biophysics* **261,** 357–367.

Hatch M, Mau S. 1977. Properties of phospho*enol*pyruvate carboxykinase operative in C_4 pathway photosynthesis. *Functional Plant Biology* **4,** 207–216.

Häusler RE, Schlieben NH, Nicolay P, Fischer K, Fischer KL, Flügge UI. 2000. Control of carbon partitioning and photosynthesis by the triose phosphate/phosphate translocator in transgenic tobacco plants (*Nicotiana tabacum* L.). I. Comparative physiological analysis of tobacco plants with antisense repression and overexpression of the triose phosphate/phosphate translocator. *Planta* **210,** 371–382.

Herzig S, Raemy E, Montessuit S, Veuthey JL, Zamboni N, Westermann B, Kunji ERS, Martinou JC. 2012. Identification and functional expression of the mitochondrial pyruvate carrier. *Science* **337,** 93–96.

Hibberd JM, Sheehy JE, Langdale JA. 2008. Using C_4 photosynthesis to increase the yield of rice—rationale and feasibility. *Current Opinion in Plant Biology* **11,** 228–231.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. *Genome Research* **9,** 868–877.

Kanai R, Edwards GE. 1999. The biochemistry of C_4 photosynthesis. In: Sage RF, Monson RK, eds. *C4 plant biology*. UK: Academic Press, 49–87.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Research* **12,** 656–664.

Kinoshita H, Nagasaki J, Yoshikawa N, Yamamoto A, Takito S, Kawasaki M, Sugiyama T, Miyake H, Weber APM, Taniguchi M. 2011. The chloroplastic 2-oxoglutarate/malate transporter has dual function as the malate valve and in carbon/nitrogen metabolism. *The Plant Journal* **65,** 15–26.

Koch K. 2004. Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. *Current Opinion in Plant Biology* **7,** 235–246.

Ku MSB, Edwards GE. 1975. Photosynthesis in mesophyll protoplasts and bundle sheath cells of various types of C_4 plants. 4. Enzymes of respiratory metabolism and energy utilizing enzymes of photosynthetic pathways. *Zeitschrift für Pflanzenphysiologie* **77,** 16–32.

Ku MSB, Spalding MH, Edwards GE. 1980. Intracellular localization of phosphoenolpyrvatue carboxykinase in leaves of C_4 and CAM plants. *Plant Science Letters* **19,** 1–8.

Li PH, Ponnala L, Gandotra N, *et al*. 2011. The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* **42,** 1060–1067.

Majeran W, Friso G, Ponnala L, *et al*. 2010. Structural and metabolic transitions of C_4 leaf development and differentiation defined by microscopy and quantitative proteomics in maize. *The Plant Cell* **22,** 3509–3542.

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

Maurino VG, Weber APM. 2013. Engineering photosynthesis in plants and synthetic microorganisms. *Journal of Experimental Botany* **64,** 743–751.

Muhaidat R, McKown AD. 2013. Significant involvement of PEP-CK in carbon assimilation of C_4 eudicots. *Annals of Botany* **111,** 577–589.

Ohnishi J, Kanai R. 1983. Differentiation of photorespiratory activity between mesophyll and bundle sheath cells of C_4 plants 1. Glycine oxidation by mitochondria. *Plant and Cell Physiology* **24,** 1411–1420.

Palmieri L, Picault N, Arrigoni R, Besin E, Palmieri F, Hodges M. 2008. Molecular identification of three *Arabidopsis thaliana* mitochondrial dicarboxylate carrier isoforms: organ distribution, bacterial expression, reconstitution into liposomes and functional characterization. *Biochemical Journal* **410,** 621–629.

Pick TR, Bräutigam A, Schlüter U, *et al*. 2011. Systems analysis of a maize leaf developmental gradient redefines the current C_4 model and provides candidates for regulation. *The Plant Cell* **23,** 4208–4220.

Pratt RD, Ferreira GC, Pedersen PL. 1991. Mitochondrial phosphate transport—import of the H^+/P_i symporter and role of the presequence. *Journal of Biological Chemistry* **266,** 1276–1280.

R Development Core Team. 2012. *R: a language and environment for statistical computing*. Vienna, Austria.

Renne P, Dressen U, Hebbeker U, Hille D, Flugge UI, Westhoff P, Weber APM. 2003. The Arabidopsis mutant *dct* is deficient in the plastidic glutamate/malate translocator DiT2. *The Plant Journal* **35,** 316–331.

Sage RF. 2004. The evolution of C_4 photosynthesis. *New Phytologist* **161,** 341–370.

Sage RF, Christin PA, Edwards EJ. 2011. The C_4 plant lineages of planet Earth. *Journal of Experimental Botany* **62,** 3155–3169.

Schliesky S, Gowik U, Weber APM, Bräutigam A. 2012. RNA-seq assembly—are we there yet? *Frontiers in Plant Science* **3,** 220.

Schomburg I, Chang A, Placzek S, *et al*. 2013. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Research* **41,** D764–D772.

Sommer M, Bräutigam A, Weber APM. 2012. The dicotyledonous NAD malic enzyme C_4 plant *Cleome gynandra* displays age-dependent plasticity of C_4 decarboxylation biochemistry. *Plant Biology* **14,** 621–629.

Sonnhammer ELL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28,** 405–420.

Sowinski P, Bilska A, Baranska K, Fronk J, Kobus P. 2007. Plasmodesmata density in vascular bundles in leaves of C_4 grasses grown at different light conditions in respect to photosynthesis and photosynthate export efficiency. *Environmental and Experimental Botany* **61,** 74–84.

Sowinski P, Szczepanik J, Minchin PEH. 2008. On the mechanism of C_4 photosynthesis intermediate exchange between Kranz mesophyll and bundle sheath cells in grasses. *Journal of Experimental Botany* **59,** 1137–1147.

Stitt M, Heldt HW. 1985. Generation and maintenance of concentration gradients between the mesophyll and bundle sheath in maize leaves. *Biochimica et Biophysica Acta* **808,** 400–414.

Strand A, Zrenner R, Trevanion S, Stitt M, Gustafsson P, Gardestrom P. 2000. Decreased expression of two key enzymes in the sucrose biosynthesis pathway, cytosolic fructose-1,6-bisphosphatase and sucrose phosphate synthase, has remarkably different consequences for photosynthetic carbon metabolism in transgenic Arabidopsis thaliana. *The Plant Journal* **23,** 759–770.

Taniguchi Y, Taniguchi M, Nagasaki J, Kawasaki M, Miyake H, Sugiyama T. 2003. Functional analysis of chloroplastic dicarboxylate transporters in maize. *Plant and Cell Physiology* **44,** S64–S64.

Toledo-Silva G, Cardoso-Silva CB, Jank L, Souza AP. 2013. De novo transcriptome assembly for the tropical grass *Panicum maximum* Jacq. *PLoS One* **8,** e70781.

Usuda H, Edwards GE. 1980. Localization of glycerate kinase and some enzymes for sucrose synthesis in C_3 and C_4 plants. *Plant Physiology* **65,** 1017–1022.

Vicentini A, Barber JC, Aliscioni SS, Giussani LM, Kellogg EA. 2008. The age of the grasses and clusters of origins of C_4 photosynthesis. *Global Change Biology* **14,** 2963–2977.

Walker R, Trevanion S, Leegood R. 1995. Phospho*enol*pyruvate carboxykinase from higher plants: purification from cucumber and evidence of rapid proteolytic cleavage in extracts from a range of plant tissues. *Planta* **196,** 58–63.

Walker RP, Acheson RM, Tecsi LI, Leegood RC. 1997. Phospho*enol*pyruvate carboxykinase in C_4 plants: its role and regulation. *Australian Journal of Plant Physiology* **24,** 459–468.

Wang Y, Bräutigam A, Weber APM, Zhu XG. 2014. Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis. *Journal of Experimental Botany* **65** (in press).

Weber A, Menzlaff E, Arbinger B, Gutensohn M, Eckerskorn C, Flügge UI. 1995. The 2-oxoglutarate/malate translocator of chloroplast envelope membranes: molecular cloning of a transporter containing a 12-helix motif and expression of the functional protein in yeast cells. *Biochemistry* **34,** 2621–2627.

Weber APM, Bräutigam A. 2013. The role of membrane transport in metabolic engineering of plant primary metabolism. *Current Opinion in Biotechnology* **24,** 256–262.

Weber APM, von Caemmerer S. 2010. Plastid transport and metabolism of C_3 and C_4 plants—comparative analysis and possible biotechnological exploitation. *Current Opinion in Plant Biology* **13,** 257–265.

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

Westhoff P, Gowik U. 2004. Evolution of C_4 phospho*enol*pyruvate carboxylase. Genes and proteins: a case study with the genus Flaveria. *Annals of Botany* **93,** 13–23.

Westhoff P, Herrmann RG. 1988. Complex RNA maturation in chloroplasts. *European Journal of Biochemistry* **171,** 551–564.

Wingler A, Walker RP, Chen ZH, Leegood RC. 1999. Phospho*enol*pyruvate carboxykinase is involved in the decarboxylation of aspartate in the bundle sheath of maize. *Plant Physiology* **120,** 539–545.

Yoshimura Y, Kubota F, Ueno O. 2004. Structural and biochemical bases of photorespiration in C_4 plants: quantification of organelles and glycine decarboxylase. *Planta* **220,** 307–317.