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

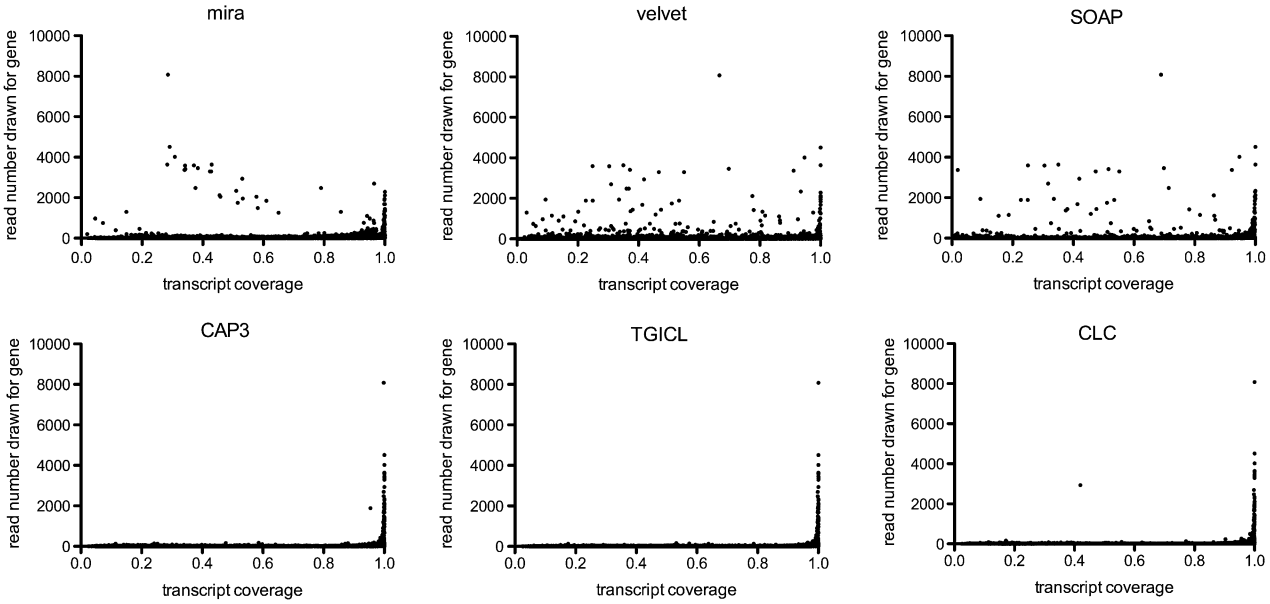

Fig. 1. Transcript coverage for the artificial library assemblies with perfect reads. For each contig, the corresponding *Arabidopsis* transcript was determined and the coverage, i.e. the percentage of bases from the transcript covered by the contig, was determined. For each transcript, the coverage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

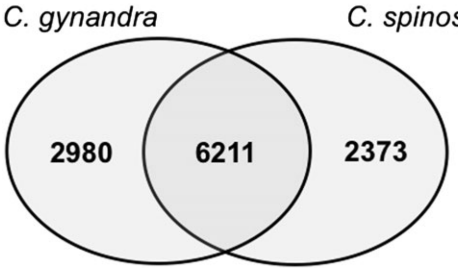

Fig. 2. Venn diagram of the mapping results of the contigs from *C. spinosa* and *C. gynandra* libraries. After the assembly of the *C. gynandra* and *C. spinosa* reads, for each contig the corresponding *Arabidopsis* transcript was determined. A total of 6211 *Arabidopsis* transcripts were matched by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

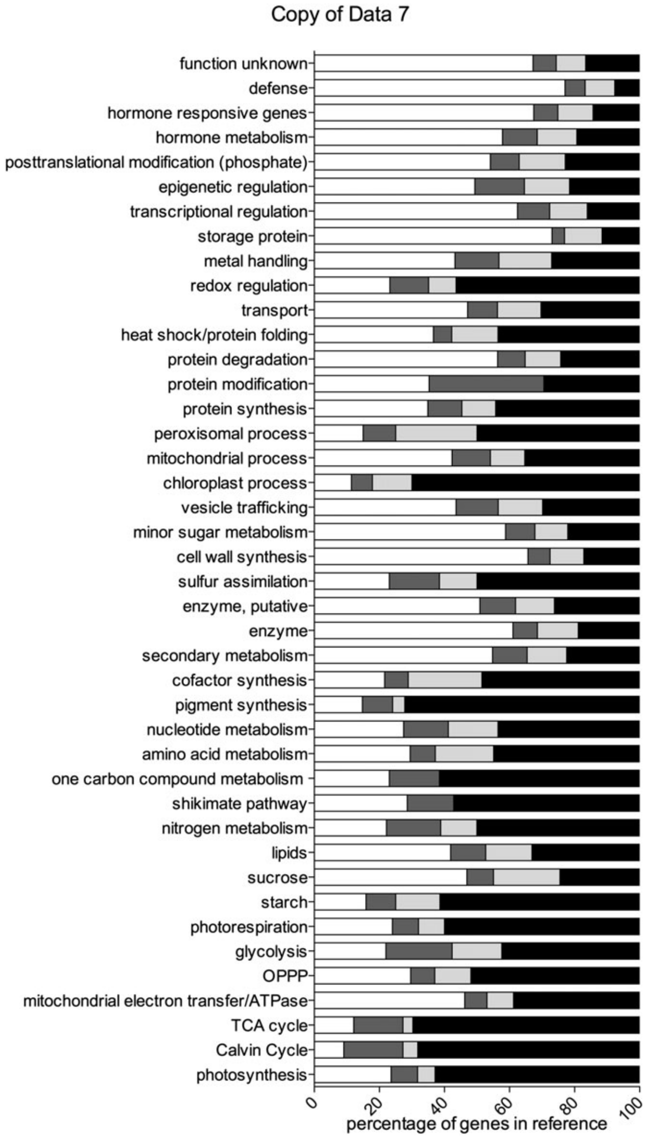

Fig. 3. Pathway representation analysis of the contig mappings to *A. thaliana*; white, not detected in either species; dark grey, detected in *C. spinosa*; light grey, detected in *C. gynandra*; black, detected in both species. For each *Arabidopsis* protein-coding transcript it was determined whether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

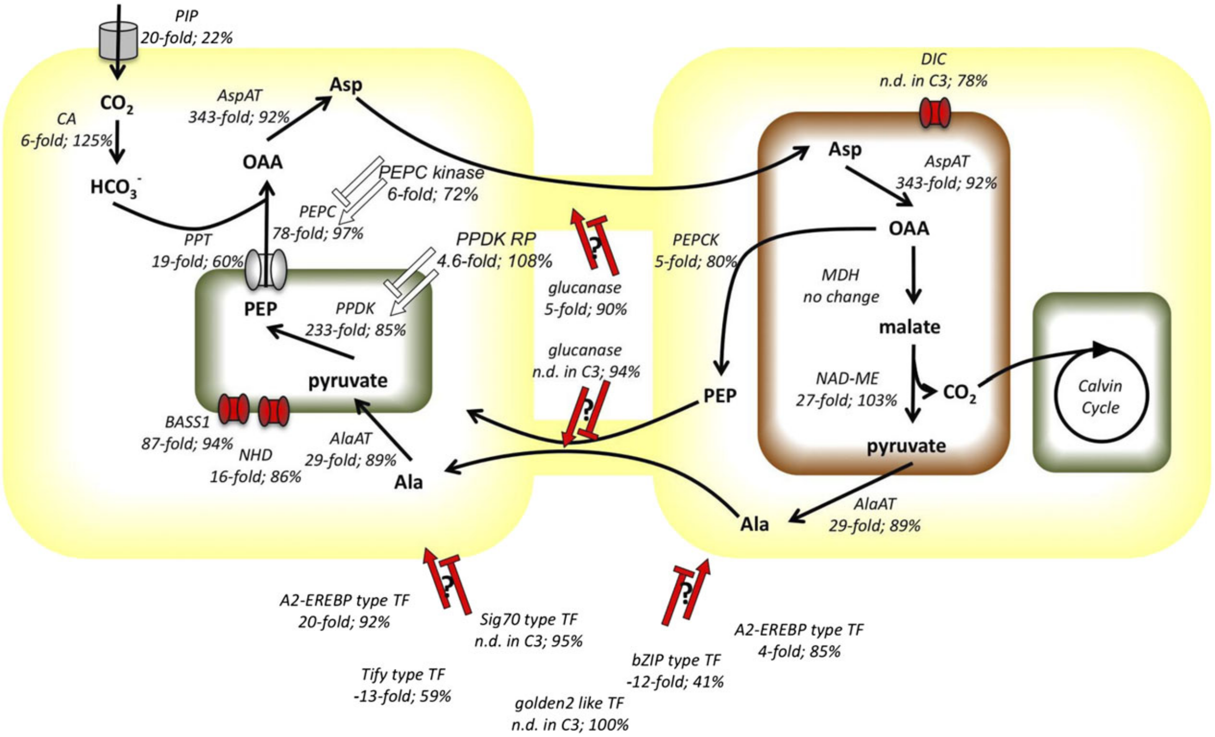

Fig. 4. Schematic representation of the C_4 cycle in *C. gynandra*. For each C_4 gene, the expression fold change compared with *C. spinosa* was extracted from Bräutigam *et al.* (2011) and compared with the contig coverage relative to the closest *Arabidopsis* homologue (in %). Organelles are colour-coded, 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.