American Journal of Botany

American Journal of Botany 99(2): 257–266. 2012.

Designing a transcriptome next-generation sequencing project for a nonmodel plant species[1]

Susan R. Strickler, Aureliano Bombarely, and Lukas A. Mueller[2]

Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, New York 14853 USA

The application of next-generation sequencing (NGS) to transcriptomics, commonly called RNA-seq, allows the nearly complete characterization of transcriptomic events occurring in a specific tissue. It has proven particularly useful in nonmodel species, which often lack the resources available for sequenced organisms. Mainly, RNA-seq does not require a reference genome to gain useful transcriptomic information. In this review, the application of RNA-seq to nonmodel plant species will be addressed. Important experimental considerations from presequencing issues to postsequencing analysis, including sample and platform selection, and useful bioinformatics tools for assembly and data analysis, are covered. Methods of assembling RNA-seq data and analyses commonly performed with RNA-seq data, including single nucleotide polymorphism detection and analysis of differential expression, are explored. In addition, studies that have used RNA-seq to elucidate nonmodel plant transcriptomics are highlighted.

Key words: 454; assembly; Illumina; next-generation sequencing; NGS; nonmodel organism; plant; RNA-Seq; transcriptomics.

Next-generation sequencing (NGS) and new complementary computational tools have allowed for high-throughput sequencing and assembly to become more commonplace. NGS can produce billions of short reads in parallel, generally 50–800 base pairs (bp) depending on technology. Sequence data that may have once taken years to generate may now be produced in a matter of days or even hours. NGS is replacing capillary sequencing in many applications due to its lower cost per base pair of DNA and its lack of a subcloning requirement. The enormous amount of data generated has enabled exciting prospects in a number of scientific disciplines, which has changed the way problems are approached in many areas of molecular biology. In particular, the study of transcriptomics has been greatly advanced by this technology.

At this time, RNA-seq, the sequencing of a transcriptome using NGS, is one of the most popular topic areas in NGS, as shown by the SEQanswers' search tag cloud (http://seqanswers.com/forums/search.php). Other methods of studying gene expression such as microarrays and serial analysis of gene expression (SAGE) are being replaced in many applications with RNA-seq. RNA-seq can show the repertoire of expressed sequences found in a particular tissue at a specific time point, even rare transcripts, due to the great depth of sequencing. In this way, it can produce a nearly complete picture of transcriptomic events in a biological sample. The data are versatile and can be used to characterize genes (Novaes et al., 2008; Alagna et al., 2009; Barakat et al., 2009; Dassanayake et al., 2009; Wang et al., 2009; Brautigam et al., 2011a), reveal information on novel transcripts (Denoeud et al., 2008), and look at gene expression, single-nucleotide polymorphisms (SNPs) (Novaes et al., 2008; Alagna et al., 2009), alternative splicing (Wang et al., 2008), and structural variation (Maher et al., 2009). RNA-seq is practical in nonmodel species when little to no genetic tools and sequence data may be available and resources may be limited, since the focus of sequencing is restricted to the coding regions, rather than the entire genome. In addition, RNA-seq can be easier than whole genome assembly in some respects, since coding regions typically have less repetitive elements and a higher GC content. Several studies have used transcriptomic approaches in nonmodel plant systems (Novaes et al., 2008; Alagna et al., 2009; Barakat et al., 2009; Dassanayake et al., 2009; Trick et al., 2009; Wang et al., 2009; Riggins et al., 2010; Angeloni et al., 2011; Der et al., 2011; Franssen et al., 2011). In light of the popularity of RNA-seq and its utility in nonmodel systems, design strategy and methods for its use in a nonmodel plant species will be examined in this review.

RNA-SEQ EXPERIMENTAL DESIGN

In this section, experimental presequencing planning will be outlined to ensure RNA-seq can address the questions of interest. RNA-seq is a versatile tool, but it may be difficult to implement in some studies, particularly when working with complicated genomes with no reference species, as will be discussed. In addition, analysis and interpretation of the large amount of data generated can be challenging and proper methodology is important.

Sample choice and treatment—Sample choice is an important first step in any transcriptomic study. Material must be chosen that will generate relevant data. This requires consideration of background information about the species of interest, determining which tissues and developmental stages, treatments, and controls will be used for RNA extraction, the amount of sequencing necessary, and proper RNA treatment prior to sequencing.

Species background information—Both variant alleles of a gene and gene duplications can complicate transcript assembly by making it difficult to distinguish between sequencing error, heterozygosity, and duplicate genes. Model species are usually selfing plants; therefore, highly homozygous lines are available. However, nonmodel species may be outcrossers, which

[1] Manuscript received 29 June 2011; revision accepted 26 October 2011.

[2] Author for correspondence (e-mail: lam87@cornell.edu)

doi:10.3732/ajb.1100292

can lead to higher levels of heterozygosity. Heterozygosity complicates SNP detection, since it can be difficult to distinguish the difference between a true SNP and a sequencing error, although taking the quality score and the frequency of the SNP into account can help resolve this matter. In addition, most angiosperms are descendants of ancient genome duplication events (Soltis et al., 2009). For example, the eudicots share a whole genome triplication event, genomic sequence from grape suggests an ancestral hexaploidization event has occurred in many angiosperm phyla (Jaillon et al., 2007) and additional genome duplication events have occurred in *Arabidopsis* (Bowers et al., 2003), rice and sorghum (Tang et al., 2010), and soybean (Schmutz et al., 2010). This high prevalence of polyploidy and paralogous gene duplications can add to the difficulty of proper assembly (Pop and Salzberg, 2008; Miller et al., 2010). An assumption of transcript assembly is that enough sequence diversity is present between duplicated genes to correctly sort reads into the correct representative gene model for each gene. In the case of recently duplicated paralogs, this assumption may not hold true, making it difficult to differentiate reads that belong to a family of duplicated genes, instead clustering them into one representative gene model. One study in maize was able to deal with the effects of heterozygosity and gene or genome duplications on assembly by using the higher variability in the 3′ UTR of mRNA to resolve gene and allele-specific transcripts (Eveland et al., 2007). By anchoring 454 sequencing to the 3′ end of transcripts, several unique transcripts were resolved including a group of cellulose synthases, some previously uncharacterized members of a histone H1 gene family, and two nearly identical auxin-repressed, dormancy-associated (Arda) genes (Eveland et al., 2007).

The level of divergence between the species of interest and the closest model plant species should also be considered. Divergence determines the methods that can be used to assemble the transcripts. The existence of a closely related reference species allows the reads to be mapped onto the reference, using it as a template for read assembly. This is known as reference-guided assembly. A method of determining whether a sequenced organism may be an appropriate reference was investigated with *Pachycladon enysii* using *A. thaliana* as a reference (Collins et al., 2008). By running mapping simulations with varying mismatch and read length parameters in ELAND, we can determine the optimal match specificity for the reads in the data set (Collins et al., 2008). Unfortunately, this approach requires that the short read data be generated a priori, making it less than ideal for projects lacking information on sequence similarity to the closest reference. The reference genome need not be complete, as demonstrated by a study to detect SNPs in *Brassica napus*, which was successful in mapping Solexa reads to a collection of contigs assembled from various species of *Brassica* (Trick et al., 2009). However, it is important to ascertain the accuracy of the reference assembly. A reference assembly may be inconsistent and lack proper annotations and tools. Oftentimes, the species of interest is too divergent to take advantage of preexisting data and tools, making de novo assembly the only feasible option.

Tissue treatment and selection—A specific treatment may be required prior to tissue collection, such as when one is interested in differential expression of genes during abiotic stress or pathogen response. In addition, the proper plant tissue and plant life stage to address the experimental questions should be used. Laser capture microdissection (LCM) can be useful in obtaining very specific cell types from microscopic regions within a tissue (Emmert-Buck et al., 1996; Ohtsu et al., 2006). This method was used to collect a sample from the shoot apical meristem of maize for the generation of 454 ESTs and SNP mining (Emrich et al., 2006; Barbazuk et al., 2007). It is important to note, often this technique requires an RNA amplification step (Emrich et al., 2006), which may not be ideal in differential expression studies. In some projects, the goal may be to capture as many different transcripts as possible, in which case a variety of tissues, treatments, and stages of development should be used in an attempt to obtain a diverse representation of transcripts from as many genes as possible.

Ideally, experimental design should include the use of biological replicates to determine within sample variation, especially when the goal of the study is to detect differential expression between groups (Auer and Doerge, 2010). Additionally, some studies have found that there can be variation between replicate Illumina samples due to batch effects, errors occurring before application to the flow cell, and lane effects, which occur in the flow cell and during base calling (Rougemont et al., 2008; Balwierz et al., 2009; Chepelev et al., 2009; Auer and Doerge, 2010). Both biological and technical replicates can be performed with the use of barcoding, as will be discussed in the next section.

Sequencing efficiently—An RNA-seq experiment can be designed to optimize the amount and type of data generated. For example, it can provide data on global gene expression representation or, alternatively, focus on a subset of the transcriptome. In some cases, such as the analysis of a gene family, it may be more cost effective to do targeted sequencing rather than sequencing of the entire transcriptome. In a technique known as hybrid capture, RNA bait is used to capture a subpopulation of RNA for sequencing using either array- or solution-based hybridization (Mamanova et al., 2010). However, when using this method it is difficult to select genes that have high homology to other transcripts, so it may not be useful in all cases.

High-throughput sequencing can be made more efficient by adding barcodes, sample-specific sequences, to transcripts originating from different groups, allowing the samples to be mixed during the sequencing process and then recovered into their respective groups later through bioinformatic means (Smith et al., 2010). This procedure is known as multiplexing. Another useful application is the tagging of replicate samples with barcodes, allowing the implementation of a balanced blocked experimental design to minimize batch or lane effects during the sequencing process (Auer and Doerge, 2010). This use is especially important when detecting differential gene expression using Illumina sequencing, since uneven coverage due to tag biases can be corrected assuming the bias is consistent within the sample (Auer and Doerge, 2010). Barcodes are normally located at the 5′ end of the sequence, so they avoid the drop in quality that tends to occur at the 3′ end of a read. When using barcoding, it is important to note that the sequence obtained from the actual transcript will be shorter, since the tag will be sequenced as well.

Post-RNA extraction treatments—Depending on the questions to be answered, sequencing libraries may be used as is or require normalization or amplification. Normalization of sequencing libraries adjusts for overrepresentation of highly abundant transcripts. If the library will be used for gene expression analysis normalization should be avoided, although it can be useful in transcriptome representation studies since it may allow for higher coverage of underrepresented transcripts (Ekblom and Galindo, 2011). Similarly, while amplification of the library may be useful in cases where the starting material is suboptimal, it should be avoided in expression studies since it can change RNA population ratios, making downstream detection of differential expression unreliable.

Choosing a platform—At this time, two main NGS technologies exist that have been used successfully for transcriptomic studies of nonmodel plants, Roche/454 (Margulies et al., 2005) and Solexa/Illumina (San Diego, California, USA). Both technologies have the option of producing sequence from both ends of a DNA molecule. This type of sequencing is known as paired-end sequencing on both platforms, but the orientation of these paired reads differ between Roche/454 and Solexa/Illumina. Roche/454 paired-end reads are in the "forward-forward" orientation with respect to how they map to the genome, whereas the Solexa/Illumina paired-end reads are in the "forward-reverse" orientation. Since the size of the DNA molecule from which the reads originated is known, the distance between the reads can be inferred, making paired-end reads useful in determining splice junctions (Au et al., 2010) and giving greater sequence specificity.

The transcriptomes of several nonmodel plant species have been assessed using the Roche/454 platform (Novaes et al., 2008; Alagna et al., 2009; Barakat et al., 2009; Dassanayake et al., 2009; Wall et al., 2009; Wang et al., 2009; Guo et al., 2010; Riggins et al., 2010; Angeloni et al., 2011; Franssen et al., 2011; Swarbreck et al., 2011). Since Solexa/Illumina generates reads shorter than 454, de novo assembly can be more difficult using this platform (Pop and Salzberg, 2008), although coverage is deeper. Indeed, likely due to the greater difficulty in assembly, few nonmodel plant transcriptome studies have used the Illumina platform (Collins et al., 2008; Trick et al., 2009; Mizrachi et al., 2010) (Table 1). A simulation study found that a combination of 454-FLX and Illumina may assist de novo assembly, based on NGS data from *Eschscholzia californica*, *Persea americana*, and *Arabidopsis thaliana* (Wall et al., 2009). As a result of this study, a NGS result simulator was developed that predicts various assembly statistics when using different technologies, as well as total cost and related measures (http://fgp.huck.psu.edu/NG_Sims/ngsim.pl). The incorporation of Sanger expressed sequence tags (ESTs) is also a common approach in transcriptomics (Guo et al., 2010; Swarbreck et al., 2011) and aids in assembly due to their longer length. Several new technologies are in the works, promising longer read length and higher quality data, which will help mitigate the difficulty of assembly. These include IonTorrent (Life Technologies, Carlsbad, CA, USA), PacBio (Pacific Biosciences, Menlo Park, CA, USA), and Helicos (Helicos Biosciences, Cambridge, MA, USA) (see Introduction to this issue for a review).

DATA PREPROCESSING AND ASSEMBLY

Once the sequencing output is obtained, the raw reads should be assessed for quality and contamination. Cleaned and filtered reads are then assembled, either by a de novo method or by mapping to a reference sequence, and the resulting assembly is evaluated for accuracy. In this section, the steps in data preprocessing, assembly, and methods of determining assembly quality will be addressed in detail.

Sequence preprocessing—An assessment of the unprocessed reads is critical to check for sequence biases and contamination. Biases in data composition are known to occur in NGS and can be determined by checking base calls, *k*-mers, and distribution of *k*-mers (Schröder et al., 2010). Some other important measures to consider are overall quality of the reads, length, duplication level, and overabundant sequences. Additionally, the raw reads may contain the adaptor and/or linker sequence used in the sequencing reaction, that need to be removed before assembly. Some assembly tools, for example gsAssembler (454 Life Sciences) and Mira (Chevreux et al., 2004), have the ability to screen for this type of contamination. If barcoding was used to distinguish RNA populations, the different populations must be sorted before assembly, and the barcodes should be removed. Tools such as NovoBarCode (Novocraft, Selangor, Malaysia), TagCleaner (Schmieder et al., 2010), and the FASTX-toolkit (Hannon Laboratory, Cold Spring Harbor, New York, USA) in the Galaxy package (to be discussed below) are useful for dealing with barcode processing (Table 2). Biases in the distribution of barcodes may also be checked to look for overrepresentation of a particular tagged group. Poly-A-stretches should generally be removed or masked as well. Several programs are useful in gathering read complexity information (Table 2). The most efficient programs have the ability to filter and preprocess the reads, trim reads based on quality or *k*-mer, and remove reads that should not be part of the assembly. Raw reads should be checked for other sources of contamination. Contamination can come from a variety of sources such as chloroplast, mitochondria, and microbial populations in the original plant sample. Stand-alone tools, such as Deconseq (Schmieder and Edwards, 2011), provide this function (Table 2).

Transcriptome assembly—Tools developed specifically for NGS assembly are a necessity, due to the large volume of data, short read length, and different error rate than capillary sequencing. When choosing an assembler, it is important to consider the amount of data in relation to the computational resources available. Powerful computers are necessary for assembly of complex transcriptomes and large sequence data sets. Some assemblers, to be discussed, have been optimized for efficiency, allowing them to better handle large data sets. The sequencing platform and operating system of the assembly computer are essential considerations in choosing a compatible assembly tool. It is also wise to choose a program with good documentation and support as well as providing an output format easily used by downstream tools. As previously discussed, when assembling a transcriptome, two different approaches can be used: de novo and reference-guided. In the following, a selection of popular assemblers that have proven useful in transcriptome assembly are discussed. For a full updated listing, see the SEQanswers wiki (http://seqanswers.com/wiki/Software).

De novo assembly/clustering—If there is no appropriate reference available, as is often the case in nonmodel organism studies, a de novo assembly is the only option for sequence assembly. In de novo assembly, the reads are assembled into contigs without the guidance of a reference sequence. In 454 sequencing, the longer reads are clustered into groups that may represent a full-length transcript. De novo assembly has the potential to allow transcripts not represented in the genome, due to alternative splicing or mis-annotation, to be discovered. However, it can be difficult to correctly assemble alternatively spliced variants of the same gene.

TABLE 1. Description of RNA-seq studies in nonmodel plant species performed to date.

Organism	Reference	Platform	Assembler	Type of assembly	Type of study
Amaranthus tuberculatus	Riggins et al., 2010	454	CAP3 (Huang, 1999), EGassembler (Masoudi-Nejad et al., 2006)	de novo	Transcriptome characterization, comparative gene expression
Artemesia annua	Wang et al., 2009	454	TGICL (Pertea et al., 2003), CAP3 (Huang, 1999)	de novo	Transcriptome characterization
Wild oat (*Avena barbata*)	Swarbreck et al., 2011	454	MALIGN (J. Chapman, unpublished), CAP3 (Huang, 1999)	de novo	Transcriptome characterization, comparative gene expression
Brassica napus	Trick et al., 2009	Illumina	MAQ (http://maq.sourceforge.net/maq-man.shtml)	mapping	SNP detection
Chestnut (*Castanea dentata*)	Barakat et al., 2009	454	Newbler/gsAssembler (454 Life Sciences)	de novo	Transcriptome characterization, comparative gene expression
Cucumber (*Cucumis sativus*)	Guo et al., 2010	454	iAssembler (http://bioinfo.bti.cornell.edu/tool/iAssembler/), Splan (Gotoh, 2008), BLAT (Kent, 2002)	de novo, mapping	Transcriptome characterization, comparative gene expression
Eucalyptus grandis	Novaes et al., 2008	454	Newbler/gsAssembler (454 Life Science), Paracel Transcript Assembler (Paracel, Pasadena, CA), GS Reference Mapper (454 Life Science)	de novo, mapping	Transcriptome characterization, SNP identification
Eucalyptus grandis × *E. urophylla*	Mizrachi et al., 2010	454	Velvet (Zerbino and Birney, 2008), Mosaik (Marth Laboratory), BWA (Li and Durbin, 2010)	de novo, mapping	Transcriptome characterization, comparative gene expression
Looking-glass mangrove (*Heritiera littoralis*)	Dassanayake et al., 2009	454	Newbler/gsAssembler (454 Life Sciences), Phrap	de novo	Transcriptome characterization, comparative gene expression
Olive (*Olea europaea* cv. Coratina)	Alagna et al., 2009	454	ParPEST (D'Agostino et al., 2005)	de novo	Transcriptome characterization, comparative gene expression
Olive (*O. europaea* cv. Tendellone)	Alagna et al., 2009	454	ParPEST (D'Agostino et al., 2005)	de novo	Transcriptome characterization, comparative gene expression
Pachycladon enysii	Collins et al., 2008	Illumina	Velvet (Zerbino and Birnery, 2008), ELAND	de novo, mapping	Reference-guided assembly using diverged reference
Garden pea (*Pisum sativum*)	Franssen et al., 2011	454	MIRA (Chevreux et al., 2004), TGICL(Pertea et al., 2003), BWA-SW (Li and Durbin, 2010)	de novo, mapping	Transcriptome characterization, comparative gene expression
Bracken fern (*Pteridium aquilinum*)	Der et al., 2011	454	MIRA (Chevreux et al., 2004)	de novo	Transcriptome characterization
Red mangrove (*Rhizophora mangle*)	Dassanayake et al., 2009	454	Newbler/gsAssembler (454 Life Sciences), Phrap	de novo	Transcriptome characterization, comparative gene expression
Scabiosa columbaria	Angeloni et al., 2011	454	CLC bio	de novo	Transcriptome characterization, SNP identification

Many assemblers have been developed based on various algorithms to assemble the short reads generated by NGS (Table 2). The most popular assembler tools for nonmodel transcriptome NGS data use an overlap layout methodology. This algorithm is able to handle the longer reads of 454. The gsAssembler (454 Life Sciences) is the most widely used transcriptomics assembler of this type and implements overlap layout consensus, meaning it assembles a consensus transcript for different alleles. Isotigs often represent alternatively spliced transcripts. Although slower than gsAssembler, Mira uses an overlap-based approach and is helpful when incorporating EST data into an assembly (Chevreux et al., 2004). Mira is able to handle a true hybrid assembly where both ESTs and 454 sequence are used and also has excellent documentation. CAP3 (Huang, 1999) is also used in 454 assembly, although it may have difficulty handling the large number of reads generated from this technology. Another approach for assembly, more commonly used in whole genome assembly of short reads, is manipulation of de Bruijn graphs where reads are represented as short words, or *k*-mers (Pevzner et al., 2001). These assembly tools are not as commonly used in transcriptomics because they were developed mainly for whole genome assembly. Velvet is one of the few that has been implemented in nonmodel plant RNA-seq (Collins et al., 2008) (Table 1). Also, Trinity, an assembly pipeline developed specifically for shorter reads, may prove to be useful in nonmodel plant sequence assembly with Illumina reads from transcript data (Grabherr et al., 2011). Adjustable parameters in the assembly tool may also be optimized to obtain the best assembly. Using varying *k*-mer values and comparing contig translations against a reference proteome, was found to substantially improve de novo assembly results using short reads (Surget-Groba and Montoya-Burgos, 2010). Benchmarks from work comparing representatives of both overlap-based and de Bruijn graph assemblers suggested that TGICL (Pertea et al., 2003), an

Table 2. A selection of software useful in sequence assembly and analysis.

Tool	Purpose	Platform	Description
CAP3 (Huang, 1999)	de novo assembly	454	Useful for clustering reads
CLCbio Genomics Workbench (CLC Bio)	de novo assembly	454, Illumina	Provides many other tools; GUI interface
gsAssembler (454 Life Sciences)	de novo assembly	454	Uses Newbler; GUI-based
MIRA (Chevreux et al., 2004)	de novo assembly, mapping	454, Illumina	Performs true hybrid assemblies
TGICL (Pertea et al., 2003)	de novo assembly	454	Clustering pipeline
Trinity (Grabherr et al., 2011)	de novo assembly	Illumina	Requires paired reads
BLAT (Kent, 2002)	reference-guided assembly/ mapping	454	BLAST-Like Alignment Tool; aligns mRNA to DNA
Bowtie (Langmead et al., 2009)	reference-guided assembly/ mapping	Illumina	Can handle paired-end reads
BWA (Li and Durbin, 2009)	reference-guided assembly/ mapping	Illumina	Fast, accurate, short read aligner; can handle paired-end reads
BWA-SW (Li and Durbin, 2010)	reference-guided assembly/ mapping	454	Fast, gapped alignment of long reads
GSNAP (Wu and Nacu, 2010)	reference-guided assembly/ mapping	Illumina	SNP-tolerant detection of long indels and splice variants
MAQ (http://maq.sourceforge.net/)	reference-guided assembly/ mapping	Illumina	Can be used in SNP detection
Mosaik (http://bioinformatics.bc.edu/marthlab/ Mosaik)	reference-guided assembly/ mapping	454, Illumina	Can handle paired-end reads
Velvet (Zerbino and Birney, 2008)	de novo assembly	454, Illumina	Requires paired reads
FastQC (http://www.bioinformatics.bbsrc.ac.uk/ projects/fastqc/)	preprocessing	Illumina	Determines composition and quality of reads
FASTX Toolkit (http://hannonlab.cshl.edu/ fastx_toolkit/)	preprocessing	454, Illumina	Performs, filtering, trimming, and masking of reads
NovoBarCode	preprocessing	454, Illumina	Sorting and removal of barcodes
Prinseq (http://prinseq.sourceforge.net/)	preprocessing	454, Illumina	Filtering and trimming of reads
SeqTrim (Falgueras et al., 2010)	preprocessing	454, Illumina	Trims reads based on several parameters
SolexaQA (Cox et al., 2010)	preprocessing	Illumina	Generates read statistics Trims reads based on qualities
Tagcleaner (Schmieder et al., 2010)	preprocessing	454	Removes barcodes sequences from reads, includes a script to split reads based on barcode
BEDtools (Quinlan and Hall, 2010)	postprocessing	454, Illumina	Tools for comparing genome features, such as coverage of reads across a contig
Picard (http://picard.sourceforge.net)	postprocessing	Illumina	Manipulates sam/bam files
sff_extract (http://bioinf.comav.upv.es/ sff_extract/)	postprocessing	454	Processes sff files
Deconseq (Schmieder and Edwards, 2011)	contamination screening	454, Illumina	Removes sequence contamination from the read data set
iAssembler (http://bioinfo.bti.cornell.edu/tool/ iAssembler/)	processing pipeline	454	Uses MIRA and CAP3 to cluster 454 and Sanger ESTs
Gap4 (http://staden.sourceforge.net/)	viewer/editor	454, Illumina	Useful for editing erroneous contigs in an assembly
Tablet (Milne et al., 2010)	viewer	454, Illumina	Accepts a wide variety of input formats
TopHat (Trapnell et al., 2009)	splice junction mapping	Illumina	Aligns reads without relying on splice junction annotations
Cufflinks (Trapnell et al., 2010)	expression	Illumina	Can be used to detect differential expression of isoforms
GATK (DePristo et al., 2011)	SNP detection	454, Illumina	Well-documented SNP discovery tool Provides other useful features such as quality score recalibration
FreeBayes (https://github.com/ekg/freebayes)	SNP detection	454, Illumina	Improved version of PolyBayes for SNP detection
SAMtools (Li, Handsaker et al., 2009)	postprocessing, SNP detection	454, Illumina	Performs manipulation of sam/bam files
Varscan (Koboldt et al., 2009)	SNP detection	454, Illumina	Detects SNPs and indels

overlap assembly pipeline, and CLC (CLCBio), a commercial program, may produce the best transcriptome assemblies based on simulation studies and comparison to SOAPdenovo, Velvet, and MIRA with data from Cleome (Brautigam et al., 2011b).

Many contigs representative of one gene model are often produced in de novo assembly due to the presence of variant alleles, sequencing errors, and alternative splicing of transcripts. Additionally, sometimes contigs representing different regions of the same gene are not properly joined as a result of poor connection-supporting reads. Merging these contigs with a program such as CAP3 can be useful in obtaining gene models that represent the full-length gene as demonstrated by the de novo assembly of *Pteridium aquilinum* (Der et al., 2011).

Mapping assembly—If an appropriate, closely related reference is available, the reads can be mapped using either the reference sequence genome or coding sequences as a template. Mapping reads to the genome has the potential to be problematic when reads flank exon boundaries within a gene, so appropriate tools must be used. Alternatively, the reads may be mapped to a reference transcriptome, but in this case all splice variants of an alternatively spliced gene may not be represented.

Usually the preferred method is to use a tool capable of gapped alignment against the genome, followed by further analysis of splicing.

A popular algorithm used in short read mapping tools is the Burrows-Wheeler Transform (BWT) for string matching that allows for speed and efficiency (Burrows and Wheeler, 2011). BWT is implemented in the widely used aligners, Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) and Bowtie (Langmead et al., 2009). Since introns may interfere with the mapping of some reads that flank splice junctions, tools have been developed, such as TopHat (Trapnell et al., 2009) and SpliceMap (Au et al., 2010), for dealing with these reads (Table 2). Additionally, GSNAP is an aligner able to deal with long indels and splicing and also has the capability to map reads to a reference space rather than a single reference sequence (Wu and Nacu et al., 2010). By using a reference space, which accounts for all SNPs known to be found in the reference, SNPs are not penalized in the alignment as a mismatch (Wu and Nacu et al., 2010). For the longer reads of 454, some variant of the Smith–Waterman alignment (Smith and Waterman, 1981) algorithm is often implemented. BWA-SW (Li and Durbin, 2010) and Mosaik (Marth Laboratory) are some of the faster tools available for longer reads. Novoalign (Novocraft Technologies) is more sensitive than some of the fast aligners (http://novocraft.com/wiki/faq1). BLAT (Kent, 2002) is also still used in many cases although it is considerably slower than aligners designed specifically for shorter reads (Li and Durbin, 2010).

Quality control—Once an assembly is produced, measures of quality control to assess its accuracy and completeness are critical. Unfortunately, no commonly accepted methods currently exist, although some quality metrics have been proposed. For example, a reference set of transcripts containing a gradation of transcripts of known abundance and length could be used to assess accuracy, completeness, contiguity, chimerism, and variant resolution (Martin and Wang, 2011). At this time, a reference data set that fits the requirements is hard to find, and no tools have been developed to calculate these metrics although by combining known tools and scripts new tools can be developed. To determine contiguity in a de novo assembly, we can use the contigs as a BLAST query against an annotated genome to estimate gene coverage. Transcripts known to be unusually long can be used as an indicator of contiguity as well. Bedtools (Quinlan and Hall, 2010) is a useful program to determine coverage of features in a reference-guided assembly to an annotated genome. De novo assemblies should also be checked for chimeric contigs by using a tool such as BLAST (Altschul et al., 1997) to search against an annotated genome and identify unique regions that match two different gene models. A manual inspection of some of the alignments also can be useful. Viewers, such as Tablet (Milne et al., 2010), can be used to visually inspect alignments to ensure reads align normally.

TRANSCRIPTOMICS DOWNSTREAM ANALYSIS: WHAT TO DO AND HOW TO DO IT

Tools for NGS downstream analysis are in constant development, and it is likely that a program already exists for a user's analysis needs. A sampling of useful software for NGS analysis will be discussed in the following section.

***Transcriptome characterization and annotation*—**A common use of NGS is transcriptome characterization to gather a representation of the genes of a species of interest, especially if it is a nonmodel system where knowledge of the gene repertoire is not yet available. For example, this type of study has been performed in *Pachycladon enysii* (Collins et al., 2008), *Artemisia annua* (Wang et al., 2009), and olive (Alagna et al., 2009). Comparative transcriptomics is another approach. The first study to attempt this in a system with no previously developed genetics tools, compared the transcriptome of two species of mangrove (Dassanayake et al., 2009). A similar study was conducted to compare the transcriptome of closely related C_3 and C_4 species of Cleome (Bräutigam et al., 2011a).

Gene annotation is usually performed by using BLAST to find significant matches to annotated genes. Annotation by sequence similarity matches has some limitations, especially in plants where the number of manually annotated genes is lower than other species, and the manual annotations are generally associated with *Arabidopsis*. One of the most important limitations is that found in reference data sets, such as GenBank Reference Proteins (refseq_protein) where most of the plant genes are from species recently sequenced and without any useful annotation. For example, the first two results of a blastx homology search using the tomato gene Solyc10g081650 (a carotenoid isomerase) as a query against the refseq_protein database are two predicted proteins for *Populus trichocarpa* and *Arabidopsis lyrata* that have no associated additional information. Some solutions to this limitation are the use of combined annotations from different data sets or the use of tools that integrate different annotation approaches. Blast2Go (Conesa et al., 2005; Conesa and Götz, 2008; Götz et al., 2008) is sequence annotation software with an intuitive user graphic interface that combines different approaches. The software provides options such as annotation using GenBank BLAST and InterProScan and assigning gene ontology terms to each locus. In cases where there is no closely related reference species, annotation using predicted amino acid sequence to allow for divergence, has proven useful (Surget-Groba and Montoya-Burgos, 2010). In reference-guided alignments, existing annotation may be used, if available, to annotate the transcripts.

Gene annotation can be integrated with other information, such as metabolic pathways. Blast2Go integrates the metabolic pathways annotations using KEGG pathways visualization (Götz et al., 2008), but the information gained from this tool is limited to the enzyme lists created in the previous annotation steps. Pathway Tools is a software system to create, visualize and analyze organism-specific metabolic pathways databases (Karp et al., 2010). The nonintuitive usage is compensated by a wide range of functions supplied in system, such as metabolic pathway reconstruction, pathway hole filling, and the possibility to manually edit sequence and pathway annotations.

***SNPs*—**Detection of variation is a common application of RNA-seq. This variation can then be used to generate markers, allowing for a greater focus on expressed regions that may have a higher likelihood of phenotypic effect. SNP detection was a main goal in studies in *Eucalyptus grandis* (Novaes et al., 2008), *Brassica napus* (Trick et al., 2009), and *Scabiosa columbaria* (Angeloni et al., 2011). SNPs are also useful for looking at the evolution of genes such as the analysis of selection by Dn/Ds (Novaes et al., 2008). It is possible to construct microarrays based on SNP data to quickly detect polymorphism in many individuals in a population without further large-scale sequencing

(Clark et al., 2007). Also, multiplexing of samples can aid in efficient SNP identification. Tools, such as SAMtools (Li and Handsaker et al., 2009) and GATK (DePristo et al., 2011), are available to detect SNPs in reference-guided assemblies (Table 2). Gigabayes is useful for detecting SNPs in 454 data (Hillier et al., 2008). Again, when detecting SNPs in heterozygotes or duplicated genes, techniques must be used to determine true variants from sequencing error or mis-assemblies.

Comparative gene expression—Expression analysis is another important application of RNA-Seq. By looking at changes in gene expression between tissues, over time, or by treatments, a greater understanding of the genes critical in certain responses may be gained. For example, in an effort to gain insight on what qualities make *Amaranthus tuberculatus* such a successful weed, plants were treated with herbicides and cold stress to determine genes that may be involved in these responses (Riggins et al., 2010). In another case, comparative analysis of genes expressed in two species of chesnut, *Castanea dentata* and *Castanea mollissima*, infected with the fungus that causes chesnut blight, identified genes that may be involved in fungal resistance in *C. mollissima* (Barakat et al., 2009).

RNA-seq can also be used to quantify transcript levels in a tissue and is becoming the standard method to measure expression. It has proven to be accurate and sensitive, without the problem of background signal from nonspecific binding found in array-based measures of expression (Hoen et al., 2008). If a genomic reference sequence is available, expression can be quantified based on read counts by mapping to the reference sequence using a tool such as BWA (Li and Durbin, 2009). Then a tool such as TopHat (Trapnell et al., 2009) is used to map remaining reads to splice junctions. Cufflinks can then group transcripts into gene models and detect transcript abundance (Trapnell et al., 2010). An alternative approach taken by Scripture is to assembly the reads ab initio to reconstruct the transcriptome and then map the assembled reads to a reference genome (Guttman et al., 2010). For detection of differential expression, read counts must be normalized to account for varying sequencing depths between lanes of the flow cell. This can be performed by using the reads per kilobase of exon model per million mapped reads (RPKM) (Mortazavi et al., 2008). A newer alternative for reporting transcript abundances in RNA-seq experiments is to calculate the fragments per kilobase of exon per million fragments mapped (FPKM) as produced by CuffDiff (Trapnell et al., 2010). In this case, the transcript abundances are measured as normalized expected fragments, allowing for measurement of read counts from platforms that produce one or more reads per single source molecules (Trapnell et al., 2010). As such, FPKM is particularly useful for paired-end data and can be used with even a greater number of reads per molecule once the technology exists (Trapnell et al., 2010). RNA-seq can also be used to detect expression of alternatively spliced variants using tools such as MISO (Katz et al., 2010), Cufflinks (Trapnell et al., 2010), and SpliceMap (Au et al., 2010). In the case of many nonmodel systems, a genomic reference does not exist and models of expressed features must be built de novo or from a related species (Trick et al., 2009).

Transcript abundance and differential expression analysis follows a similar protocol when 454 reads are used. Typically, in nonmodel plant transcripomics where no reference genome exists, the reads are assembled de novo, and the number of reads per contig is used as an indicator of expression (Barakat et al., 2009; Alagna et al., 2009). Alternatively, they can be mapped to previously generated EST sequences or, in the case of an existing reference, the genes are mapped to the genome (Guo et al., 2010) or to predicted gene models (Swarbreck et al., 2011). A hybrid sequencing approach was demonstrated in *Tragopogon* that used 454 sequencing to assemble a reference transcriptome for *T. dubius* that was used to map Illumina reads from *T. miscellus*, *T. dubius*, and *T. pratensus* (Buggs et al., 2010). In 454, differential expression is normally detected by comparing the abundance of a gene in different libraries and performing a statistical test based on a log likelihood ratio method to ascertain the significance of differences (Stekel and Falciani, 2000). Any number of libraries can be used in this approach, and the test statistic generated is used to estimate the genes that have the most variable expression across the libraries (Stekel and Falciani, 2000).

Serial analysis of gene expression (SAGE), a method of generating tags that are used to measure gene expression, has benefited from NGS technology, resulting in an improved method known as superSAGE (Matsumura et al., 2010). A study in chickpea used superSAGE to detect genes that are differentially expressed between drought-stressed and nonstressed controls (Molina et al., 2008). Data gained from RNA-seq has also been used to construct microarrays to look at expression in a specific tissue (Bellin et al., 2009). In this study, 454 sequencing of a library made to gather a representative sample of genes expressed in *Vitis vinifera* berry was used to develop an array of the transcriptome that was used to look at gene expression in the berry (Bellin et al., 2009). Microarrays constructed from NGS data could be a useful and cost-effective method of looking at changes in gene expression from many biological samples.

SHARING WITH THE COMMUNITY: FORUMS AND DATA

Lastly, once the RNA-seq analyses have been performed, the data typically will be shared in some way through publications and databases. Due to the huge amounts of data generated by NGS, storage of data for community use can be problematic. Some journals, such as *Plant Physiology*, request that the researcher sends the large-scale data sets to permanent public repositories. The Sequence Read Archive (SRA) is one of the suggested databases (Leinonen et al., 2010a), although a recent announcement gave it a limited lifetime (http://www.ncbi.nlm.nih.gov/About/news/16feb2011). An alternative is the European Nucleotide Archive (ENA), at the European Bioinformatic Institute (EBI) (Leinonen et al., 2010b). Another option that has been proposed is the use of noncentralized repositories. Perhaps species-specific databases, such as TAIR (Rhee, 2003), Gramene (Liang et al., 2007), SGN (Bombarely et al., 2011), GDR (Jung et al., 2008), CuGenDB (http://www.icugi.org), Soybase (Grant et al., 2010), MaizeGDB (Lawrence et al., 2008), and Dendrome (Wegrzyn et al., 2008), could host NGS archives sharing a common search portal through web services.

Much more than just NGS data can be shared. The popularity of NGS has led to an increase in the production of new methodologies for assembly, annotation, and analysis as discussed above. Social tools such us web forums, blogs, and Facebook groups are being used to share knowledge about NGS topics. SEQanswers, a web portal to discuss questions about NGS, is a particularly useful site that allows NGS users to interact and discuss various related issues (http://seqanswers.com/).

COST/TIME AND RESOURCES

It is important to determine how much data are needed, for example, the number and types of replicates needed and how the data will be processed and managed. In some cases, a single lane of sequencing may be enough to address the questions of interest. As previously mentioned, another cost-reducing strategy involves the use of multiplexing. These options may be considered for laboratories with limited funding for a sequencing project.

Ultimately, someone must be responsible for the analysis of these data. If this analysis is to be done in-house, appropriate infrastructure with adequate CPU and memory must be available for the resource-demanding task of sequence assembly.

Additionally, many of the assemblers and tools for data analysis are Linux-based and can require a bit of computational expertise. Working with established databases and bioinformatics laboratories can help ameliorate some of the computational strain of transcripomics. Additionally, web-based resources, such as Galaxy (http://main.g2.bx.psu.edu/) and iPlant (http://www.iplantcollaborative.org/), are available for the upload and manipulation of NGS data. Galaxy provides a user friendly framework that integrates a large number of tools to manipulate and analyze NGS data, such as BWA, Bowtie, TopHat, and Cufflinks (Goecks et al., 2010). iPlant is a collaborative project that provides access to a world-class physical cyberinfrastructure with comprehensive hardware, such as cluster computation and massive data storage capabilities, as well as extensively used open source software tools and support through multidisciplinary teams (http://www.iplantcollaborative.org/about).

FUTURE PERSPECTIVES

Despite the many impressive feats that RNA-seq has helped accomplish, there is still room for improvement. Read lengths are less than ideal, and for this reason, NGS cannot yet completely replace Sanger sequencing. Most assembly programs are not written to efficiently allocate computational resources, so parallel processing needs to be better implemented. While tool output is converging on the sam format, further standarization of file formats is necessary. Additionally, problems in assembly and data analysis exist, such as the lack of a standard for quality control of the final assembly and the preferential detection of long transcripts in differential expression studies (Bullard et al., 2010). Currently, several efforts exist to remedy the NGS problems. To find new and better ways of assembling NGS data, the Assemblathon (http://assemblathon.org/) was organized, and we hope that other various efforts will be put forth to perfect assembly tools. Third-generation sequencing technologies are starting to offer several new platforms that entice with claims of longer sequence length, shorter run times, and greater accuracy, which will also ease the process of assembly. Future development of NGS technology and tools will no doubt allow NGS to continue to transform biology.

LITERATURE CITED

Alagna, F., N. D'Agostino, L. Torchia, M. Servili, R. Rao, M. Pietrella, G. Giuliano, et al. 2009. Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* 10: 399.

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

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

Au, K. F., H. Jiang, L. Lin, Y. Xing, and W. H. Wong. 2010. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research* 38: 4570–4578.

Auer, P. L., and R. W. Doerge. 2010. Statistical design and analysis of RNA sequencing data. *Genetics* 185: 405–416.

Balwierz, P. J., P. Carninci, C. O. Daub, J. Kawai, Y. Hayashizaki, W. Van Belle, C. Beisel, and E. van Nimwegen. 2009. Methods for analyzing deep sequencing expression data: Constructing the human and mouse promoterome with deepCAGE data. *Genome Biology* 10: R79.

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

Barbazuk, W. B., S. J. Emrich, H. D. Chen, L. Li, and P. S. Schnable. 2007. SNP discovery via 454 transcriptome sequencing. *Plant Journal* 51: 910–918.

Bellin, D., A. Ferrarini, A. Chimento, O. Kaiser, N. Levenkova, P. Bouffard, and M. Delledonne. 2009. Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species. *BMC Genomics* 10: 555.

Bombarely, A., N. Menda, I. Y. Tecle, R. M. Buels, S. Strickler, T. Fischer-York, A. Pujar, et al. 2011. The Sol Genomics Network (solgenomics.net): Growing tomatoes using Perl. *Nucleic Acids Research* 39: D1149–D1155.

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

Bräutigam, A., T. Mullick, S. Schliesky, and A. P. M. Weber. 2011a. 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., T. Mullick, S. Schliesky, and A. P. M. Weber. 2011b. Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C_3 and C_4 species. *Journal of Experimental Botany* 62: 3093–3102.

Buggs, R. J. A., S. Chamala, W. Wu, L. Gao, G. D. May, P. S. Schnable, D. E. Soltis, et al. 2010. Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology* 19: 132–146.

Bullard, J. H., E. Purdom, K. D. Hansen, and S. Dudoit. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11: 94.

Burrows, M., and D. J. Wheeler. 1994. A block-sorting lossless data compression algorithm. *Technical report* 124. Digital Equipment Corp., Palo Alto, California, USA.

Chepelev, I., G. Wei, Q. Tang, and K. Zhao. 2009. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Research* 37: e106.

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

Clark, R. M., G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, et al. 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317: 338–342.

Collins, L. J., P. J. Biggs, C. Voelckel, and S. Joly. 2008. An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome informatics. *International Conference on Genome Informatics* 21: 3–14.

Conesa, A., and S. Götz. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* 2008: 1–12.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676.

Cox, M. P., D. A. Peterson, and P. J. Biggs. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11: 485.

D'Agostino, N., M. Aversano, and M. L. Chiusano. 2005. ParPEST: A pipeline for EST data analysis based on parallel computing. *BMC Bioinformatics* 6: S9.

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

Denoeud, F., J. Aury, C. Da Silva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, et al. 2008. Annotating genomes with massive-scale RNA sequencing. *Genome Biology* 9: R175.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43: 491–498.

Der, J. P., M. S. Barker, N. J. Wickett, C. W. dePamphilis, and P. G. Wolf. 2011. De novo characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. *BMC Genomics* 12: 99.

Ekblom, R., and J. Galindo. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107: 1–15.

Emmert-Buck, M. R., R. F. Bonner, P. D. Smith, R. F. Chuaqui, Z. Zhuang, S. R. Goldstein, R. A. Weis, and L. A. Liotta. 1996. *Laser capture microdissection. Science* 274: 998–1001.

Emrich, S. J., W. B. Barbazuk, L. Li, and P. S. Schnable. 2006. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research* 17: 69–73.

Eveland, A. L., D. R. McCarty, and K. E. Koch. 2007. Transcript profiling by 3′-untranslated region sequencing resolves expression of gene families. *Plant Physiology* 146: 32–44.

Falgueras, J., A. J. Lara, N. Fernández-Pozo, F. R. Canton, G. Pérez-Trabado, and M. G. Claros. 2010. SeqTrim: A high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics* 11: 38.

Franssen, S. U., R. P. Shrestha, A. Bräutigam, E. Bornberg-Bauer, and A. P. Weber. 2011. Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. *BMC Genomics* 12: 227.

Goecks, J., A. Nekrutenko, and J. Taylor, The Galaxy Team. 2010. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11: R86–R98.

Gotoh, O. 2008. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Research* 36: 2630–2638.

Götz, S., J. M. García-Gómez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* 36: 3420–3435.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29:644-652.

Grant, D., R. T. Nelsen, S. B. Cannon, and R. C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research* 38: D843–D846.

Guo, S., Y. Zheng, J. Joung, S. Liu, Z. Zhang, O. R. Crasta, B. W. Sobral, et al. 2010. Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genomics* 11: 384.

Guttman, M., M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* 28: 503–510.

Hillier, L. W., G. T. Marth, A. R. Quinlan, D. Dooling, G. Fewell, D. Barnett, P. Fox, et al. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* 5: 183–188.

Hoen, P. T., Y. Ariyurek, and H. Thygesen. 2008. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research* 36: e141.

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

Jaillon, O., J. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463–467.

Jung, S., M. Staton, T. Lee, A. Blenda, R. Svancara, A. Abbott, and D. Main. 2008. GDR (Genome Database for Rosaceae): Integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Research* 36: D1034–D1040.

Karp, P. D., S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, et al. 2010. Pathway Tools version 13.0: Integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics* 11: 40–79.

Katz, Y., E. T. Wang, E. M. Airoldi, and C. B. Burge. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* 7: 1009–1015.

Kent, W. J. 2002. BLAT—The BLAST-Like Alignment Tool. *Genome Research* 12: 656–664.

Koboldt, D. C., K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, et al. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25: 2283–2285.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10: R25.

Lawrence, C. J., L. C. Harper, M. L. Schaeffer, T. Z. Sen, T. E. Seigfried, and D. A. Campbell. 2008. MaizeGDB: The maize model organism database for basic, translational, and applied research. *International Journal of Plant Genomics* .

Leinonen, R., R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tarraga, Y. Cheng, I. Cleland, et al. 2010b. The European Nucleotide Archive. *Nucleic Acids Research* 39: D28–D31.

Leinonen, R., H. Sugawara, and M. Shumway [on behalf of the International Nucleotide Sequence Database Collaboration]. 2010a. The Sequence Read Archive. *Nucleic Acids Research* 39: D19–D21.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589–595.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Liang, C., P. Jaiswal, C. Hebbard, S. Avraham, E. S. Buckler, T. Casstevens, B. Hurwitz, et al. 2007. Gramene: A growing plant comparative genomics resource. *Nucleic Acids Research* 36: D947–D953.

Maher, C. A., C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, et al. 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458: 97–101.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–380.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, et al. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7: 111–118.

MARTIN, J. A., AND Z. WANG. 2011. Next-generation transcriptome assembly. *Genetics* 12: 671–682.

MASOUDI-NEJAD, A., K. TONOMURA, S. KAWASHIMA, Y. MORIYA, M. SUZUKI, M. ITOH, M. KANEHISA, ET AL. 2006. EGassembler: Online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. *Nucleic Acids Research* 34: W459–W462.

MATSUMURA, H., K. YOSHIDA, S. LUO, E. KIMURA, T. FUJIBE, Z. ALBERTYN, R. A. BARRERO, ET AL. 2010. High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. *PLoS ONE* 5: e12010.

MILLER, J. R., S. KOREN, AND G. SUTTON. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* 95: 315–327.

MILNE, I., M. BAYER, L. CARDLE, P. SHAW, G. STEPHEN, F. WRIGHT, AND D. MARSHALL. 2010. Tablet–next generation sequence assembly visualization. *Bioinformatics* 26: 401–402.

MIZRACHI, E., C. A. HEFER, M. RANIK, F. JOUBERT, AND A. A. MYBURG. 2010. De novo assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* 11: 681.

MOLINA, C., B. ROTTER, R. HORRES, S. M. UDUPA, B. BESSER, L. BELLARMINO, M. BAUM, ET AL. 2008. SuperSAGE: The drought stress-responsive transcriptome of chickpea roots. *BMC Genomics* 9: 553.

MORTAZAVI, A., B. WILLIAMS, K. MCCUE, AND L. SCHAEFFER. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5: 621–628.

NOVAES, E., D. R. DROST, W. G. FARMERIE, G. J. PAPPAS, D. GRATTAPAGLIA, R. R. SEDEROFF, AND M. KIRST. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9: 312.

OHTSU, K., H. TAKAHASHI, P. S. SCHNABLE, AND M. NAKAZONO. 2006. Cell type-specific gene expression profiling in plants by using a combination of laser microdissection and high-throughput technologies. *Plant & Cell Physiology* 48: 3–7.

PERTEA, G., X. HUANG, F. LIANG, V. ANTONESCU, R. SULTANA, S. KARAMYCHEVA, Y. LEE, ET AL. 2003. TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics* 19: 651–652.

PEVZNER, P. A., H. TANG, AND M. S. WATERMAN. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences, USA* 98: 9748–9753.

POP, M., AND S. L. SALZBERG. 2008. Bioinformatics challenges of new sequencing technology. *Trends in Genetics* 24: 142–149.

QUINLAN, A. R., AND I. M. HALL. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.

RHEE, S. Y. 2003. The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Research* 31: 224–228.

RIGGINS, C. W., Y. PENG, C. N. STEWART JR., AND P. J. TRANEL. 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 Management Science* 66: 1042–1052.

ROUGEMONT, J. A., C. AMZALLAG, L. ISELI, I. FARINELLI, I. XENARIOS, AND F. NAEF. 2008. Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* 9: 431.

SCHMIEDER, R., AND R. EDWARDS. 2011. Fast identification and removal of sequence contamination from genomics and metagenomic datasets. *PLoS ONE* 6: e17288.

SCHMIEDER, R., Y. W. LIM, F. ROHWER, AND R. EDWARDS. 2010. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* 11: 341.

SCHMUTZ, J., S. B. CANNON, J. SCHLUETR, J. MA, T. MITROS, W. NELSON, D. O. HYTEN, ET AL. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178–183.

SCHRÖDER, J., J. BAILEY, T. CONWAY, AND J. ZOBEL. 2010. Reference-Free Validation of Short Read Data. *PLoS ONE* 5: e12681.

SMITH, A. M., L. E. HEISLER, R. P. ST ONGE, E. FARIAS-HESSON, I. M. WALLACE, J. BODEAU, A. N. HARRIS, ET AL. 2010. Highly-multiplexed barcode sequencing: An efficient method for parallel analysis of pooled samples. *Nucleic Acids Research* 38: e142.

SMITH, T. F., AND M. S. WATERMAN. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147: 195–197.

SOLTIS, D. E., V. A. ALBERT, J. LEEBENS-MACK, C. D. BELL, A. H. PATERSON, C. ZHENG, D. SANKOFF, ET AL. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336–348.

STEKEL, D. J., AND F. FALCIANI. 2000. The comparison of gene expression from multiple cDNA libraries. *Genome Research* 10: 2055–2061.

SURGET-GROBA, Y., AND J. I. MONTOYA-BURGOS. 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Research* 20: 1432–1440.

SWARBRECK, S. M., E. A. LINDQUIST, D. D. ACKERLY, AND G. L. ANDERSEN. 2011. Analysis of leaf and root transcriptomes of soil-grown *Avena barbata* plants. *Plant & Cell Physiology* 52: 317–332.

TANG, H., J. E. BOWERS, X. WANG, AND A. H. PATERSON. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences, USA* 107: 472–477.

TRAPNELL, C., L. PACHTER, AND S. L. SALZBERG. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.

TRAPNELL, C., L. PACHTER, AND S. L. SALZBERG. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28: 511–515.

TRICK, M., Y. LONG, J. MENG, AND I. BANCROFT. 2009. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnology Journal* 7: 334–346.

WALL, P. K., J. LEEBENS-MACK, A. S. CHANDERBALI, A. BARAKAT, E. WOLCOTT, H. LIANG, L. LANDHERR, ET AL. 2009. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* 10: 347.

WANG, E. T., R. SANDBERG, S. LUO, I. KHREBTUKOVA, L. ZHANG, C. MAYR, S. F. KINGSMORE, ET AL. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470–476.

WANG, W., Y. WANG, Q. ZHANG, Y. QI, AND D. GUO. 2009. Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* 10: 465.

WEGRZYN, J. L., J. M. LEE, B. R. TEARSE, AND D. B. NEALE. 2008. Tree Genes: A forest tree genome database. *Internationl Journal of Plant Genomics*. doi:.

WU, T. D., AND S. NACU. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26: 873–881.

ZERBINO, D. R., AND E. BIRNEY. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18: 821–829.