Discovering New Biology through Sequencing of RNA[1]

Andreas P.M. Weber*

Institute of Plant Biochemistry, Cluster of Excellence on Plant Science, Heinrich-Heine-Universität, D–40231 Duesseldorf, Germany

ORCID ID: 0000-0003-0970-4672 (A.P.M.W.).

Sequencing of RNA (RNA-Seq) was invented approximately 1 decade ago and has since revolutionized biological research. This update provides a brief historic perspective on the development of RNA-Seq and then focuses on the application of RNA-Seq in qualitative and quantitative analyses of transcriptomes. Particular emphasis is given to aspects of data analysis. Since the wet-lab and data analysis aspects of RNA-Seq are still rapidly evolving and novel applications are continuously reported, a printed review will be rapidly outdated and can only serve to provide some examples and general guidelines for planning and conducting RNA-Seq studies. Hence, selected references to frequently update online resources are given.

Sequencing of RNA (RNA-Seq) is a recent technique that emerged shortly after next-generation sequencing (NGS) was invented approximately 10 years ago and since has revolutionized biological research in the 21st century. The major advance and basis of NGS is the application of sequencing-by-synthesis technology, which entails real-time monitoring of de novo DNA biosynthesis by imaging methods and reading out the sequence of newly synthesized DNA molecules upon iterative addition of the four different nucleotides. This is in contrast to sequencing after synthesis, which is based on the physical separation of differently sized DNA molecules generated by the chain termination inhibitor method in polyacrylamide gels or by capillary electrophoresis after completion of the sequencing reaction (Sanger et al., 1977).

Most of the current sequencing-by-synthesis technologies are based on the immobilization of a denatured, single-stranded sequencing template on a surface, either a glass slide or nano beads. Immobilization on a surface allows for repeated cycles of reagent delivery to the immobilized DNA molecule, which permits solid-phase oligonucleotide primer-initiated synthesis of a new DNA strand, using repetitive and iterative cycles of addition of the nucleotides A, C, G, and T. High-resolution imaging is used to detect the incorporation of the nucleotide, either during or after nucleotide incorporation, followed by iterative additional rounds of nucleotide incorporation. The sequence is then eventually deduced from the imaging data.

The first successful NGS approach that gained wide acceptance by the community was 454 sequencing, a massively parallel pyrosequencing approach (Margulies et al., 2005). 454 Sequencing is based on the detection of pyrophosphate released during de novo synthesis of a new DNA strand by DNA polymerase, which allows real-time measurements of DNA biosynthesis (Ronaghi et al., 1998). Pyrophosphate released during DNA synthesis is converted to ATP by the action of sulfurylase, followed by generation of a luminescent light signal from ATP, using firefly luciferase. The major advance in 454 technology was combining pyrosequencing with immobilization of the DNA template to nano beads to allow for solid-phase DNA pyrosequencing. The immobilized DNA template is amplified by emulsion PCR and then combined with beads carrying immobilized sulfurylase and firefly luciferase enzymes, followed by loading into picotiter glass plates that are subsequently inserted into the sequencing machine. A reagent delivery system then iteratively floods the plates with nucleotides, DNA polymerase, and oxyluciferin. Inorganic pyrophosphate released during incorporation of a nucleotide into a newly synthesized DNA strand is converted into a light signal via sulfurylase/luciferase, which is recorded by a high-resolution and very sensitive camera system. Remaining inorganic pyrophosphate is destroyed by a wash cycle with apyrase, then a new round of nucleotide incorporation occurs. Initially, this method delivered approximately 250,000 reads with approximately 100-nucleotide (nt) read length, which was a massive progress in throughput over established Sanger sequencing methods. Later versions of this technology provided long reads of 400-nt lengths and over 1 million reads per run. Initial applications of the new sequencing technology included the sequencing of ancient genomic DNA, for example, from Neanderthals (Green et al., 2006; Noonan et al., 2006) and the wooly mammoth (Poinar et al., 2006).

In the meantime, 454 pyrosequencing has been mainly superseded by Illumina sequencing, which combined chain termination technology with immobilization of the sequencing template on a glass surface, an extension of in situ fluorescence sequencing (Mitra et al., 2003; Shendure et al., 2005). In this technology,

[1] This work was supported by the Deutsche Forschungsgemeinschaft (grant nos. WE2231/8–2, WE22319–2, IRTG 1525, and EXC 1028).

* Address correspondence to andreas.weber@hhu.de.

www.plantphysiol.org/cgi/doi/10.1104/pp.15.01081

DNA molecules immobilized on a glass surface are amplified by bridge amplification, followed by synthesis of new DNA strands using four differently colored fluorescently labeled chain terminators (Mardis, 2008). After each cycle of DNA synthesis, the newly incorporated nucleotides are detected by fluorescence color imaging, followed by removal of the fluorophore and the blocked 3′ terminus of the terminal nucleotide. Then follow iterative new rounds of nucleotide incorporation, imaging, fluorophore removal, and 3′ end deblocking. Initially, this method allowed for read lengths of 25 nt, whereas the most recent versions of the technology enable read lengths of 300 nt on MiSeq machines and 150 nt on the HiSeq instruments (Illumina Inc.). One cycle on an HiSeq instrument delivers up to 5 billion reads, which is sufficient for approximately 500 RNA-Seq reactions, assuming 10 million reads are required per sample to achieve saturating coverage. Due to its enormous throughput, as of September 2015, Illumina is currently the dominant technology in the RNA sequencing market.

Although single-molecule direct sequencing of DNA molecules was demonstrated more than 10 years ago (Braslavsky et al., 2003) and was later applied in a proof-of-concept study to the quantification of the yeast (*Saccharomyces cerevisiae*) transcriptome by single-molecule RNA sequencing (Lipson et al., 2009), the attempt to commercially introduce this technology by Helicos was unsuccessful. Pacific Biosciences (PacBio) has developed a commercially successful platform for single-molecule real-time sequencing that provides very long read length, but currently does not provide sufficient read numbers for quantitative transcriptomics. It has to be mentioned, though, that the PacBio single-molecule long-read technology is extremely helpful for the de novo generation of reference transcriptomes.

In the absence of a commercially viable direct RNA-Seq method, to date, sequencing of RNA is based on the conversion of RNA into DNA molecules by reverse transcription, followed by amplification of the DNA template using liquid- and/or solid-phase PCR methods. That is, when we speak of RNA-Seq, we really mean sequencing of reverse-transcribed RNA, which is an important difference since the process of reverse transcription and amplification might introduce bias into the analysis, such as suppression of sequences having a higher G/C contents or containing long homopolymer stretches. That said, RNA-Seq has now mostly superseded previous technologies for transcriptome analysis, such as serial analysis of gene expression and microarrays, for a number of reasons:

(1) RNA-Seq is not dependent on prior sequence knowledge (i.e. it can be applied to any system from which RNA can be isolated in sufficient quality and quantity). In contrast, the design of microarrays depends on prior sequence information, be it from genome sequencing or sequencing of expressed-sequence tags.
(2) RNA-Seq provides a direct measure of RNA abundance in contrast to microarrays, which provide relative fluorescence intensities. Hence, it is rather difficult to compare the results of microarrays between laboratories, whereas this is more straightforward with RNA-Seq data.
(3) RNA-Seq enables simultaneous sequence discovery and quantitation.
(4) RNA-Seq provides a dynamic range at least 2 orders of magnitude larger than microarrays, which allows for the quantitation of low-abundance transcripts in the presence of highly abundant transcripts, given sufficient depth of sequencing.
(5) RNA-Seq allows for the detection of sequence variants, which enables analysis of allele-specific expression in heterozygous individuals and the detection of sequence variants between individuals.
(6) Recent instruments enable highly multiplexed sequencing of hundreds of bar-coded RNA-Seq samples in a single run, which makes RNA-Seq relatively economic.

According to an ISI Web of Science search in July 2015, the first publications containing the keyword RNA-sequencing appeared in 2008, and since then, close to 7,000 manuscripts containing this keyword have been published. However, the first manuscripts on RNA-Seq not yet using this term had been published before, for example, the pioneering manuscripts on the transcriptomes of prostate cancer cell lines (Bainbridge et al., 2006), *Medicago truncatula* (Cheung et al., 2006), maize (*Zea mays*; Emrich et al., 2007), and Arabidopsis (*Arabidopsis thaliana*; Weber et al., 2007). The two latter studies benchmarked RNA-Seq data against previous expressed-sequence tag and microarray work and concluded that transcriptome analysis by sequencing methods will soon replace these previous methods.

APPLICATIONS OF RNA-SEQ

As outlined above, RNA-Seq almost always involves the conversion of RNA to DNA by reverse transcription before sequencing. This sets the frame and requirements for RNA-Seq: pretty much any RNA sample that can be isolated with sufficient quality and purity to allow for subsequent reverse transcription to DNA is suitable for analysis by RNA-Seq. For most steps in preparing RNA-Seq libraries, commercial kits and reagent sets with detailed and reliable protocols are available, and the actual sequencing reactions are frequently conducted by central facilities or commercial suppliers. Hence, this update will mostly focus on steps preceding RNA-Seq library preparation and on post-sequencing analysis. For a detailed primer on differential gene expression analysis by RNA-Seq covering

aspects of experimental planning, library preparation, and details of data analysis, the reader is referred to the recent reviews by Külahoglu and Bräutigam (2014) and Griffith et al. (2015).

QUALITATIVE ANALYSIS OF RNA-SEQ DATA: ASSEMBLY OF TRANSCRIPTOMES FROM RNA-SEQ READS

In general, two major types of analyses are conducted on RNA-Seq data: assembly of reads into contiguous sequences (contigs) and mapping of reads to a reference either to obtain an account of transcript amounts or to verify/modify gene models or discover splice or sequence variants.

Of these two types of analyses, the assembly of RNA-Seq data into contigs, in particular, de novo assembly from short reads without a guiding reference, is still problematic (Schliesky et al., 2012). In principle, two different assembly strategies exist: overlap-based assemblers such as CAP3 (Huang and Madan, 1999) and De Bruijn graph-based assemblers, such as Velvet/Oases and Trinity. In our experience, overlap-based assemblers tend to produce good assemblies and a relatively low number of high-quality contigs. However, overlap-based assemblers are computationally expensive and not applicable to large numbers of short reads. De Bruijn graph-based assemblers are computationally efficient but tend to produce inflated numbers of contigs from short-read sequence data, in particular for highly expressed transcripts (Bräutigam et al., 2011b; Schliesky et al., 2012). A large number of studies comparing different assembly strategies have been published, and it is difficult to distill a straightforward recommendation on which algorithm to use for short-read assemblies. As a first try, established tools such as Trinity (Grabherr et al., 2011; Haas et al., 2013) and Velvet/Oases (Schulz et al., 2012) will provide a good start, in particular given the very good tutorials and manuals available for these assemblers. Perhaps more critical than the algorithm for assembly is the experimental design: here, less data might be better than more. That is, instead of trying to de novo assemble a reference transcriptome from multiple replicated short-read samples, it is highly recommended to generate a separate long-read paired-end sequencing run on a library consisting of RNAs isolated from a broad range of different cell types or tissues. That is, the quality and coverage of the reference transcriptome is improved by generating a mixed library including a balanced amount of RNAs from tissues with different functions, such as leaves, stems, roots, flower organs, and developing seeds. In the case of Illumina sequencing, best assemblies are obtained from long paired-end reads, for example, 2× 300-nt reads on a MiSeq instrument. The recently developed full-length transcript sequencing method by PacBio in principle circumvents the assembly step since it provides full-length sequences of single complementary DNA molecules, albeit with low sequence accuracy. Either sufficient sequencing depth for error correction or error correction using short reads obtained with other sequencing technologies is needed to generate the accurate sequence of the full-length transcript (Sharon et al., 2013; Tilgner et al., 2014). Assembly of reference transcriptomes from short reads is hampered by sequence variants (as expected from heterozygous individuals and allopolyploids), which is not the case with single-molecule full-length sequencing. Hence, it is recommended that this technology be taken into consideration when planning RNA-Seq experiments on species without a sequenced reference genome. It is likely that the extra cost for generation of additional libraries and sequencing runs will be amortized by more straightforward downstream data analysis and lower bioinformatics costs.

QUANTITATIVE ANALYSIS OF RNA-SEQ DATA: ESTIMATING TRANSCRIPT AMOUNTS FROM RNA-SEQ READS

Quantifying transcript amounts using RNA-Seq data requires aligning of the RNA-Seq reads to a reference (genome or reference transcriptome), counting the reads per feature, followed by differential gene expression analysis. Again, as for contig assembly from RNA-Seq reads, multiple programs and algorithms are available for these tasks. For RNA-Seq data coming from a species with a sequenced genome, the choice of the reference for read mapping is straightforward. However, in nonmodel species without sequenced genomes, several choices are available. Either the reads are mapped to a reference transcriptome generated from this species or they are mapped to the genome (or reference transcriptome) from a related species. Neither approach is perfect since a reference transcriptome might be incomplete, and hence a number of reads might not be mappable. A related genome reference might lack genes that are present in species of choice, which are hence not detected in the mapping, and the mapping efficiency might be low due to sequence divergence. A further complication comes into play if transcript amounts are to be compared between species (and not between cell types, tissues, etc. of a single species). Gene family sizes might differ between the compared species, and possible bias in mapping efficiencies might exist for individual genes. Also, the issue of calling true orthologs makes cross-species comparisons more complicated. Informatics methods using machine learning approaches have been developed to overcome these issues, for example, to compare gene expression data across a broad data set spanning more than 140 million years of evolutionary separation (Aubry et al., 2014a). This method makes use of a unique method for orthology assignments, thereby improving the abundance estimates for de novo assembled transcripts, even across large evolutionary distances. An alternative approach (which is frequently used in our lab) is cross-species mapping of reads to a

common reference genome, for example, Arabidopsis (Bräutigam et al., 2011a, 2014; Gowik et al., 2011; Külahoglu et al., 2014; Mallmann et al., 2014). In this approach, reads are mapped in the protein space (i.e. after translation of read sequences into all six reading frames and then mapping them to a protein database, such as the Arabidopsis proteome, using the BLAT tool; Kent, 2002). Mapping in protein space enables mapping of reads to a relatively distant reference proteome since protein sequences show lower rates of divergence than nucleotide sequences and the BLAST-like alignment tool BLAT allows for more mismatches than NGS mapping tools such as Bowtie or Tophat. The main caveats of this approach are that genes that are not present in the mapping reference will not be called, and that a mapping bias might exist if one of the mapped species is closer to the reference than the other. On the upside, downstream data analysis, such as functional assignments, is facilitated by making use of a well-annotated reference, such as the Arabidopsis genome.

Once a decision on the mapping reference has been made, the next steps are trimming and quality control of reads (for example, using the FASTX toolkit), mapping of reads to the reference, followed by calling of differentially expressed genes. A large number of protocols, manuals, and tutorials are available for this; hence the details are not discussed here. Instead, a few good starting points are listed in Table I. Following the Bioconductor rnaseqGene example or the iPlant Collaborative RNA-Seq tutorial, which within the Discovery Environment is using Tophat for mapping of reads to the reference and CuffLinks/CuffDiff for calling differentially expressed genes (see Table I for Web links) will provide sufficiently detailed instructions to enable independent analysis of own data sets. RobiNA (see Table I for link) provides a user-friendly graphical interface to the R/Bioconductor packages typically used in the analysis of RNA-Seq data and enables straightforward downstream functional analysis of differentially expressed genes using the MapMan tool (Usadel et al., 2009; Lohse et al., 2012). It is emphasized that it is crucial to understand the differences between various methods for aligning reads and calling differentially expressed genes, such as edgeR, DESeq, or CuffDiff. Using different algorithms will lead to different lists of differentially expressed genes; scientific reasoning is required to interpret these differences and make the right choice for analysis of own data sets. When performing large numbers of statistical tests, as is the case in differential expression analysis, correction for false-discovery rates must be performed, which depending on the method used will influence the power to detect true positives and the number of false positives. It is hence important to understand the concepts of the correction methods used to interpret the outcome and to choose the most appropriate method. A helpful and brief discussion of this aspect is given in Krzywinski and Altman (2014). The Web provides comprehensive information and documentation, as well as helpful blogs and tutorials. The SEQanswers wiki (see Table I) is recommended as a starting point for more information (Li et al., 2012).

GENE EXPRESSION ANALYSIS BY RNA-SEQ: A PROXY FOR TRANSCRIPTIONAL ACTIVITY AND PROTEIN AMOUNTS?

RNA-Seq is most frequently used to quantify RNA steady-state amounts. The goal of this type of analysis is obtaining a quantitative account of transcript amounts in organisms, organs, tissues, or specific cell types, frequently comparing transcript amounts between different samples, such as cell types, mutants and wild type, or response to certain treatments. Typically in these studies, total RNA is extracted from the sample of choice, either enriched for poly-adenylated mRNAs or depleted from ribosomal RNA, and then subjected to sequencing. Although this approach is highly successful in quantifying transcript amounts and in identifying differentially expressed genes, a valid point of critique is that high transcript amounts do not necessarily reflect the rate of gene expression or protein amounts. Both

Table I. *Useful resources for RNA-Seq*

Collection of frequently updated online resources and starting points for experimental protocols and tutorials for data analysis.

Web Site	URL	Description
RNA-seqlopedia	http://rnaseq.uoregon.edu	Comprehensive overview on all aspects of RNA-Seq, from experimental design to data analysis
SEQanswers	http://seqanswers.com/forums/	Online community on all aspects of NGS
RNA-Seq bioinformatics tools wiki	https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools	Comprehensive, frequently updated, and annotated collection of RNA-Seq bioinformatics tools
FASTX Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/	Collection of tools for preprocessing of FASTX/FASTQ files
RobiNA	http://mapman.gabipd.org/web/guest/robin	User-friendly open source graphical interface to RNA-Seq data analysis
RNA-Seq Tutorial	https://pods.iplantcollaborative.org/wiki/pages/viewpage.action?pageId=10659468	RNA-Seq tutorial by the iPlant collaborative wiki
RNA-Seq analysis in the Cloud	https://github.com/griffithlab/rnaseq_tutorial/wiki	Comprehensive tutorial on RNA-Seq in the Cloud, with step-by-step instructions (Griffith et al., 2015)
Bioconductor rnaseqGene	http://www.bioconductor.org/help/workflows/rnaseqGene/	A detailed workflow for differential gene expression analysis using DESeq2

points of critique can be addressed by variations of the RNA-Seq theme: gene expression rates can be estimated by native elongating transcript sequencing, which is based on immunoprecipitation of RNA polymerase II, followed by RNA-Seq of the 3′-end of RNA protected by the active site of polymerase II (Mayer et al., 2015). In addition, this approach also provides insights into the regulation of transcriptional activity at the level of polymerase II posttranslational modification by phosphorylation (Nojima et al., 2015). An alternative strategy to assess transcriptional and posttranscriptional regulation is exon-intron split analysis, which exploits the deep sequence coverage to detect intronic reads that can be used as proxy for nascent transcript amounts. A recent study showed that changes in intronic read counts directly measure changes in transcriptional activities (Gaidatzis et al., 2015). By using this approach across a range of different experimental conditions, it becomes possible to distinguish transcriptional and posttranscriptional effects on steady-state RNA levels directly from RNA-Seq data. It has to be noted that appropriate quality controls are required to distinguish true intronic reads coming from premRNAs from reads resulting from contamination with genomic DNA. In addition, reference gene models defining intron-exon borders are required.

Approximations on translation of transcripts to proteins can be obtained by translating ribosome affinity purification in combination with RNA-Seq (Reynoso et al., 2015). By affinity purification of 80S ribosomes, mRNAs associated with the ribosome are pulled down and can then be quantified by RNA-Seq, providing a snapshot of mRNAs that are likely to be actively translated. Using cell-specific promoters to drive the expression of ribosomal subunits enables insights into cell-specific translatomes (Zanetti et al., 2005; Mustroph et al., 2009). For example, driving bundle sheath-specific expression of a FLAG-tagged ribosomal protein L18 in Arabidopsis showed that this cell type in C_3 plants plays a specific role in sulfur metabolism and transport as well as in the biosynthesis of glucosinolates and in trehalose metabolism (Aubry et al., 2014b).

Importantly, though, at least for maize leaves, a good correlation between transcript and protein abundance was found (Ponnala et al., 2014). Although control mechanisms such as protein and transcript stability as well as translational control were found to have significant effects, the mRNA amount was shown to be the major factor influencing protein abundance (Ponnala et al., 2014). Hence, as a first approximation, quantification of mRNA does provide an estimate of relative protein abundance, with high mRNA amounts correlating with high protein amounts. It is also worth mentioning that, despite massive progress in mass spectrometric analysis of proteomes, quantification of mRNAs can be conducted with higher throughput and sensitivity at lower cost than measuring protein amounts by mass spectrometry. For species with unknown genomes, RNA-Seq actually is a prerequisite for proteomic analyses since high-throughput proteomics depends on sequence databases for peptide identification (Bräutigam et al., 2008a, 2008b; Schulze et al., 2012). In the long term, concerted measurements of RNA transcription and decay rates, rates of protein translation and degradation, as well as transcription factor binding and chromatin state will be needed to obtain a comprehensive picture of the intricate interplay of multiple factors involved in regulating transcript and protein abundance.

RNA-SEQ AS ENABLING TOOL IN NONMODEL SPECIES

Although long-read sequencing technologies such as PacBio's SMRT technology are now facilitating genome sequencing and assembly, large and complex plant genomes are still difficult to sequence and assemble. RNA-Seq provides a relatively fast and economic tool for gene discovery and for gene expression quantification in species without a sequenced genome, which enables exciting new insights into plant metabolism, crop domestication, and development. This strategy has been particularly successful in the discovery of unknown enzymes and regulators in metabolic pathways, for example, in plant-specialized metabolism, such as medicinally relevant monoterpene indole alkaloids from Asterids (Góngora-Castillo et al., 2012) or sesquiterpenes in tomato (*Solanum lycopersicum*; Schilmiller et al., 2010). Novel components of xyloglucan biosynthesis have been discovered by conducting an RNA-Seq time series of seed development in *Tropaeolum majus* (Jensen et al., 2012), and comparative time-resolved RNA-Seq of seed development in four different plant species revealed communalities and specifics of glycerolipid biosynthesis in oil seeds (Troncoso-Ponce et al., 2011). Identification of genes relevant to the function of C_4 photosynthesis was achieved by comparative RNA-Seq of related C_3 and C_4 species (Bräutigam et al., 2011a, 2014; Gowik et al., 2011; Aubry et al., 2014a; Külahoglu et al., 2014; Wang et al., 2014), of developmental time series within one species (Li et al., 2010; Wang et al., 2013), and comparative analysis of cell-specific expression patterns between species (Aubry et al., 2014b; John et al., 2014). For example, comparative RNA-Seq of mature leaves of C_3 and C_4 plant species has led to the discovery of the gene encoding the plastidial sodium:pyruvate transporter that is required for the biosynthesis of phospho*enol*pyruvate in the stroma of mesophyll cell chloroplasts (Furumoto et al., 2011). The above-mentioned studies also provided a large number of candidate genes that might be involved in controlling metabolic or anatomical aspects of the C_4 trait. As an example for the latter, comparison of leaf developmental time series between non-Kranz husk leaves and Kranz-type foliar leaves of maize revealed a regulatory network containing the transcriptional regulators SCARECROW and SHORTROOT that is involved in patterning Kranz anatomy (Wang et al., 2013; Fouracre et al., 2014).

RNA-Seq can also be applied to better understand the molecular mechanisms and genetic consequences of crop plant domestication and breeding and thereby provide novel leads for crop improvement and the design of breeding and prebreeding programs. For example, it was shown that during the domestication of common bean in Mesoamerica, drastic changes in the pattern and structure of gene expression occurred, with overall lower diversity of gene expression patterns and a general down-regulation of gene expression levels in the domesticated variants as compared with the ancestral species. In this case, the loss of genetic diversity during domestication was directly associated with a reduced diversity of gene expression patterns (Bellucci et al., 2014). Comparative analysis of gene expression during cotton fiber development in four wild and five cultivated accessions of cotton revealed, among other important findings, that human selection during domestication has led to a prolonged duration of fiber elongation. Also, the wild accessions allocate a larger part of their transcriptional investment to stress-response pathways, whereas the domesticated species allocate more to growth-related processes (Yoo and Wendel, 2014). Again, comparison of wild ancestors with cultivated accessions showed a dramatic and wide-ranging rewiring of the transcriptome as a consequence of domestication (Yoo and Wendel, 2014). Transcriptomic comparison of wild and cultivated tomato accessions identified hundreds of thousands of polymorphic positions between ancestral and domesticated variants. By including in these comparisons wild ancestors adapted to a highly diverse range of habitats, including the desert-adapted *Solanum pennellii*, it became possible to distinguish effects of natural and artificial selection at a genomic scale (Koenig et al., 2013). Comparing the transcriptome of the grapevine (*Vitis vinifera*) 'Tannat' cultivar with that of the cv Pinot Noir reference identified close to 2,000 unique genes that are not present in the cv Pinot Noir reference genome. Functional annotation of these genes revealed an expansion of genes encoding enzymes involved in polyphenol biosynthesis (Da Silva et al., 2013). Berries of cv Tannat produce unusually high amounts of polyphenolic compounds, some of which have been associated with longevity and promotion of vascular health in humans (Corder et al., 2006). Quantitative gene expression analysis during berry development showed that the cv Tannat-specific polyphenol biosynthesis genes contributed strongest to the overall transcriptional investment into polyphenol biosynthesis, indicating that the specific and potentially health-promoting properties of the cv Tannat berries are a consequence of a unique gene set in this cultivar (Da Silva et al., 2013).

Key to the success of these and other studies is the selection of species or cultivar, tissues, developmental stage, or cell types to compare as well as the procedures for (statistical) data analysis used to extract the relevant information from the large data sets obtained by RNA-Seq. That is, the more prior knowledge is available on the system of choice, the better the experimental design and eventually the outcome of the RNA-Seq experiment will be. In addition, the RNA-Seq data were frequently contextualized with anatomical data (for example, microscopy of developing leaves), metabolic and enzymatic data, or proteomic data. Meta-analyses of RNA-Seq data with other data domains facilitate the discovery of genes of interest by correlative approaches, such as weighted gene correlation network analysis ((Langfelder and Horvath, 2008, 2012) or linear models (Brady et al., 2015). Therefore, as outlined above, exploring and understanding the procedures for data analysis before designing an RNA-Seq experiment is highly recommended because the requirements of the data analysis routines will influence the range of parameters to be measured in the experiment. In most cases it will be very difficult, if not impossible, to obtain the relevant data after the RNA-Seq experiment is conducted, so good planning is key to success.

CAVEATS IN RNA-SEQ

As with any experimental approach, the quality and reliability of data obtained in RNA-Seq experiments are influenced by a large number of variables that need to be controlled to avoid erroneous results. Recent large-scale studies on RNA-Seq (Kratz and Carninci, 2014; SEQC/MAQC-III Consortium, 2014) have shown that substantial variation of resulting data exists when identical samples are run in different laboratories or on different instruments, and the procedures of library preparation and sample processing influence the outcome. It is thus important to implement good experimental practice, such as randomized-block design of experiments, randomization of samples during processing, and minimizing the number of hands involved in each step of sample preparation as well as running all samples in the same facility on the same instrument, ideally in a single run to avoid between-run variations. RNA is very sensitive to degradation, and differential degradation of the RNA samples will severely affect the outcome. Hence, controlling for RNA quality (purity and intactness) at all steps of the procedure is essential. Another common source of error is remaining genomic DNA in the sample, which will lead to skewed results. Again, samples need to be carefully controlled for DNA contamination and repeatedly treated with DNase to remove the contamination, if needed. Although not yet commonly used, particularly in complex experiments with large numbers of samples, it is advisable to include spiked internal references to control for variations in sample processing and sequencing and to facilitate data postprocessing, such as normalization. Once data have been obtained, clustering by sample and principle component analysis should be used to verify that samples cluster by treatments, not by experimenter or other unintended experimental variables.

PERSPECTIVES

The last decade has seen a decline in the costs of DNA sequencing by at least 5 orders of magnitude, and it is expected that costs will decline even further. Also, new protocols have been developed that dramatically reduce the cost of the preparation of sequencing libraries (Hou et al., 2015). These reductions in costs make larger-scale experiments possible, for example, generating quantitative transcriptomes of hundreds or thousands of genetically diverse individuals of one species, such as Arabidopsis ecotypes or structured mapping populations. This will allow for establishing associations between genetic variation and variation in transcript amounts and the identification of cis- and trans-factors determining transcript amounts through quantitative genetics. RNA-Seq enables molecular analyses that were previously precluded by a lack of sequence information, for example, proteomic analyses of inner and outer chloroplast envelope membranes that to date can only be isolated from species without sequenced genomes, such as spinach (*Spinacia oleracea*) and pea (*Pisum sativum*; Gutierrez-Carbonell et al., 2014). Bottlenecks exist in data processing, storage, and analysis, the latter part frequently being the slowest in RNA-Seq projects. It is thus crucial that, in addition to training in wet-bench skills, instruction in large-scale data analysis becomes an integral part of undergraduate and graduate training curricula (Wingreen and Botstein, 2006).

Received August 13, 2015; accepted September 9, 2015; published September 9, 2015.

LITERATURE CITED

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

Aubry S, Smith-Unna RD, Boursnell CM, Kopriva S, Hibberd JM (2014b) Transcript residency on ribosomes reveals a key role for the *Arabidopsis thaliana* bundle sheath in sulfur and glucosinolate metabolism. Plant J **78:** 659–673

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, et al (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics **7:** 246

Bellucci E, Bitocchi E, Ferrarini A, Benazzo A, Biagetti E, Klie S, Minio A, Rau D, Rodriguez M, Panziera A, et al (2014) Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. Plant Cell **26:** 1901–1912

Brady SM, Burow M, Busch W, Carlborg Ö, Denby KJ, Glazebrook J, Hamilton ES, Harmer SL, Haswell ES, Maloof JN, et al (2015) Reassess the *t* test: interact with all your data via ANOVA. Plant Cell **27:** 2088–2094

Braslavsky I, Hebert B, Kartalov E, Quake SR (2003) Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci USA **100:** 3960–3964

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

Bräutigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Mass J, Lercher MJ, et al (2011a) An mRNA blueprint for C_4 photosynthesis derived from comparative transcriptomics of closely related C_3 and C_4 species. Plant Physiol **155:** 142–156

Bräutigam A, Mullick T, Schliesky S, Weber APM (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. J Exp Bot **62:** 3093–3102

Bräutigam A, Schliesky S, Külahoglu C, Osborne CP, Weber APM (2014) Towards an integrative model of C_4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C4 species. J Exp Bot **65:** 3579–3593

Bräutigam A, Shrestha RP, Whitten D, Wilkerson CG, Carr KM, Froehlich JE, Weber APM (2008b) Low-coverage massively parallel pyrosequencing of cDNAs enables proteomics in non-model species: comparison of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. J Biotechnol **136:** 44–53

Cheung F, Haas BJ, Goldberg SMD, May GD, Xiao Y, Town CD (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. BMC Genomics **7:** 272

Corder R, Mullen W, Khan NQ, Marks SC, Wood EG, Carrier MJ, Crozier A (2006) Oenology: red wine procyanidins and vascular health. Nature **444:** 566

Da Silva C, Zamperin G, Ferrarini A, Minio A, Dal Molin A, Venturini L, Buson G, Tononi P, Avanzato C, Zago E, et al (2013) The high polyphenol content of grapevine cultivar Tannat berries is conferred primarily by genes that are not shared with the reference genome. Plant Cell **25:** 4777–4788

Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res **17:** 69–73

Fouracre JP, Ando S, Langdale JA (2014) Cracking the Kranz enigma with systems biology. J Exp Bot **65:** 3327–3339

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

Gaidatzis D, Burger L, Florescu M, Stadler MB (2015) Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. Nat Biotechnol **33:** 722–729

Góngora-Castillo E, Childs KL, Fedewa G, Hamilton JP, Liscombe DK, Magallanes-Lundback M, Mandadi KK, Nims E, Runguphan W, Vaillancourt B, et al (2012) Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. PLoS One **7:** e52506

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

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol **29:** 644–652

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, et al (2006) Analysis of one million base pairs of Neanderthal DNA. Nature **444:** 330–336

Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL (2015) Informatics for RNA sequencing: a web resource for analysis on the cloud. PLoS Comput Biol **11:** e1004393

Gutierrez-Carbonell E, Takahashi D, Lattanzio G, Rodríguez-Celma J, Kehr J, Soll J, Philippar K, Uemura M, Abadía J, López-Millán AF (2014) The distinct functional roles of the inner and outer chloroplast envelope of pea (Pisum sativum) as revealed by proteomic approaches. J Proteome Res **13:** 2941–2953

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc **8:** 1494–1512

Hou Z, Jiang P, Swanson SA, Elwell AL, Nguyen BKS, Bolin JM, Stewart R, Thomson JA (2015) A cost-effective RNA sequencing protocol for large-scale gene expression studies. Sci Rep **5:** 9570–9575

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

Jensen JK, Schultink A, Keegstra K, Wilkerson CG, Pauly M (2012) RNA-Seq analysis of developing nasturtium seeds (Tropaeolum majus):

identification and characterization of an additional galactosyltransferase involved in xyloglucan biosynthesis. Mol Plant **5:** 984–992

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

Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res **12:** 656–664

Koenig D, Jiménez-Gómez JM, Kimura S, Fulop D, Chitwood DH, Headland LR, Kumar R, Covington MF, Devisetty UK, Tat AV, et al (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proc Natl Acad Sci USA **110:** E2655–E2662

Kratz A, Carninci P (2014) The devil in the details of RNA-seq. Nat Biotechnol **32:** 882–884

Krzywinski M, Altman N (2014) Points of significance: comparing samples - part II. Nat Methods **11:** 355–356

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

Külahoglu C, Denton AK, Sommer M, Maß J, Schliesky S, Wrobel TJ, Berckmans B, Gongora-Castillo E, Buell CR, Simon R, et al (2014) Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C_3 and C_4 plant species. Plant Cell **26:** 3243–3260

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9:** 559

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

Li JW, Robison K, Martin M, Sjödin A, Usadel B, Young M, Olivares EC, Bolser DM (2012) The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. Nucleic Acids Res **40:** D1313–D1317

Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR, et al (2010) The developmental dynamics of the maize leaf transcriptome. Nat Genet **42:** 1060–1067

Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M (2009) Quantification of the yeast transcriptome by single-molecule sequencing. Nat Biotechnol **27:** 652–658

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res **40:** W622-7

Mallmann J, Heckmann D, Bräutigam A, Lercher MJ, Weber AP, Westhoff P, Gowik U (2014) The role of photorespiration during the evolution of C_4 photosynthesis in the genus Flaveria. eLife **3:** e02478–e02478

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet **9:** 387–402

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:** 376–380

Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R, Stamatoyannopoulos JA, Churchman LS (2015) Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell **161:** 541–554

Mitra RD, Shendure J, Olejnik J, Edyta-Krzymanska-Olejnik, Church GM (2003) Fluorescent in situ sequencing on polymerase colonies. Anal Biochem **320:** 55–65

Mustroph A, Zanetti ME, Jang CJH, Holtan HE, Repetti PP, Galbraith DW, Girke T, Bailey-Serres J (2009) Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis. Proc Natl Acad Sci USA **106:** 18843–18848

Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ (2015) Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. Cell **161:** 526–540

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, et al. (2006) Sequencing and analysis of Neanderthal genomic DNA. Science **314:** 1113–1118

Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RDE, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, et al (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science **311:** 392–394

Ponnala L, Wang Y, Sun Q, van Wijk KJ (2014) Correlation of mRNA and protein abundance in the developing maize leaf. Plant J **78:** 424–440

Reynoso MA, Juntawong P, Lancia M, Blanco FA, Bailey-Serres J, Zanetti ME (2015) Translating Ribosome Affinity Purification (TRAP) followed by RNA sequencing technology (TRAP-SEQ) for quantitative assessment of plant translatomes. Methods Mol Biol **1284:** 185–207

Ronaghi M, Uhlén M, Nyrén P (1998) A sequencing method based on real-time pyrophosphate. Science **281:** 363–365, 365

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA **74:** 5463–5467

Schilmiller AL, Miner DP, Larson M, McDowell E, Gang DR, Wilkerson C, Last RL (2010) Studies of a biochemical factory: tomato trichome deep expressed sequence tag sequencing and proteomics. Plant Physiol **153:** 1212–1223

Schliesky S, Gowik U, Weber APM, Bräutigam A (2012) RNA-seq assembly: are we there yet? Front Plant Sci **3:** 220

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics **28:** 1086–1092

Schulze WX, Sanggaard KW, Kreuzer I, Knudsen AD, Bemm F, Thøgersen IB, Bräutigam A, Thomsen LR, Schliesky S, Dyrlund TF, et al (2012) The protein composition of the digestive fluid from the venus flytrap sheds light on prey digestion mechanisms. Mol Cell Proteomics **11:** 1306–1319

SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol **32:** 903–914

Sharon D, Tilgner H, Grubert F, Snyder M (2013) A single-molecule long-read survey of the human transcriptome. Nat Biotechnol **31:** 1009–1014

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science **309:** 1728–1732

Tilgner H, Grubert F, Sharon D, Snyder MP (2014) Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proc Natl Acad Sci USA **111:** 9869–9874

Troncoso-Ponce MA, Kilaru A, Cao X, Durrett TP, Fan J, Jensen JK, Thrower NA, Pauly M, Wilkerson C, Ohlrogge JB (2011) Comparative deep transcriptional profiling of four developing oilseeds. Plant J **68:** 1014–1027

Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M (2009) A guide to using MapMan to visualize and compare omics data in plants: a case study in the crop species, maize. Plant Cell Environ **32:** 1211–1229

Wang L, Czedik-Eysenberg A, Mertz RA, Si Y, Tohge T, Nunes-Nesi A, Arrivault S, Dedow LK, Bryant DW, Zhou W, et al (2014) Comparative analyses of C_4 and C_3 photosynthesis in developing leaves of maize and rice. Nat Biotechnol **32:** 1158–1165

Wang P, Kelly S, Fouracre JP, Langdale JA (2013) Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz anatomy. Plant J **75:** 656–670

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

Wingreen N, Botstein D (2006) Back to the future: education for systems-level biologists. Nat Rev Mol Cell Biol **7:** 829–832

Yoo MJ, Wendel JF (2014) Comparative evolutionary and developmental dynamics of the cotton (Gossypium hirsutum) fiber transcriptome. PLoS Genet **10:** e1004073

Zanetti ME, Chang IF, Gong F, Galbraith DW, Bailey-Serres J (2005) Immunopurification of polyribosomal complexes of Arabidopsis for global analysis of gene expression. Plant Physiol **138:** 624–635