 doi:10.1093/bfgp/elp057

Exploring plant transcriptomes using ultra high-throughput sequencing

Lin Wang, Pinghua Li and Thomas P. Brutnell

Advance Access publication date 3 February 2010

Abstract

Ultra high-throughput sequencing (UHTS) technologies offer the potential to interrogate transcriptomes in detail that has traditionally been restricted to single gene surveys. For instance, it is now possible to globally define transcription start sites, polyadenylation signals, alternative splice sites and generate quantitative data on gene transcript accumulation in single tissues or cell types. These technologies are thus paving the way for whole genome transcriptomics and will undoubtedly lead to novel insights into plant development and biotic and abiotic stress responses. However, several challenges exist to making this technology broadly accessible to the plant research community. These include the current need for a computationally intensive analysis of data sets, a lack of standardized alignment and formatting procedures and a relatively small number of analytical software packages to interpret UHTS outputs. In this review we summarize recent findings from UHTS and discuss potential opportunities and challenges for broad adoption of these technologies in the plant science community.

Keywords: *NextGen sequencing; ultra high-throughput sequencing; transcriptome profiling; deep sequencing; RNA-seq*

THE RAPIDLY EVOLVING TRANSCRIPTOMICS TECHNOLOGIES

The transcriptome encompasses the set of transcripts from a cell or a population of cells, which include protein-coding mRNAs and non-coding small RNAs (e.g. ribosomal, tRNA, miRNA). Traditionally, transcriptome profiling, or transcriptomics, has focused on quantifying gene expression [1, 2]. With the advent of Ultra high-throughput sequencing (UHTS) technologies, it is now possible to obtain highly resolved structural information of RNA populations on a high-throughput platform. This includes mapping transcript initiation and termination sites, splice junctions and post-transcriptional modifications [3]. Such information will lead to a better understanding of the functional elements within the genome and the discovery of novel developmental or environmental regulatory networks.

Transcriptomic studies are often limited by the number of genes that can be surveyed simultaneously. From the 1990s to early 2000s, many analytical methods were developed for high-throughput profiling of the gene space including differential display [4], serial analysis of gene expression (SAGE) [5], microarray [6], cDNA- amplified fragment length polymorphism (AFLP) [7] and massively parallel signature sequencing (MPSS) [8]. Among these tools, hybridization-based microarrays became the dominate platform and has been routinely used to analyze transcriptional changes in many species [6, 9]. Notably for plant biologists, the ATH1 Genome Array developed by Affymetrix® has been extensively used for transcriptional studies in Arabidopsis [10]. Microarray platforms were also used to characterize transcriptomes for other plant species, including maize [11, 12], rice [13, 14], barley [15, 16], soybean [17, 18] and tomato [19, 20].

Corresponding author. Thomas P. Brutnell, Boyce Thompson Institute, Cornell University, 1 Tower Road, Ithaca, NY 14853, USA. Tel: +1-607-254-8656; Fax: +1-607-254-1242; E-mail: tpb8@cornell.edu

Lin Wang is a postdoctoral fellow at the Boyce Thompson Institute for Plant Research. His current research focuses on functional genomics of maize C4 photosynthesis and circadian regulated gene expression using systems biology approaches.

Pinghua Li is a postdoctoral fellow at the Boyce Thompson Institute for Plant Research. Her current research centers on understanding the maize leaf transcriptome.

Tom Brutnell is an Associate Scientist at the Boyce Thompson Institute for Plant Research. His current research focuses on using UHTS technologies to investigate transcriptional networks in C3 and C4 photosynthetic grasses.

While microarray-based transcriptomic studies are fruitful, the hybridization-based technology has a few intrinsic limitations. First, the dynamic range of microarrays is restricted by factors such as the probe density/availability, the intensities of fluorescent dyes and the sensitivity of scanning instruments. As a result, typical microarray platforms have a limited dynamic range of two to three orders of magnitude [3, 6]. However, quantitative RT-PCR analysis has shown that the expression of some genes can vary up to several thousand-fold, particularly those of small RNAs [21]. Second, the sensitivity of microarrays is reduced by non-specific cross-hybridization, which can mask isoform expression and inflate the expression of rare transcripts. Moreover, developing new microarray platforms is usually labor and time-intensive, and commonly requires knowledge of the target genome. Thus, the microarray technology is generally limited to species with a sequenced genome or well-characterized transcriptome (e.g. extensive cDNA sequence) and is therefore considered a closed-architecture environment. That is, the arrays are built on the knowledge of the genome at the time. For example, the ATH1 array is likely missing up to 3000 protein-coding genes and all miRNAs that are present in the Arabidopsis genome (based on Affymetrix ATH1 summary and TAIR9 annotation from http://www.affymetrix.com and http://www.arabidopsis.org, respectively). This is in contrast to an open-architecture environment where surveys of transcript accumulation are not limited by genome annotation.

As an alternative to the microarray-based approaches, direct measures of gene expression can be obtained through sequencing. Several of these open-architecture methods have been used including random expressed sequence tags (ESTs) sequencing, SAGE and MPSS [22]. One significant advantage of sequence-based transcriptomics is the potential to precisely quantify the abundance of any transcript, drastically increasing the dynamic range of the experiment [23]. Another advantage is that they are not limited by the availability of a sequenced genome. ESTs can be synthesized, sequenced and annotated from any genome, providing a platform for gene discovery. However, these early iterations of sequencing-based approaches have not been widely adapted due to the relatively low throughput and high cost associated with Sanger sequencing platforms.

In recent years, the development of UHTS technologies have dramatically increased the throughput of sequence generation and decreased the overall cost. Currently, UHTS is offered by several companies; these include Roche/454 Life Science that utilizes a pyrosequencing platform; Illumina and bridge-PCR-based Solexa sequencing; Applied Biosystems Inc and Sequencing by Oligo Ligation and Detection (SOLiD); and most recently Helicos and their single-molecule sequencing (for more detailed reviews of next-generation sequencing technologies, see [24–26]). The significantly increased throughput of UHTS relies on the generation of short sequence reads (30–400 bps) of thousands to millions of DNA molecules in parallel [27]. These newly developed 'ultrahigh-throughput' sequencing technologies promise to provide a much more detailed view of plant transcriptomes and to revolutionize the way eukaryotic transcriptomes are analyzed [3]. In this review, we focus on the Illumina platform as the example of UHTS technology.

TRANSCRIPTOME SEQUENCING

UHTS (also referred to as NextGen, RNA-seq) refers to the deep-sequencing of RNA pools. While UHTS often refers to deep sequencing of mRNAs, any RNA population can be analyzed. Two methods are typically used to capture and sequence RNA pools (Figure 1). In both methods mRNA pools are enriched by capturing the molecules through the polyadenylated tails, and a ribosomal RNA removal step is often added before or after the mRNA purification. In one method, mRNA-enriched pools are then fragmented into roughly equal lengths and then reverse-transcribed using random hexamers to generate a cDNA library. Alternatively, RNA is reverse transcribed using an oligo-dT adapter and the resulting cDNA is fractionated. The former method has the advantage of more uniform representation across the coding region of the transcript, but may result in the under-representation 5′ and 3′ sequences. The later method provides good coverage of 3′ sequences but biases against the body of the transcript [3]. The cDNAs are then fitted with adaptors at one or both ends through a ligation step(s). It is desirable to add these adaptors during the single-strand stage (RNA or cDNA) synthesis step in order to retain strand specificity in the final sequence reads [27]. The tagged cDNA library is subsequently amplified

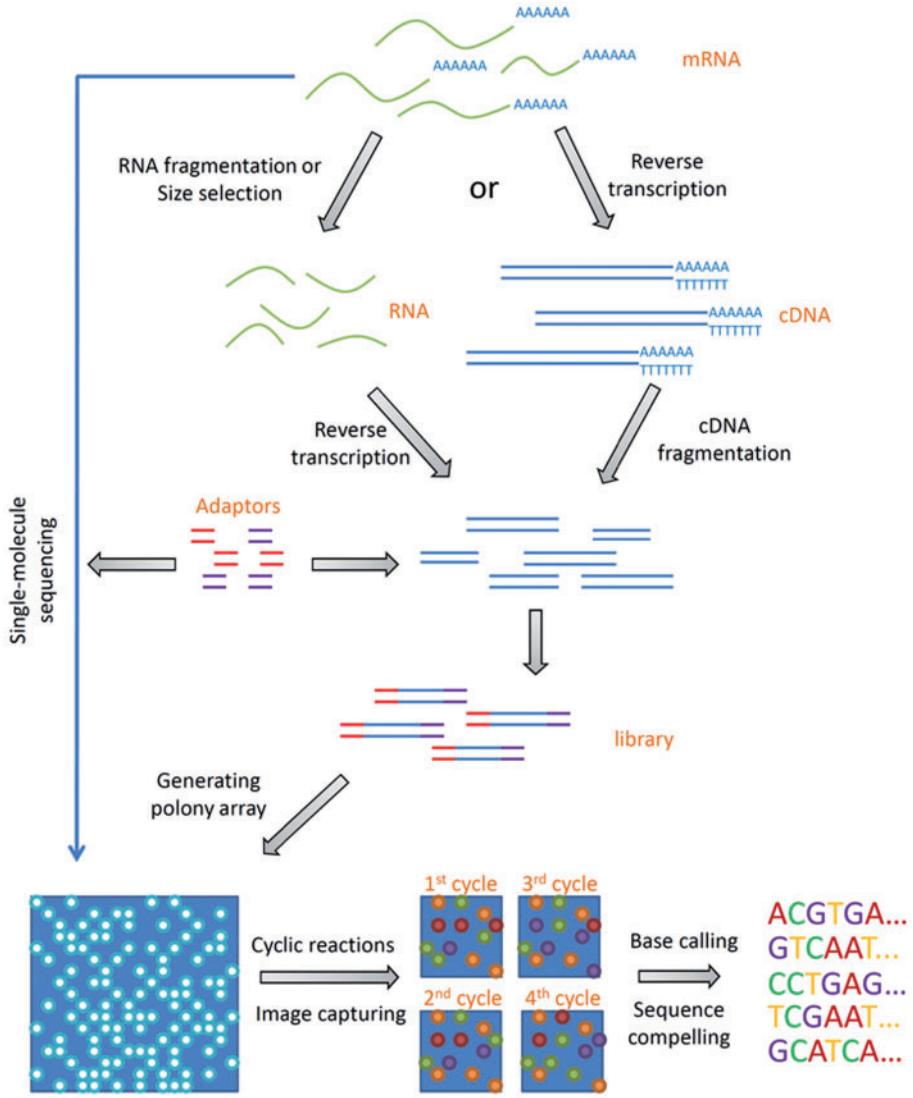

Figure 1: Overview of RNA-seq experimental procedures. For a typical RNA-seq experiment, mRNA is isolated and reverse-transcribed (RT) into cDNA libraries with homogeneous lengths. This is achieved by either RNA or cDNA fragmentation. Recently, single-molecule capture methods have been developed (left) that obviate the need for a RT step. In the case of smRNA studies, total RNA is fractionated on acrylamide gels and smRNAs excised. Adaptors at one or both ends of the RNA are added prior to cDNA amplification and library construction. For the Solexa platform, cDNA molecules are anchored onto a polony array (flow cell) surface, which are then subjected to PCR amplification. Images are taken after each cycle for base calling and sequence generation. Currently for the Illumina platform, ~120 million single or paired-end reads of 32–80 nt are generated on a single flow cell (8 lanes/flow cell) which is then processed further depending on the research goals.

through PCR before being sequenced. For the more recently developed single-molecule approach, the PCR-amplification step is entirely eliminated to further reduce the amplification-based biases and to increase the throughput [28, 29]. Procedures of smRNA sequencing are very similar to those of mRNA-seq, substituting a size-selection step for small RNA molecules for the RNA fragmentation step [27]. 'Digital gene expression (DGE)' or 'Digital tag profiling' is another method of transcript profiling that is qualitatively similar to SAGE analysis, in that single transcripts are identified through a 3′ end tag (http://www.illumina.com). In theory, this method could allow for a vast increase in throughput as each molecule would be represented by a single read rather than hundreds of reads for cDNA-based methods. However, in practice, we have found that the technique suffers a number of shortcomings

including a dependence on a relatively small tag (21 and 20 bps when using *Nla*III and *Dpn*II, respectively) resulting in a large number of redundant placements in complex genomes, large differences in library populations if multiple restriction enzymes are used to generate the 3′ tags, and low correlations between cDNA-based methods, where gene transcripts are represented by multiple alignments.

All RNA-seq projects to date have utilized a reference genome and a number of algorithms have been developed to perform alignments including BWA [30], ELAND [31], SOAP [32], MAQ [33], BOWTIE [34], PASS [35] and RMAP [36]. A generic alignment format, SAM, for storing read alignments generated by these programs has also been proposed that captures much of the metadata associated with a given RNA-seq project [37]. Subsequently, quantification of transcripts is achieved by counting the density of the reads that are mapped to the exon regions of a specific gene, often correcting for transcript length. For smRNAs, the reads may not need to be directly mapped to the genome if the sequence is highly conserved and has been previously characterized. A common output for mRNA expression values using the Illumina platform is reads per kilobase of exon per million mapped reads (RPKM) and for small RNA's, reads per million reads (RPM) [31, 38]. After normalization, this information is then used to calculate the expression level of a transcript or the abundance of a smRNA species. Moreover, additional information such as alternative splicing, or multiple initiation and termination sites can be captured when sequences are mapped to the genome. Programs such as TopHat [39] and Supersplat (http://supersplat.cgrb.oregonstate.edu) have been developed for such tasks. However, to date, a standard method for quantitative analysis of transcript isoforms does not exist as RPKM values are calculated for gene models. However, efforts are being made in our lab and others to build a pipeline for such tasks. When strandedness is captured in library construction, it is also possible to identify sense and antisense transcripts. For species in which a genome sequence is not yet available, it should be possible to use EST or GSS databases as templates for scaffold assembly. However, to date, there are limited reports demonstrating the feasibility of this approach. Given the rapid progress in genome sequencing technology and the projected lower costs, it is likely that most researchers will be able to utilize partial or complete genome sequences to perform their assemblies in the near future.

Because of the improved throughput and lower cost, RNA-seq is increasingly being regarded as the new standard for transcriptomics. Studies to detect rare miRNAs/mRNAs and splicing events that have been proven difficult by using traditional methods have been revisited with the help of UHTS technologies [31, 40–42]. Several RNA-seq studies have been performed in Arabidopsis, maize, barley, *Medicago* and tomato [43–48]. Below, we discuss two areas in plant biology that have benefited extensively from UHTS—small RNA studies and transcriptome profiling.

ANALYSING SMALL RNA

Small RNA typically refers to those non-coding RNA (ncRNA) that modulate gene expression and have been implicated in many aspects of the life cycle such as development/differentiation [49, 50], phytohormone signaling [51], genome maintenance [51, 52] and adaptation to environmental challenges [53–55]. UHTS technologies have provided an ideal platform for high-throughput smRNA studies in plants. One of the first smRNA studies using UHTS was conducted in Arabidopsis [44]. Using Illumina/Solexa sequencing, Lister and colleagues surveyed the 'smRNAome' from DNA methylation- and demethylation-deficient mutants to investigate RNA-directed DNA methylation. From the 15–30 bps fraction of total RNA, over 2.6 million smRNAs were identified, consisting of a majority of 24 and 21 nt smRNAs. Roughly half of the smRNAs have multiple targeting sites while the other half align uniquely to the Arabidopsis genome. This study identified the presence of 'multi-targeting' smRNAs that play a major role in RNA-mediated DNA methylation. This study also revealed a role of ta-siRNAs in *de novo* DNA methylation and the self-reinforcing nature between DNA-methylation and smRNA biogenesis [44]. A similar survey of maize smRNA was also performed using the Illumina/Solexa platform [45]. Small RNA samples were isolated and sequenced from wild type maize and *mop1-1*, a mutant allele of RNA-dependent RNA polymerase 2 (RDR2). This study identified a similar smRNA pattern to that observed in the Arabidopsis *rdr2* mutant, for which the 24 nt heterochromatin siRNAs were significantly reduced resulting in an elevated miRNA and

ta-siRNA content. Interestingly, 22 nt heterochromatin siRNAs, that are depleted in Arabidopsis *rdr2* plants, remain relatively constant in the maize *mop1-1* mutant, suggesting the existence of an alternative heterochromatic siRNA biogenesis pathway in maize [45]. Other UHTS-based studies were carried out in *Solanum lycopersicum* [48], *Medicago truncatula* [43, 56] and *Brachypodium distachyon* [57]. From these studies, several new classes of species-specific miRNAs were identified, including tomato miRNA that are correlated to fruit ripening, *Medicago* miRNA involved in nodulation and *Brachypodium* miRNA associated with stress response. Another example of an UHTS-based smRNA study is the survey of *cis*- and *trans*-natural antisense siRNAs in *Oryza sativa* [58]. Combined with computational prediction, Zhou and colleagues sampled smRNA populations in control and salt/drought stressed rice seedlings used 454-pyrosequencing. The majority of the identified *cis*-natural antisense siRNAs are stress-specific, suggesting they may be involved in regulating the stress responses [58]. Taken together, these studies have convincely demonstrated the robustness and effectiveness of UHTS-based technologies in smRNA research. As the cost of UHTS decreases and sequencing capacities increase, it is likely that the 'smRNAome' will be surveyed from an increasingly larger number of plants. These datasets will help elucidate the evolutionary relationships among smRNA populations across the plant kingdom.

PROFILING THE TRANSCRIPTOME

To date, several groups have used UHTS for discovery-based studies of the mRNA populations in plants. Early analyses in *Medicago* [59], maize [60] and Arabidopsis [61] used the 454-pyrosequencing platform. The pioneering studies of Emrich [60] and Cheung [59] were the first to demonstrate the power of UHTS to elucidate a plant transcriptome. Cheung and colleagues sequenced a normalized adaptor-tagged cDNA library from *M. truncatula* [59]. From their work, close to two million unique sequences were generated from pyrosequencing and over one third of these sequences were mapped to *Medicago* BACs (Bacterial Artificial Chromosomes) and over ten thousand novel transcripts were identified [59]. In a similar study by Weber and colleagues, over 5 million ESTs were generated from Arabidopsis seedlings [61]. These ESTs were mapped to over 15 000 genes, which accounted for over 90% of transcripts that were predicted to be expressed. Over 60 previously unannotated transcripts were also discovered, providing the primary experimental support for these novel genes. Emrich and colleagues combined UHTS with laser-capture microdissection to examine the transcriptome of the maize shoot apical meristem [60]. They found close to 400 maize-specific transcripts in the meristem, providing a glimpse into the divergent and complex nature of plant transcriptomes.

Work from our group has exploited the Illumina platform to define the maize leaf transcriptome along a developmental gradient that captures photosynthetic sink and source tissues. In total we have generated ~30 million reads from each of four developmental zones and mapped reads to the maize genome. We have also performed a deep sequencing of leaf tip and base RNA pools (~94 million reads total) and were able to provide evidence for over 1700 gene models that lacked EST support [62]. It is also notable that through deep sequencing of leaf RNA pools, expression was detectable for 28 560 of the predicted 32 540 maize genes (87%) demonstrating the robustness of the UHTS approach.

Another important aspect of UHTS technologies that distinguishes it from microarray studies is the ability to define alternative splicing and initiation/termination sites. So far, UHTS-based studies of transcript isoforms have been carried out in a few non-plant species such as yeast [63], human [64–66], mouse [67, 68] and zebrafish [69]. In these studies, thousands of previously unidentified transcript isoforms were discovered. It was shown that over 90% of human 'multi-exon' genes undergo alternative splicing, with exon-skipping being the most prevalent [65]. Moreover, individual transcript isoforms preferentially accumulate under different conditions or in different tissue types, suggesting that alternative splicing is an intrinsic mechanism that contributes to the increased cellular and functional complexity in higher eukaryotes [64]. Alternative splicing is also prevalent in many plant species. A few studies have been conducted to survey the alternative splicing in plants including Arabidopsis [70, 71], rice [72], maize [73, 74] and moss [75]. One interesting finding is that exon-skipping is far less common in plants than in animals, averaging <10% of total splicing events [73].

In contrast, up to 80% of transcript isoforms in plants arise from alternative 3′-splicing sites and intron retention events. With the unprecedented sequencing depth provided by UHTS technology, it is evident that rare transcript isoform can be captured providing the most detailed view of the plant transcriptome to date.

Quantification of gene expression, or expression profiling, is another important aspect of transcriptomic studies. While UHTS-based studies that focus solely on expression profiling in plants are still limited, several analyses have been carried out in animal systems. For example, Han and colleagues analyzed the transcriptome from mouse embryonic and neonatal cortex cells [67]. From over 27 million sequence reads that were generated by Illumina/Solexa sequencing, over 16 000 genes were assayed in two developmental stages and a total of 3758 genes were identified as differentially expressed, including many novel neurogenesis-related genes. This study provides the blueprint of gene expression profiles during the early stages of mouse brain development. A similar expression profiling study was performed in zebrafish by Hegedus and colleagues to analyze the transcriptomic response to mycobacterium infection [69]. Using Illumina's DGE system, a total of 5049 significantly changed tag entities were detected corresponding to over 1500 UniGene transcripts (http://www.ncbi.nlm.nih.gov/unigene). A majority of these transcripts originated from the sense-strand of the zebrafish DNA, suggesting that the antisense regulation may not be a major factor in regulating the defense response [69]. Consistent with their experimental design, the majority of the genes that preferentially accumulated were those of immunity-related functions. Interestingly, genes encoding proteolytic enzymes and some ATPases were also more abundant, suggesting the importance of protein turnover and energy consumption in zebrafish immunity [69].

One of the main goals of our maize leaf transcriptomic study is to understand the profiles of gene expression along the developmental gradient and their relationship to differentiation of C4 photosynthesis. For this, we have used UHTS to quantify gene expression in four developmental zones and in bundle sheath and mesophyll cells. From over 27 000 expressed genes, we have identified several groups of differentially expressed genes, that cluster by functional annotation such as photosynthesis system, sugar metabolism, hormone signaling, redox responses, cell wall biosynthesis and many classes of transcription factors. Figure 2 lists some examples of functional enrichment from the differentially expressed genes along the leaf gradient. These findings suggest that the differentiation of mesophyll/bundle sheath cells and development of C4 apparatus is an extremely dynamic process resulting in the partitioning of many biochemical functions between two morphologically distinct cell types.

CHALLENGES AND OPPORTUNITIES

As a newly developed technology, UHTS transcriptomics faces many unique challenges. The most apparent obstacle arises from the complexity of the sequence data that is generated by the UHTS platforms. To process, interpret and visualize these large datasets, it is necessary to develop efficient and sophisticated algorithms and pipelines. As outlined in Figure 3, processing the data often involves several steps—sequence acquisition and base calling, filtering raw reads, aligning reads to a reference genome, determining statistical significance, calculating gene coverage, normalization and estimating gene expression. Methods are currently being developed to streamline these procedures, which may create standards for future UHTS-based studies.

The first customizable step in the processing pipeline is mapping the short reads to the reference genome or assembling them into contigs. In the absence of a sequenced genome, assembling short reads into contigs is extremely challenging. Thus, it is likely that most UHTS transcriptome profiling will be restricted to organisms with sequenced genomes, at least in the short term. For species where a complete genome sequence is not available, one alternative is to first assemble a pseudo-transcriptome using 454 technology to generate contigs that capture a majority of the gene space. These contigs can then be used as a scaffold for gene quantification using a high throughput sequencing platform such as Illumina/Solexa or ABI SOLiD.

Determining transcript structure also presents a number of challenges. To examine RNA isoforms generated through alternative splicing requires that reads be mapped to exon–exon junctions. This can be challenging when read lengths are small (e.g. 32 nt). One solution is to create a database with all well-established transcript isoforms to facilitate the

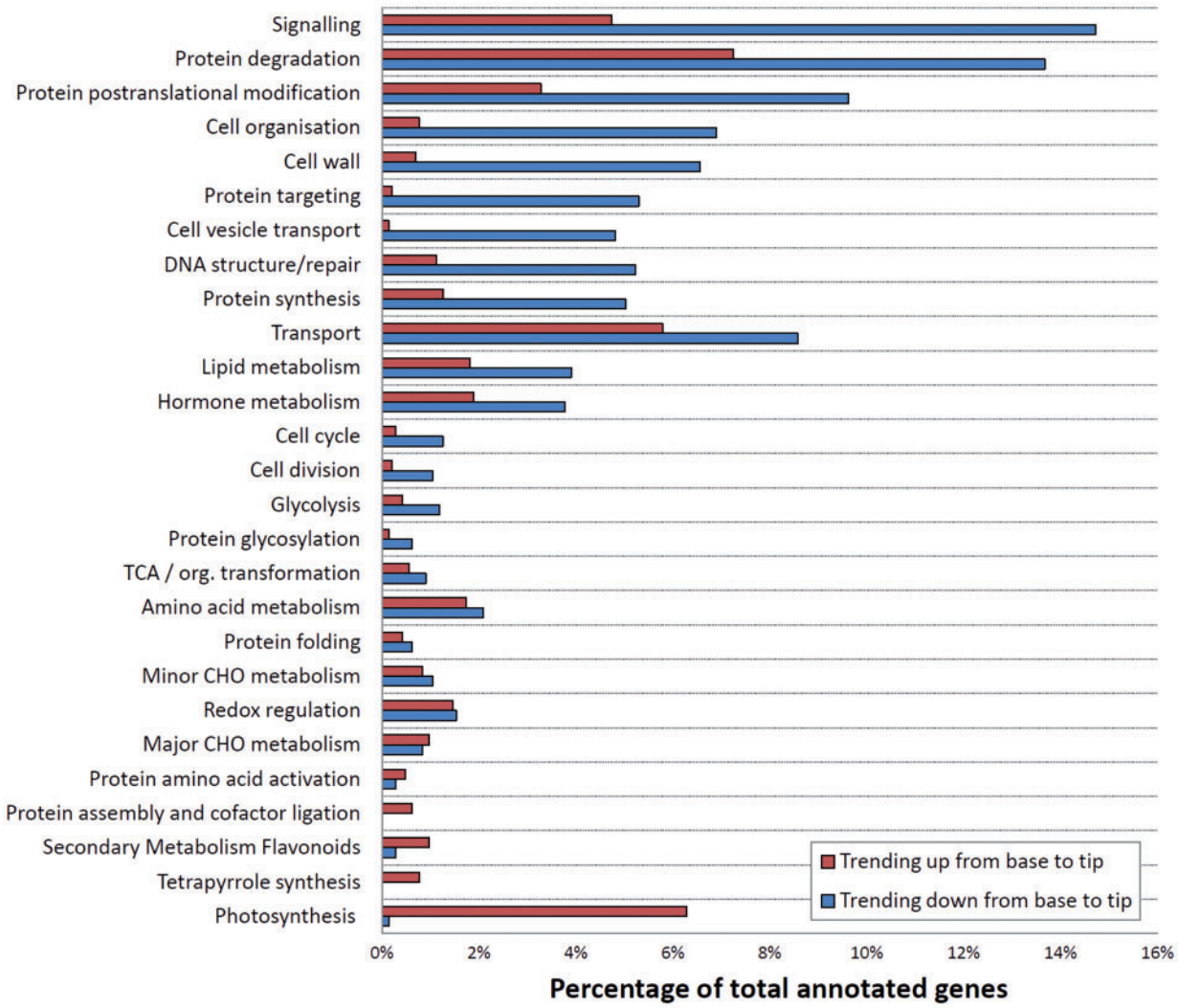

Figure 2: Summary of maize leaf transcriptome. Genes that are detected along the leaf developmental gradient were categorized based on their expression trends. The two major trends are detected: one cluster of genes accumulates gradually from leaf base to tip; the other cluster decreases in abundance from base to tip. The functional classes are derived from Mapman categories [86] and are represented as a percentage of the total number of genes. The functional categories are ordered so that the most significant differences between trends are shown at the top and the bottom.

mapping the transcript crossing multiple exons [40]. Junction fragments can then be defined as those reads that map uniquely to an exon–exon junction. Often these junction fragments are excluded from RPKM analysis [31], but could contribute significantly to the expression value for a given gene. Thus, RPKM values should be refined to incorporate junction fragments. This is particularly important for genes with multiple exons as their expression values could be underestimated if junction fragments are excluded. Aligning short reads to the genome can be further complicated by the existence of highly conserved gene families in eukaryotic genomes. A sequence read that matches to a conserved region shared by several gene family members may cause ambiguity of its placement. While such uncertainly can be partially negated by generating longer reads or pair-end reads, it is nevertheless difficult to process transcript isoforms that originate from the same gene. One possibility is to exclude these reads from the analysis, but with large and complex genomes, this could result in discarding the vast majority of reads. Another solution is to assign such reads proportionally based on the number of adjacent unique reads [31, 76]. However, this approach is deemed less useful when reads are generated from repetitive regions with high copy numbers [3]. A few statistical methods have also been designed to specifically address the

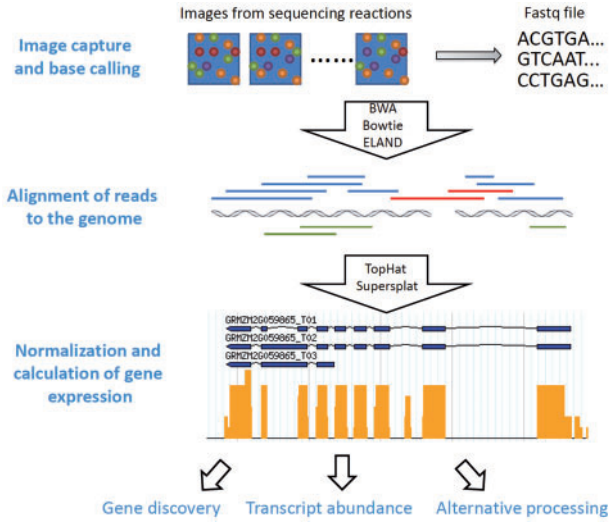

Figure 3: Overview of RNA-seq data analysis. Image capture and base pair calls are commonly performed by proprietary software that is developed by individual UHTS companies. The raw output files of this software contain the sequences of short reads, usually also accompanied by quality scores. Publically available software/algorithms such as BWA [30], Bowtie [34] and ELAND [31] are able to align these reads to the reference genome (blue lines represent read that are mapped to exons, red as mapped to introns, green as mapped to antisense strand). Alternative splicing event can be analyzed with software such as TopHat [39] and Supersplat (http://supersplat .cgrb.oregonstate.edu). Results generated from the pipeline can be used to discover novel transcripts, monitor transcriptome dynamics and explore transcript structures.

'multi-matching' problem [77, 78]. Sequencing error is another important factor that affects the ability to align reads to the genome. It has been estimated that current Illumina Genome Analyzer, yields up to 1% erroneous reads, while the single-molecule sequencing platforms misread at even higher frequencies [29, 79]. Such inaccuracy may also adversely affect the discovery of SNPs (single nucleotide polymorphisms), which is a focus of some recent genomic studies [80–82].

Normalization procedures to estimate gene expression is another area of active debate. One of the proposed methods assumes sequence placement follows a Poisson distribution over the genome. This enables the normalization of transcript reads over gene length and total number of mapped reads as the RPKM to represent gene expression values [31]. Another underlying assumption of this approach is that there is no systematic variance among samples which may cause the total read counts to vary. While this may be a safe assumption for experiments that use similar RNA templates and amplification procedures, it is not applicable when the biological templates or library construction procedures differ substantially. An alternative normalization model has been proposed by Balwierz and colleagues for the analysis of cap analysis of gene expression (CAGE) and RNA-seq data when sequence tags were found to follow the power-law distribution [83, 84]. A more objective normalization method using synthetic RNA spiking controls, originally developed for microarray analysis [38], may also prove effective for UHTS.

Another challenge of using UHTS is to overcome the artificial biases that are introduced during the experimental procedures. Nagalakshmi and colleagues first demonstrated that such biases exist for different fragmentation methods, showing that RNA fragmentation enriches the reads in the transcript body while cDNA fragmentation enriches reads at the transcript ends [63]. While most of mRNA-seq protocols now employ RNA-fragmentation to maximize the representation of exons, important information at the 3′ and 5′-ends of the transcript may be missed. Biases can also be introduced through PCR-amplification steps, where factors such as primer quality and template GC content may affect the efficiency of PCR leading to artificial enrichment of certain groups of transcripts. Recently developed 'single-molecule' sequencing technologies may make such amplification steps obsolete, potentially providing a more accurate estimate of transcript abundance [79]. Biases also exist in the statistical power of determining whether a gene is differentially expressed. Oshlack and colleagues [85] have reevaluated some published data, which demonstrated that the differences in statistical power rely on the chances of a read to hit a certain gene. Longer genes intrinsically accumulate more reads over shorter ones, which create the differences of statistical power in calculating significance. It is notable that such a discrepancy cannot be completely eliminated even when reads are normalized by the gene length, because a difference in variance still exists. Interestingly, such variances in statistical power do not exist in microarray-based analyses, suggesting it may be complementary to UHTS-based transcriptomics studies [85].

The transcriptome coverage is also an important issue for mRNA-seq. While in principle close to

100% of transcripts can be captured by adequate sequencing depth, the associated cost of an exhaustive approach can be inhibitory. An early study in yeast demonstrated that 30 million 35-base pairs reads can capture up to 90% of all genes from cultures that were grown under one condition [63]. In a much more complex maize genome, we have shown that transcripts for over 87% of annotated genes are detectable from 94 million reads derived from leaf RNA samples. With the UHTS technology being steadily improved and the sequencing costs dropping, the issue of transcriptomic coverage may soon be less of a concern.

Despite these challenges, UHTS-based transcriptomics approaches promise 'never-before' opportunities to explore plant transcriptomes. As improvements to the sequencing chemistry, sequencing hardware and software and statistical methods of analysis continue to progress, the expectations for transcriptomics studies will continue to increase. It has been speculated that the cost of sequencing a complete genome or transcriptome will not be a limiting factor in the foreseeable future. This may allow experiments that were deemed as economically unfeasible in the past to be routinely performed. For instance, an EMS mutagenesis that generates a point mutation with an interesting phenotype may be quickly mapped and characterized through a dual whole genome DNA/RNA-seq profiling approach. In summary, UHTS-based approaches have clearly demonstrated their advantages over previously developed methods and are becoming the new standard for transcriptomics studies.

Key Points

- UHTS technologies have dramatically increased the throughput and dramatically decreased cost of sequence generation. As a result, they are increasingly being regarded as the standard method for transcriptomic studies.
- UHTS-assisted transcriptome profiling studies have provided novel insight into the diversity of small RNA species, the complexity of transcript structures and the dynamics of gene expression at the genomic level.
- Because of the sudden burst of data-generating capacity, UHTS faces many unique challenges including data storage, processing and interpretation. Many novel bioinformatic toolsets and pipelines are being developed to address these needs.

FUNDING

National Science Foundation (grant number IOS-0701736 to T.P.B.).

References

1. Schulze A, Downward J. Navigating gene expression using microarrays–a technology review. *Nat Cell Biol* 2001;**3**: E190–5.
2. Stears RL, Martinsky T, Schena M. Trends in microarray analysis. *Nat Med* 2003;**9**:140–5.
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**: 57–63.
4. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992;**257**:967–71.
5. Velculescu VE, Zhang L, Vogelstein B, *et al*. Serial analysis of gene expression. *Science* 1995;**270**:484–7.
6. Schena M, Shalon D, Davis RW, *et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;**270**:467–70.
7. Bachem CW, van der Hoeven RS, de Bruijn SM, *et al*. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant J* 1996;**9**:745–53.
8. Brenner S, Johnson M, Bridgham J, *et al*. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 2000;**18**:630–4.
9. Donson J, Fang Y, Espiritu-Santo G, *et al*. Comprehensive gene expression analysis by transcript profiling. *Plant Mol Biol* 2002;**48**:75–97.
10. Busch W, Lohmann JU. Profiling a plant: expression analysis in Arabidopsis. *Curr Opin Plant Biol* 2007;**10**:136–41.
11. Zhu Y, Fu J, Zhang J, *et al*. Genome-wide analysis of gene expression profiles during ear development of maize. *Plant Mol Biol* 2009;**70**:63–77.
12. Strable J, Borsuk L, Nettleton D, *et al*. Microarray analysis of vegetative phase change in maize. *Plant J* 2008;**56**:1045–57.
13. Hazen SP, Pathan MS, Sanchez A, *et al*. Expression profiling of rice segregating for drought tolerance QTLs using a rice genome array. *Funct Integr Genomics* 2005;**5**:104–16.
14. Kim SH, Bhat PR, Cui X, *et al*. Detection and validation of single feature polymorphisms using RNA expression data from a rice genome array. *BMC Plant Biol* 2009;**9**:65.
15. Delp G, Gradin T, Ahman I, *et al*. Microarray analysis of the interaction between the aphid Rhopalosiphum padi and host plants reveals both differences and similarities between susceptible and partially resistant barley lines. *Mol Genet Genomics* 2009;**281**:233–48.
16. Hansen M, Friis C, Bowra S, *et al*. A pathway-specific microarray analysis highlights the complex and co-ordinated transcriptional networks of the developing grain of field-grown barley. *J Exp Bot* 2009;**60**:153–67.
17. O'Rourke JA, Nelson RT, Grant D, *et al*. Integrating microarray analysis and the soybean genome to understand the soybeans iron deficiency response. *BMC Genomics* 2009; **10**:376.
18. Brechenmacher L, Kim MY, Benitez M, *et al*. Transcription profiling of soybean nodulation by *Bradyrhizobium japonicum*. *Mol Plant Microbe Interact* 2008;**21**:631–45.
19. Auge GA, Perelman S, Crocco CD, *et al*. Gene expression analysis of light-modulated germination in tomato seeds. *New Phytol* 2009;**183**:301–14.

20. Jiang F, Zheng X, Chen J. Microarray analysis of gene expression profile induced by the biocontrol yeast *Cryptococcus laurentii* in cherry tomato fruit. *Gene* 2009;**430**:12–6.
21. Wang X. A PCR-based platform for microRNA expression profiling studies. *RNA* 2009;**15**:716–23.
22. Kamoun S, Hraber P, Sobral B, *et al*. Initial assessment of gene diversity for the oomycete pathogen Phytophthora infestans based on expressed sequences. *Fungal Genet Biol* 1999;**28**:94–106.
23. Matsumura H, Reuter M, Kruger DH, *et al*. SuperSAGE. *Methods Mol Biol* 2008;**387**:55–70.
24. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45.
25. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008;**9**:387–402.
26. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol* 2009;**25**:195–203.
27. Simon SA, Zhai J, Nandety RS, *et al*. Short-read sequencing technologies for transcriptional analyses. *Annu Rev Plant Biol* 2009;**60**:305–33.
28. Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 2009;**27**:847–52.
29. Lipson D, Raz T, Kieu A, *et al*. Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* 2009;**27**:652–8.
30. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**: 1754–60.
31. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
32. Li R, Li Y, Kristiansen K, *et al*. SOAP: short oligonucleotide alignment program. *Bioinformatics* 2008;**24**:713–4.
33. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008;**18**:1851–8.
34. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
35. Campagna D, Albiero A, Bilardi A, *et al*. PASS: a program to align short sequences. *Bioinformatics* 2009;**25**:967–8.
36. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* 2008;**9**:128.
37. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**: 2078–9.
38. Fahlgren N, Sullivan CM, Kasschau KD, *et al*. Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA* 2009;**15**:992–1002.
39. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;**25**: 1105–11.
40. Wilhelm BT, Marguerat S, Watt S, *et al*. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 2008;**453**:1239–43.
41. Marioni JC, Mason CE, Mane SM, *et al*. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**:1509–17.
42. Adams MD, Kelley JM, Gocayne JD, *et al*. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 1991;**252**:1651–6.
43. Szittya G, Moxon S, Santos DM, *et al*. High-throughput sequencing of Medicago truncatula short RNAs identifies eight new miRNA families. *BMC Genomics* 2008;**9**:593.
44. Lister R, O'Malley RC, Tonti-Filippini J, *et al*. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 2008;**133**:523–36.
45. Nobuta K, Lu C, Shrivastava R, *et al*. Distinct size distribution of endogeneous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. *Proc Natl Acad Sci USA* 2008;**105**:14958–63.
46. Wicker T, Narechania A, Sabot F, *et al*. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* 2008;**9**:518.
47. Qi X, Bao FS, Xie Z. Small RNA deep sequencing reveals role for Arabidopsis thaliana RNA-dependent RNA polymerases in viral siRNA biogenesis. *PLoS One* 2009;**4**:e4971.
48. Moxon S, Jing R, Szittya G, *et al*. Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. *Genome Res* 2008;**18**:1602–9.
49. Garcia D, Collier SA, Byrne ME, *et al*. Specification of leaf polarity in Arabidopsis via the trans-acting siRNA pathway. *Curr Biol* 2006;**16**:933–8.
50. Chuck G, Candela H, Hake S. Big impacts by small RNAs in plant development. *Curr Opin Plant Biol* 2009;**12**:81–6.
51. Teixeira FK, Heredia F, Sarazin A, *et al*. A role for RNAi in the selective correction of DNA methylation defects. *Science* 2009;**323**:1600–4.
52. Slotkin RK, Vaughn M, Borges F, *et al*. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* 2009;**136**:461–72.
53. Sunkar R, Zhu JK. Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell* 2004;**16**: 2001–19.
54. Ruiz-Ferrer V, Voinnet O. Roles of plant small RNAs in biotic stress responses. *Annu Rev Plant Biol* 2009;**60**:485–510.
55. McManus MT. Small RNAs and immunity. *Immunity* 2004; **21**:747–56.
56. Lelandais-Briere C, Naya L, Sallet E, *et al*. Genome-wide medicago truncatula small RNA analysis revealed novel microRNAs and isoforms differentially regulated in roots and nodules. *Plant Cell* 2009;**21**:2780–94.
57. Zhang J, Xu Y, Huan Q, *et al*. Deep sequencing of Brachypodium small RNAs at the global genome level identifies microRNAs involved in cold stress response. *BMC Genomics* 2009;**10**:449.
58. Zhou X, Sunkar R, Jin H, *et al*. Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in *Oryza sativa*. *Genome Res* 2009;**19**:70–8.
59. Cheung F, Haas BJ, Goldberg SM, *et al*. Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics* 2006;**7**:272.
60. Emrich SJ, Barbazuk WB, Li L, *et al*. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res* 2007;**17**:69–73.
61. Weber AP, Weber KL, Carr K, *et al*. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol* 2007;**144**:32–42.

62. Schnable PS, Ware D, Fulton RS, *et al*. The B73 maize genome: complexity, diversity, and dynamics. *Science* 2009;**326**:1112–5.
63. Nagalakshmi U, Wang Z, Waern K, *et al*. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008;**320**:1344–9.
64. Pan Q, Shai O, Lee LJ, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008;**40**:1413–5.
65. Sultan M, Schulz MH, Richard H, *et al*. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 2008;**321**:956–60.
66. Wang ET, Sandberg R, Luo S, *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008; **456**:470–6.
67. Han X, Wu X, Chung WY, *et al*. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proc Natl Acad Sci USA* 2009;**106**:12741–6.
68. Tang F, Barbacioru C, Wang Y, *et al*. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**: 377–82.
69. Hegedus Z, Zakrzewska A, Agoston VC, *et al*. Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. *Mol Immunol* 2009;**46**:2918–30.
70. Wang BB, Brendel V. Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci USA* 2006; **103**:7175–80.
71. Nagasaki H, Arita M, Nishizawa T, *et al*. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene* 2005;**364**:53–62.
72. Campbell MA, Haas BJ, Hamilton JP, *et al*. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics* 2006;**7**:327.
73. Barbazuk WB, Fu Y, McGinnis KM. Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res* 2008;**18**:1381–92.
74. Haberer G, Young S, Bharti AK, *et al*. Structure and architecture of the maize genome. *Plant Physiol* 2005;**139**:1612–4.
75. Rensing SA, Lang D, Zimmer AD, *et al*. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* 2008;**319**:64–9.
76. Cloonan N, Forrest AR, Kolle G, *et al*. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 2008;**5**:613–9.
77. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 2009;**25**:1026–1032.
78. Zheng S, Chen L. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Res* 2009;**37**:e75.
79. Ozsolak F, Platt AR, Jones DR, *et al*. Direct RNA sequencing. *Nature* 2009;**461**:814–8.
80. Van Tassell CP, Smith TP, Matukumalli LK, *et al*. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* 2008;**5**: 247–52.
81. McMullen MD, Kresovich S, Villeda HS, *et al*. Genetic properties of the maize nested association mapping population. *Science* 2009;**325**:737–40.
82. Smith DR, Quinlan AR, Peckham HE, *et al*. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res* 2008;**18**:1638–42.
83. Balwierz PJ, Carninci P, Daub CO, *et al*. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol* 2009;**10**:R79.
84. Shiraki T, Kondo S, Katayama S, *et al*. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci USA* 2003;**100**:15776–81.
85. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 2009;**4**:14.
86. Thimm O, Blasing O, Gibon Y, *et al*. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J* 2004;**37**:914–39.