Next-generation sequencing technologies for gene expression profiling in plants

Mukesh Jain

Advance Access publication date 6 December 2011

Abstract
Next-generation sequencing (NGS) provides a better approach to gene expression profiling with several advantages. The power of NGS along with novel molecular techniques and computational tools allow the researchers to perform the gene expression profiling to reveal transcriptional complexity of an organism and answering several biological questions. Although many studies for gene expression profiling related to various aspects have been performed in animal systems revealing unprecedented levels of complexity of transcriptomes, their use is still limited in plant biology. This review describes the use of NGS technologies with respect to gene expression profiling, bioinformatics challenges associated with data analysis and advances made so far in the plant biology research. We anticipate many more studies in recent future, which will surely advance our understanding of the complexity of plant genomes.

Keywords: gene expression profiling; next-generation sequencing; RNA-seq; transcriptome

INTRODUCTION
Sequencing-based approaches provide a better alternative to gene expression analysis as compared to hybridization-based methods. Serial analysis of gene expression (SAGE) was the first reported sequencing-based high-throughput method for gene expression profiling followed by massively parallel signature sequencing [1, 2]. However, despite several advantages over other methods, they have not been widely employed as that of microarray. The development of next-generation sequencing (NGS) technologies has highlighted the importance of sequencing-based approaches in gene expression profiling. The NGS technologies are used for whole-transcriptome sequencing (RNA-seq) or in conjunction with SAGE protocols for gene expression studies. The unprecedented level of sensitivity and high-throughput nature make NGS technologies as the method of choice for gene expression analysis [3–8]. RNA-seq is quickly superceding the microarray based approaches for studying gene expression, splice isoforms and novel transcribed regions. The absolute measurement of gene expression using RNA-seq provides greater quantitative and qualitative insight, and accuracy than microarray [8–11]. RNA-seq combined with appropriate bioinformatics tools provides a better approach to study gene expression dynamics on a global scale in different cellular and biological contexts. Although it offers several advantages, but also poses statistical (availability of user-friendly and reliable software programs/statistical methods) and computational (availability of sufficient computational and storage resources) challenges, which are still limiting the adoption of this technology.

The gene expression profiling using RNA-seq is based on the assumption that the depth of coverage of a sequence is proportional to the expression of corresponding gene of interest. Although a few studies reported bias in the representation of certain transcript sequences [12–15], several studies revealed RNA-seq as a powerful tool for studying gene expression.
expression and has clear advantages over other existing methods [16–18]. RNA-seq is also used to construct the complete transcriptome of an organism either by reference-based or de novo assembly [19, 20], which has been reviewed elsewhere in this special issue. Here, I discuss the different NGS technologies available for gene expression studies, challenges in RNA-seq data analyses and provide an overview of the studies performed in plants using this approach. This review will help the plant biologists in understanding various applications of the NGS technologies in gene expression studies.

RNA-SEQ AND NGS TECHNOLOGIES

In general, a RNA-seq experiment includes the conversion of a population of mRNA into a library of randomly broken cDNA fragments with adaptors ligated to one or both ends. This is followed by sequencing with or without PCR amplification from one end (single-end) or both ends (paired-end). The read length varies from 30 to 400 bp depending on the sequencing technology used. Three major NGS technologies are available for RNA-seq, including pyrosequencing-based Roche 454 GS-FLX, sequencing-by-synthesis based Illumina Genome Analyzer/HiSeq and sequencing-by-ligation based Life Technologies SOLiD (Sequencing by Oligo Ligation and Detection). These sequencing technologies and their use in various applications have been extensively reviewed elsewhere [21–24] and is beyond the scope of this review. With reference to their use in gene expression profiling, all the three technologies have their pros and cons. Roche 454 offers the advantage of longer reads (average of 400 bp), which are more likely to be uniquely mapped as compared to those of short reads from Illumina and SOLiD technologies. However, the number of reads generated per run is much less as compared to other two technologies. The higher number of reads generated by Illumina and SOLiD technologies makes them more cost-effective and provides the advantage of higher dynamic range of transcript expression detection and more accurate gene expression measurement. It has been shown that a very high sequencing depth is required to detect the rare transcripts or isoforms, which is achievable by using short-read platforms [25]. In addition, the recent improvements in the Illumina and SOLiD sequencing technologies provide sufficiently large reads (up to 150 bp), which will overcome the disadvantage of shorter read length. Although Illumina and SOLiD offer similar advantages and disadvantages, the use of Illumina platform has superseded the SOLiD. The advantage of Illumina technology has been demonstrated to be comparable to the microarray but superior in detection of low-expressed genes, alternative splice variants and novel transcripts [10].

In addition to above three most commonly used NGS technologies; more advanced single-molecule sequencing (SMS) technologies have also emerged very recently. These emerging technologies include Life Technologies semiconductor based Ion Torrent, Helicos solid-phase-based Genetic Analysis Platform and Pacific Biosciences single-molecule real-time sequencing-based PacBio RS [26–30]. A few other SMS technologies are also under development [29]. The advantages of direct RNA sequencing using Helicos SMS technology has been shown to achieve bias-free comprehensive understanding of yeast transcriptome [28]. Although SMS technologies certainly provide advancement over NGS technologies, their promise in diverse applications still needs to be proved.

DATA ANALYSIS AND CHALLENGES

Although RNA-seq offers unprecedented level of sensitivity, it poses vastly increased bioinformatics challenges in data handling (storage and processing) and analysis to extract biologically relevant information. A typical gene expression analysis experiment using RNA-seq involves four steps in data analysis (Figure 1). First step is the quality control of the data in which poor-quality reads are eliminated or trimmed from ends to accelerate subsequent data analysis. Secondly, the most important step is accurate mapping of sequencing reads to the corresponding reference genome/transcriptome [31]. This preliminary but crucial step is most computationally intensive and depends on the type of available sequence data (read length, amount of data and data format). In general, the reads that span exon junctions and map to multiple locations in the reference make the mapping step complicated. In addition, sequencing errors and polymorphisms also present problems to the read mapping. The situation is further complicated by alternative splicing and
presence of repetitive sequences in higher eukaryotes especially plants. Although some solutions to these problems (obtaining longer reads, removal of low-quality reads, trimming sequence reads, compilation of splice junction library and probabilistic assignment of the reads that map to multiple locations etc.) have been proposed, they further need to be addressed carefully by developing computationally better methods. Several programs, including Eland, SOAP, MAQ, SSAHA2, SHRiMP, Stampy, TopHat, RNA-MATE and Bowtie, are available for mapping reads to the reference implementing different alignment algorithms [4, 20, 31–36]. Although basic features of these programs are common, some of them offer unique features as well, such as use of base quality to improve alignment accuracy, identification of splice junctions and mapping of strand-specific RNA-seq data, etc. The use of most of mapping programs is limited to the identification and quantification of know exons/exon splice junctions. However, a few programs, including TopHat, MapSplice, SpliceMap and GSNAp are also available, which have the ability to identify new splice junctions in addition to the known ones [35, 37–39]. Further, many of these programs support mapping of paired-end reads, which provides increased specificity and is better for detection and quantification of splice variants.

In the third step, a normalized read count is assigned to each gene/transcript, which represents its true abundance in the sample. In general, the total number of mapped reads for a given gene/transcript is roughly proportional to both its abundance and length. Reads per kilobase per million of mapped reads (RPKM) has been proposed as a quantitative normalized measure for comparing the different genes within same sample or differential gene expression across different biological samples/conditions [4]. Different algorithms for the calculation of RPKM values have been implemented in different software packages. However, it has been demonstrated that RPKM method does not provide a completely unbiased measure of the expression level due to different transcript length and/or overrepresentation of 5’- or 3’-ends of the transcripts [13, 40]. As a last step, a statistical test is applied to identify the genes/transcripts exhibiting differential gene expression. Many statistical methods, including likelihood ratio test, Z-test, Fisher exact test and empirical Bayesian method based on Poisson or binomial distribution models have been proposed to analyze the significance of differential expression [5, 10, 41–47]. Several program/software tools have been developed for differential gene expression analysis based on different statistical algorithms, programming language, computational platforms and data size in last few years. Cufflinks, ALEXA-seq, DESeq, DEGseq, Myrna, MMSEQ, rQuant, edgeR and ERANGE are the few examples of softwares developed for differential gene expression profiling [47–52]. These programs provide several common and unique features to facilitate the measurement of gene expression on different types of data sets. The utility of most of these programs is limited to pairwise sample comparisons. The recently proposed empirical Bayes method, baySeq, can handle the analysis of experimental designs involving
multiple sample groups in addition to the pairwise comparisons and has been shown to perform better than many other methods [44].

The quantification of abundance of multiple transcript isoforms of same gene and multiple genes within same gene family is challenging. A few programs, such as ALEXA-seq, MMSEQ and Cufflinks support the estimation of isoform abundance based on different methods [48, 51, 53]. However, these methods have certain advantages and disadvantages and thus needs to be considered carefully [20]. Further, the evaluation of various statistical methods has showed that normalization and differential expression analysis strategies affect the gene expression results substantially and the issue becomes more important for lowly expressed genes [40]. In fact, main difference between test statistics is of their ability to handle low read counts [40]. Sometimes, the expression information of genes expressed at very low levels but are otherwise functionally significant may be lost. Therefore, the selection of appropriate statistical methods for normalization and differential expression analysis, which greatly impact the accuracy of results, is very important.

**RNA-SEQ AND INSIGHTS INTO PLANT BIOLOGY**

RNA-seq technology has been applied to various organisms to reveal the entire transcriptional landscape of gene activity and many other novel aspects [4, 25, 54, 55]. However, its use has been very limited in the plant biology research. Only a few proof-of-concept studies have been performed to reveal the transcriptional complexity in plants. Here, we provide an overview of the RNA-seq based gene expression studies performed in plants, which provide novel insights into the various biological aspects.

The analysis of transcript abundance based on Illumina sequencing in maize revealed the differential expression of a very high fraction of genes (64.4%) along the leaf development gradient [56]. The results provided the evidence for dynamic reprogramming of transcriptome with transcripts for basic cellular metabolism at the leaf base to transcripts for secondary cell wall biosynthesis and C4 photosynthetic development towards the leaf tip, which can serve as resource for a systems approach to understand the photosynthetic development in plants [56]. The transcriptional complexity during berry development in grapevine has been characterized by presenting an exhaustive overview of the gene expression dynamics and identification of splice variants using RNA-seq by Illumina sequencing [57]. Novel insights into molecular mechanisms of plant sex determination process were gained via the transcriptome sequencing of flower buds from two near-isogenic lines of cucumber representing different sex types [58]. An integrated transcriptome atlas of the model legume crop, soybean, has been generated, which resulted in the identification of tissue-specific genes [59]. Further, this expression data has been utilized for comparative analyses of gene expression from other legumes, *Medicago truncatula* and *Lotus japonicus*. The analysis revealed the events of functional conservation, sub-functionalization or neo-functionalization for orthologous genes and genes showing synteny among these species [59]. In another independent study, more than 177 genes involved in the agronomically important trait, seed filling process, have been identified in soybean using RNA-seq [60]. In a recent study, we have also identified the differentially expressed transcripts in different tissue samples of another legume, chickpea, using massively parallel pyrosequencing [18]. This study reported the identification of differentially expressed genes in a tissue-by-tissue comparison and tissue-specific transcripts. Furthermore, tissue-specific chickpea transcripts exhibiting tissue-specific expression were also identified [61]. The transcriptional complexity and evolution of meiosis in *Arabidopsis* has been unraveled in two independent studies taking the advantage of RNA-seq technology [62, 63].

Pyrosequencing data provided the evidence for active transcription of at least 648 predicted gene loci in *Arabidopsis*, for which no expression evidence was available previously [16]. Further, at least 60 loci were identified in the *Arabidopsis* genome, for which expression evidence was provided by pyrosequencing and likely represent protein-coding genes, but were not annotated as genes previously [16]. The function of these putative novel genes remains to be determined. The transcriptional complexity in rice has also been unraveled via sequencing of mRNA from various tissues in two subspecies. This study reported the identification of more than 15 000 novel transcriptionally active regions and 3464 differentially expressed genes [64]. Evidences for extended untranslated regions (UTRs) of the current *Arabidopsis* and rice gene models have also
been provided [64, 65]. A novel sequence-based approach using Roche 454 technology focused on sequencing unique fragments at 3'-UTRs of genes to distinguish closely related transcripts such as members of same gene family and quantify their expression [66]. The tag-based digital gene expression profiling of wild-type and RAMOSA3 (RA3) gene mutant maize plants using Illumina NGS technology helped in the identification of putative targets and defining mechanism of action of RA3 gene, which controls the determinacy of axillary meristems [67]. This study provided novel insights into the genetic control of branching, which is a very important aspect for crop improvement programs. Further, a few studies have also been performed to understand the metabolic processes and pathways in plants. Illumina-based digital gene expression profiling has been used to study the molecular processes regulating lycopene accumulation in red-flesh mutant of sweet orange [68]. Recently, Hao and colleagues [69] demonstrated that taxane biosynthetic genes are preferentially expressed in roots as compared to leaf and stem and proposed the utility of NGS technologies in defining metabolic pathways. The deep sequencing of wild-type and snrk2.1 mutants of Chlamydomonas revealed the massive changes in cellular physiology and metabolic pathways critical to cell survival in response to sulphur depletion [11]. Further, the characterization of gene expression data and its integration with epigenome using Illumina technology for two rice subspecies and their reciprocal hybrids presented a comprehensive overview of the transcriptional and epigenetic trends in heterotic rice crosses [70]. This study provided a resource for studying the genetic and epigenetic basis of gene action in different genetic backgrounds leading to phenotypic variability. The application of NGS technologies in transcriptomics of plant pathogens have also been reviewed recently [71].

Alternative splicing and RNA editing are very important mechanisms, which enhances the transcriptome plasticity and proteome diversity of an organism. In animals, RNA-seq has been exploited to show the alternative splicing of up to 95% of the genes and has been associated with diseases [53, 72, 73]. In plants, the sequencing of transcriptome at single-base resolution showed the alternative splicing of ~42 and 33–48% of the intron-containing genes in Arabidopsis and rice, respectively [64, 74, 75], which is much higher than the previous estimation based on Sanger sequencing [76, 77]. Many of the alternatively spliced isoforms were having premature stop codon, which targeted their degradation via nonsense mediated mRNA decay machinery or regulated unproductive splicing and translation mechanism to regulate the gene expression [74]. In another study, the use of RNA-seq in the global identification of RNA editing sites has been demonstrated in plants [78]. The study suggested the use of both Illumina and SOLiD technologies together greatly improve the detection of RNA editing. Further, RNA-seq has also led to the identification of several events of transcript fusions and evidence for their functionality has been shown [75].

**FUTURE PERSPECTIVES**

The most important challenge in using NGS technologies for studying gene expression is in the downstream computational analysis and in the bioinformatics infrastructure required for large amount of data generated and correct data interpretation. Besides the requirement of huge storage space and processing capability, the quality control of sequencing data and alignment to reference are equally challenging. The large data sets may allow more accurate estimation of abundance of transcript levels, but will increase data deluge. Finally, the data analysis, visualization and correct interpretation require significant level of expertise in bioinformatics. The availability of complete information about a RNA-seq-based gene expression profiling experiment and uniformity across various data sets is also equally important. Therefore, guidelines need to be formulated for the minimal requirements for publication of NGS technologies based gene expression data in journals and online databases.

Besides several challenges in RNA-seq data analysis, we anticipate its use in more studies in plant biology. As revealed from several studies in animals and few proof-of-concept studies in plants, it is clear that the transcriptional complexity is considerably higher than previously anticipated in eukaryotes. The single-base resolution provided by RNA-seq has the potential to revise several aspects of gene/genome annotation in plants. I expect to see their use in studies that evaluate transcriptome dynamics during plant development and in response to various environmental stimuli, including abiotic and biotic factors. The use of RNA-seq to study gene
regulation via gene expression studies will be very interesting, which may provide novel insights into the gene regulatory network governing the biological processes. In addition, RNA-seq may also be used to study several novel aspects in plants related to transcriptional and post-transcriptional gene regulation such as cell-type specific transcriptome, alternative initiation codon, switch in the dominant transcription start site or splice form according to the biological context, alternative polyadenylation, antisense loci, determination of RNA secondary structure, RNA editing and gene fusions, etc. The integration of gene expression data with other data sets such as epigenome and small RNA will help understanding of how various biological processes work together in a cellular context. The recently emerged direct RNA sequencing method will be very useful for various transcriptome-related studies in plant biology, which require very small quantities of RNA/small RNA samples. It is anticipated that new emerging sequencing technologies along with availability of better data analysis tools will surely advance gene expression analyses studies to gain novel insights in plant biology.

Key points
- Transcriptome sequencing using NGS technologies provides better alternative for the gene expression studies.
- Novel transcripts and transcript isoforms and their differential expression can be identified using RNA-seq.
- Several software tools are available for the analysis of gene expression using RNA-seq and analysis requirement needs to be considered carefully.
- RNA-seq based proof-of-concept studies have provided some novel insights into various biological processes in plants.
- More diverse studies on different aspects of transcriptome analyses in plants are anticipated in future.

Acknowledgements
M.J. is an Innovative Young Biotechnologist Awardee from the Department of Biotechnology, Government of India. Thanks are due to Dr R. Garg for suggestions and valuable comments.

FUNDING
The Department of Biotechnology, Government of India, New Delhi under the Next Generation Challenge Programme on Chickpea Genomics and core grant from the National Institute of Plant Genome Research (NIPGR).

References


