A next-generation approach to the characterization of a non-model plant transcriptome*

*Mukesh Jain***

Next-generation sequencing (NGS) technologies provide a revolutionary tool with numerous applications in transcriptome studies. The power of NGS technologies to address diverse biological questions has already been proved in many studies. One of the most important applications of NGS is the sequencing and characterization of transcriptome of a non-model species using RNA-seq. This application of NGS technologies can be used to dissect the complete expressed gene content of an organism. In this article, I illustrate the use of NGS technologies in transcriptome characterization of a non-model species taking example of chickpea from our recent studies.

Keywords: Expressed sequence tags, next-generation sequencing technologies, non-model species, transcriptome.

TRANSCRIPTOME is the complete set of all transcripts, including mRNA and non-coding RNAs, in a cell. The study of transcriptome is essential to understand and interpret the functional complement (gene content) of the genome of an organism. The lack of genome sequence information limits gene discovery in non-model species. The generation of expressed sequence tags (ESTs) derived from protein-coding mRNA sequences is considered as the most useful approach for gene discovery. However, the Sanger sequencing-based generation of ESTs is costly and time-consuming. With the advent of next-generation sequencing (NGS) technologies such as Roche 454, Illumina and Life Technologies SOLiD, gene discovery via RNA sequencing (RNA-seq) has become rapid and cost-effective. These sequencing technologies provide very high throughput by generating million(s) of reads in a single sequencing run of few hours to few days.

NGS technologies provide a revolutionary tool with numerous applications that can address different biological questions. The NGS technologies and their various applications have been comprehensively reviewed[1–4]. Briefly, Roche 454 is a pyrosequencing-based sequencing technology, which generates more than one million long reads with average length of about 400 bp in a single run of a few hours. Roche 454 was the first high-throughput sequencing platform introduced commercially in 2004. The initial model (GS20) had a read-length of 100 bp, but the new model (GS-FLX) produces larger reads with average length of 400 bp. Although the raw base accuracy of Roche 454 platform is more than 99%, it is error-prone for homopolymers. Illumina is a short-read sequencing platform based on sequencing-by-synthesis principle and generates several million reads of desired length up to 150 bp. Illumina Genome Analyser (GA) model launched in 2006 was capable of generating tens of millions of reads of 32 bp length. However, latest models, Illumina GAIIx and Hiseq, can produce hundreds of millions of reads of up to 150 bp. The raw base accuracy in Illumina sequencing is greater than 99.5%. SOLiD is also a short-read sequencing platform based on sequencing-by-ligation principle, which generates millions of reads of length up to 75 bp with a raw base accuracy of 99.94%. These technologies allow the use of DNA/RNA fragments directly for sequencing without the requirement of their insertion into a vector, thus removing the costly and time-consuming steps of Sanger sequencing. The cost of sequencing varies with different platforms. In terms of data output, Roche 454 technology is the most expensive among the three NGS platforms. However, the cost is much less than that of the automated Sanger sequencing. Further, the sequencing cost for a given platform is dependent on various factors like number of samples to be sequenced in a single run, read-length and read-type (single-/paired-end), etc. A comparative account of various features of Sanger and NGS technologies is provided in Table 1. These technologies offer the potential to interrogate the transcriptional complexity of an organism efficiently at single-base resolution via RNA-seq. Various applications of NGS technologies include transcriptome characterization, identification of novel transcripts/transcript isoforms, measurement of gene expression, identification of single-nucleotide polymorphisms and

*Based on a talk delivered during the 22nd Mid-year Meeting of the Indian Academy of Sciences, Bangalore held at the Indian Institute of Science, Bangalore from 8 to 9 July 2011.

**Mukesh Jain is in the National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110 067, India.
e-mail: mjain@nipgr.res.in

Table 1. Features of the sequencing technologies used for transcriptome applications

Sequencing platform	ABI 3730 × L	Roche 454 GS-FLX	Illumina GA/HiSeq	Life technologies SOLiD
Sequencing method	Automated Sanger	Pyrosequencing	Sequencing-by-synthesis	Sequencing-by-ligation
Template amplification method	Cloning	Emulsion PCR	Bridge PCR	Emulsion PCR
Read-length (bp)	Up to 1200	300–600	Up to 150	Up to 75
Accuracy (%)	99*	> 99	> 99.5	> 99.94
Throughput	~ 0.08 Mbp/run	~ 500 Mbp/run	up to 600 Gbp/run	> 100 Gbp/run
Run time	1 h	10 h	Up to 14 days**	Up to 10 days**
Cost	High	Moderate	Low	Low
Sensitivity	Low	Moderate	High	High
Bioinformatics infrastructure requirements	Low	High	Very high	Very high

*For reads 100–700 bp long. **Run time depends on read-length and read-type (single-/paired-end).

Box 1. Potential applications of next-generation sequencing (NGS) technologies in transcriptome studies

- Transcriptome characterization.
- Detection of novel transcribed regions.
- Detection of antisense transcripts.
- Detection and quantification of alternative splicing/transcript isoforms.
- Detection of alternative initiation codon/polyadenylation sites.
- Identification of alternative promoters/UTRs.
- Identification of non-coding RNAs.
- Gene expression studies.
- Sequence variations/detection of polymorphisms.
- Allele-specific expression.
- Detection of gene fusions.
- Determination of RNA secondary structure.
- Detection of RNA editing.
- Genome annotation.

defining gene structure (genome annotation). A comprehensive list of potential applications of NGS technologies in transcriptome studies is given in Box 1. The power of NGS technologies along with appropriate computational tools have already been proved in various studies. However, despite the availability of such technologies, EST sequencing and gene discovery have been limited in some organisms. This is mainly because of availability of resources and lack of expertise in data analysis tools.

Sequencing and characterization of transcriptome of non-model species using RNA-seq is one of the most important applications of NGS technologies[5]. One of the most important challenges in NGS is *ab initio* construction of transcriptome of an organism for which the genome sequence is not available. Recently, NGS technologies have facilitated the transcriptome characterization of several non-model species without a sequenced genome, such as mangroves[6], chestnut[7], *Artemisia*[8], olive[9], tea[10], living fossil tree *Ginkgo*[11], microalgae *Dunaliella tertiolecta*[12], gymnosperm *Taxus*[13], vegetable carrot[14], and legumes *Pisum sativum*[15] and *Cicer arietinum*[16,17]. These transcriptome studies helped in gene discovery and provided novel insights into various unique species-specific biological processes/pathways. Although Illumina produces several orders of magnitude more sequence data at a fraction of the cost of the Roche 454 platform, the latter remains dominant as the platform of choice for transcriptome sequencing. This is because the short-read sequencing poses a great challenge in the *de novo* assembly for non-model organisms. However, a few studies have now demonstrated that high-throughput short-read sequencing can also be used for *de novo* construction of transcriptomes and need not be restricted to the model organisms[13,14,16].

Chickpea is an important legume crop having high nutritional value and the unique ability to fix atmospheric nitrogen. Despite its economic value, fewer efforts have been made to generate genomic resources for chickpea. Recently, we sequenced, assembled and characterized the transcriptome of chickpea using two NGS technologies, Roche 454 and Illumina[16,17]. The transcriptome of chickpea was sequenced from various tissue samples to present the gene content in chickpea[16,17]. An overview of the strategy used for the characterization of chickpea transcriptome is described below and presented diagrammatically in Figure 1. A total of about 107 million high-quality short reads were obtained by sequencing three libraries constructed from root, shoot and mixed tissue samples using Illumina platform[16]. An optimized *de novo* assembly of these reads using Oases program generated a total of 53,409 non-redundant transcripts of an average length of 523 bp. To further improve the quality of chickpea transcriptome, six libraries constructed from different tissue samples (root, shoot, mature leaf, flower bud, young pod and mixed tissue) were sequenced using Roche 454 technology[17]. The sequencing resulted in the generation of about two million high-quality reads of 372 bp average length. Further, the *de novo* assembly of these reads was optimized using different assembly programs. Although merged assembly has been suggested to give better results in an earlier study[18], it did not give good results with the chickpea dataset. In fact, merged assembly generated redundant and chimeric transcripts and thus was not considered. However, a hybrid assembly of the primary assemblies of Roche 454 (obtained using Newbler v2.3) and short-read (obtained using Oases) datasets

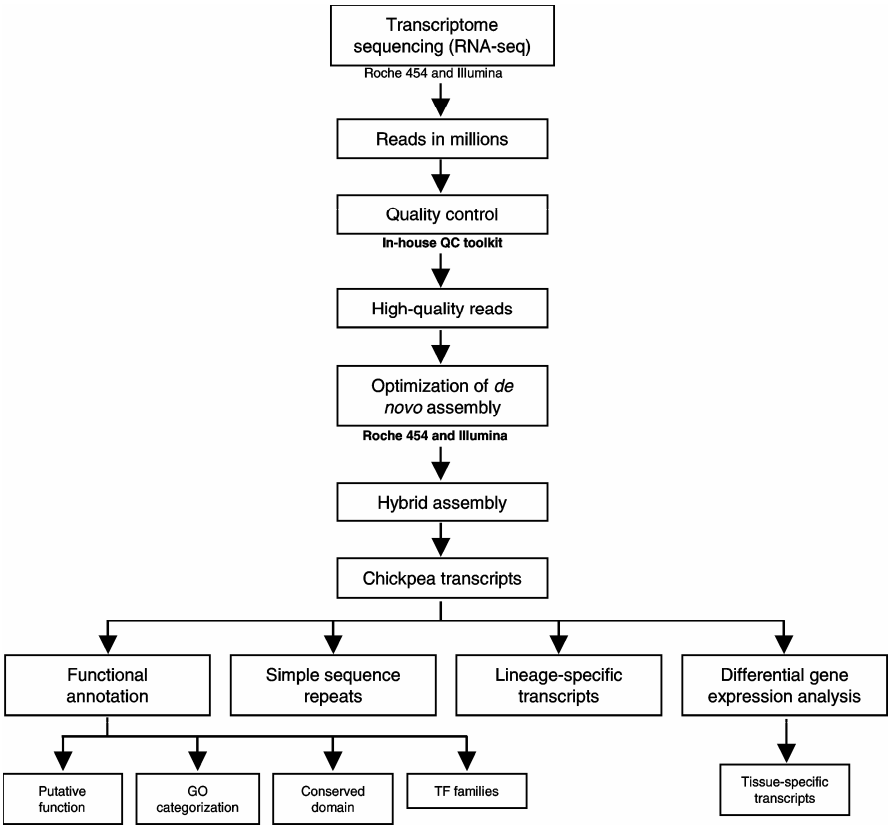

Figure 1. An overview of the strategy followed for sequencing and characterization of chickpea transcriptome by Garg *et al.*[16,17]. The sequencing of chickpea transcriptome was performed using Illumina and Roche 454 platforms, which resulted in millions of reads. After several quality control (QC) steps using in-house developed toolkit, high-quality reads were obtained, which were used for the optimization of *de novo* assembly using various programs. After optimization of *de novo* assemblies of Illumina and Roche 454 datasets, a hybrid assembly of the two primary assemblies results in the final optimal assembly and a total of 34,760 chickpea transcript sequences were obtained. These transcript sequences were subjected to functional annotation (assignment of putative function, gene ontology (GO) categorization, and identification of conserved domains and transcription factors (TF)), identification of microsatellites (simple sequence repeats) and lineage-specific transcripts, and differential gene expression analysis to identify the tissue-specific transcripts.

using TGICL program gave the best assembly results based on several assessment criteria[17]. This assembly generated a total of 34,760 transcripts of an average length 1,020 bp representing 35.5 Mb of total transcriptome size.

Further, functional annotation of the chickpea transcripts was performed to assign putative biological function(s). Putative function was assigned to 73.2% of the transcripts based on their similarity with *Arabidopsis* proteins or proteins/domains from other public databases and at least one GOSlim term was assigned to 71.9% of the chickpea transcripts. In addition, 5.3% transcripts encoding for putative transcription factors were identified. Two sets of lineage-specific transcripts, legume-specific (highly conserved in legumes, but not detected in non-legumes) and chickpea-specific (putative novel transcripts lacking significant similarity with any sequence from other plants available in public databases), were also identified. Further, many of the lineage-specific transcripts exhibited tissue-specific expression in chickpea[19]. The study of these transcripts will be important to gain insights into species-specific functions and evolutionary processes. More than 4000 simple sequence repeats (SSRs) classified into di-, tri-, tetra-, penta- and hexa-nucleotide repeats were identified in about 10% of the transcripts, which could be developed as functional markers for various mapping and breeding purposes.

One of the major applications of NGS technologies is the measurement of gene expression and differential gene expression analysis, which provides several advantages over other hybridization-based methods, including higher sensitivity and greater dynamic range of gene expression detection[5,20,21]. Utilizing the Roche 454 sequencing data from different libraries, genes differentially/preferentially expressed in root, shoot, mature leaf, flower bud and young pod tissues of chickpea were identified using a normalized count of reads corresponding to each transcript represented in different tissue samples[17]. The results of differential gene expression could be validated successfully using quantitative real-time PCR analysis. Finally, a public database, Chickpea Transcriptome Database (http://www.nipgr.res.in/ctdb.html), was developed, which allows data mining via a variety of search options.

Box 2. Salient highlights of the NGS-based chickpea transcriptome characterization studies[16,17,19]

- Chickpea transcriptome was sequenced at a very high-depth using two NGS platforms; short-read Illumina and long-read Roche 454.
- *De novo* assembly optimization of Illumina and Roche 454 data was performed using various assembly programs and parameters.
- Hybrid assembly of Illumina and Roche 454 data gave better results than individual and merged assemblies.
- A total of 34,760 transcripts of an average length of 1,020 bp representing 4.8% (35.5 Mb) of the genome sequence were generated.
- At least 4,111 simple sequence repeats were identified in the chickpea transcriptome.
- Functional characterization revealed the representation of diverse genes in the transcript dataset.
- At least 1,851 transcription factor encoding genes were identified in chickpea.
- Chickpea transcripts expressed in tissue-specific manner were identified.
- Lineage-specific genes were identified and their tissue-specific expression was revealed.
- A public web resource (CTDB) was developed.

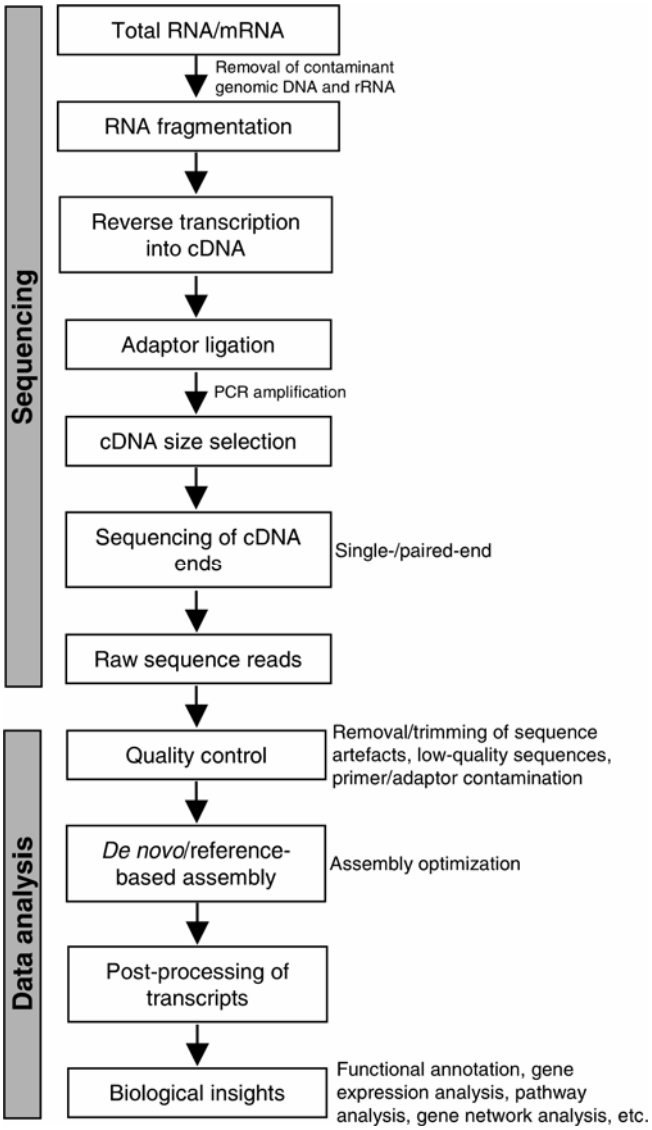

Figure 2. Workflow of a typical RNA-seq experiment. The sequencing step involves extraction of total RNA/mRNA, which after removal of genomic DNA/rRNA contamination is fragmented into smaller fragments. After fragmentation, mRNA is reverse transcribed into cDNA followed by ligation of adaptors and PCR amplification. Finally, cDNA fragments of a particular size range are selected and sequenced using NGS technologies at one (single-end) or both (paired-end) ends to generate raw reads. The data analysis step involves quality control of raw data to remove/trim sequence artefacts followed by optimization of *de novo*/reference-based assembly using various programs and parameters. After optimization of the assembly, transcript sequences are processed for error correction and analysed to obtain various biological insights.

These studies[16,17,19] provided a comprehensive genomic data resource for chickpea, which will surely facilitate further research in chickpea and other legumes. For instance, the salient highlights of our studies on chickpea transcriptome sequencing and characterization are given in Box 2. A similar strategy can be utilized successfully to characterize the gene content of any non-model plant species of interest.

The steps involved in a typical transcriptome sequencing and characterization experiment using NGS technologies are shown in Figure 2. In general, with respect to transcriptome sequencing and characterization using NGS technologies, we need to consider several issues. First, one needs to design the experiment carefully in terms of tissue samples to be selected to get the most possible representation of transcripts, sequencing platform, sequencing depth, read-type (single-end or paired-end) and read-length. The sequencing cost should also be considered while designing the experiment. Once the experiment is designed, sequencing may be performed accordingly, which is relatively easy as several commercial sequencing service providers are available. After the sequencing, raw data quality control is essential to filter poor-quality sequences, which facilitates easier downstream data processing and better biological interpretation of the results. The next step of sequence assembly is the most important and challenging in the characterization of the transcriptome. Although, several assembly programs based on different algorithms with different features have been developed, the most suitable program and its parameters need to be optimized for each dataset. Further, different strategies may also be employed for the optimal assembly output. Several assessment criteria (assembly statistics, number of reads assembled, read mapping and alignment to reference) need to be considered to select the best assembly. In addition, the assembly process requires high computational resources for large datasets generated using NGS. For example, we assembled the chickpea transcriptome datasets using a server with 48 processors and 128 GB random access memory[16,17]. Therefore, the analysis requirement also needs to be considered carefully. After the optimal transcriptome assembly, the sequences are subjected to various analyses to assign putative function(s) and answer specific biological question(s).

The power of NGS technologies along with appropriate computational tools has already been proved in many studies. Recently, third-generation sequencing platforms (Helicos, Ion Torrent and PacBio)[22] have also been launched, which ensure much more data output at a lower cost. The use of these technologies is increasing rapidly and I anticipate the sequencing and characterization of transcriptomes of thousands of non-model plant species in the coming years, which could be utilized for their genetic enhancement. However, the tremendous challenges associated with data handling, analysis and presentation posed due to huge datasets will need great attention.

1. Mardis, E. R., Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet*., 2008, **9**, 387–402.
2. Simon, S. A. *et al*., Short-read sequencing technologies for transcriptional analyses. *Annu. Rev. Plant Biol*., 2009, **60**, 305–333.
3. Varshney, R. K. *et al*., Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol*., 2009, **27**, 522–530.
4. Metzker, M. L., Sequencing technologies – the next generation. *Nature Rev. Genet*., 2010, **11**, 31–46.
5. Morozova, O., Hirst, M. and Marra, M. A., Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genomics Hum. Genet*., 2009, **10**, 135–151.
6. Dassanayake, M. *et al.*, Shedding light on an extremophile lifestyle through transcriptomics. *New Phytol*., 2009, **183**, 764–775.
7. Barakat, A. *et al.*, Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biol*., 2009, **9**, 11.
8. Wang, W. *et al.*, Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics*, 2009, **10**, 10.
9. Alagna, F. *et al*., Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics*, 2009, **10**, 15.
10. Shi, C. Y. *et al.*, Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics*, 2011, **12**, 131.
11. Lin, X. *et al.*, Functional genomics of a living fossil tree *Ginkgo* based on next generation sequencing technology. *Physiol. Plant*., 2011, doi:10.1111/j.1399-3054.2011.01500.x.
12. Rismani-Yazdi, H. *et al.*, Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: pathway description and gene discovery for production of next-generation biofuels. *BMC Genomics*, 2011, **12**, 148.
13. Franssen, S. U. *et al.*, Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. *BMC Genomics*, 2011, **12**, 227.
14. Hao, da C. *et al.*, The first insight into the tissue specific *Taxus* transcriptome via illumina second generation sequencing. *PLoS One*, 2011, **6**, e21220.
15. Iorizzo, M. *et al.*, *De novo* assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. *BMC Genomics*, 2011, **12**, 389.
16. Garg, R. *et al.*, *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res*., 2011, **18**, 53–63.
17. Garg, R. *et al*., Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. *Plant Physiol*., 2011, **156**, 1678–1771.
18. Kumar, S. and Blaxter, M. L., Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics*, 2010, **11**, 571.
19. Garg, R. and Jain, M., Pyrosequencing data reveals tissue-specific expression of lineage-specific transcripts in chickpea. *Plant Signal. Behav.*, 2011, **6**, 1868–1870.
20. Marioni, J. C. *et al*., RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*., 2008, **18**, 1509–1517.
21. Mortazavi, A. *et al*., Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 2008, **5**, 621–628.
22. Schadt, E. E., Turner, S. and Kasarskis, A., A window into third-generation sequencing. *Hum. Mol. Genet.* 2010, **19**, R227–R240.

ACKNOWLEDGEMENT. The work on chickpea genomics is financially supported by the Department of Biotechnology, New Delhi.

Received 2 September 2011; revised accepted 20 October 2011