The review of transcriptome sequencing: principles, history and advances

Haotian Zhang

1Basic Medical Department, Peking University, No. 38 Xueyuan Road, China

Abstract. A transcriptome is a collection of RNA transcribed from a particular issue or cell at a certain developmental stage or functional state. Studies of transcriptomes could reveal gene function and gene structure, and promote our understandings on specific biological processes and molecular mechanisms. Transcriptome sequencing, or RNA-seq technologies, allow efficient transcriptome measurement and empower transcriptome research, by subjecting RNA-derived to high-throughput sequencing analysis. In recent years, a variety of improvements on RNA-seq methods have emerged, spanning from the acquisition of target RNA, the fragmentation of RNA, the synthesis of cDNA, and the sequencing methods, to meet the needs of different occasions and conditions. In addition, single cell RNA-seq method has been developed to better characterize the transcriptomes of various cell types in biological tissues and reveal the heterogeneity of gene expression between cells. This article focuses on the principle, development, and application of RNA-seq technologies, and elaborates on various up-to-date RNA-seq improving methods.

1. Introduction
With the advent of the post-genomic era, various omics techniques such as transcriptomics, proteomics, and metabolomics have emerged, of which transcriptomics is the first and most widely used technology [1]. Transcriptomics studies transcriptome. A transcriptome is the sum of all RNAs transcribed from particular tissue or cell at a certain developmental stage or functional state, including messenger RNA (mRNA) and non-coding RNA (ncRNA). mRNA, considered as ‘bridges’, precisely regulates the transmission of genetic information from DNA to protein [2], while ncRNA regulates gene expression, protein synthesis, and different cellular activities at different levels [3]. Therefore, understanding of transcriptome promotes the study of the activities of cells, tissues and organisms. Transcriptome sequencing, also known as RNA-Seq, is a recently developed technique that quantifies transcriptome in given samples and thus promotes transcriptome analysis [4].

2. The history transcriptome sequencing
In 1970s, Walter Fiers and his coworkers sequenced the complete transcriptome of bacteriophage MS2, which is 3,569 nucleotides long [5], pioneered the study of transcriptome sequencing. However, because of the single-stranded structure, RNA is very unstable and is easily degraded by ubiquitous RNases in cells. Accurate sequencing of larger transcriptome is rigorous.

With the discovery of reverse transcriptase [6], it is possible to convert the mRNA and ncRNA to stabilizer DNA. The reverse synthesized DNA are named complementary DNA (cDNA). Facilitated by the DNA sequencing system founded by the British chemist Sanger in 1975 [7], and the cDNA PCR
method that exponentially amplifies cDNA creatively applied by Iscove [8], it became possible to sequence RNA by sequencing the cDNA it synthesized.

Based on these pioneering work, microarray (chip) technology came into being [9]. Microarray technology utilizes the principle of molecular hybridization to mount hundreds of known partial sequence DNA probes on a solid support slide or nylon membrane using an automated instrument Arrayer. A large number of genes can be quantitatively detected through one hybridization. With reverse synthesized cDNA, transcriptome of different biological samples can rapidly be sequenced [10], making microarray a powerful and systematic way to fully explore gene function. However, the design and functional application of the technology are mostly based on known genes, and the analysis of unknown genes is still limited. In addition, it is difficult for the chip to detect multiple transcripts formed by alternative splicing.

In recent years, another new technology, RNA-Seq, has been flourishing. RNA-Seq is based on next generation sequencing (NGS) [11]. NGS, or high-throughput sequencing, is represented by Roche's 454 technology, Illumina's Solexa technology, and ABI's SOLiD technology [12]. It is characterized by shorter reads and stronger ability in sequencing thousands of or even millions of DNA molecules in parallel with shorter time and lower price in contrast to the first-generation sequencing (Table 1).

RNA-seq is one of the most important applications of next-generation sequencing technology, and one of the most important tools for transcriptome study. It overcomes the drawbacks of microarray and brings deeper insight for transcriptome research.

Table 1. Comparison of different sequencing techniques

| Technology          | Principle                  | Signal     | Through-put | Resolution | Sequencing Read | Initial amount of nucleic acids |
|---------------------|----------------------------|------------|-------------|------------|----------------|---------------------------------|
| DNA Sequencing      | First Generation Sequencing| Dideoxychain-termination method(Sanger Sequencing) | Digital signal | Low        | Single bp    | Long                            |
|                     | Next Generation Sequencing | High through-put Sequencing | Digital signal | High       | Single bp    | Short                           |
| Microarray          | Hybridization              | Fluorescence signal | High        | Several-10 0bp | Long           | High                            |
| RNA Sequencing      | High through-put Sequencing| Digital signal | High        | Single bp  | Short         | Low                             |

3. Use of transcriptome sequencing

RNA-seq promotes the understanding of gene expression under different conditions, and allows for the discovery of new genes, transcription patterns, and RNAs [13-22], which helps to understand cell function and metabolic mechanism.

First, RNA-seq allows for comparison of gene expression levels between different samples (differential gene expression analysis). Before the advent of deep sequencing technology, the main means of measuring the expression levels of different genes was the microarray. However, hybridization techniques have limited sensitivity, are difficult to detect low-abundance targets, and cannot capture small changes in the expression level of the target gene [14]. Therefore RNA-Seq is more accurately than the microarray. In principle, RNA-Seq has the potential to determine the absolute number of each molecule in a cell population and directly compare the results between experiments. In cancer research, RNA-seq is widely used to help with the better understanding the pathogenesis of cancer. For example, Tuch et al. compared the gene expression of oral squamous cell carcinoma and normal cells and found that a group of differentially expressed genes in cancer cells with cell adhesion and differentiation functions, revealing that allele-specific deletions and duplications can induce cancer [15]. In another study, RNA-seq technology was used to explore the differences in gene expression of 69 individuals with lymphoblasts. The study found that more than a thousand genetic
variants affect overall expression levels or spliced genes and demonstrated the high efficiency of high-throughput sequencing in detecting differential expression of genes [16].

Second, RNA-seq promotes new gene discovery. The annotations to transcripts in existing databases may not be comprehensive. The outputs of RNA-seq could be self-assembled without known genome annotations, which allow the identification of new genes. For example, in 2014, a team of researchers identified four new genes relative to the flexibility of reproduction in response to the different salinity for Oreochromismossambicus using RNA-seq technology [17].

Third, RNA-Seq also demonstrates its great potential in identifying sequence differences such as fusion gene identification and coding sequence polymorphism studies. Alternative splicing allows one gene to produce multiple mRNA transcripts, and different mRNAs may be translated into different proteins with different functions. In eukaryotes, alternative splicing is ubiquitous. The mRNA precursors (pre-mRNA) formed by gene transcription can be formed differently during the splicing process by removing different intron regions or retaining different exon regions. In RNA-seq, the sequences of all transcripts, including the sequences spanning the splice junction regions, can be detected with appropriate sequencing depth. The depth describes the ratio of the total number of bases (bp) to genome size (Genome) obtained by sequencing, which is one of the indicators for evaluating the amount of sequencing. For instance, an RNA-Seq analysis of Arabidopsis thaliana by Sergei et al [18] found that at least about 42% of genes containing introns were subjected to variable splicing.

Lastly, an important aspect of transcriptomics research is the discovery and analysis of ncRNA (non-coding RNA). Studies showed that at least 93% of the human genome is transcribed into RNA [19]. With less than 2% as protein encoding region [20], the remaining 91% of the genome could be transcribed into non-protein-encoded RNA molecule, i.e. ncRNA. Some of the ncRNAs regulate gene expression at multiple levels by binding to DNA, RNA, and proteins [21], while far more ncRNA and their functions remain to be revealed. With RNA-seq technique, the researchers from National Research Council identified 2046 genes and 47 long ncRNAs that are first affected in hippocampal region CA1 in LOAD (late-onset Alzheimer’s disease) patients. To be noted, the long ncRNAs are severely deregulated in the hippocampus of patients. These discoveries suggested the importance of long ncRNA in Alzheimer’s disease, and the molecular changes may reveal new directions for there search of the disease [22].

4. Advances in transcriptome sequencing

The workflow to prepare total RNA from biological samples for RNA-seq is described in Figure 1. 1) Target RNAs were extracted from total RNA. Poly(A)-tails are unique features of mRNA. For mRNA extraction, poly(T) oligo nucleotides are commonly used. ncRNA without poly(A)-tails are extracted with size selection or ribosome RNA depletion [23,24,25]. Next, 2) the resulting RNA was randomly broken into fragments by physical fragmentation (sonication), enzymatic methods (transposases, restriction endonucleases such as RNAse III) and chemical fragmentation (heat, Mg2+ or Ca2+), and 3) the cDNA fragment are synthesized from the RNA fragment by random primers and reverse transcriptases. 4) Double-stranded cDNA are then synthesized with Ribonuclease (RNase H, DNA polymerase I or T4 polynucleotide kinase). Finally, 5) the cDNA fragment is end-repaired and ligated to the sequencing adapter to obtain the final cDNA library for sequencing. This commonly used Transcriptome sequencing process is called TruSeq.
During transcription, only one of the two DNA strands is transcribed to RNA, and we describe it as the direction the RNA is transcribed. With TruSeq method, the direction information is lost inevitably. However, for new gene identification and functional analysis, knowing the direction of the transcript is particularly important. Thus, it is necessary to preserve the RNA direction information. Recently, several RNA-seq sample preparation methods that could retain direction information during library preparation have been reported in the literature [26-28], such as QuantSeq and SLAMseq. QuantSeq is a 3’ mRNA-Seq library preparation protocol. Due to its simplicity, cost-effectiveness and excellent data quality, QuantSeq is the preferred method for RNA-Seq gene expression analysis [29]. It screens poly(A) tails by oligo dT primers during reverse transcription, while the second strand is synthesized by random primers. The random primers introduce the Illumina specific adaptors sequencing when synthesizing cDNA, so the specificity of strand is maintained well. SLAMseq, an alkylation-based RNA metabolic sequencing method based on linked thiol (SH), enables simultaneous quantitative analysis of nascent and existing RNA in total RNA sample [30]: the cells are cultured with S4U-containing medium to label the nascent RNA, and the reverse transcriptase will pair with guanine (G) instead of adenine (A) at the position of the reduced S4U-modified nucleotide.

Optimizations are also made for other steps of RNA-seq. In the step of RNA fragmentation, a new method can use the Tn5 transposase to break up and label RNA instead of using nucleases, and thus develop the Nextera kit [31]. The transposase can simultaneously catalyze the fragmentation of RNA and the incorporation of adaptors, thereby reducing the step and time of library preparation, and the insertion sites of the T5 transposase have a smaller intrinsic bias than the mechanical fragmentation and endonuclease fragmentation [32]. The BRB-seq (Bulk RNA Barcoding and sequencing) technology recently released has been improved on the basis of the Nextera kit. It is able to process large batches of samples quickly and economically while ensuring the accuracy of the results by further optimizing the adaptor for RNA-seq [33].

5. Transcriptome sequencing on single cell level
Recent studies showed that gene expression of similar cells can be heterogeneous, which is determined by different derived genomes, cell cycle and micro-environment [34]. However, the heterogeneity is lost with the conventional RNA-seq method, as it obtains the commonality of a large amount of cellular gene expression information. In order to dissect these cellular heterogeneities, it is necessary to perform gene expression analysis at the single cell level [35].

Single-cell RNA-Seq technology provides information on the expression of nearly 10,000 genes in a single cell, providing a powerful tool for identifying transcriptome features of various cell types in biological tissues and for revealing the heterogeneity of gene expression between cells [36]. Transcriptome analysis of individual cells is derived from single-cell cDNA amplification method reported by Brady et al. [37] in 1990. In 2009, Tang et al. [38] first reported single-cell RNA-Seq...
technology based on high-throughput sequencing. Since then, single-cell RNA-Seq technology has developed rapidly. Here, we summarized several prevalent single-cell RNA-seq approaches in use.

Lack of cell population isolation maker and amplification bias caused by transcript abundances variation in different cell groups are two barriers faced by Single-cell RNA-Seq technology provides information on the expression of nearly 10,000 genes in a single cell, providing a powerful tool for identifying transcriptome features of various cell types in biological tissues and for revealing the heterogeneity of gene expression between cells[36]. Transcriptome analysis of individual cells is derived from single-cell cDNA amplification method reported by Brady et al.[37] in 1990. In 2009, Tang et al.[38] first reported single-cell RNA-Seq technology based on high-throughput sequencing. Since then, single-cell RNA-Seq technology has developed rapidly. Here, we summarized several prevalent single-cell RNA-seq approaches in use.

Lack of cell population isolation maker and amplification bias caused by transcript abundances variation in different cell groups are two barriers faced by single-cell RNA-seq compared to conventional RNA-seq. The STRT(single-cell tagged reverse transcription) technology isolates single cells to 96-well plates prior to reverse transcription, and introduces cell-specific barcodes to the 3’ end of cDNA of each single cell by exploiting the reverse transcriptase template-switching mechanism[39] with distinct helper oligos. Thus, after sequencing, a bi-directional cell map that reflects the information of the cell population, functional subpopulations, and individual cells is achieved. The method is able to perform analysis at the individual and population levels without the need for known markers and cell types[40].

However, with STRT method, the 3’ terminus of the mRNA is preferential amplified and 5’ terminus is often incomplete during the reverse transcription. The problem could be solved with Smart-Seq method [41]. Smart-Seq method adapts the terminal transferase of MMLV (moloney murine leukemia virus) to add Cytosine to the 3’ end of the cDNA once reverse transcription proceeds to the 5’ end of the RNA, thus to generate extension templates, and promote the reverse transcription of 5’ end. After template conversion, the resulting cDNA contains the entire sequence of the 5’ end of the mRNA followed by amplification.

SCRB-seq (single-cell RNA barcoding and sequencing) is one of the most used and cost-efficient scRNA-seq technologies [42], which requires flow cytometry or other cell sorting methods to separate single cells into microwells [43]. This method is similar to Smart-seq, except that SCRB-seq integrates the specific cellular, a unique nucleic acid sequence, for labeling individual cells to track them through space and time[44] to resolve the source of amplified molecules and quantify transcripts more accurately by reducing processing steps.

While all the single-cell RNA-Seq methods above require separation and lysis of cells from tissues to obtain RNA, Church team reported a Fluorescent in situ Transcriptome sequencing (FISSEQ) method to perform in situ single-cell transcriptome analysis. The technique performs reverse transcription and cDNA amplification in cells immobilized on a glass slide, so that each RNA molecule forms a DNA amplification product sphere of 200-400 nm in diameter in situ. High-throughput sequencing analysis of these DNA amplification product spheres are then performed by Solid Link Sequencing [45]. Comparing to other sequencing techniques, the in situ single-cell transcriptome sequencing method retains the positional information of the cells, which might provide further inspiration for transcriptome research.

Reference

[1] D.J. Lockhart, E.A. Winzeler, Nature, 405, 827-836 (2000)
[2] V. Costa, C. Angelini, I.D. Feis, A.Ciccodicola, J Biomed Biotechnol, 2010, 853916 (2010)
[3] S. Kumar, S.K. Razzaq, A.D. Vo, et al. Wiley InterdiscipRev RNA, 7, 811-823 (2016)
[4] Z. Wang, M. Gerstein, M. Snyder, Nat Rev Genet, 10, 57–63 (2009)
[5] W.Fiers, R.Contreras, et al.Nature, 260, 500-507 (1976)
[6] H.M. Temin, S. Mizutani, Nature, 226, 1211–3 (1976)
[7] F. Sanger, A.R. Coulson, Journal of Molecular Biology, 94, 441–448 (1975)
[8] G. Brady, M. Barbara, N.N. Iscove, Methods in Molecular and Cellular Biology, 2(1990)
[9] T.W. Brian, R.L. Josette, Methods, 48, 249-257 (2009)
[10] T.A. Clark, C.W. Sugnet, M.J. Ares, Science, 296, 907-910 (2002)
[11] M. Margulies, M. Egholm, W.E. Altman, et al. Nature, 437, 376-380 (2005)
[12] Z. Wang, M. Gerstein, M. Snyder, Nat Rev Genet, 10, 57-63 (2009)
[13] A. Mortazavi, B.A. Williams, K. McCue, et al. Nat Methods, 5, 621-628 (2008)
[14] J. Shendure, Nat Methods, 5, 585-587 (2008)
[15] B.B. Tuch, R.R. Laborde, X. Xu, et al. PloS One, 5, e9317- (2010)
[16] J.K. Pickrell, J.C. Marioni, A.A. Pai, et al. Nature, 464, 768-772 (2010)
[17] J.C. Avarre, R. Dugué, P. Alonso, et al. Mol Ecol Resour, 14, 139-149 (2014)
[18] S.A. Filichkin, H.D. Priest, S.A. Givan, et al. Genome Research, 20, 45-58 (2010)
[19] E. Birney, J.A. Stamatoyannopoulos, A. Dutta, et al. Nature, 447, 799–816 (2007)
[20] M. Clamp, B. Fry, M. Kamal, X.H. Xie, et al. Nat Methods, 5, 621-628 (2008)
[21] J. Shendure, Nat Methods, 5, 585-587 (2008)
[22] B.B. Tuch, R.R. Laborde, X. Xu, et al. PloS One, 5, e9317- (2010)
[23] Z. Wang, M. Gerstein, M. Snyder. Nat Rev Genet, 10, 57–63 (2009)
[24] K. Ellen, L. Barbara, et al. Plos one, 6, e28240 (2011)
[25] None, Nature Methods, 2, 151–152 (2005)
[26] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, et al. Nucleic Acids Res, 37, e123 (2009)
[27] T.T. Perkins, R.A. Kingsley, M.C. Fookes, et al. PLoS Genet, 5, e1000569 (2009)
[28] L. Mamanova, R.M. Andrews, K.D. James, et al. Nat Methods, 7, 130-132 (2010)
[29] M. Matthias, E. Anja, N. Tobias, U. Christian, J. Julian, et al. Science, 360, 800-805 (2018)
[30] S. Picelli, A. Bjorklund, B. Reinius, S. Sagasser, G. Winberg, R. Sandberg, Genome Res, 24, 2033-2040 (2014)
[31] A. Adey, H.G. Morrison, et al. Genome Biol, 11, R119 (2010)
[32] T. Barrett, S.E. Wilhite, P. Ledoux, et al. Nucleic Acids Res, 41, D991-D995 (2013)
[33] A. Anita, M. Caterina. Sci Rep, 8, 4282 (2018)
[34] Z. Wang, M. Gerstein, M. Snyder. Nat Rev Genet, 10, 57–63 (2009)
[35] K. Ellen, L. Barbara, et al. Plos one, 6, e28240 (2011)
[36] None, Nature Methods, 2, 151–152 (2005)
[37] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, et al. Nucleic Acids Res, 37, e123 (2009)
[38] T.T. Perkins, R.A. Kingsley, M.C. Fookes, et al. PLoS Genet, 5, e1000569 (2009)
[39] L. Mamanova, R.M. Andrews, K.D. James, et al. Nat Methods, 7, 130-132 (2010)
[40] M. Matthias, E. Anja, N. Tobias, U. Christian, J. Julian, et al. Science, 360, 800-805 (2018)
[41] S. Picelli, A. Bjorklund, B. Reinius, S. Sagasser, G. Winberg, R. Sandberg, Genome Res, 24, 2033-2040 (2014)
[42] A. Adey, H.G. Morrison, et al. Genome Biol, 11, R119 (2010)
[43] T. Barrett, S.E. Wilhite, P. Ledoux, et al. Nucleic Acids Res, 41, D991-D995 (2013)
[44] A. Anita, M. Caterina. Sci Rep, 8, 4282 (2018)
[45] Z. Wang, M. Gerstein, M. Snyder. Nat Rev Genet, 10, 57–63 (2009)
[46] K. Ellen, L. Barbara, et al. Plos one, 6, e28240 (2011)
[47] None, Nature Methods, 2, 151–152 (2005)
[48] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, et al. Nucleic Acids Res, 37, e123 (2009)
[49] T.T. Perkins, R.A. Kingsley, M.C. Fookes, et al. PLoS Genet, 5, e1000569 (2009)
[50] L. Mamanova, R.M. Andrews, K.D. James, et al. Nat Methods, 7, 130-132 (2010)
[51] M. Matthias, E. Anja, N. Tobias, U. Christian, J. Julian, et al. Science, 360, 800-805 (2018)
[52] S. Picelli, A. Bjorklund, B. Reinius, S. Sagasser, G. Winberg, R. Sandberg, Genome Res, 24, 2033-2040 (2014)
[53] A. Adey, H.G. Morrison, et al. Genome Biol, 11, R119 (2010)
[54] T. Barrett, S.E. Wilhite, P. Ledoux, et al. Nucleic Acids Res, 41, D991-D995 (2013)
[55] A. Anita, M. Caterina. Sci Rep, 8, 4282 (2018)