Advancing transcriptome platforms

Shuobo Shi*

Department of Biochemical Engineering, School of Chemical Engineering and Technology, Key Laboratory of Systems Bioengineering, Ministry of Education, Tianjin University, 92 Weijin Road, Tianjin 300072, P. R. China

**Author for correspondence (Tel: +86-22-2740-6770; Fax: +86-22-2740-6770; E-mail:shishuobo@hotmail.com)*

Keywords: Microarray, MPSS, RNA-Seq, SAGE, Transcriptome platforms

Abstract

During the last decade of years, remarkable technological innovations have emerged that allow the direct or indirect determination of transcriptome at unprecedented scale and speed. Studies using these methods have already altered our view of the extent and complexity of transcript profiling, which has advanced from one-gene-at-a-time to a holistic view of the genome. Here, we outline the major technical advances in transcriptome characterization, including the most popular used hybridization-based platform, the well accepted tag-based sequencing platform, and the recently developed RNA-Seq (RNA sequencing) based platform. Especially, the next-generation technologies make a revolution in assessing the entire transcriptome via

the recent RNA-Seq technology.

Introduction

Transcriptome is defined as the set of all messenger RNA (mRNA) molecules, or transcripts, produced in one or a population of cells at a certain time. The term can also be applied to the specific subset of transcripts present in a particular cell type. It offers a more holistic approach to interpret the functional elements of the genome and therefore builds a foundation for the global understanding of biological processes.

Various high-throughput systems have been developed to deduce and quantify transcriptome. These methods can be divided into three classes: (1) Hybridization-based platform, typically involving fluorescently labeled cDNA microarrays or commercial high-density oligo microarrays; (2) Tag-based sequencing platform, including SAGE or MPSS; (3) RNA-Seq based transcriptome platform. Currently, each of these methods has inherent advantages and disadvantages, often related to expense, technical difficulty, specificity, and reliability (Table 1). The appropriate method should be chosen to fit the objectives in each experiment. However, recent studies implicated that RNA-Seq approach has clear advantages over existing approaches and is opening a new vista in revolutionizing the manner in which transcriptome is analyzed (Blow 2009; Mortazavi et al. 2008; Wang et al. 2009).

Hybridization-based transcriptome platform

Microarrays, or gene chips, which allow the simultaneous monitoring of the expression of thousands of genes, have become the most popular platform among scientists for performing global gene expression analysis. Looking back, this technique evolved from the classical Southern-blotting approaches (Southern 1975). Then, researchers began to work on a reciprocal technique, in which the known genes or fragments were immobilized on a substrate, and the solution to be queried was labeled and hybridized to the surface (Bains and Smith 1988; Drmanac et al. 1989). The first use of microarray as a global approach to create a transcript profiling was reported in 1995 (Schena et al. 1995), with some 1000 cDNAs printed on a glass slide. In 1997, a complete genome (*Saccharomyces cerevisiae*) on a microarray was published (DeRisi et al. 1997; Lashkari et al. 1997).

The essence of microarray is the parallel hybridization of complementary nucleic acid strands (Fig. 1). The specific targets (the labeled nucleic acid from a given cell) could be simultaneously hybridized with the thousands of probes (DNA fragments of a gene) immobilized on a solid surface (nylon membranes, glass slides, etc). The signal intensity of the hybridized probe is proportional to the gene expression level and hence it serves as an estimate of the expression level of each specific gene. According to the nature of the probe, there are two main types of DNA microarrays: cDNA microarray (Schena et al. 1995) and oligonucleotide microarray (Lockhart et al. 1996). Probes for cDNA microarrays are usually obtained from genomic clones or polymerase chain reaction (PCR) amplification, and then these sequences are spotted using an automated microarray spotter. Probes for oligonucleotide microarrays are short sequences designed to match parts of the sequence of known or predicted open reading frames, and these sequences are synthesized in situ onto the array surface.

Since it appears, microarray has produced a vast amount of data. MIAME (Minimum Information about a Microarray Experiment) (Brazma et al. 2001) and MAQC (MicroArray Quality Control) (Shi et al. 2008; Shi et al. 2006) projects have been created to advance the field of transcriptome by establishing and assessing standard guidelines for comparison of these data. Now, Microarray derived data must adhere to MIAME standards and should be deposited in a public repository such as Gene Expression Omnibus (GEO) (Edgar et al. 2002), ArrayExpress Database (Brazma et al. 2003), and Stanford Microarray Database (Ball et al. 2005). The tools to analyze microarray data are also widely available and developed, making analysis less complicated and easier. Indeed, microarray has been a reliable and stable technology, which is well accepted by scientists.

Now, one main trend in DNA microarray development is high-density of microarray (tiling arrays) that more than 1 million probes can be mounted onto every square cm by in situ synthesis. This could provide massively parallel approaches for the characterization of all expressed exons. However, any unexpected sequences will still be missed. Minimization of reaction system is another trend in DNA microarray development, as sometimes large amounts of material can be difficult to be obtained. The decreasing minimum starting requirement of input RNA or DNA would make it possible to study the expression profile of even a single cell.

Listed in table 1, the fundamental reliance of microarray on nucleic-acid hybridization results in several inherent limitations. There may be potentially confounding effects of cross-hybridization among highly related sequences. The sensitivity of microarrays is still low despite of the improvements in signal detection (Stears et al. 2000). In addition, the method is actually semi-quantitive, as the method can detect only twofold or greater changes, with bad technical reproducibility in slight changes of gene expression. The number of DNA probes that fit on a microarray is limited, putting constraints on the provided information. Moreover, comparing expression levels across different experiments is often difficult and it requires complicated normalization methods.

Tag-based sequencing transcriptome platform

In parallel to the remarkable advances in hybridization based techniques, tag-based sequencing technologies for measuring absolute abundance of gene expression have made significant advances. Large-scale quantitative expression technologies involve the collection and sequencing of short sequence tags from a given RNA sample, and use the abundance of these sequence tags to determine the abundance of each transcript. As an open system, the methods are complementary to standard microarrays, and they can be utilized to discover novel transcripts expressed in a cell. Laborious and costly cloning and sequencing steps have far greatly limited their use. However, the sequencing steps have greatly improved with the introduction of deep sequencing technology, enabling the simultaneous sequencing of up to millions of different DNA molecules.

Serial analysis of gene expression, SAGE

Serial analysis of gene expression (SAGE) is a method that could both qualitatively and quantitatively evaluate the expression of thousands of genes simultaneously by sequencing the tagged cDNA fragments (Velculescu et al. 1995). As an improvement of traditional EST sequencing (Adams et al. 1991), which is the first sequencing-based method to detect gene expression, SAGE provides a highly reliable identification of gene expression at a much less cost than EST sequencing by minimizing the amount of information collected per transcript.

The procedure of SAGE involves several steps (Fig. 1), including total RNA isolation, mRNA purification, cDNA synthesis, tags generation by enzyme digestion, concatenation, amplification and sequencing of tags, and it only rests on two principles: First, a short nucleotide sequence tag of 10-14 bp, released from the cDNA by a specific Type IIS restriction enzyme (tagging enzyme) (e.g. *Nla*III and *Bsm*FI) digestion, contains enough information to uniquely identify a transcript (Pennisi 2000); Second, these tags could be serially connected to form a long stretch of DNA molecule, and sequencing reaction would be needed only once. The frequency of detection of each tag represents the quantification and identification of the transcripts detected.

Now SAGE is a popular used method to characterize transcripts profiles because it can be performed in individual labs and it has generated data sets that have proven valuable for the annotation of complex genomes (Saha et al. 2002; Wang 2007). Correspondingly, several SAGE databases have been constructed, including SAGE-map (http://www.ncbi.nlm.nih.gov/SAGE/), SAGE-net (http://www.sagenet.org/), grainSAGE database (http://www.scu.edu.au/research/cpcg/igfp/index.php), and Genzyme’s SAGE database (Commercial, available from Celera Genomics and Compugen).

The conventional SAGE has been further optimized. LongSAGE (Saha et al. 2002), CAGE (Shiraki et al. 2003) and SuperSAGE (Matsumura et al. 2003) were developed by using different tagging enzymes to release longer tags, which would provide higher specificity for transcript identification. Recently, a so-called SuperSAGE array was developed by spotting SuperSAGE tags as probes onto the oligonucleotides array (Matsumura et al. 2006). The SuperSAGE array, regardless of the availability of genome sequence, combines the advantage of the highly quantitative SuperSAGE expression analysis with the high-throughput microarray technology, and produces highly reproducible hybridization signals. SAGE-Lite (Datson et al. 1999), MicroSAGE (Peters et al. 1999) and RL-SAGE (Gowda et al. 2004) were reported to reduce the requirements of total RNA. With the help of new sequencing technologies, SAGE based transcriptome platform has been developed to GIS-PET (Ng et al. 2005), DeepSAGE (Nielsen et al. 2006), 5′-RATE (Gowda et al. 2006), and digital transcriptome subtraction (DTS) (Feng et al. 2007). In addition, it is pronounced that future SAGE-derived experiments will only require the generation of di-tags, as the new sequencing technologies can directly use these as templates for sequencing (Hanriot et al. 2008; Vega-Sanchez et al. 2007). The advancing SAGE-derived technologies facilitated detection of rare transcripts and acquirement of reliable, cost-effective, holistic understanding of the whole transcript profiling with less labour and cost. These innovations are much better suited to explore the depth and complexity of transcriptome.

At present, a few disadvantages remain with the technique (Table 1). Due to the relatively short sequences generated, a major limitation of SAGE is that the accuracy of the gene origin determined for some SAGE tags cannot be guaranteed. This will influence the understanding of the dynamic transcripts processes (Pleasance et al. 2003; Wang 2007). In addition, although the generation of sequence tags is without prior knowledge of nucleotide and ideal for the discovery of novel transcripts, a reference genome to determine gene identity is needed. The SAGE process involves enzyme digestions, PCR amplifications, cloning and colony propagations, which could result in a quantitative bias for different tags (Margulies et al. 2001; Siddiqui et al. 2006). Furthermore, the rare mRNAs also exceed the detection region of SAGE due to its limited sequencing capability (Wang 2007).

Massively parallel signature sequencing, MPSS

Developed at what is now Illumina, Inc. (originally Lynx Therapeutics, Hayward,CA), MPSS is another tag-based methodology that has been used for deep transcriptome analysis and genome annotation (Brenner et al. 2000). MPSS has similar advantages and disadvantages to SAGE. However, the unique feature of MPSS is the bead-based sequencing technology, which generates longer and more tags that make gene identification more accurate, sensitive, and fast. Hence, MPSS potentially provides a greater and more accurate coverage of transcriptome than SAGE.

MPSS is based on the *in vitro* cloning of millions of cDNA fragments and hybridization-ligation based parallel sequencing (Fig. 1). First, cDNA is synthesized followed by digestion with the tagging enzyme *DpnII*, thus leaving a signature or tag sequence. These fragments are cloned into a special plasmid carrying the random 32-bp adapter (barcode), followed by PCR amplification. The amplification products are then linked to the corresponding microbeads by adapter hybridization, followed by high throughput sequencing in a flow cell. The procedure is completely parallel, facilitating more than 1 million tags sequencing at once. A tag sequence of around 16-20 bp is obtained per bead. The 16-20 bp tags have a higher specificity in the complex genome (Meyers et al. 2004b; Saha et al. 2002), and they are used for quantitative measurements of gene expression in a manner similar to SAGE.

With respect to SAGE, there are several noteworthy advantages of the technology. First, this method, allowing the identification of millions of tag-sequencing events in one run, is unprecedented parallel and significantly surpasses the largest SAGE applications that only cover hundreds of thousands of tags. Second, the method significantly increases tag length compared with conventional SAGE, and it was expected to improve the prospects for unique genome and transcriptome tag mapping. Furthermore, MPSS is faster and less laborious because of its bead-based highly throughput sequencing method.

This approach has been widely shown to be effective in plant and animal cells for measuring gene expression levels, and it has proven instrumental in characterizing the complexity of transcriptome. The first application of MPSS to the study of a plant genome was done in *Arabidopsis* (Hoth et al. 2002; Meyers et al. 2004a; Meyers et al.

2004b). Except that, its applications rapidly spread across many other plants, and a series of plant MPSS databases have been constructed (http://mpss.udel.edu/). MPSS has also been used to assay the transcript profiling of animal cells such as human (Freed et al. 2008; Liu et al. 2006; Oudes et al. 2005), mouse (Peters et al. 2007; Wei et al. 2005) and *Drosophila* (Lee et al. 2005; Torres et al. 2008).

However, it should be noted, due to the nucleotides bias, that some transcripts are lost in the course of sequencing and the tag library construction (Lawrence et al. 2007; Meyers et al. 2004a). The major disadvantage is that MPSS is so complex and it may appear inaccessible to the broad scientific community, as the technology is, until recently, only available from a company, Illumina, Inc. The now gradually discontinued MPSS technique had been updated and replaced by a new platform that uses a sequencing-by-synthesis approach, known as Illumina's Genome Analyzer (www.Illumina.com).

RNA-Seq based transcriptome platform

As a newer and more comprehensive platform to map transcriptome, RNA-Seq (RNA-Sequencing) approach is developed by direct ultra high-throughput sequencing of cDNAs using one of the NGS (next-generation sequencing) methods (Table 2). The principle of quantitative estimates of gene expression in RNA-Seq approach came from that of EST, SAGE, or MPSS.

When it comes to the analysis of transcriptome, RNA-Seq approach sequences fulllength cDNA libraries, and the depth of sequencing required for analysis of rare transcripts is much greater (Wang et al. 2009). First, the mRNA transcript pool is converted to construct a cDNA library (Fig. 1). Then the cDNA library is randomly sheared. The resulting individual DNA fragments, with or without amplification, are sequenced by massively parallel sequencing methods. Subsequently, the resulting sequence reads are individually mapped to the source genome, or assembled de novo without the genomic sequence to produce a genome-scale transcription map. Alternatively, another application of RNA-Seq approach focuses on capturing the information-rich 3’-untranslated region (UTR) of messenger RNAs (mRNAs) (Eveland et al. 2008), and it is not designed to discriminate different splice variants or to detect mutations within the expressed exons.

In theory, any NGS technology (Table 2) can be used for RNA-Seq method, and Illumina's Genome Analyzer, Roche's 454 sequencer and Applied Biosystems' SOLiD System have already used for this purpose. The recent commercial availability of Helicos Biosciences tSMS system, which has not yet been used for published RNA-Seq studies, is also appropriate and offers the additional advantage of avoiding the requirement for amplification of target cDNA. In the case of transcriptome sequencing, more informative read length is always an advantage (Torres et al. 2008) and the read length is determined by the adopted NGS technology. Longer read length (>200 bp) could be achieved with 454-based platform, initially described by Margulies et al. (Margulies et al. 2005). Despite potential drawbacks in read length, the short read sequencing technologies (Genome Analyzer and SOLiD System) are much better suited in the fully sequenced species (or their close relatives), in which the specificity would be guaranteed to allow effective annotation and assembly of complete transcriptome (Hudson 2008). Moreover, the two sequencing technologies generate sequence data at a faster rate and a cheaper per-nucleotide cost than 454 sequencer (Table 2).

RNA-Seq method has clear advantages over existing approaches (Table 1). First, the longer signatures of RNA-Seq method can be unambiguously mapped to unique transcripts of the genome. Second, the method allows evaluating global splicing patterns, detecting novel expressed exons and identifying transcript sequence polymorphisms in a given sample. Third, the reproducibility of the approach has been shown to be extremely high for both technical and biological replicates. Thus, the presence and amount of each transcript can be compared with that in any other sequenced sample, now or in the future. Finally, RNA-Seq method requires less RNA sample and avoids the bias formed in the cloning steps. There is even no need for amplification step armed with the single molecule sequencing (e.g. Helicos Biosciences tSMS system). Moreover, RNA-Seq can provide information on transcripts that are expressed at very low levels, limited only by the total number of reads that are generated.

RNA-Seq method has demonstrated its enormous potential for transcriptome studying in life science. Several recent papers described the application of RNA-Seq method to acquire the transcriptomes of mammals (Cloonan et al. 2008; Mortazavi et al. 2008; Sultan et al. 2008), yeast (Nagalakshmi et al. 2008; Yassour et al. 2009) and plants

(Lister et al. 2008; Weber et al. 2007). Meanwhile, its bright perspectives are currently being summarized and commented (Blow 2009; Wang et al. 2009).

Nevertheless, RNA-Seq method also has its weaknesses (Table 1): (1) The need for bioinformatics infrastructure is a pressing challenge; (2) Sample preparation is significantly more complicated and time consuming than that of microarray analysis; (3) The technology is currently costly in terms of the equipments and the reagents; (4) Compared with traditional Sanger genome sequencing, higher sequencing error rates were reported in next generation sequencing technologies (Moore et al. 2006; Wheat 2008). Currently, the biggest challenge for researchers looking at RNA-Seq approach is probably the bioinformatics challenges, including the methods to store, retrieve and process the vast volume of sequence data, development of algorithms to reduce errors in image analysis and remove low-quality reads. Current RNA-Seq methods are not yet mature and well suitable; there are opportunities for improvement of its effectiveness.

Summary and perspectives

The hybridization-based transcriptome platform has been the method of choice for transcriptome profiling for more than a decade, with a lower workload and a relatively lower cost. The platform suffers from limitations such as background noise and cross hybridization. In addition, this technology can only provide information about the transcripts that are included on the array. Therefore, it would be obsolete for determining transcriptome in the future.

The tag-based transcriptome platform (SAGE or MPSS) has provided key information on transcripts in the past, but splice isoforms are generally indistinguishable from each other. Meanwhile, the platform suffers from a poor coverage of each transcript and potentially ambiguous mapping because of the short read length. Its use will decline for profiling of transcriptome.

Although RNA-Seq method is still in its infancy, it has clear advantages over previously developed transcriptome platforms. With the trend of declining sequencing costs, more researchers would prefer RNA-Seq method because of the added power and quality that involve determining the structure and dynamics of transcripts. Simultaneously, it should be noted that it is a long way to go before RNA-Seq method reaches the level of adoption that microarrays have.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (NSFC-20536040), the National Project of Key Fundamental Research (2007CB707802), and the Development Project of Science and Technology of Tianjin (05YFGZGX04500).

Reference

Adams MD, Kelley JM, Gocayne JD et al (1991) Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. Science 252:1651-1656

Bains W, Smith G (1988) A novel method for nucleic acid sequence determination. J Theor Biol 135:303-307

Ball CA, Awad IAB, Demeter J et al (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res 33:D580-D582

Blow N (2009) Transcriptomics: The digital generation. Nature 458:239-242

Brazma A, Hingamp P, Quackenbush et al (2001) Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. Nat Genet 29:365-371

Brazma A, Parkinson H, Sarkans U et al (2003) ArrayExpress - a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31:68-71

Brenner S, Johnson M, Bridgham J et al (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630-634

Cloonan N, Forrest ARR, Kolle G et al (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5:613-619

Datson NA, van der Perk-de Jong J, van den Berg MP et al (1999) MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. Nucleic Acids Res 27:1300-1307

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680-686

Drmanac R, Labat I, Brukner I et al (1989) Sequencing of megabase plus DNA by hybridization: theory of the method. Genomics 4:114-128

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207-210

Eveland AL, McCarty DR, Koch KE (2008) Transcript profiling by 3'-untranslated region sequencing resolves expression of gene families. Plant Physiol 146:32-44

Feng HC, Taylor JL, Benos PV et al (2007) Human transcriptome subtraction by using short sequence tags to search for tumor viruses in conjunctival carcinoma. J Virol 81:11332-11340

Freed WJ, Chen J, Backman CM et al (2008) Gene expression profile of neuronal progenitor cells derived from hESCs: activation of chromosome 11p15.5 and comparison to human dopaminergic neurons. PLoS ONE 3:e1422- e1433

Gowda M, Jantasuriyarat C, Dean RA et al (2004) Robust-LongSAGE (RL-SAGE): A Substantially Improved LongSAGE Method for Gene Discovery and Transcriptome Analysis. Plant Physiol 134:890-897

Gowda M, Li H, Alessi J et al (2006) Robust analysis of 5'-transcript ends (5'-RATE): a novel technique for transcriptome analysis and genome annotation. Nucleic Acids Res 34:e126- e134

Hanriot L, Keime C, Gay N et al (2008) A combination of LongSAGE with Solexa sequencing is well suited to explore the depth and the complexity of transcriptome. BMC Genomics 9:418-426

Hoth S, Morgante M, Sanchez JP et al (2002) Genome-wide gene expression profiling in *Arabidopsis thaliana* reveals new targets of abscisic acid and largely impaired gene regulation in the *abi1-1* mutant. J Cell Sci 115:4891-4900

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 8:3-17

Lashkari DA, DeRisi JL, McCusker JH et al (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 94:13057-13062

Lawrence H, Vattipally S, Mai V et al (2007) Deep analysis of cellular transcriptomes–LongSAGE versus classic MPSS. BMC Genomics 8:333-346

Lee S, Bao JY, Zhou GL et al (2005) Detecting novel low-abundant transcripts in *Drosophila*. RNA 11:939-946

Lister R, O'Malley R, Tonti-Filippini J et al (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133:523-536

Liu Y, Shin S, Zeng X et al (2006) Genome wide profiling of human embryonic stem cells (hESCs), their derivatives and embryonal carcinoma cells to develop base profiles of U.S. Federal government approved hESC lines. BMC Dev Biol 6:20-35

Lockhart DJ, Dong HL, Byrne MC et al (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675-1680

Margulies EH, Kardia SL, Innis JW (2001) Identification and prevention of a GC content bias in SAGE libraries. Nucleic Acids Res 29:E60-E67

Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380

Matsumura H, Bin Nasir KH, Yoshida K et al (2006) SuperSAGE array: the direct use of 26-base-pair transcript tags in oligonucleotide arrays. Nat Methods 3:469-474

Matsumura H, Reich S, Ito A et al (2003) Gene expression analysis of plant host-pathogen interactions by SuperSAGE. Proc Natl Acad Sci USA 100:15718-15723

Meyers BC, Tej SS, Vu TH et al (2004a) The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. Genome Res 14:1641-1653

Meyers BC, Vu TH, Tej SS et al (2004b) Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. Nat Biotechnol 22:1006-1011

Moore MJ, Dhingra A, Soltis PS et al (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol 6:17-29

Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621-628

Nagalakshmi U, Wang Z, Waern K et al (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344-1349

Ng P, Wei CL, Sung WK et al (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2:105-111

Nielsen KL, Hogh AL, Emmersen J (2006) DeepSAGE - digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic Acids Res 34:e133- e137

Oudes AJ, Roach JC, Walashek LS et al (2005) Application of affymetrix array and massively parallel signature sequencing for identification of genes involved in prostate cancer progression. BMC Cancer 5:86-97

Pennisi E (2000) Human genome project - And the gene number is...? Science 288:1146-1147

Peters DG, Kassam AB, Yonas H et al (1999) Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. Nucleic Acids Res 27:e39- e44

Peters LM, Belyantseva IA, Lagziel A et al (2007) Signatures from tissue-specific

MPSS libraries identify transcripts preferentially expressed in the mouse inner ear. Genomics 89:197-206

Pleasance ED, Marra MA, Jones SJM (2003) Assessment of SAGE in transcript identification. Genome Res 13:1203-1215

Saha S, Sparks AB, Rago C et al (2002) Using the transcriptome to annotate the genome. Nat Biotechnol 20:508-512

Schena M, Shalon D, Davis RW et al (1995) Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray. Science 270:467-470

Shi LM, Perkins RG, Fang H et al (2008) Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. Curr Opin Biotechnol 19:10-18

Shi LM, Reid LH, Jones WD et al (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24:1151-1161

Shiraki T, Kondo S, Katayama S et al (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 100:15776-15781

Siddiqui AS, Delaney AD, Schnerch A et al (2006) Sequence biases in large scale gene expression profiling data. Nucleic Acids Res 34:e83-e92

Southern EM (1975) Detection of Specific Sequences among DNA Fragments Separated by Gel-Electrophoresis. J Mol Biol 98:503-517

Stears RL, Getts RC, Gullans SR (2000) A novel, sensitive detection system for high-density microarrays using dendrimer technology. Physiol Genomics 3:93-99

Sultan M, Schulz M, Richard H et al (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956-959

Torres TT, Metta M, Ottenwalder B et al (2008) Gene expression profiling by massively parallel sequencing. Genome Res 18:172-177

Vega-Sanchez ME, Gowda M, Wang GL (2007) Tag-based approaches for deep transcriptome analysis in plants. Plant Sci 173:371-380

Velculescu VE, Zhang L, Vogelstein B et al (1995) Serial Analysis of Gene-Expression. Science 270:484-487

Wang SM (2007) Understanding SAGE data. Trends Genet 23:42-50

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57-63

Weber APM, Weber KL, Carr K et al (2007) Sampling the *Arabidopsis* Transcriptome with Massively Parallel Pyrosequencing. Plant Biol 144:32-42

Wei C, Miura T, Robson P et al (2005) Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state. Stem Cells 23:166-185

Wheat CW (2008) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica DOI 10.1007/s10709-008-9326-y

Yassour M, Kaplan T, Fraser HB et al (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci USA

106:3264-3269

Table 1 Comparison of platforms used in transcriptome analysis

	Advantages	Disadvantages

Hybridization-based	1. High specificity;
	2. Low cost, rapid;
	3. Ease of sample preparation;
	4. Flexibility in extent to be analyzed;
	5. Mature informatics and statistics;
		a. Constrained by gene discovery and prediction procedures;
		b. Difficulties to compare data from different experiments and to obtain absolute quantity of mRNA expressed;
		c. Cross-hybridization;
		d. High background noise;
		e. Sequence dependent (only for cDNA microarrays);

Tag-based	6. Identification of novel transcripts; 7. Quantitative, inter-laboratory comparable; 8. Sensitive, low background; 9. Sequence independent;	f. Biased sampling; g. Ambiguity in identifying transcripts caused by the short length of tags; h. Dependence on reference sequence database; i. Expensive and labourious work; j. Complex sample preparation;
RNA-Seq	6,7,8,9 10. Avoiding the need for cloning; 11. Determination of RNA	f, i, j k. Limited bioinformatics.

	splicing and sequence polymorphisms; 12. Longer signatures, more accurate annotation; 13. Low input RNA.

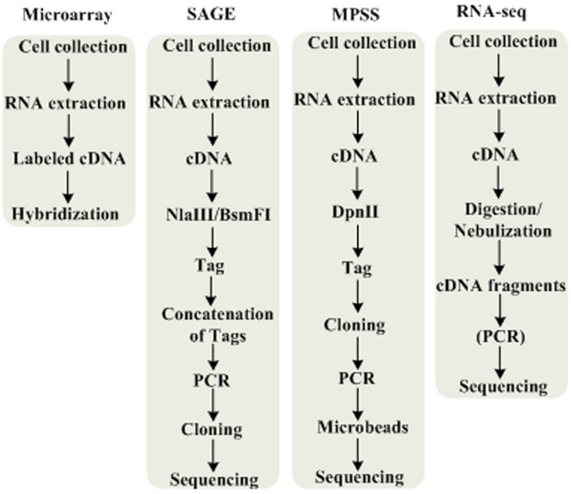
Table 2 The next generation sequencing technologies available

Technology	Read length	Bp per day	Cost per Mb	Supplier	Commercial Day
454 sequencer	200-400bp	480-1000Mb	~$60	Roche Applied Science (http://www.roche-applied-science.com/)	2005
Genome Analyzer	32-40	300-500Mb	~$2	Illumina, Inc. (http://www.illumina.com/)	2007
ABI-SOLiD sequencer	35	120-400Mb	~$2	Applied Biosystems (http://www.appliedbyosystems.com)	2007
tSMS	20-55	1.1-2 Gb	~$1	Helicos Biosciences (http://helicosbio.com/)	2008

Figure Captions

Fig. 1 Schematic illustration of the procedures of the main transcriptome platforms

1**Fig. 1**


Microarray
Cell collection
RNA extraction
Labeled cDNA
Hybridization
SAGE
Cell collection
RNA extraction
cDNA
NlaIII/BsmFI
Tag
Concatenation of Tags
PCR
Cloning
Sequencing
MPSS
Cell collection
RNA extraction
cDNA
DpnII
Tag
Cloning
PCR
Microbeads
Sequencing
RNA-seq
Cell collection
RNA extraction
cDNA
Digestion/ Nebulization
cDNA fragments
(PCR)
Sequencing

2