SCIENCE CHINA
Life Sciences

SPECIAL ISSUE: Non-coding RNAs
• REVIEW •

October 2013 Vol.56 No.10: 960–967
doi: 10.1007/s11427-013-4557-2

Transcriptomics: Advances and approaches

DONG ZhiCheng* & CHEN Yan

Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China

Received August 14, 2013; accepted September 6, 2013

Transcriptomics is one of the most developed fields in the post-genomic era. Transcriptome is the complete set of RNA transcripts in a specific cell type or tissue at a certain developmental stage and/or under a specific physiological condition, including messenger RNA, transfer RNA, ribosomal RNA, and other non-coding RNAs. Transcriptomics focuses on the gene expression at the RNA level and offers the genome-wide information of gene structure and gene function in order to reveal the molecular mechanisms involved in specific biological processes. With the development of next-generation high-throughput sequencing technology, transcriptome analysis has been progressively improving our understanding of RNA-based gene regulatory network. Here, we discuss the concept, history, and especially the recent advances in this inspiring field of study.

transcriptomics, next-generation sequencing (NGS), non-coding RNA, RNA-seq

Citation: Dong Z C, Chen Y. Transcriptomics: Advances and approaches. Sci China Life Sci, 2013, 56: 960−967, doi: 10.1007/s11427-013-4557-2

After the completion of the Human Genome Project as well as many other model or non-model organism genomes [1,2], the sequences of genome as genetic information carriers are available. However, to resolve the black box between genome expression and cell function remains challenging. According to central dogma, transcription is the first and key regulatory step of gene expression. Transcriptomics has become an inspiring field of life science research in the post-genome era [3], for the following reasons: (i) Transcriptome analysis reflects the dynamics of genome expression. Although most cells share the same set of genes, their transcription patterns are highly temporal and spatial specific, which leads to different cell types and/or functions. (ii) Transcriptomics study supports the proteomics research. Transcriptome analysis not only helps to explain the inconsistency of the coding gene number with the number of proteins translated, but also is the starting point for the study of translational regulation. (iii) Structural and functional studies of nonprotein-coding RNA (ncRNA) expand the scope of transcriptomics. Recently, a large number of RNA species, transcribed from nonprotein-coding genomic regions, have been discovered with important roles in gene regulation [4–6]. In addition to protein coding genes, both prokaryotic and eukaryotic genomes contain nonprotein-coding sequences [7,8]. Transcriptomics studies have shown that the vast majority of eukaryotic genome is transcribed into RNAs [9–11]. For example, more than 93% of the human genome is transcribed into RNAs [9], among which only 2% is from the coding region [7,12]. (iv) Methodology innovation, especially of the next-generation sequencing (NGS) technology, has allowed a higher throughput and resolution level of transcriptome studies and generated more data with biological meanings [13,14].

1 Transcriptome

1.1 A brief history

Transcriptomics is the study of RNA, single-stranded nucleic acid, which was not separated from the DNA world

*Corresponding author (email: zhicheng_dong@scbg.ac.cn)

until the central dogma was formulated by Francis Crick in 1958, i.e., the idea that genetic information is transcribed from DNA to RNA and then translated from RNA into protein [15–17]. In 1961, Jacob and Monod [18] proposed a model that the protein-coding gene is transcribed into a special short-lived intermediate associated with the ribosome, which was designated as messenger RNA (mRNA). In 1958, together with the central dogma of molecular biology, the "adaptor" hypothesis was indicated by Crick to explain how mRNA template directs the protein synthesis [17,19]. In this hypothesis, Crick predicted that each amino acid was first attached to its own "adaptor" which could fit onto the mRNA template by base-pairing and thus carry the amino acid to the specific site of the RNA template. A short, stable RNA, transfer RNA (tRNA), was identified as the predicted "adaptor" [20]. Shortly, ribosomal RNA (rRNA) involved in protein synthesis was purified [16].

In 1977, Sharp [21] and Roberts [22] showed that the mRNA sequence of adenovirus displayed discontinuous distribution in the genome, and therefore first suggested that a typical eukaryotic gene consists of exons, the protein-coding sequence, and introns, the non-coding sequence; the protein-coding sequence was interrupted by the non-coding sequence. During RNA splicing, the introns are cut out from the primary transcripts and degraded, while the exons are reassembled into different mature messenger RNAs (mRNAs) (alternative splicing). The discovery of the split gene was a complete surprise and had revolutionized our understanding of the architecture of genes.

Since the late 1970s, Altman and Cech [23–26] revealed respectively that RNA can function as a catalyst. In 1982, Kruger [26] put forward the "ribozyme" concept, demonstrating that RNA could act as both genetic material (like DNA) and a biological catalyst (like protein enzymes).

In the early 1990s, it was observed by a number of scientists independently that RNA inhibited gene expression in plants and fungi with unknown mechanism [27–29]. In 1998, Fire and Mello [30,31] found that double-stranded RNAs (dsRNAs) could recognize specific mRNA sequence and then led to the degradation of the target mRNAs, which was known as RNA interference (RNAi). Further studies indicated that the actual molecules that directly caused RNAi were short dsRNA fragments of 21–25 base pair, called small interfering RNA (siRNA) [32–34].

In recent years, the species of RNA are increasing rapidly with the development of NGS technology [13,14]. The expanding universe and functional characterization of non-coding RNAs (ncRNA) have been making our understanding of the RNA world more comprehensive and in-depth. Consequently, the research content of transcriptomics has been expanded.

1.2 RNA category

The history of RNA research indicates that there is a great variety of RNAs, which have, besides structural roles, important and previously underestimated regulatory roles in many cellular processes. According to their products, RNAs can be largely divided into two classes: protein-coding RNAs and nonprotein-coding RNAs (ncRNAs) [5]. Protein-coding RNAs are also known as mRNAs.

In a general sense, ncRNAs refer to all RNAs that are not translated into functional proteins. Based on their functions, ncRNAs can be divided into two categories: house-keeping and regulatory ncRNAs. House-keeping ncRNAs usually play structural and catalytic roles, including tRNAs and rRNAs involved in translation, small nuclear RNAs (snRNAs) involved in mRNA splicing, small nucleolar RNAs (snoRNAs) involved in rRNA splicing, guide RNAs (gRNAs) involved in RNA editing, and others [5,16].

Many ncRNAs play regulatory roles in a diverse variety of biological processes. Depending on the length, regulatory ncRNAs can be divided into small ncRNAs and long ncRNAs (lncRNAs). Small ncRNAs, 17–35 nt in length, include microRNAs (miRNAs), siRNA and Piwi-interacting RNA (piRNA). miRNAs, 22 nt in length, derive from pri-miRNA containing hairpin structures [35]. The hairpins are processed subsequently by RNase III Drosha and Dicer to form mature miRNAs [35,36]. Pairing with their target mRNAs, miRNAs inhibit gene expression by translational repression or to promote mRNA degradation [37]. miRNAs have been found to play crucial regulatory roles in many biological processes, such as development, biological stress response, and cell behavior [36]. siRNAs and piRNAs are small RNAs produced through different pathways, which mainly act in gene silencing of transposons and repetitive sequences to maintain genomic stability [38–41].

lncRNAs, with a length of more than 200 nt, lack open reading frame and are usually RNA polymerase II transcripts [42]. lncRNAs can be antisense, sense, intergenic, bidirectional, and intronic transcripts [42], which may regulate protein-coding gene expression in different ways: Transcriptional regulation can occur through lncRNA-protein interaction to inhibit the activity of transcription factors or RNA polymerase II directly, or by lncRNA helping to recruit regulatory protein factors of chromatin structure to influence transcription indirectly; lncRNA may affect mRNA stability at the posttranscriptional level. Like small ncRNAs, lncRNAs play essential roles in many biological processes [42,43]. Currently, the challenge of lncRNA research is to discover and quantify different lncRNAs in different tissues or under different physiological conditions, and then determine their biological functions and mechanisms of action.

2 Transcriptomics

Transcriptome is the whole set of RNAs transcribed by the genome from a specific tissue or cell type at a developmen-

tal stage and/or under a certain physiological condition [8,44]. After the genome has been sequenced, transcriptome analysis allows us to understand the expression of genome at the transcription level, which provides information on gene structure, regulation of gene expression, gene product function, and genome dynamics. Transcriptome analysis will further reveal the regulation network of biological processes and eventually give some guidance in disease diagnosis, clinical therapy, and crop improvement.

2.1 Quantifying the transcript

Transcriptional response of the genome varies in different tissues or under different physiological conditions or environmental stimuli. To discover differentially expressed genes was one of the earliest goals of transcriptome analysis. Expressed sequence tag based method (EST, SAGE), hybridization based gene microarray or chip technology, and NGS based RNA-sequencing (RNA-seq) technology were developed to scan the transcriptome quickly and obtain the differentially expressed genes [45–51]. Many key genes from various developmental, physiological, or pathological processes were identified by these means.

2.2 Defining the gene structure and RNA metabolism

Gene is usually defined as a genetic unit. When talking about the transcription, we also refer to gene as a transcription unit [52]. From the gene to the functionally mature RNA, multiple steps of transcription and post-transcriptional processing take place in the cells [52]. A transcription procedure consists of transcription initiation (forming the transcription initiation complex), elongation, pausing (transcription complexes stop right downstream the transcription start site), and termination [16,52]. Combining the traditional biochemical and molecular biology technology with NGS, transcription events were observed globally in higher throughput and more precise level (Table 1) [53].

To understand the RNA transcript structure and its promoter, mapping the transcription start site (TSS) is required. Taking the advantage of cap structure of mRNA 5′ end, cap analysis of gene expression (CAGE) method was developed to sequence the 5′ ends using Sanger sequencing [54,55]. This method was improved when NGS took the place of Sanger sequencing. Paired-end analysis of TSSs (PEAT), deepCAGE, nanoCAGE and CAGEscan revealed precisely the TSS of each gene [58–60]. Similarly, precision nuclear run-on and sequencing (PRO-cap) method allows detection of TSS of nascent RNAs [61].

Digital quantification by regular RNA-seq only represents a certain RNA species' steady-state, which does not reflect the dynamic process of RNA metabolism. To determine the biogenesis rate of RNA, genome-wide nuclear run-on and sequencing (GRO-seq) and the improved version PRO-seq succeeded in monitoring the nascent mRNA globally at very high resolution (single nucleotide for PRO-seq) [61,62]. These two methods, combining the RNA-seq and nuclear run-on assay, provide not only the rates of transcription initiation and elongation but also the RNA polymerase pausing positions. Kwak et al. [61] found that transcription pausing and elongation activation happen widely in the *Drosophila* genome.

2.3 Studying the post-transcriptional processing

The maturing of RNAs is a series of steps such as 5′ capping, splicing, 3′ cleavage and adding polyA [52].

In the human genome, protein-coding genes are less than 30000, but in human cells more than 80000 different proteins can be produced, which is mainly because that RNA precursor of one gene will generate different mature RNA molecules by post-transcriptional processing such as alternative splicing [77]. By pair-end sequencing, improving read length and depth, the majority of coding genes with introns were found to have more than one isoforms [78]. Hence, it is important to consider the abundance of each

Table 1 RNA-seq based methods

NGS method	Conventional method	Description
RNA-seq [45–47]	EST [48], SAGE [50], microarray [51] CAGE [54,55]	Quantify and characterize the transcriptome
Small RNA-seq [56,57]	miRNA microarray	Characterize small non-coding RNA
deepCAGE [58], nanoCAGE [59], PEAT [60], CAGEscan [59], PRO-cap [61]	CAGE [54,55], 5′ RACE	Map the 5′ end of mRNA/transcription start site
PRO-seq [61], GRO-seq [62]	Nuclear run on	Detect nascent RNA
3P-seq [63], PAS-seq [64]	3′ RACE	Detect alternative polyadenylation
BRIC-seq [65,66]	NA	Measure half life of RNA transcripts
PAR-CLIP [67], Argonaute HITS-CLIP [68], Argonaute CLIP-seq [69]	NA	Detect the Argonaute associate RNA to predict miRNA targets
PARE-seq [70], degradome-seq [71–73]	Modified 5′ RACE	Map the 5′ end of RNA degradation products to predict miRNA targets
STRT [74], SMART-seq [75]	SMART/template switch	Single cell or low RNA input transcriptome analysis
CEL-seq [76]	*In vitro* transcription based cDNA amplification	Single cell or low RNA input transcriptome analysis

isoform rather than calculate the sum of various isoforms, when quantifying a certain gene expression.

RNA editing is a post-transcriptional processing, where RNA sequence alteration is introduced, such as uridine insertion and deletion, A-to-I shift. These alterations may lead to changes of amino acid sequence in protein, splicing sites within RNA precursor, or seed sequence of miRNAs [79]. By comparing the RNA-seq results to the reference genome sequence [80], Park et al. [81] found 500–3000 RNA editing events in certain cell type after filtering out the polymorphisms and somatic mutations.

mRNA 3′ end processing involves endonucleolytic cleavage and adding multiple adenosines. The mRNA products of many genes have more than one 3′ cleavage or polyadenylation sites, known as alternative polyadenylation (APA) [82]. APA may affect the length of protein coding sequence or 3′ untranslated region, thereby regulating mRNA translation efficiency and/or its half-life. RNA-seq based method, PAS-seq and 3P-seq, together with specific bioinformatics analysis revealed that APA is an evolutionarily conserved mechanism of gene regulation [63,64,82].

Degradation is an important step in the metabolism of RNA, Tani et al. [71,72] developed the 5′-bromo-uridine immunoprecipitation chase-deep sequencing analysis (BRIC-seq) method to determine half-life of RNAs by sequencing the pulse-labeling RNA, finding many short half-life non-coding RNAs.

2.4 Discovering and characterizing the non-coding RNA

Tremendous progress has been made in characterizing regulatory non-coding RNAs recently. Unbiased transcriptome analysis allowed the discovery of numerous previously unknown RNA transcripts.

Short non-coding RNA usually includes miRNA, siRNA, and piRNA, with lengths shorter than 35 nt. The short non-coding RNA cDNA library construction started with the purification of 15–35 base RNAs by denaturing polyacrylamide gel, followed by ligation of 3′ and 5′ adapter, and finished by reverse transcription [56,57]. At present, more than 20000 of miRNA genes have been cloned from ~200 species [83]. Therefore, it is more challenging to identify miRNA's target mRNA in order to elucidate its function. Two methods emerged based on the fact that miRNA interacts with and cleave its target mRNA through Argonaute protein [84]. Argonaute cross-linking immunoprecipitation and sequencing (CLIP-seq), designed to immunoprecipitate the Argonaute-RNA complex, allows to sequence the Argonaute associated RNA [67–69]. While, degradome-seq or parallel analysis of RNA end (PARE) methods succeed to sequence the 5′ ends of the target mRNA cleavage products by miRNA [70–73]. Along with the bioinformatics analysis, miRNA:mRNA complex regulatory network can be reconstituted [85].

lncRNAs, especially antisense transcripts, were recovered by the strand-specific RNA-seq [61,62,86–88]. This modification on RNA-seq provides the direction information of transcripts sequenced and therefore allows distinguishing antisense ncRNA from sense coding transcripts. Strand-specific sequencing of polyA RNA or rRNA minus RNA fraction found that the human genome is able to express more than 10000 lncRNA [89]. Defining the differential expression in different tissues and/or under different physiological conditions will help to elucidate the function expression of lncRNAs. How to find the regulating targets for each lncRNA will be another challenge.

2.5 Checking the genome by RNA-seq

Huge amount of transcriptome data were generated from medical research, especially in cancer research. Besides the comparison of gene expression in normal and pathological conditions, further changes in genome sequence can be learned, such as somatic mutations in disease tissues, including mutation, insertion and deletion [90]. Gene fusion often indicates the genomic rearrangements, such as translocation, deletion, and inversion. Gene fusion revealed by RNA-seq may provide extra functional hints compared with that by genome sequencing [91,92].

Although the genome sequencing cost continues to decline, there are still a lot of non-model organisms of interests lacking genome reference sequence. Through the pair-end, long read length RNA-seq and *de novo* assembly [93], both quantification and transcripts structure information will be provided for an unsequenced genome. Such transcriptome analysis also helps to annotate the genome to be or being sequenced.

3 Technology for transcriptome analysis

Techniques have been evolved for almost 20 years, from the initial expression sequencing tag (EST) strategy to gene chips, and now the RNA-seq. To analyze the transcriptome becomes cost effective with higher throughput, better sensitivity, and less starting RNA [13,14].

3.1 EST and microarray

Sanger sequencing of EST or cDNA library provided information for genome annotation in the early days of genome research [48]. Due to the limitations on throughput and cost, it is impossible to achieve transcriptome quantitative analysis using EST methods. With serial analysis of gene expression (SAGE) [50] and CAGE [54,55], respectively, multiple 3′ and 5′ cDNA ends were concatenated to be one clone. Therefore, multiple sequence tags can be recovered from one Sanger sequencing reaction, which overcomes those limits and makes quantitative analysis possible.

However, due to the high cost of Sanger sequencing and the difficulty to map the short sequence (~20 bp) tags to genome, CAGE and SAGE were replaced by DNA microarray shortly.

DNA microarray or chip method is based on nucleic acid hybridization. Fluorescent labeled cDNAs incubate with oligonucleotide probes on the chip, then the abundance of RNA is determined by measuring fluorescence density [49, 51]. High-density gene chip allowed relatively low cost gene expression profiling. Specific microarrays were designed according to the purpose of the experiment, such as arrays to detect different isoforms from alternative splicing [94]. In addition, the genome tiling array is an unbiased design, without prior knowledge of genome transcription information, using a set of overlapping oligonucleotide probes for the detection of whole genome expression with the resolution up to a few nucleotides [95–97]. However, for large genomes, tiling array is expensive. Another limiting factor of hybridization methodology is high background, because it is unable to distinguish RNA molecules sharing high sequence similarity [98].

3.2 RNA-seq

Compared with Sanger sequencing, the core of NGS is massive parallel sequencing. Development of nanotechnology makes it possible to sequence hundreds of thousands of DNA molecules simultaneously [13,99]. The prototype of NGS is massive parallel signature sequencing (MPSS) [99], which applies four rounds of restriction enzyme digestion and ligation reactions to determine the nucleotide sequence of cDNA ends generating a 17–20 bp sequence as the fingerprint of a corresponding RNA. MPSS is used to digitize the quantitative transcriptome with the capacity to produce more than 100000 signatures at a time. However, due to the nature of digestion and ligation reactions, a large fraction of the sequence signatures obtained is not long enough to be unique fingerprints of RNA molecules.

Overcoming the limits of MPSS, Illumina, Roche, Lifescientific, and other companies developed their own platforms with considerable improvement on the throughput, reading length, and sequencing accuracy [13]. Based on these platforms, the RNA-seq methodology became the most convenient and cost effective tool for transcriptome analysis [14]. Briefly, total or part of RNA transcripts (e.g., polyA RNA or small RNA fraction) are purified and reversely transcribed into cDNAs, which are subjected to massive parallel sequencing. By analyzing millions to billions of 25–500 bp sequence tags from massive parallel sequencing, the transcriptome can be studied qualitatively and quantitatively [45]. In addition, RNA-seq is an approximately unbiased way, even without prior knowledge of genomic information. Because of the single-base resolution, RNA-seq's background noise is very low compared to hybridization-based technology. Linear detection range of RNA by RNA-seq spans several orders, which is at least one order higher than the DNA chip [45].

3.3 Advance of RNA-seq

In order to accurately reveal the transcriptome of complex biological tissue or precious sample, low RNA input (even single cell) RNA-seq techniques have been developed by direct RNA sequencing or RNA amplification method [100–103]. Tang et al. [104] reported the first single cell transcriptome analysis, where the authors reversely transcribed polyA RNA from a single mouse blastomere lysate. The cDNA was subjected to PCR amplification with the primers annealing to the anchoring sequences introduced during the generation of the 1st and 2nd strand cDNA. The STRT (single-cell tagged reverse transcription) [74] and SMART-seq (switching mechanism at the 5' end of the RNA transcript sequencing) [75] methods took the advantage of the extra few cytosines added by MMLV reverse transcriptase to amplify the cDNAs transcribed from a single cell. The *in vitro* transcription amplification strategy was applied in the latest CEL-seq (cell expression by linear amplification and sequencing) [76], which introduced the T7 promoter into the 5' end of the 1st strand cDNA and directed a linear and less biased amplification. Besides the spiking RNA molecules, barcoding strategy was implemented in single cell RNA-seq to overcome the bias brought from the cDNA amplification [105,106], in which each single RNA molecule was uniquely barcoded. Therefore, multiple reads of a specific RNA molecule with the same barcode would be considered as amplification redundancy and counted only once.

Over the past few years, RNA-seq technology has been widely used in the transcriptomics study not only because of its advantages over previously prevalent methodologies but also because of its fast evolvement. In order to fulfill different research purposes, the preparation of the cDNA library has been modified to different forms, such as strand-specific RNA-seq. Combined with conventional molecular biology and biochemistry methods (Table 1), RNA-seq was applied to study different aspects of the transcriptome, such as deepCAGE and CAP-Seq to map TSS, GRO-seq and PRO-seq to detect nascent RNA, small RNA-seq, degradome-seq and Argounate CLIP-seq to characterize miRNA targets, and so on.

4 Bioinformatics analysis

It becomes increasingly obvious that bioinformatics analysis is a significant part of transcriptome research. The challenge facing bioinformatics analysis is almost as big as the experimental procedure including RNA purification, cDNA library construction, and high-throughput sequencing. The difficulty of analysis comes from not only massive amounts

of data, errors introduced by sequencing experiments, but also unimaginable complexity of the transcriptome.

A typical RNA-seq data analysis can be summarized as follows: (i) Perform quality control for raw RNA-seq data. Low-quality sequence tags produced from library construction or sequencing process are trimmed away by the software provided by the sequencing platform. (ii) For cases with a reference genome, map millions of short reads to the reference genome, determine the position of each RNA transcript in the genome, calculate the expression level of each transcript, and then find differentially expressed genes across the samples. All the above processes are carried out by corresponding pipelines, including representatives of Bowtie to map reads to the reference genome [107], Tophat to identify splice junctions [108], cufflinks to test for differential expression [109], and so on. For cases without a reference genome, *de novo* transcriptome assembly is performed from short RNA-seq reads [93], and then all assembled contigs are subjected to functional annotation, which requires extensive computer resources.

Further specific analysis will be performed to answer certain questions involved in transcriptomics, such as analysis of RNA editing and ncRNAs, discovery of novel transcripts, and correlation of transcriptome data to available genomic or epigenetic data.

5 Summary

Over the past 10 years, great achievements have been made in understanding the transcriptome of cells due to the genomics research needs and omics technology innovation. Particularly, the implementation of NGS based RNA-seq technology changed our view on the transcription of genome and its regulation. Meanwhile, previously impossible high throughput transcriptome analyses for basic or applied research become routine and cost effective. However, transcriptomics study faces several challenges: RNA-seq technology needs to be improved to overcome the bias introduced by library construction or RNA amplification and to reduce the cost for low input RNA-seq; experimental design and bioinformatics analysis remain to be optimized to efficiently and accurately characterize the transcriptome; how to reconstruct the complex RNA-based gene regulatory networks using transcriptome data, in order to help us fully understand the biological processes under different physiological conditions at cell, tissue, individual, and population levels. In short, functional studies of transcriptome will increasingly impact the fields from molecular biology to clinical applications and so on.

This work was supported by grants from the National Natural Science Foundation of China (31271318), Natural Science Foundation of Guangdong (S2012010008912), Foundation of Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences.

1 Venter J C, Adams M D, Myers E W, et al. The Sequence of the human genome. Science, 2001, 291: 1304–1351
2 Hamilton J P, Buell C R. Advances in plant genome sequencing. Plant J, 2012, 70: 177–190
3 Lockhart D J, Winzeler E A. Genomics, gene expression and DNA arrays. Nature, 2000, 405: 827–836
4 Carthew R W, Sontheimer E J. Origins and mechanisms of miRNAs and siRNAs. Cell, 2009, 136: 642–655
5 Eddy S R. Non-coding RNA genes and the modern RNA world. Nat Rev Genet, 2001, 2: 919–929
6 Mattick J S. The functional genomics of noncoding RNA. Science, 2005, 309: 1527–1528
7 Taft R J, Pheasant M, Mattick J S. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays, 2007, 29: 288–299
8 Willingham A T, Gingeras T R. TUF love for "junk" DNA. Cell, 2006, 125: 1215–1220
9 Carninci P, Yasuda J, Hayashizaki Y. Multifaceted mammalian transcriptome. Curr Opin Cell Biol, 2008, 20: 274–280
10 Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. Science, 2005, 309: 1559–1563
11 Shabalina S A, Spiridonov N A. The mammalian transcriptome and the function of non-coding DNA sequences. Genome Biol, 2004, 5: 105
12 Green E D, Chakravarti A. The human genome sequence expedition: Views from the "base camp". Genome Res, 2001, 11: 645–651
13 Mardis E R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 2008, 9: 387–402
14 Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10: 57–63
15 Kohler R E. The eighth day of creation: Makers of the revolution in biology. Isis, 1997, 88: 730–731
16 Gesteland R F, Cech T R, Atkins J F. The RNA World, 2nd Ed: The Nature of Modern RNA Suggests a Prebiotic RNA World. New York: Cold Spring Harbor Laboratory Press, 1999
17 Crick F H. On protein synthesis. Symp Soc Exp Biol, 1958, 12: 138–163
18 Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol, 1961, 3: 318–356
19 Crick F H. The origin of the genetic code. J Mol Biol, 1968, 38: 367–379
20 Hoagland M B, Stephenson M L, Scott J F, et al. A soluble ribonucleic acid intermediate in protein synthesis. J Biol Chem, 1958, 231: 241–257
21 Berget S M, Moore C, Sharp P A. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA, 1977, 74: 3171–3175
22 Chow L T, Gelinas R E, Broker T R, et al. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell, 1977, 12: 1–8
23 Stark B C, Kole R, Bowman E J, et al. Ribonuclease P: An enzyme with an essential RNA component. Proc Natl Acad Sci USA, 1978, 75: 3717–3721
24 Cech T R. The generality of self-splicing RNA: Relationship to nuclear mRNA splicing. Cell, 1986, 44: 207–210
25 Guerrier-Takada C, Gardiner K, Marsh T, et al. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell, 1983, 35(3 Pt 2): 849–857
26 Kruger K, Grabowski P J, Zaug A J, et al. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. Cell, 1982, 31: 147–157
27 Ecker J R, Davis R W. Inhibition of gene expression in plant cells by expression of antisense RNA. Proc Natl Acad Sci USA, 1986, 83: 5372–5376
28 Napoli C, Lemieux C, Jorgensen R. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes *in trans*. Plant Cell, 1990, 2: 279–289

29 Romano N, Macino G. Quelling—transient inactivation of gene-expression in neurospora-crassa by transformation with homologous sequences. Mol Microbiol, 1992, 6: 3343–3353
30 Fire A, Xu S, Montgomery M K, et al. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature, 1998, 391: 806–811
31 Montgomery M K, Xu S Q, Fire A. RNA as a target of double-stranded RNA-mediated genetic interference in *Caenorhabditis elegans*. Proc Natl Acad Sci USA, 1998, 95: 15502–15507
32 Elbashir S M, Harborth J, Lendeckel W, et al. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature, 2001, 411: 494–498
33 Hamilton A J, Baulcombe D C. A species of small antisense RNA in posttranscriptional gene silencing in plants. Science, 1999, 286: 950–952
34 Zamore P D, Tuschl T, Sharp P A, et al. RNAi: Double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell, 2000, 101: 25–33
35 Kim V N. microRNA biogenesis: Coordinated cropping and dicing. Nat Rev Mol Cell Biol, 2005, 6: 376–385
36 Krol J, Loedige I, Filipowicz W. The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet, 2010, 11: 597–610
37 Brodersen P, Voinnet O. Revisiting the principles of microRNA target recognition and mode of action. Nat Rev Mol Cell Biol, 2009, 10: 141–148
38 Lippman Z, Martienssen R. The role of RNA interference in heterochromatic silencing. Nature, 2004, 431: 364–370
39 Mello C C, Conte D Jr. Revealing the world of RNA interference. Nature, 2004, 431: 338–342
40 Peng J C, Lin H. Beyond transposons: The epigenetic and somatic functions of the Piwi-piRNA mechanism. Curr Opin Cell Biol, 2013, 25: 190–194
41 Saxe J P, Lin H. Small noncoding RNAs in the germline. Cold Spring Harb Perspect Biol, 2011, 3: a002717
42 Zhang H, Chen Z, Wang X, et al. Long non-coding RNA: A new player in cancer. J Hematol Oncol, 2013, 6: 37
43 Wilusz J E, Sunwoo H, Spector D L. Long noncoding RNAs: Functional surprises from the RNA world. Genes Dev, 2009, 23: 1494–1504
44 Jacquier A. The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs. Nat Rev Genet, 2009, 10: 833–844
45 Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 2008, 320: 1344–1349
46 Wilhelm B T, Marguerat S, Watt S, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature, 2008, 453: 1239–1243
47 Mortazavi A, Williams B A, Mccue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods, 2008, 5: 621–628
48 Boguski M S, Tolstoshev C M, Bassett D E. Gene discovery in dbEST. Science, 1994, 265: 1993–1994
49 Schena M, Shalon D, Davis R W, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995, 270: 467–470
50 Velculescu V E, Zhang L, Vogelstein B, et al. Serial analysis of gene expression. Science, 1995, 270: 484–487
51 Lashkari D A, DeRisi J L, McCusker J H, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA, 1997, 94: 13057–13062
52 Lewin B, Krebs J E, Goldstein E S, et al. Lewin's genes X. Sudbury: Jones and Bartlett, 2011
53 Ozsolak F, Milos P M. RNA sequencing: Advances, challenges and opportunities. Nat Rev Genet, 2011, 12: 87–98
54 Shiraki T. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA, 2003, 100: 15776–15781
55 Carninci P. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet, 2006, 38: 626–635
56 Lu C, Tej S S, Luo S, et al. Elucidation of the small RNA component of the transcriptome. Science, 2005, 309: 1567–1569
57 Lu C, Meyers B C, Green P J. Construction of small RNA cDNA libraries for deep sequencing. Methods, 2007, 43: 110–117
58 Valen E. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res, 2009, 19: 255–265
59 Plessy C. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. Nat Methods, 2010, 7: 528–534
60 Ni T. A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat Methods, 2010, 7: 521–527
61 Kwak H, Fuda N J, Core L J, et al. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science, 2013, 339: 950–953
62 Core L J, Waterfall J J, Lis J T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science, 2008, 322: 1845–1848
63 Jan C H, Friedman R C, Ruby J G, et al. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. Nature, 2011, 469: 97–101
64 Shepard P J, Choi E A, Lu J, et al. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA, 2011, 17: 761–772
65 Tani H, Mizutani R, Salam K A, et al. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res, 2012, 22: 947–956
66 Imamachi N, Tani H, Mizutani R, et al. BRIC-seq: A genome-wide approach for determining RNA stability in mammalian cells. Methods, 2013, doi: 10.1016/j.ymeth.2013.07.014
67 Hafner M, Landthaler M, Burger L, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell, 2010, 141: 129–141
68 Chi S W, Zang J B, Mele A, et al. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature, 2009, 460: 479–486
69 Zisoulis D G, Lovci M T, Wilbert M L, et al. Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. Nat Struct Mol Biol, 2010, 17: 173–179
70 German M A, Pillay M, Jeong D H, et al. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol, 2008, 26: 941–946
71 Addo-Quaye C, Eshoo T W, Bartel D P, et al. Endogenous siRNA and miRNA targets identified by sequencing of the *Arabidopsis* degradome. Curr Biol, 2008, 18: 758–762
72 Zhou M, Gu L, Li P, et al. Degradome sequencing reveals endogenous small RNA targets in rice (*Oryza sativa* L. ssp. indica). Front Biol, 2010, 5: 67–90
73 Wu L, Zhang Q, Zhou H, et al. Rice microRNA effector complexes and targets. Plant Cell, 2009, 21: 3421–3435
74 Islam S, Kjallquist U, Moliner A, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res, 2011, 21: 1160–1167
75 Ramskold D, Luo S, Wang Y C, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol, 2012, 30: 777–782
76 Hashimshony T, Wagner F, Sher N, et al. CEL-Seq: Single-cell RNA-Seq by multiplexed linear amplification. Cell Rep, 2012, 2: 666–673
77 Sultan M. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008, 321: 956–960
78 Carninci P. Is sequencing enlightenment ending the dark age of the transcriptome? Nat Methods, 2009, 6: 711–713
79 Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem, 2010, 79: 321–349
80 Li J B. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science, 2009, 324: 1210–1213

81 Park E, Williams B, Wold B J, et al. RNA editing in the human ENCODE RNA-seq data. Genome Res, 2012, 22: 1626–1633
82 Shi Y. Alternative polyadenylation: New insights from global analyses. RNA, 2012, 18: 2105–2117
83 Kozomara A, Griffiths-Jones S. miRBase: Integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res, 2011, 39: D152–D157
84 Hammell C M. The microRNA-argonaute complex: A platform for mRNA modulation. RNA Biol, 2008, 5: 123–127
85 Yang J H, Li J H, Shao P, et al. starBase: A database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res, 2011, 39: D202–D209
86 Mamanova L. FRT-seq: Amplification-free, strand-specific transcriptome sequencing. Nat Methods, 2010, 7: 130–132
87 Lipson D. Quantification of the yeast transcriptome by single-molecule sequencing. Nat Biotechnol, 2009, 27: 652–658
88 Parkhomchuk D. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res, 2009, 37: e123
89 Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res, 2012, 22: 1775–1789
90 Cloonan N. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods, 2008, 5: 613–619
91 Maher C A. Transcriptome sequencing to detect gene fusions in cancer. Nature, 2009, 458: 97–101
92 Korbel J O. Paired-end mapping reveals extensive structural variation in the human genome. Science, 2007, 318: 420–426
93 Grabherr M G, Haas B J, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol, 2011, 29: 644–652
94 Clark T A, Sugnet C W, Ares M Jr. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science, 2002, 296: 907–910
95 Stolc V, Samanta M P, Tongprasit W, et al. Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. Proc Natl Acad Sci USA, 2005, 102: 4453–4458
96 Bertone P, Stolc V, Royce T E, et al. Global identification of human transcribed sequences with genome tiling arrays. Science, 2004, 306: 2242–2246
97 Cheng J, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science, 2005, 308: 1149–1154
98 Draghici S, Khatri P, Eklund A C, et al. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet, 2006, 22: 101–109
99 Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol, 2000, 18: 630–634
100 Dafforn A. Linear mRNA amplification from as little as 5 ng total RNA for global gene expression analysis. Biotechniques, 2004, 37: 854–857
101 Lo Y M. Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. Nature Med, 2007, 13: 218–223
102 Amit I. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science, 2009, 326: 257–263
103 Ozsolak F, Platt A R, Jones D R, et al. Direct RNA sequencing. Nature, 2009, 461: 814–818
104 Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods, 2009, 6: 377–382
105 Shiroguchi K, Jia T Z, Sims P A, et al. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci USA, 2012, 109: 1347–1352
106 Kivioja T, Vaharautio A, Karlsson K, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods, 2012, 9: 72–74
107 Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, 10: R25
108 Trapnell C, Pachter L, Salzberg S L. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25: 1105–1111
109 Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010, 28: 511–515