SCIENCE CHINA
Life Sciences

• REVIEWS •

December 2011 Vol.54 No.12: 1121–1128
doi: 10.1007/s11427-011-4255-x

Overview of available methods for diverse RNA-Seq data analyses

CHEN Geng[1], WANG Charles[2] & SHI TieLiu[1,3*]

[1]*Center for Bioinformatics and Computational Biology, Institute of Biomedical Sciences, School of Life Science, East China Normal University, Shanghai 200241, China;*
[2]*Functional Genomics Core, Beckman Research Institute, City of Hope Comprehensive Cancer Center, Duarte, CA 91010, USA;*
[3]*Shanghai Information Center for Life Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Received October 6, 2011; accepted November 16, 2011

RNA-Seq technology is becoming widely used in various transcriptomics studies; however, analyzing and interpreting the RNA-Seq data face serious challenges. With the development of high-throughput sequencing technologies, the sequencing cost is dropping dramatically with the sequencing output increasing sharply. However, the sequencing reads are still short in length and contain various sequencing errors. Moreover, the intricate transcriptome is always more complicated than we expect. These challenges proffer the urgent need of efficient bioinformatics algorithms to effectively handle the large amount of transcriptome sequencing data and carry out diverse related studies. This review summarizes a number of frequently-used applications of transcriptome sequencing and their related analyzing strategies, including short read mapping, exon-exon splice junction detection, gene or isoform expression quantification, differential expression analysis and transcriptome reconstruction.

next generation sequencing, transcriptome, RNA-Seq data analysis, transcriptomics

Citation: Chen G, Wang C, Shi T L. Overview of available methods for diverse RNA-Seq data analyses. Sci China Life Sci, 2011, 54: 1121−1128, doi: 10.1007/s11427-011-4255-x

RNA-Seq is a very powerful technology for transcriptomics studies. It enables us to investigate the gene activities of organisms at different tissues, different stages, and/or under different conditions. Compared with microarrays, RNA-Seq could capture almost all of the expressed transcripts for a snapshot of cells in theory, while microarrays rely on prior information that cannot detect novel splicing variants, novel genes, and novel transcripts. In addition, RNA-Seq has low background noise and high sensitivity, requires less RNA sample, and is becoming more cost-effective with the rapid advancements in the technology [1,2]. Those advantages of RNA-Seq provide us the abilities to illustrate the complexity of transcriptome more comprehensively and generate an unprecedented global view of the transcriptome for various species [3].

To date, RNA-Seq has been applied to a number of species for various research, such as inferring alternative splicing [4,5], quantifying the expression of genes and transcripts [6,7], detecting gene fusions [8,9], revealing long noncoding RNAs (lncRNAs) [10], and identifying single nucleotide variants (SNVs) in expressed exons [11]. Although RNA-Seq has brought tremendous benefits to these studies, it also faces many challenges from both sequencing technologies themselves and bioinformatics analyses of the data. In detail, RNA-Seq has biases in library construction, and strand-specific libraries are still not easy to be produced but are important for determining the orientation of transcripts [1]. Furthermore, RNA-Seq generates a large amount of data, and the read length is generally short and sequencing errors exist in the reads. These aspects challenge the corresponding methods and algorithms to effectively process the large amount of RNA-Seq data.

*Corresponding author (email: tlshi@sibs.ac.cn)

Reference genome sequences are crucial for accurately conducting various RNA-Seq studies, because they provide the templates for reads-mapping. The related annotations on the reference sequences can guide algorithms to optimize analysis of the results. Since the current sequencing technologies are mainly used on model organisms and common species involved in research, many other organisms remain to be sequenced, and lack available reference genomes. In addition, despite that the genomes of some organisms have been sequenced, their reference genomes still leave gaps that have not been filled and/or their reference genomes are not well annotated. For those organisms that have relatively complete and have high quality reference genomes, we can directly map the RNA-Seq reads onto the reference and carry out diverse transcriptomics studies. However, for those organisms without reference genomes or their reference genomes are uncompleted, other methods are required to accomplish related research.

In this review, we present an overview of currently available methods that can be used to carry out diverse transcriptomics studies using transcriptome sequencing data, including short read mapping, exon-exon splicing junctions detection, genes or isoforms expression quantification, differential expression analysis, and transcriptome reconstruction (Figure 1). Considering that some species have the built reference genomes, but most of remaining organisms still have no corresponding available references, we also provide related suggestions with different strategies to achieve the corresponding research goals.

1 Applications of RNA-Seq data

1.1 Short read mapping

Transcriptome sequencing reads are usually first mapped to the genome or the transcriptome sequences, and read alignment is a basic and crucial step for the mapping-first based analytical methods. The complexities of genome sequences have direct influences on the mapping accuracy of short reads. The prokaryote genomes are small and their genomic sequences are not as complex relative to eukaryotes. However, mammalian genomes are usually very large and contain many repetitive and homologous sequences. These high sequence similarities are big challenges for short read mapping. Furthermore, the reads from the splice junctions need to be split into segments across the introns and then mapped onto the reference genome sequences. However, the exons and introns are very different in length and these differences create difficulties in developing well-performing mapping algorithms across genomes. Given that the introns are either too short or too long, it would take more computational time to search the true boundaries of

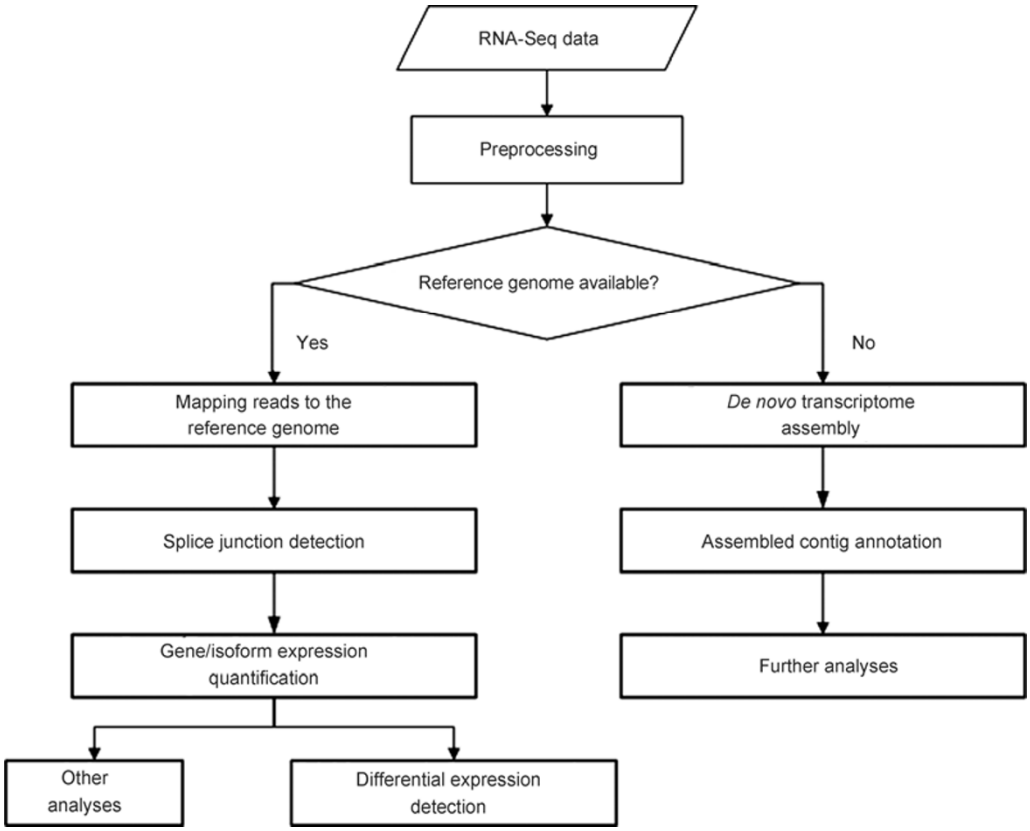

Figure 1 Regular analysis procedure of RNA-Seq data. Transcriptome sequencing data usually can be analyzed using two different strategies according to the corresponding high-quality-assembled reference genome available or not.

them and map the segments correctly. If the exons are shorter than the read length, then the reads with those exons will need to be split into multiple segments during mapping, which further complicates the process. Moreover, for the reads that are 35–400 bp long, the sequencing errors in the reads and the large amount of reads add the difficulties and ambiguities for their alignments. Accordingly, mapping these short read sequences rapidly and accurately is crucial to effectively process the RNA-Seq data and accomplish various analytical tasks.

Short read mappers for RNA-Seq could be divided into unspliced and spliced ones. Unspliced read mappers are suitable for aligning reads against known transcript databases to quantify the gene or isoform expression. Spliced read mappers are usually used to align reads onto reference genomes, allowing large gaps in consideration of introns. Those spliced read aligners first align the reads to the reference genomes using unspliced aligners, then split the unmapped reads into shorter segments, and map them independently to cross possible introns. They are often used to infer exon-exon splice junctions and will be introduced in the next section. Currently, two classic approaches are widely used in the unspliced short read mappers (Table 1): Hash Look-up Table Algorithms and Burrows-Wheeler Transform (BWT)-based methods [12–23]. Hash-based implementations (such as Maq [12], ZOOM [13], RMAP [14], SeqMap [15], and SOAP [16]) can be further differentiated into two classes based on the memory usage. One type of the memory usages depends on the size of reads and the read length, and the other relies on the size of genome and the seed length. BWT-based approaches can significantly reduce the memory desired and accelerate the mapping speed significantly (such as Bowtie [17], SOAP2 [18], and BWA [19]). Both Hash-based and BWT-based strategies can be used to process short reads, but they have some differences in performance due to their different ways of aligning short reads. These differences include memory usage, time consumed (or speed), read length support, number of mapped reads, and alignment accuracy. In practice, using BWT methods to index the reference genome can reduce memory usage and obtain a higher speed for mapping, while Hash-based approaches might achieve better mapping sensitivity and accuracy.

When mapping the short reads to reference sequences, many factors should be considered. Due to sequencing errors, some nucleotides in the reads might be incorrect and will influence the read mapping. A pre-processing is needed to remove those low-quality bases or reads. Although many short read aligners allow mismatches, only a few of them support gapped alignment (it is important for considering the insertions and deletions). In addition, several pieces of software consider the quality of the bases during read alignment while others do not. Another big challenge is that paralogous gene families, repetitive sequences, and the high sequence similarity between alternative spliced isoforms from the same gene will cause mapping ambiguities, and will result in some reads to be aligned to more than one mapping locations. These factors will affect the further analysis, such as splice junction detection, and gene or isoform expression quantification. Therefore, addressing these read mapping difficulties is crucial for the mapping-first related studies. Some approaches have been proposed to handle multi-mapped reads such as allocating them in proportion to the number of uniquely mapped reads [24], and using generative statistical model and associated inference methods to address the computational issue of read mapping uncertainty [25].

1.2 Exon-exon splicing junction detection

Alternative splicing is very common in the gene transcriptional process of eukaryotes, and is very important for the genomes to generate various RNAs (both protein-coding and non-protein-coding) to ensure the related organisms function normally [10,26]. As of now, only a few model organisms have relatively well-annotated exon-exon splicing junctions, and the genomes of the majority of species are still not sequenced or well-annotated. However, even for those well-annotated model organisms, their gene annotations on their reference genomes are also incomplete. Trapnell *et al.* [7] detected thousands of previously unannotated transcripts by analyzing the RNA-Seq data from mouse myoblast cell line. Studies by Guttman *et al.* [10] revealed more than a thousand large intergenic noncoding RNAs

Table 1 Tools for short read mapping

Name	Website	Strategy	Ref.
Bowtie	http://bowtie.cbcb.umd.edu	BWT-based	[17]
BWA	http://bio-bwa.sourceforge.net/	BWT-based	[19]
Soap2	http://soap.genomics.org.cn/soapaligner.html	BWT-based	[18]
Maq	http://maq.sourceforge.net/	Hash-based	[12]
RMAP	http://rulai.cshl.edu/rmap/	Hash-based	[14]
SeqMap	http://biogibbs.stanford.edu/Bjiangh/SeqMap/	Hash-based	[15]
SHRiMP	http://compbio.cs.toronto.edu/shrimp/	Hash-based	[20]
SSAHA2	http://www.sanger.ac.uk/resources/software/ssaha2/	Hash-based	[21]
SOAP	http://soap.genomics.org.cn/soap1/	Hash-based	[16]
ZOOM	http://www.bioinfor.com/	Hash-based	[13]

from the transcriptome sequencing data of mouse embryonic stem cells. In addition, the detected splice junctions between exons are crucial for further inferring the isoforms generated from genes and quantifying the expression of genes and/or isoforms. Therefore, accurate detection of the splice junctions between exons is extremely important for further analyses.

RNA splicing causes the main challenge to correctly map the reads that cover splice junctions to reference sequences. To identify the splice junctions between exons, the software must support spliced mapping for reads, because the reads across the splice junctions need to be split into smaller segments, and then mapped to different exons by crossing-checking with possible introns. Several pieces of software for detecting the splice junctions have been developed as shown in Table 2 [27–33]. TopHat [27] aligns RNA-Seq reads to genomes using bowtie [17] and then predicts the splice junctions between exons according to the mapping results. Because most introns have a "GT-AG" pattern, to ensure the accuracy and save time, TopHat only reports alignments across "GT-AG" introns for reads short than 75 bp. TopHat will also search the "GC-AG" and "AT-AC" introns with longer reads. The method of SpliceMap [28] does not rely on any existing annotation of gene structures and is capable of detecting novel splice junctions with high accuracy. MapSplice [29] is another efficient software that can quickly detect splice junctions with high sensitivity and specificity, and it does not depend on splice site features or intron length. Recently, SOAPsplice [30] has also been developed to robustly detect the splice junctions without using any information of known splice junctions. The software could be used for *de novo* prediction of the splice junctions and used to study the mechanisms of alternative splicing. Since these strategies all need to first map the RNA-Seq reads to the reference genome, they are only applicable to those organisms with available reference sequences.

Table 2 A list of software for splice junction detection

Name	Website	Ref.
HMMSplicer	http://derisilab.ucsf.edu/index.php?software=105	[31]
MapSplice	http://www.netlab.uky.edu/p/bioinfo/MapSplice	[29]
SOAPsplice	http://soap.genomics.org.cn/soapsplice.html	[30]
SpliceMap	http://www.stanford.edu/group/wonglab/SpliceMap/	[28]
SplitSeek	http://solidsoftwaretools.com/gf/project/splitseek/	[32]
Supersplat	http://supersplat.cgrb.oregonstate.edu/	[33]
TopHat	http://tophat.cbcb.umd.edu/	[27]

1.3 Gene and isoform expression quantification

Before RNA-Seq technologies, microarrays were the dominating technologies for investigating the gene expression profiles. However, when quantifying the expression of genes, microarrays are limited to the gene level. By contrast, RNA-Seq can estimate gene expression at both the gene and the isoform levels. Many multi-exon genes would generate multiple isoforms during their expression and different isoforms could play different roles. To comprehensively understand the intricate transcriptome, it is necessary to study the genes at the isoformic level. Our previous work has shown that expression study at the isoformic level enables us to explore the alternative splicing mechanisms in more detail and interpret the complexity of gene expression more comprehensively [34]. Furthermore, RNA-Seq can be used to detect the unannotated genes and isoforms for any species, while microarrays depend on prior information and can only quantify known genes. Those advantages of RNA-Seq make it very useful for annotating the genes of newly sequenced genomes and detecting novel genes or isoforms for organisms whose gene annotations are incomplete.

Up until now, much software is available for gene expression analysis based on the RNA-Seq data (Table 3). Some is designed for quantifying the expression of known genes or isoforms and some others do not need the prior gene structure annotation information [7,10,35–39]. Cufflinks [7] assembles the alignments into a parsimonious set of transcripts and then estimates the relative abundances of these transcripts based on how many reads are mapped onto them. Cufflinks can predict novel genes and isoforms according to the read mapping results on the reference genome. Scripture [10] can ab initio reconstruct the transcriptome and quantify the transcript expression. MISO (Mixture of Isoforms) [36] is a probabilistic framework and uses the inferred assignment of reads to isoforms to estimate the abundances of those isoforms. ALEXA-Seq [35] is a method for alternative expression analysis and also can quantify the expression of isoforms. Besides these algorithms, there are also other software that can be used for the gene expression analysis (Table 3). Users can choose the corresponding

Table 3 Software for gene or isoform expression quantification

Name	Website	Ref.
ALEXA-Seq	http://www.alexaplatform.org/alexa_seq/index.htm	[35]
Cufflinks	http://cufflinks.cbcb.umd.edu/	[7]
IsoInfer	http://www.cs.ucr.edu/~jianxing/IsoInfer.html	[37]
MISO	http://genes.mit.edu/burgelab/miso/	[36]
MMSEQ	http://bgx.org.uk/software/mmseq.html	[38]
rSeq	http://www-personal.umich.edu/~jianghui/rseq/	[39]
Scripture	http://www.broadinstitute.org/software/scripture/?q=home	[10]

software to carry out their analyses according to their needs and research goals

The accuracy of the quantification of gene or isoform expression is largely determined by the mapping results of RNA-Seq reads. Reference genome sequences usually have many repetitive and homologous sequences, and those sequences will cause mapping ambiguities for a portion of the reads. Moreover, assigning these reads across the splice junctions to the correct positions on the reference genome is difficult. Considering these aspects, the best way to precisely quantify the gene or isoform expression is to directly map the RNA-Seq reads to the transcriptome sequences. However, the transcriptome is complex and it is hard to construct an absolute and complete transcript database for an organism, even for the well-studied species like humans or mice. However, if we only want to investigate the expression profiles of known transcripts, directly mapping the transcriptome sequencing reads onto those known transcripts to quantify their expression levels is the best choice.

1.4 Differential expression analysis

Under different conditions, eukaryotic genes will express a number of different and distinct isoforms to meet the organism's need. If we want to assess the expression changes of genes or isoforms between two different states or two samples, we can carry out differential expression analysis to detect the different expressed genes or isoforms. The cost of RNA-Seq is rapidly reducing, and its advantages over microarrays make it increasingly more popular in gene and isoform expression studies. Additionally, RNA-Seq can be used to detect both differentially expressed genes and isoforms, while microarrays are limited for differentially expressed genes. Since multi-exon genes can encode different functional isoforms, this is an important factor to consider when selecting the proper technologies for research. Although it is still relatively more costly to sequence multiple samples than microarrays, RNA-Seq will inevitably and eventually replace microarrays.

For RNA-Seq, the expression level for genes or transcripts is related to the number of reads mapped on them, while for microarrays this is reflected by the fluorescence level recovered after its hybridization process. If the observed difference or change in read count for a gene or transcript between two different experimental conditions is statistically significant, this gene or transcript could be regarded as differentially expressed in RNA-Seq data. However, several RNA-Seq biases should be taken into account when carrying out differential expression analyses, such as sequencing depth, count distribution among samples, and the length of genes or transcripts. Normally, the higher the sequencing depth, the higher the counts will be. Meanwhile, the count distribution among samples can also have differences. Moreover, the read counts for the related transcripts are proportional to the transcript length times the expression level of corresponding RNA. These RNA-Seq biases should be considered in determining the truly differentially expressed genes or isoforms.

Increasingly more strategies are designed to use RNA-Seq data to detect those differentially expressed tags from the investigated gene or transcript sets under different conditions (Table 4). Those methods can be divided into two categories according to their use or disuse of parametric models [7,40–48]. Parametric approaches are based on known probability distributions, such as Binomial, Poisson, and Negative Binomial. By contrast, non-parametric approaches have no assumptions about the data distribution. Recently, Tarazona *et al*. [40] proposed a powerful non-parametric method NOISeq that models the noise distribution from actual data and can be robust against the sequencing depth changes. Their testing results show that it is more flexible than most existing parametric approaches (baySeq [41], DESeq [42], edgeR [43]) against changes of sequencing depth. DESeq, edgeR and baySeq use the Negative Binomial (NB) distribution, and Tarazona *et al*. have demonstrated that these methods show high sequencing depth dependency while NOISeq does not.

1.5 Transcriptome reconstruction

The transcriptome is the total RNAs produced in one or a population of cells, which includes various protein-coding and noncoding RNAs. To obtain the whole transcriptome of an organism, RNA-Seq is a sensible and practical choice. At

Table 4 Available tools for differential expression analysis

Name	Website	Ref.
baySeq	http://www.bioconductor.org/packages/2.8/bioc/html/baySeq.html	[41]
Cuffdiff	http://cufflinks.cbcb.umd.edu/	[7]
DEGseq	http://bioinfo.au.tsinghua.edu.cn/software/degseq/	[44]
DESeq	http://www-huber.embl.de/users/anders/DESeq/	[42]
EdgeR	http://www.bioconductor.org/packages/release/bioc/html/edgeR.html	[43]
GPSeq	http://www-rcf.usc.edu/~liangche/software.html	[45]
Myrna	http://bowtie-bio.sourceforge.net/myrna/index.shtml	[46]
NOISeq	http://bioinfo.cipf.es/noiseq/doku.php?id=start	[40]
ASC	http://www.stat.brown.edu/Zwu/research.aspx	[47]
GENE-Counter	http://changlab.cgrb.oregonstate.edu/?q=node/view/527	[48]

present, there are mainly two classes of strategies for reconstructing the transcriptome (Table 5) [7,10,49–53]. First, there is the 'genome-guided' approach which first maps all the transcriptome sequencing reads to the reference genome, and assembles the aligned reads into transcripts or fragments according to the read mapping information. Programs such as Cufflinks [7] and Scripture [10] use this genome-guided approach. Cufflinks and Scripture both use the spliced reads directly to reconstruct the transcriptome and they have similar computational requirements. Although they are built on conceptually similar assembly graphs, they have differences in processing the graph into transcripts. Cufflinks's process is based on maximum precision while Scripture's is based on maximum sensitivity [49]. The genome-guided method needs a relatively completed and high-quality reference genome that has been established and available for the investigated organism. Another approach to reconstruct the transcriptome is 'genome-independent' approach, which does not need a reference genome, and it directly assembles the reads into transcripts. Programs such as Velvet [50], Trans-ABySS [51], Trinity [52], and Oases (unpublished) are based on this genome-independent approach. It is interesting to note that Velvet can be used for both *de novo* assembling genome and transcriptome. The *de novo* assembly software mainly uses the de Bruijn graphs to model the overlapping subsequences of *k*-mers from the reads. Then it applies a series of algorithms to parse the de Bruijn graph and finally assembles the reads into contigs or scaffolds.

Generally speaking, the genome-guided methods are better suited for species that have high-quality-assembled reference genomes available, and the genome-independent methods can be used for any species, whether they have available reference sequences or not. If one gene was expressed and its transcripts were sequenced, the sequencing reads from that gene could be aligned to the corresponding position the gene locates. This expressed gene would be detected by the genome-guided approaches, regardless of which level this gene is expressed. However, the genome sequences (especially for mammalian genomes) usually contain many repetitive and homologous sequences, and the isoform sequences encoded by the same gene are very similar. These factors will result in ambiguities in the step of reads-mapping from the genome-guided methods, and also lead to assembly collapses for those genome-independent strategies. In addition, the genome-independent methods can mainly reconstruct those transcripts that were expressed at moderate or high levels, but it is difficult to obtain those transcripts expressed at low levels due to the limitation of algorithms, unless the sequencing depth is large.

Whether a genome-guided or genome-independent approach should be adopted largely depends on the research goals, the availability, quality and completeness of the reference genome for that organism. If an organism has a high-quality and relatively complete reference genome, the genome-guided method is the best choice for its gene expression analysis. However, for those organisms that have no available reference genomes, which is still the majority of known species, the genome-independent method is the more reasonable choice. It is worth noting that the repetitive sequences, the limitation of sequencing technologies and assembly algorithms are all major challenges for genome assembling in the genome-independent method. Moreover, even for well-studied model organisms, their reference genomes might be still incomplete and contain gaps and misassembled regions. We have revealed that a significant number of human genes are missing from the human reference genome and expressed with human brain tissues and 10 mixed cell lines in our previous study [54]. Consequently, to construct a complete transcriptome, *de novo* assembly strategy is vital for capturing those transcripts that cannot be obtained from the genome-guided methods due to the incompleteness or misassembly of the reference genome sequences. Hence, combing these two class methods together would enable us to construct a more comprehensive transcriptome for any organism.

2 Conclusion

There are diverse applications for RNA-Seq, and for each application, there are usually a number of available software that can be chosen. However, the software may also have certain parameters that need to be optimized according to the data properties (single-end or paired-end, stranded or non-stranded, etc.) and the characteristics of the organism to be analyzed. Choosing suitable software to carry out related studies and selecting the optimal parameters for the software are both very important and they both directly influ-

Table 5 Transcriptome reconstruction tools

Name	Availability	Category	Ref.
Cufflinks	http://cufflinks.cbcb.umd.edu/	Genome-guide	[7]
Scripture	http://www.broadinstitute.org/software/scripture/?q=home	Genome-guide	[10]
Velvet	http://www.ebi.ac.uk/~zerbino/velvet/	Genome-independent	[50]
Trans-ABySS	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss	Genome-independent	[51]
Trinity	http://trinityrnaseq.sourceforge.net/	Genome-independent	[52]
Oases	http://www.ebi.ac.uk/~zerbino/oases/	Genome-independent	No
Rnnotator	Need contact Virginia de la Puente at vtdelapuente@lbl.gov	Genome-independent	[53]

ence the results. Suitable software and good parameter setting can help us gain better results and achieve our research goals. Moreover, the algorithms used in different software for the same application also have various differences in their design and would possess different advantages on the same dataset. Therefore, it is hard to claim which software is the best or most appropriate on account of that different software has different strengths and different datasets have different features. Accordingly, it is necessary to test the software and different parameters to find an effective way to generate better results before making the final decision. Initial testing could help us find the better and more efficient strategy and significantly improve the analyzing results.

The sequencing technologies and bioinformatics algorithms can influence the analyzing results from different aspects. Although sequencing technologies are undergoing fast development and the algorithms for various applications are also rapidly improved to meet the demands of research, they still have their limitations and drawbacks. In the process of sequencing, the sample preparation step might bring in contaminants and the library construction step might lose sources and fail to capture all the targets. These uncertainties can increase the data noises and lead to incomplete information. Additionally, the sequencing technologies also have bias in sequencing and the bioinformatics algorithms have their own limitations, which can also raise the difficulties in analyzing the data and result in negative results. Undoubtedly, the improvements of sequencing technologies and corresponding analysis algorithms will greatly benefit the data interpretations and facilitate our cognition of the transcriptomes for various species.

In the future, the cost of sequencing will continue to reduce and more powerful algorithms will continue to be developed, which will enable researchers to investigate diverse transcriptomes from different organisms more easily and comprehensively. Furthermore, these changes will also provide us great opportunities to investigate the functions of noncoding RNAs (both short and long) which have been considered transcriptional noise in the past, but in fact might have unknown functions. As investigation of different transcriptomes continues, these contingent research results will enrich our knowledge and even change our previous views about the transcriptome. These new findings will definitely facilitate various related studies and improve our understanding of life.

This work was supported by the National Basic Research Program of China (Grant Nos. 2010CB945401, 2007CB108800), National Natural Science Foundation of China (Grant Nos. 30870575, 31071162, 31000590), and Science and Technology Commission of Shanghai Municipality (Grant No. 11DZ2260300).

1 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10: 57–63
2 Marguerat S, Bahler J. RNA-seq: from technology to biology. Cell Mol Life Sci, 2010, 67: 569–579
3 Ozsolak F, Milos P M. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet, 2011, 12: 87–98
4 Sultan M, Schulz M H, Richard H, *et al*. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008, 321: 956–960
5 Gan Q, Chepelev I, Wei G, *et al*. Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. Cell Res, 2010, 20: 763–783
6 Mortazavi A, Williams B A, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods, 2008, 5: 621–628
7 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
8 Maher C A, Kumar-Sinha C, Cao X, *et al*. Transcriptome sequencing to detect gene fusions in cancer. Nature, 2009, 458: 97–101
9 Pflueger D, Terry S, Sboner A, *et al*. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res, 2011, 21: 56–67
10 Guttman M, Garber M, Levin J Z, *et al*. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol, 2010, 28: 503–510
11 Chepelev I, Wei G, Tang Q, *et al*. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res, 2009, 37: e106
12 Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res, 2008, 18: 1851–1858
13 Lin H, Zhang Z, Zhang M Q, *et al*. ZOOM! Zillions of oligos mapped. Bioinformatics, 2008, 24: 2431–2437
14 Smith A D, Xuan Z, Zhang M Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics, 2008, 9: 128
15 Jiang H, Wong W H. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics, 2008, 24: 2395–2396
16 Li R, Li Y, Kristiansen K, *et al*. SOAP: short oligonucleotide alignment program. Bioinformatics, 2008, 24: 713–714
17 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
18 Li R, Yu C, Li Y, *et al*. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 2009, 25: 1966–1967
19 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009, 25: 1754–1760
20 Rumble S M, Lacroute P, Dalca A V, *et al*. SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol, 2009, 5: e1000386
21 Ning Z, Cox A J, Mullikin J C. SSAHA: a fast search method for large DNA databases. Genome Res, 2001, 11: 1725–1729
22 Trapnell C, Salzberg S L. How to map billions of short reads onto genomes. Nat Biotechnol, 2009, 27: 455–457
23 Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods, 2009, 6: S6–S12
24 Faulkner G J, Forrest A R, Chalk A M, *et al*. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics, 2008, 91: 281–288
25 Li B, Ruotti V, Stewart R M, *et al*. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics, 2010, 26: 493–500
26 Black D L. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem, 2003, 72: 291–336
27 Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25: 1105–1111
28 Au K F, Jiang H, Lin L, *et al*. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res, 2010, 38: 4570–4578

29 Wang K, Singh D, Zeng Z, *et al*. MapSplice: accurate mapping of RNA-Seq reads for splice junction discovery. Nucleic Acids Res, 2010, 38: e178
30 Huang S, Zhang J, Li R, *et al*. SOAPsplice: genome-wide ab initio detection of splice junctions from RNA-Seq data. Front. Gene, 2011, 2: 46
31 Dimon M T, Sorber K, DeRisi J L. HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. PLoS ONE, 2010, 5: e13875
32 Ameur A, Wetterbom A, Feuk L, *et al*. Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol, 2010, 11: R34
33 Bryant D W Jr., Shen R, Priest H D, *et al*. Supersplat—spliced RNA-seq alignment. Bioinformatics, 2010, 26: 1500–1505
34 Chen G, Yin K, Shi L, *et al*. Comparative analysis of human protein-coding and noncoding RNAs between brain and 10 mixed cell lines by RNA-Seq. PLoS ONE, 2011, 6: e28318
35 Griffith M, Griffith O L, Mwenifumbo J, *et al*. Alternative expression analysis by RNA sequencing. Nat Methods, 2010, 7: 843–847
36 Katz Y, Wang E T, Airoldi E M, *et al*. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods, 2010, 7: 1009–1015
37 Feng J, Li W, Jiang T. Inference of isoforms from short sequence reads. J Comput Biol, 2011, 18: 305–321
38 Turro E, Su S Y, Goncalves A, *et al*. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol, 2011, 12: R13
39 Jiang H, Wong W H. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics, 2009, 25: 1026–1032
40 Tarazona S, Garcia-Alcalde F, Dopazo J, *et al*. Differential expression in RNA-seq: A matter of depth. Genome Res, 2011, 21: 2213–2223
41 Hardcastle T J, Kelly K A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics, 2010, 11: 422
42 Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol, 2010, 11: R106
43 Robinson M D, McCarthy D J, Smyth G K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010, 26: 139–140
44 Wang L, Feng Z, Wang X, *et al*. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics, 2010, 26: 136–138
45 Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res, 2010, 38: e170
46 Langmead B, Hansen K D, Leek J T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol, 2010, 11: R83
47 Wu Z, Jenkins B D, Rynearson T A, *et al*. Empirical bayes analysis of sequencing-based transcriptional profiling without replicates. BMC Bioinformatics, 2010, 11: 564
48 Cumbie J S, Kimbrel J A, Di Y, *et al*. GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences. PLoS ONE, 2011, 6: e25279
49 Garber M, Grabherr M G, Guttman M, *et al*. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods, 2011, 8: 469–477
50 Zerbino D R, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res, 2008, 18: 821–829
51 Robertson G, Schein J, Chiu R, *et al*. *De novo* assembly and analysis of RNA-seq data. Nat Methods, 2010, 7: 909–912
52 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
53 Martin J, Bruno V M, Fang Z, *et al*. Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics, 2010, 11: 663
54 Chen G, Li R, Shi L, *et al*. Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. BMC Genomics, 2011, 12: 590