NUCLEIC ACID THERAPEUTICS
Volume 22, Number 4, 2012

DOI: 10.1089/nat.2012.0367

RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation

Yongjun Chu and David R. Corey

Introduction

Detailed characterization of cellular RNA facilitates the design of nucleic acid therapeutics and interpretation of experimental data. Since the first reports of next generation sequencing (NGS) technology-based RNA sequencing (RNA-seq) (Nagalakshmi et al., 2008; Wilhelm, et al., 2008), rapid advances in experimental protocol development and data acquisition and analysis have brought comprehensive sequencing of cellular RNA within reach of most laboratories (Martin and Wang, 2011; Ozsolak and Milos, 2011). While RNA-seq is becoming widely used for laboratory research and clinical studies, the field is still new and strategies for successfully applying RNA-seq to a given experimental problem are often obscure. The goal of this perspective is to introduce some of the choices confronted during RNA sequencing and analysis.

What is RNA-seq?

Twenty years ago, obtaining hundreds of bases of sequence information from a slab gel marked a productive day. Ten years ago, obtaining thousands of bases from a core facility was routine. Today, RNA-seq can reveal the identities of most RNA species inside a cell, providing tens to hundreds of millions of sequence "reads" and information on billions of individual bases. Using this mass of data to gain valid insights, however, requires investing the time to develop a sophisticated understanding of bioinformatics and statistics. It is easy to initiate a project, but it is difficult to obtain and interpret data to adequately answer experimental questions.

RNA sequencing (RNA-seq) is the application of any of a variety of next-generation sequencing techniques (also known as deep sequencing because of their potential for high coverage) to study RNA. It does not usually mean sequencing RNA molecules directly—the actual sequencing step is generally the same for RNA-seq and for DNA sequencing—but the library preparation and the analysis are quite different. RNA-seq library preparation usually includes reverse transcription. In some cases the direction of the transcription is lost during strand amplification, but methods are available for defining transcript direction. Data analysis of RNA-seq may include transcript assembly, alternatively spliced transcript, or novel transcript discovery and transcript quantitation.

Selecting Sequencing Platforms

Before starting a RNA-seq experiment, one must first choose a sequencing platform. The data obtained from the different RNA sequencing platforms vary, and this variation can affect how experiments are interpreted. Protocols for sample preparation differ and choosing the right platform for a given application is a prerequisite to achieve experimental success.

Several NGS platforms are commercially available and more are under active development (Metzker, 2010). Most are based on sequencing-by-synthesis technology, with a DNA polymerase or ligase as the key component. Roche 454, Illumina, Helicos, and PacBio (Pacific Biosciences) use a DNA polymerase to drive their sequencing reaction, while SOLiD (Life Technologies) and Complete Genomics use a DNA ligase. The sequencing platforms can be further categorized as either single molecule-based (sequencing a single molecule, such as Helicos and PacBio) or ensemble-based (sequencing of multiple identical copies of a DNA molecule, such as Illumina and SOLiD).

The selection of a sequencing platform depends on the experimental goals. For example, the sample preparation protocol for Helicos sequencing is relatively simple and might be preferred if the amount of RNA sample is limiting. Helicos also avoids a polymerase chain reaction (PCR) amplification step, giving a direct reflection of RNA expression levels. These characteristics are typical for all single molecule sequencing (SMS) platforms. Generally, non-SMS platforms use amplification steps. With care these amplification-based protocols can provide relative expression levels for most RNAs but require more controls (to avoid PCR over-amplification) and more laborious computational analysis (taking into consideration potential sequence-directed PCR amplification biases).

Single-molecule-based platforms such as Helicos have an inherently high error rate (~5%), dominated by insertions and deletions. A higher error rate makes it more difficult to match sequencing reads with a reference genome and lowers the number of usable reads. If a low sequencing error rate is needed, Illumina or SOLiD are often the best choices (<1%), with mismatches as the major type of errors. The advantage of low error rates is particularly important for microRNA (miRNA) sequencing. Because of the relatively small sizes of miRNAs (ranging from 15 to 27 nt, with most 20 to 22 nt long

Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas.

on average), high error rates cause many raw reads to be lost at the alignment stage. Illumina and SOLiD also offer a greater sequencing depth—higher sequencing capacity renders lowly expressed transcripts to be detected more easily.

Transcriptome assembly is necessary to transform individual reads into sequences of entire mRNAs or noncoding transcripts. The longer the individual reads, the simpler it is to assemble transcripts unambiguously. Currently, Roche 454 and PacBio are the 2 commercialized sequencing platforms that best characterize longer reads, although the paired-end (data describe both 3′ and 5′ ends of the original RNA species prior to amplification) sequencing approach recently implemented by Illumina enables it to provide sequence information for a read that is a few hundred nucleotides long. Roche 454 and PacBio can generate reads that are up to 500 nt long and over 1000 nt long, respectively. Nanopore sequencing techniques may someday reliably sequence a DNA fragment of up to 50 kb or longer (Schneider and Dekker, 2012).

One option is to combine 2 or more RNAseq approaches. For example, a common approach to transcriptome assembly is to obtain data from Illumina/SOLiD to get adequate deep depth of short reads and then use Roche 454 to obtain long reads (Dalloul et al., 2010; Jackman et al., 2010). The short reads are assembled into contigs (contiguously mapped RNA sequences), and long reads are used as scaffolds to connect and validate contigs.

Library Preparation

Library preparation for RNA-seq involves converting cellular RNA into molecules that can be sequenced. Some abundant RNAs, such as ribosomal RNA (rRNA) can comprise up to 80% of total cellular RNA. Sequencing these RNAs can waste resources and reduce the depth of sequence coverage, resulting in less detection of lowly expressed RNA species. rRNA may be removed by an enzymatic degradation approach (such as duplex specific nuclease treatment) (Yi et al., 2012) or hybridization-based depletion methods (Chen and Duan, 2011). This helps ensure that rare transcripts are sequenced to adequate depth.

Library preparation will vary depending on the size of the library desired. RNA can be fragmented, usually by chemical hydrolysis or enzymatic digestion to a size appropriate for the chosen sequencing platform. In some cases the RNA species under investigation, such as miRNAs, are small (under 200 bases) and no fragmentation is required. In other cases the RNAs are long and must be fragmented to smaller sizes, such as ~200–250 nt long, to be suitable for sequencing by Illumina or SOLiD platforms.

Once RNAs of the appropriate size are obtained for most platforms, they are converted into complementary DNA (cDNA) by a reverse transcriptase using random primers. Adapter oligonucleotides are then ligated to the cDNA to allow amplification and enable sequencing.

For some types of Illumina or SOLiD library preparations, the use of specific adapters ligated to the 3′ and 5′ ends of RNA, prior to reverse transcription and PCR, allows identification of the direction of the original RNA strand. For example, the Illumina TruSeq small RNA sample preparation kit produces strand-specific libraries. The kit specifically selects the RNA species that have a monophosphate group at its 5′-end and a hydroxyl group at 3′-end, a typical structure for miRNAs. The Helicos Direct RNA Sequencing (DRS) technique replaces the adaptor ligation with a poly-A tailing step that modifies the RNA fragments directly. Because Helicos DRS directly sequences the RNA it directly provides information about strand specificity.

Bias can be introduced by sample preparation that will cause data to not reflect the actual composition of the sample. Source of bias include reverse transcription (enzymes sometimes not only synthesize first strand cDNA but also make the second strand), ligation (RNA–RNA or RNA–DNA ligation can be inefficient and may be more efficient at some sequences than at others), and random priming (may produce uneven coverage) (Hansen et al., 2010; Ozsolak and Milos, 2011).

RNA Immunoprecipitation and Sequencing

In live cells, most RNAs function by binding to proteins. In RNA immunoprecipitation (RIP) an antibody against a protein of interest is used to recover RNA species bound to the protein. The sequence information of RNA species bound to a specific protein is often desired. RNA immunoprecipitation and sequencing (RIP-seq) is a technique frequently used to address this issue.

A central challenge of RIP-seq is the recovery of bound RNA. If the antibody dissociates from the protein or the protein from the RNA, no signal will be observed. Conversely, if the protein associates with RNA after cells are lysed it is possible that artifactual RNA interactions will be detected. *In vivo* crosslinking is one solution to both of these problems (Chi et al., 2009). The use of *in vivo* ultraviolet (UV) crosslinking captures both transient and kinetically stable RNA-protein interactions. Subsequent treatment with endonuclease elucidates the specific binding sites within the RNA because they will be protected from digestion. RNA protein complexes are then purified by electrophoresis to reduce the non-specific RNA content.

While simple in theory, RIP-seq is challenging in practice. The crosslinking step (UV, 254 nm) is relatively inefficient and only a small amount of RNA is available for library construction. Photoactivatable nucleosides, such as 4-thio-uridine (4-SU) and 6-thio-guanosine (6-SG) (Hafner et al., 2010), can be used to increase UV-crosslinking efficiency. 4-SU incorporation leads to a mutation at the crosslinked site, such that uracil becomes cytosine in the final sequencing read. The appearance of characteristic mutations can help verify the protein binding sites for a RNA. Such mutations, however, make it more difficult to match sequence reads to a reference genome.

Data Analysis

Dealing with the large volume of RNA-seq data generated during experiments is time-consuming and challenging. For example, Hiseq2000 (Illumina) can produce up to 200 million 100-nt reads (approximately 50 GB) of data in one lane in one sequencer run. These data must be processed to not only identify matches to the transcriptome, but also for assembly into transcripts and quantitated before insights can be made into biological meaning. Duplicate or triplicate experimental datasets alleviate data variability and facilitate optimal interpretation of data.

Data is most often supplied in FASTQ format. This format contains an ID number for each read, the read sequence, and a quality score. Being familiar with UNIX working environment is ideal, and an ability to write programming scripts is helpful. There are 2 main stages for sequencing data analysis. First, one must remove sequencing artifacts and errors from the data set. Artifacts may include the ligation adaptors and low-complexity reads. There are publicly available tools that can be used to address these issues easily (Lassmann et al., 2009; Falgueras et al., 2010; Kelley et al., 2010). Sequencing errors can be removed or corrected based on the quality score. Reads containing these errors may be trimmed or corrected to improve the assembly quality. When a large genome is the duty subject, for example the human genome, extremely short reads (< 17 nt) may be filtered out prior to alignment. Short repeats will probably not be assigned to a unique region and will therefore be less definitive.

At the second stage, one aligns the processed data to a reference genome using an appropriate aligner and does downstream data analysis. The data alignment and analysis approach choice depends on the sequencing platform and particular RNA-seq application. Although most of the commercial sequencing platforms have developed their own data analysis pipelines, there are some publically available programs that can be freely downloaded and efficiently run by individual laboratories to carry out total RNA-seq data analysis. TopHat, a fast splice junction mapper for RNA-seq reads, is one of the most commonly used programs (Trapnell et al., 2009). Programs such as Cufflinks (Trapnell et al., 2010) and Scripture (Guttman et al., 2010) may be used to reconstruct the full transcripts, resolve individual variants, and even quantitate expression levels for each transcript and gene. This assembly approach can be used to discover novel transcripts that are not currently annotated.

Further downstream analysis, which may be carried out within Cufflinks package, may include differential expression analysis, in which a variety of statistical models could be used to assess the significance of the data. The final data can then be viewed in a visualization program (such as IGV; Robinson et al. 2011). TopHat, Cufflinks, and Scripture have been used extensively on data generated from Illumina platform but will not be applicable for every RNA-seq application. For data generated from other platforms or specially designed RNA-seq experiments, one may need other programs (Martin and Wang, 2011).

A general challenge to accurate analysis is the need to ensure that software intended to analyze splice junctions and mRNA boundaries not become confused by the existence of multiple isoforms of the same genes or multiple similar genes. It may be necessary to perform locus-directed experiments, such as rapid amplification of cDNA ends, quantitative PCR with strategically chosen primers, and/or targeted RNA-seq (Ozsolak and Milos, 2011) to verify data. While RNA sequencing is powerful, experimental validation will always be necessary to confirm most results.

Summary

Next generation sequencing technologies are evolving rapidly and it is likely that RNA-seq will become routine for many laboratories within the next 5 years. Sequencers are becoming smaller and more personal and are beginning to equip individual departments and laboratories. Library preparation protocols are also becoming shorter and more efficient. Single molecule sequencing will afford insights into the precise orientation of transcription. Advances in methods to acquire sequences are likely to be accompanied by equally rapid advances in computation and data analysis. For most investigators who are not computational biologists, wisely choosing available commercial resources or seeking multidisciplinary partnerships is a short path to success in RNA-seq in the foreseeable future.

Acknowledgments

This work was supported by grants from the Robert A. Welch Foundation (I-1244) and the National Institutes of Health (GM 73042). We thank David Dodd, Keith Gagnon, and Jonathan Watts for their critical review of this manuscript.

Author Disclosure Statement

No competing financial interests exist.

References

CHEN, Z., and DUAN, X. (2011). Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. Methods Mol. Biol. **733**, 93–103.

CHI, S.W., ZANG, J.B., MELE, A., and DARNELL, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature **460**, 479–486.

DALLOUL, R.A., et al. (2010). Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. PLoS Biol. **8**, e1000475.

FALGUERAS, J., et al. (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. BMC Bioinformatics **11**, 38.

GUTTMAN, M., et al. (2010). *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature Biotech. **28**, 503–510.

HAFNER, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell **141**, 129–141.

HANSEN, K.D., BRENNER, S.E., and DUDOIT, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. **38**, e131.

JACKMAN, S.D., and BIROL, I. (2010). Assembling genomes using short-read sequencing technology. Genome Biol. **11**, 202.

KELLEY, D.R., SCHATZ, M.C., and SALZBERG, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol. **11**, R116.

LASSMANN, T., HAYASHIZAKI, Y., and DAUB, C.O. (2009). TagDust: a program to eliminate artifacts from next generation sequencing data. Bioinformatics **25**, 2839–2840.

MARTIN, J.A., and WANG, Z. (2011). Next-generation transcriptome assembly. Nature Rev. Genet. **12**, 671–682.

METZKER, M.L. (2010). Sequencing technologies: the next generation. Nature Rev. Genet. **11**, 31–46.

NAGALAKSHMI, U., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320**, 1344–1349.

OZSOLAK, F., and MILOS, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nature Rev. Genet. **12**, 87–98.

ROBINSON, J.T. et al. (2011). Integrative genomics viewer. Nature Biotechnol. **29**, 24–26.

SCHNEIDER, G.F., and DEKKER, C. (2012). DNA sequencing with nanopores. Nature Biotechnol. **30**, 326–328.

TRAPNELL, C., PACHTER, L., and SALZBERG, S.L. (2009). TopHat: discovering splice junctions with RNA-seq. Bioinformatics **25**, 1105–1111.

TRAPNELL, C., et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnol. **28**, 511–515.

WILHELM, B.T., et al. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature **453**, 1239–1243.

YI, H., et al. (2012). Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. Nucleic Acids Res. **40**, e140.

Address correspondence to:
Dr. David R. Corey
Department of Pharmacology
University of Texas Southwestern Medical Center
6001 Forest Park Road
Dallas, TX 75390

E-mail: david.corey@utsouthwestern.edu

Received for publication May 11, 2012; accepted after revision June 28, 2012.