RNA-Seq: A Method for Comprehensive Transcriptome Analysis

Ugrappa Nagalakshmi,[1] Karl Waern,[1] and Michael Snyder[1]

[1]Molecular, Cellular, and Developmental Biology Department, Yale University, New Haven, Connecticut

ABSTRACT

A recently developed technique called RNA Sequencing (RNA-Seq) uses massively parallel sequencing to allow transcriptome analyses of genomes at a far higher resolution than is available with Sanger sequencing- and microarray-based methods. In the RNA-Seq method, complementary DNAs (cDNAs) generated from the RNA of interest are directly sequenced using next-generation sequencing technologies. The reads obtained from this can then be aligned to a reference genome in order to construct a whole-genome transcriptome map. RNA-Seq has been used successfully to precisely quantify transcript levels, confirm or revise previously annotated 5′ and 3′ ends of genes, and map exon/intron boundaries. This unit describes protocols for performing RNA-Seq using the Illumina sequencing platform. *Curr. Protoc. Mol. Biol.* 89:4.11.1-4.11.13. © 2010 by John Wiley & Sons, Inc.

Keywords: RNA-Seq • transcriptome • high-throughput sequencing • gene expression • annotation • cDNA library preparation

INTRODUCTION

The transcriptome is the complete set of transcripts in a cell, both in terms of type and quantity. Various technologies have been developed to characterize the transcriptome of a population of cells, including hybridization-based microarrays and Sanger sequencing–based methods (Yamada et al., 2003; Bertone et al., 2004; David et al., 2006). The advent of high-throughput sequencing–based methods has changed the way in which transcriptomes are studied. RNA sequencing (RNA-Seq) involves direct sequencing of complementary DNAs (cDNAs) using high-throughput DNA sequencing technologies followed by the mapping of the sequencing reads to the genome. It provides a more comprehensive understanding than has hitherto been possible of the complexity of eukaryotic transcriptomes in that it allows for the identification of exons and introns, the mapping of their boundaries, and the identification of the 5′ and 3′ ends of genes. It also allows the identification of transcription start sites (Tsuchihara et al., 2009), the identification of new splicing variants, and the monitoring of allele expression (unpub. observ.). Finally, it allows for the precise quantification of exon expression and splicing variants (Cloonan et al., 2008; Marguerat et al., 2008; Morin et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Shendure, 2008; Wilhelm et al., 2008; Wang et al., 2009).

This unit contains relevant protocols for RNA-Seq. The Basic Protocol describes the generation of a double-stranded cDNA library using random or oligo(dT) primers. The resulting library exhibits a bias towards the 5′ and 3′ ends of genes, which is useful for mapping the ends of genes and identifying transcribed regions. The cDNA is made from poly(A)$^+$ RNA, then fragmented by DNase I and ligated to adapters. These adapter-ligated cDNA fragments are then amplified and sequenced in a high-throughput manner to obtain short sequence reads. An Alternate Protocol describes the generation of a double-stranded cDNA library using random primers, but starting with poly(A)$^+$ RNA

Current Protocols in Molecular Biology 4.11.1-4.11.13, January 2010
Published online January 2010 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/0471142727.mb0411s89
Copyright © 2010 John Wiley & Sons, Inc.

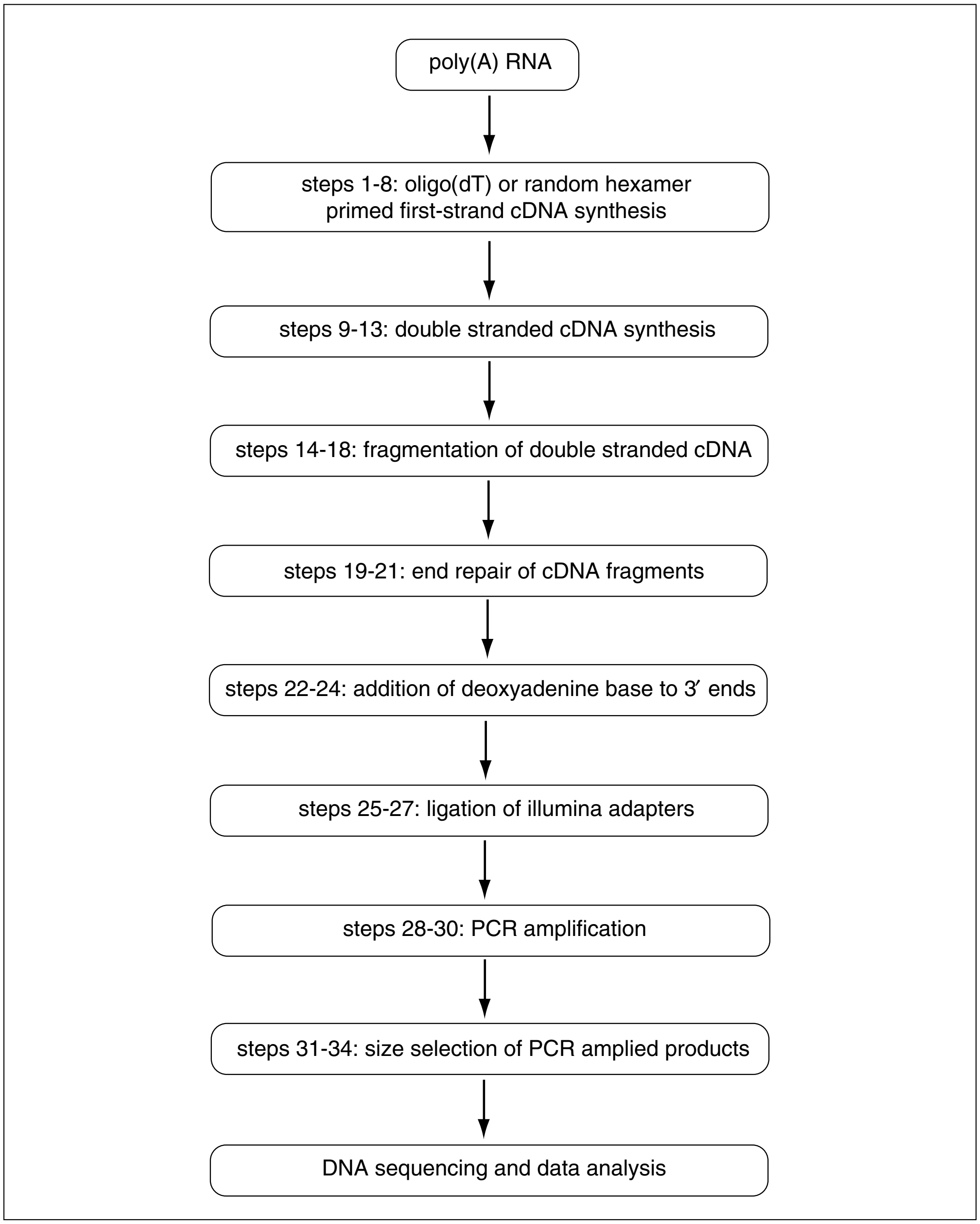
poly(A) RNA
steps 1-8: oligo(dT) or random hexamer primed first-strand cDNA synthesis
steps 9-13: double stranded cDNA synthesis
steps 14-18: fragmentation of double stranded cDNA
steps 19-21: end repair of cDNA fragments
steps 22-24: addition of deoxyadenine base to 3′ ends
steps 25-27: ligation of illumina adapters
steps 28-30: PCR amplification
steps 31-34: size selection of PCR amplied products
DNA sequencing and data analysis

Figure 4.11.1 Flow chart of steps involved in RNA-Seq method (step numbers refer to Basic Protocol).

fragmented by partial hydrolysis. This provides a more uniform representation throughout the genes, which is helpful in quantifying exon levels, but is not as good for end mapping. Sequencing is done with an Illumina Genome Analyzer. Most of the reagents required are available in kits from commercial sources. See Figure 4.11.1 for a flowchart of the steps involved in an RNA-Seq analysis.

BASIC PROTOCOL

cDNA LIBRARY PREPARATION USING FRAGMENTED DOUBLE-STRANDED cDNA

The goal of these procedures is to generate high-quality, full-length cDNA that can be fragmented and ligated to an adapter for amplification and sequencing. Generation of double-stranded cDNA from mRNA involves a number of steps: first, mRNA is converted into first-strand cDNA using reverse transcriptase with either random hexamers or oligo(dT) as primers. The resulting first-strand cDNA is then converted into double-stranded cDNA, which is fragmented and ligated to Illumina adapters for amplification and sequencing.

Materials

Total RNA
500 ng/µl oligo(dT)$_{12-18}$ primers (Invitrogen; store at −80°C)
10 mM dNTP mix (Invitrogen)
Nuclease-free, sterile H_2O
50 ng/µl random hexamer primers (Invitrogen; store at –80°C)
5× first-strand buffer (Invitrogen)
100 mM dithiothreitol (DTT)
200 U/µl SuperScript II Reverse Transcriptase (Invitrogen)
5× second-strand buffer (Invitrogen)
10 U/µl *E. coli* DNA ligase
10 U/µl *E. coli* DNA polymerase I
2 U/µl *E. coli* RNase H
5 U/µl T4 DNA polymerase (Promega)
0.5 M EDTA, pH 8.0 (*APPENDIX 2*)
QIAquick PCR Purification Kit including Buffer EB (Qiagen)
DNase I buffer (New England Biolabs)
DNase I enzyme (New England Biolabs)
End-It DNA End-Repair Kit (Epicentre Biotechnologies) including:
  10× End-Repair Buffer
  End-Repair Enzyme Mix
  10 mM ATP
  2.5 mM dNTP mix
Klenow buffer (NEB buffer 2; New England Biolabs)
Klenow fragment (3′ to 5′ exo$^-$; New England Biolabs)
1 mM dATP (prepare from 100 mM dATP; New England Biolabs); store in 25-µl single-use aliquots at –20°C
QIAquick MinElute PCR Purification Kit including Buffer EB (Qiagen)
T4 DNA ligase buffer (Promega)
Illumina Genomic DNA Sequencing Kit including:
  Illumina Adapter Mix (part no. 1000521)
  Illumina PCR primer 1.1 (part no. 1000537)
  Illumina PCR primer 2.1 (part no. 1000538)
3 U/µl T4 DNA ligase (Promega)
2× Phusion High Fidelity Master Mix (Finnzymes, cat. no. F-531; *http://www.finnzymes.us/*)
1.5% to 2% agarose gel in TAE buffer (*UNIT 2.5A*)
Qiagen Gel Extraction Kit including Buffer EB

Heat block
Thermal cycler
PCR tubes
Horizontal agarose gel electrophoresis system (*UNIT 2.5A*)
Disposable scalpels
NanoDrop spectrophotometer (Thermo Scientific)

Additional reagents and equipment for preparation of poly(A)$^+$ RNA (*UNIT 4.5*) and agarose gel electrophoresis (Support Protocol 2)

Synthesize first-strand cDNA

1. Prepare 100 ng to 1 µg poly(A)$^+$ RNA from the total RNA using an appropriate method (see *UNIT 4.5*). Keep the poly(A)$^+$ RNA at a concentration of at least 100 ng/µl.

   See Critical Parameters and Troubleshooting for more details.

2. Prepare a cocktail on ice containing the following:

   500 ng oligo(dT) or 50 to 250 ng random hexamer primers
   1 µl 10 mM dNTPs
   100 ng to 1 µg poly(A)$^+$ RNA.

   Bring the final volume to 12 µl using nuclease-free sterile water, if necessary.

3. Heat the mixture to 65°C in a heat block for 5 min and quick-chill on ice. Collect the contents of the tube by brief centrifugation.

4. Add the following (total volume should be 19 µl):

   4 µl 5× first strand buffer (1× final)
   2 µl 100 mM DTT (10 mM final)
   1 µl nuclease-free water.

   Mix by pipetting and collect contents by brief centrifugation.

   If more than one RNA sample needs to be processed to generate cDNA, one can prepare a master mix containing these components for all the RNA samples at once.

5. Incubate samples with oligo(dT)$_{12\text{-}18}$ primers for 2 min at 42°C, or with random-hexamer primers for 2 min at 25°C.

6. Add 1 µl (200 U) of SuperScript II reverse transcriptase and mix gently by flicking. Collect contents by brief centrifugation.

7. Incubate samples with oligo(dT)$_{12\text{-}18}$ primers for 50 min at 42°C. For samples with random-hexamer primers incubate for 10 min at 25°C followed by 50 min at 42°C.

8. Incubate tubes at 70°C for 15 min to inactivate the reverse transcriptase.

Synthesize double-stranded cDNA

In the following series of steps the RNA is removed from the DNA-RNA hybrid and a replacement strand is synthesized, thereby generating double-stranded cDNA.

9. Add the following reagents, in this order, to the first-strand reaction tube from step 8 (total volume should be 150 µl).

   91 µl nuclease-free H_2O
   30 µl 5× second strand buffer (1× final)
   3 µl 10 mM dNTP mix (0.2 mM final)
   1 µl 10 U/µl *E. coli* DNA ligase
   4 µl 10 U/µl *E. coli* DNA polymerase I
   1 µl 2 U/µl *E. coli* RNase H.

10. Mix well by pipetting up and down, and incubate for 2 hr at 16°C in a thermal cycler. Take care not to allow the temperature to rise above 16°C.

11. Add 2 µl of 5 U/µl T4 DNA polymerase, mix by pipetting up and down, and incubate at 16°C for an additional 5 min.

    Invitrogen recommends this step for second-strand cDNA synthesis and, while an end repair with T4 DNA polymerase will be done again in steps 19 to 21, a slight increase in mappable reads is typically obtained when this step is included (unpub. observ.).

12. Add 10 µl of 0.5 M EDTA, microcentrifuge briefly to collect solution at bottoms of tubes, and place the tubes on ice.

13. Purify the double-stranded cDNA product using Qiagen's QIAquick PCR Purification Kit. Follow the manufacturer's recommended protocol, but elute in a final volume of 25 µl of Buffer EB.

Fragment double-stranded cDNA

Double-stranded cDNA obtained in step 13 is fragmented using DNase I to generate small fragments of cDNA suitable for sequencing using an Illumina Genome Analyzer.

14. Mix 8 µl of water, 1 µl of DNase I buffer, and 1 µl of DNase I enzyme (2 U/µl) in a microcentrifuge tube.

15. Add 2 µl of this mixture to 25 µl of cDNA from step 13.

16. Add nuclease-free water to bring the total volume of 34 µl. Incubate for 10 min at 37°C and immediately transfer to a 100°C heat block and incubate for 10 min to terminate the DNase I reaction.

    Failure to do this in a timely fashion can result in completely digested cDNA. This incubation time is optimized for yeast and may need to be optimized for other organisms, particularly if the average transcript length differs significantly from yeast. See Critical Parameters and Troubleshooting for further details.

17. Purify the fragmented cDNA using the QIAquick PCR Purification Kit. Follow the manufacturer's recommended protocol, but elute in a final volume of 34 µl of Buffer EB.

18. Place the tube on ice until ready for library preparation.

Perform end repair of cDNA fragments

This protocol converts any overhangs at the cDNA ends into blunt ends using T4 DNA polymerase. The 3′ to 5′ exonuclease activity of these enzymes removes 3′ overhangs, and the polymerase activity fills in 5′ overhangs.

19. Add the following reagents to fragmented cDNA from step 18 (total volume should be 50 µl) and mix by pipetting up and down:

    5 µl 10× end-repair buffer (1× final)
    5 µl 2.5 mM dNTP mix (0.25 mM final)
    5 µl 10 mM ATP (1 mM final)
    1 µl end-repair enzyme mix.

    The standard 50-µl reaction will end-repair up to 5 µg of DNA; the reaction can be scaled up if necessary.

20. Incubate 45 min at room temperature.

21. Purify the end-repaired cDNA fragments using the QIAquick PCR Purification Kit. Follow the manufacturer's recommended protocol, but elute in a final volume of 34 µl of Buffer EB.

Add deoxyadenine base to 3′ ends

An overhanging adenine (A) base is added to the 3′ end of the blunt DNA fragments by the use of Klenow fragment. This aids the ligation of the Illumina adapters, which have a single thymine (T) base overhang at their 3′ ends.

22. Combine and mix the following components in a clean microcentrifuge tube:

    34 μl end-repaired DNA from step 21
    5 μl of Klenow buffer (NEB buffer 2)
    10 μl of 1 mM dATP (see note below)
    1 μl of Klenow fragment (3′ to 5′ exo$^-$)
    Total volume should be 50 μl.

    1 mM dATP stocks should be prepared using 100 mM dATP from NEB. Store the 1 mM dATP in 25-μl aliquots at –20°C. Thaw stocks only once for use in the above described reaction, as dATP is adversely affected by freeze-thaw cycles.

23. Incubate 30 min at 37°C in a water bath or heat block.

24. Purify using Qiagen's QIAquick MinElute PCR Purification kit. Follow the manufacturer's recommended protocol, and elute in a final volume of 10 μl of Buffer EB.

    Note that this kit uses different elution columns than the QIAquick PCR Purification kit.

Ligate Illumina adapters

This protocol ligates adapters (supplied by Illumina) to the ends of cDNA fragments.

25. Combine and mix the following components in a clean microcentrifuge tube (total volume should be 30 μl):

    10 μl purified DNA from step 24
    15 μl of T4 DNA ligase buffer
    1 μl Illumina adapter mix (diluted 1:10 to 1:50 in H_2O)
    2 μl of nuclease-free water
    2 μl of 3 U/μl T4 DNA ligase.

    Illumina recommends diluting their adapter oligo mix at a ratio of 1:10 with water before use. If a low amount of starting material was used, dilute the Illumina adapters 1:30, as excess adapters can interfere with sequencing. The adapters may have to be titrated relative to starting material; see Troubleshooting for more details.

26. Incubate for 15 min at room temperature.

27. Purify 150- to 350-bp DNA fragments using agarose gel electrophoresis (Support Protocol 2). Elute in a final volume of 23 μl Buffer EB.

    If a large starting amount of RNA was used, a QiaQuick PCR Purification Kit can be used instead of agarose gel purification. However, to ensure a higher-quality library, we recommend performing agarose gel purification to remove excess free adapters prior to Illumina sequencing. Adapters can multimerize if this step is not performed.

PCR amplification

28. For each reaction, add the following components to a PCR tube (total volume should be 50 μl):

    23 μl DNA from step 27
    1 μl Illumina PCR primer 1.1
    1 μl Illumina PCR primer 2.1
    25 μl of 2× Phusion DNA polymerase master mix.

    Mix gently by pipetting up and down. Try to avoid creation of bubbles and centrifuge briefly to collect the solution in the bottom of the tube.

29. Place the tubes in the thermal cycler and perform the following thermal cycling program:

1 cycle:	30 sec	98°C	(initial denaturation)
15 cycles:	10 sec	98°C	(denaturation)
	30 sec	65°C	(annealing)
	30 sec	72°C	(extension)
1 cycle:	5 min	72°C	(final extension).

The cycling conditions may need to be optimized, but these above are reasonable starting conditions.

30. Purify the PCR product using the QIAquick MinElute PCR Purification Kit. Follow the manufacturer's recommended protocol, but elute in a final volume of 15 μl of Buffer EB.

Note that this step again uses the MinElute version of the kit.

Size select PCR-amplified cDNA library products

Refer to Support Protocol 2.

31. Electrophorese 15 μl of PCR-amplified product from step 30 on 1.5% to 2% TAE agarose gel (Support Protocol 2).

32. Excise the bands in a range of 150 to 350 bp with a clean, disposable scalpel (Support Protocol 2).

33. Recover the cDNA library product from the gel slices by using Qiagen's Gel Extraction Kit. Follow the manufacturer's recommended protocol and include all optional steps, but elute in a final volume of 15 μl of Buffer EB.

34. Check the concentration of the cDNA library using a spectrophotometer.

A NanoDrop spectrophotometer is recommended, as only 1 to 2 μl volume is required.

The ideal concentration is 15 to 25 ng/μl. If the cDNA concentration is lower, the sequencing efficiency will be low.

cDNA LIBRARY PREPARATION USING HYDROLYZED OR FRAGMENTED RNA

ALTERNATE PROTOCOL

Sequencing using RNA fragmented by partial hydrolysis can also be done for comprehensive transcriptome analysis. This protocol describes cDNA library preparation by partially hydrolyzing the RNA before making cDNA. The cDNA is then made using random hexamers or oligo(dT) primers and sequenced using an Illumina Genome Analyzer. As with the cDNA fragmentation step in the Basic Protocol (step 16), care should be taken to avoid complete degradation during RNA fragmentation.

Additional Materials *(also see Basic Protocol)*

Poly(A)$^+$ RNA prepared from total RNA (*UNIT 4.5*)
10× RNA fragmentation buffer (Ambion)
Stop-reaction buffer (0.2 M EDTA, pH 8.0)

Additional reagents and equipment for purification of fragmented cDNA by ethanol precipitation (Support Protocol 1)

1. Prepare the following reaction mix in a nuclease-free microcentrifuge tube:

1 μl 10× RNA fragmentation buffer
100 ng to 1 μg poly(A)$^+$ RNA
nuclease-free H_2O for total volume of 10 μl.

2. Incubate the tube for 5 min in a 65°C heat block.
3. Add 1 μl of reaction stop buffer and place the tube on ice for 1 min.
4. Purify fragmented RNA by ethanol precipitation (see Support Protocol 1).
5. To prepare the cDNA library for sequencing using the Illumina Genome Analyzer, follow the Basic Protocol starting at step 2 with 100 ng to 1 μg fragmented poly(A)$^+$ RNA, but use 34 μl of Buffer EB in step 13, and skip steps 14 to 18.

SUPPORT PROTOCOL 1

PURIFICATION OF FRAGMENTED RNA BY ETHANOL PRECIPITATION

This protocol describes a basic ethanol precipitation of RNA, and is included for the sake of completeness. Note that there is a commercially available kit to purify the fragmented RNA using bead-based technology from Applied Biosystems. For labs that do not routinely handle RNA, this may be a more convenient solution.

Materials

Tube containing fragmented RNA (Alternate Protocol, step 3)
3 M sodium acetate pH 5.2
100% nuclease-free ethanol
70% nuclease-free ethanol
Nuclease-free water

1. Add the following to the tube containing fragmented RNA in step 3 of the Alternate Protocol:

   2 μl of 3 M sodium acetate pH 5.2
   60 μl of 100% nuclease-free ethanol.

2. Incubate at –80°C for 30 min.
3. Microcentrifuge the tube for 25 min at 14,000 rpm, 4°C.
4. Carefully pipet out ethanol without disturbing the RNA pellet.
5. Wash the pellet in 250 μl of 70% ethanol.
6. Microcentrifuge the pellet for 5 min at 14,000 rpm, 4°C. Pipet off the ethanol without disturbing the pellet.
7. Air dry the pellet for 5 to 10 min.
8. Resuspend the RNA pellet in 10 μl of nuclease-free water.
9. Proceed to double-stranded cDNA synthesis as described in step 5 of the Alternate Protocol.

SUPPORT PROTOCOL 2

PURIFICATION OF cDNA FRAGMENTS

The following protocol is used in the Basic Protocol at steps 27 and 31 to purify a cDNA library from an agarose gel in order to isolate and purify only the cDNA fragments of a length suitable for sequencing on an Illumina Genome Analyzer. An agarose slice is cut from the gel, melted, and purified using Qiagen's Gel Extraction kit following the manufacturer's recommended protocol.

Materials

cDNA library to be isolated
TAE buffer (*APPENDIX 2*)

Disposable scalpels
Qiagen Gel Extraction Kit

Additional reagents and equipment for agarose gel electrophoresis (*UNIT 2.5A*)

1. Prepare a 1.5% to 2% standard agarose/ethidium bromide gel using a 100-cm gel rack (*UNIT 2.5A*).

   Approximately 100 ml of agarose solution will be needed, containing 3 μl of 10 mg/ml ethidium bromide.

2. Load 15 μl of the cDNA library with 1× DNA loading buffer (*UNIT 2.5A*) into each well.

3. Electrophorese at 80 to 100 V for 60 to 90 min.

4. Stop electrophoresis and cut out the target band in a range of 150 to 350 bp with a clean, disposable scalpel.

5. Purify the gel using Qiagen's Gel Extraction Kit following the manufacturer's recommended protocol.

DNA SEQUENCING AND DATA ANALYSIS

SUPPORT PROTOCOL 3

DNA sequencing is performed according to the manufacturer's protocols. Reads mapping and bioinformatic analysis are performed as outlined in Wang et al. (2009), but a brief overview is provided here.

For most labs, the actual sequencing of the cDNA libraries will be done by a core facility. It is useful to have some general knowledge of the process, however, and manufacturers' Web sites, including Illumina's, contain overviews of how their technologies work. Discussions in more depth of the Roche, Illumina, and Applied Biosystems high-throughput sequencing platforms are also available (Mardis, 2008; Morozova and Marra, 2008). cDNA sequencing on the Illumina Genome Analyzer is done in two steps. In the first step, a cluster station is used to prepare a flow cell with up to eight samples (one per lane on the flow-cell). In the second step, the Genome Analyzer sequences the DNA bound to the flow-cell.

In the cluster station, denatured double-stranded sequencing template is loaded onto the flow cell, where the template anneals to oligos covalently bound to the surface of the flow-cell. A second strand is synthesized from these surface-bound oligos, creating a double-stranded template molecule covalently attached to the flow cell surface. These templates are denatured, the free ends of these bound templates are captured by complementary oligos on the flow cell surface, and a new second strand is synthesized, also covalently attached to the flow cell. This process is repeated to create covalently attached "clusters" of identical DNA strands. The more DNA that is loaded onto the flow-cell, the more densely packed these clusters will be. Up to a point, this increases the number of reads, but, as they cluster ever more closely, the Genome Analyzer's intensity-analysis capacity will no longer be able to differentiate between neighboring clusters, at which point little useful data is gained from the sequencing run.

The Genome Analyzer itself will take a flow cell prepared in a cluster station and sequence the DNA bound to it. The DNA strands are denatured, and a sequencing primer—complementary to a sequence in the Illumina adapter oligos attached to each template strand—is used to prime the reaction. The Genome Analyzer then performs sequencing by synthesis, adding one base pair at a time to the DNA in the clusters; each base is color coded with a fluorophore, and the Genome Analyzer's camera records the color of each cluster to determine which base was incorporated. Before commencing the next cycle, the fluorophores are cleaved off.

Software from Illumina with modules called Firecrest, Bustard, and Gerald then convert this fluorophore information to sequence data. The Gerald module can also map these sequences to a reference genome. However, for RNA-Seq it is recommend to use a custom software package that can also map gapped reads (putative introns), end tags [with an extra-genomic poly(A) run], and a slightly higher percentage of the remaining reads.

The bioinformatic analysis of RNA-Seq data (Nagalakshmi et al., 2008; Wang et al., 2009) can be done in several stages. A brief overview follows.

Sequence reads are mapped with a combination of SOAP (Li et al., 2008) and BLAT (Kent, 2002). SOAP is a very fast mapping program, and BLAT contains powerful options for mapping gapped reads. Most tags will map back to a unique place in the genome. The correct location of tags that map to repetitive regions of the genome, however, cannot be unambiguously determined. In addition, two special types of tags are searched for in this process: 3′ end tags and gapped alignments. Gapped alignments are reads which putatively span an intron. 3′ end tags are sequence reads with a non-genomic run of "A" or "T" bases, indicating that they are the site of a polyadenylation event. These are useful in determining the 3′ ends of genes. In addition, these tags help determine from which strand a given transcript was transcribed, as reads mapping to the plus strand with poly(A) tails or to the minus strand with poly(T) tails were transcribed from the plus strand. Conversely, reads mapping to the minus strand with a poly(A) tail or to the plus strand with a poly(T) tail were transcribed from the minus strand.

Subsequently, the bioinformatics pipeline is able to calculate the expression level for each base pair of the genome. In addition, it is possible to annotate the genome with information on where introns are located (via gapped alignments), 5′ ends (via sudden expression drops), and 3′ ends (via sudden expression drops and 3′ end tags). However, overlapping gene expression and low expression levels of a gene can hinder the annotation process.

COMMENTARY

Background Information

The emerging view is that eukaryotic transcriptomes are very complex, involving overlapping transcripts, transcribed intergenic regions, and abundant noncoding RNAs. In the last decade, the transcriptional complexity of the genome has been interrogated mainly with hybridization-based microarray technology (Yamada et al., 2003; Bertone et al., 2004; David et al., 2006). However, the recent advent of high-throughput sequencing technologies is revolutionizing the way complex transcriptomes can be analyzed (Morozova and Marra, 2008). The newly developed RNA-Seq method makes use of next-generation sequencing technology to directly sequence complementary DNAs generated from mRNA, in a high-throughput manner. RNA-Seq yields a comprehensive view of both the transcriptional structure and the expression levels of transcripts (Nagalakshmi et al., 2008; Wang et al., 2009).

Compared to other technologies, RNA-Seq provides a very high signal-to-noise ratio and very large dynamic range. RNA-Seq is also very reproducible, providing a high correlation across biological and technical replicates (Cloonan et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008). With enough sequenced and mapped reads, it can detect and measure rare, yet physiologically relevant, species of transcripts—even those with abundances of 1 to 10 RNA molecules per cell (Mortazavi et al., 2008). It should be noted that, at present, obtaining this level of sequencing depth is an expensive proposition, but sequencing costs can be expected to decline dramatically within the next few years. In addition, spliced transcripts can be uniquely detected through the presence of sequence reads spanning exon-exon junctions (Sultan et al., 2008). Such positive evidence for splicing is not available from microarray-based methods, which require a prior knowledge of the splice sites. RNA-Seq can also be used to determine 5′ start and 3′ termination sites of a transcript (Nagalakshmi et al., 2008). Hence, RNA-Seq is a powerful approach for analyzing the structure of the transcriptome at single-base-pair resolution, for annotating the genome, and for quantifying gene expression on a genome-wide scale.

The protocols described in this unit provide a general method for preparing a cDNA library for high-throughput sequencing. The method has proven to be effective for comprehensive analysis of the transcriptome in yeast (Nagalakshmi et al., 2008) and human (Sultan et al., 2008). The success of this procedure depends on the generation of high-quality, full-length, double-stranded cDNA from mRNA that is representative of the sequence, size, and complexity of the mRNA population. The availability of high-quality commercial kits and engineered reverse transcriptase enzyme simplifies the procedures. The Basic Protocol uses fragmented cDNA to prepare the cDNA library for sequencing and is the preferred method for mapping 5′ and 3′ ends of genes. The presence of tags containing poly(A) or poly(T) sequences allows the precise identification of 3′ ends. An abundance of reads accumulate at the 5′ends of genes and a sharp transition in signal at this end marks the 5′ gene boundary. The RNA-fragmentation method (Alternate Protocol) provides a more uniform distribution of sequence tags throughout the transcript and is useful for quantifying exon levels (Wang et al., 2009).

RNA-Seq has been used successfully in a number of organisms including *Saccharomyces cerevisiae* (Nagalakshmi et al., 2008), *Schizosaccharomyces pombe* (Wilhelm et al., 2008), *Mus musculus* (Cloonan et al., 2008; Mortazavi et al., 2008), *Homo sapiens* (Marioni, et al., 2008), *Arabidopsis thaliana* (Lister et al., 2008), and *Caenorhabditis elegans* (LaDeana et al., 2009). Other organisms include *Drosophila melanogaster* and *Bacillus halodurans* (unpub.observ.).

Critical Parameters and Troubleshooting

The quality of RNA is very important for successful cDNA library preparation; hence, care should be taken when handling RNA samples. RNA is prone to degradation by ribonucleases, so an RNase-free environment is essential, and keeping the RNA constantly on ice helps. See, e.g., Ambion's Technical Bulletin no. 159 (*http://www.ambion.com/techlib/tb/tb_159.html*) for advice on working with RNA. Also take care to ensure that there is no genomic DNA contamination in the RNA preparation. Since ribosomal RNA (rRNA) makes up about 80% of total RNA, it is very difficult to recover poly(A)$^+$ RNA that does not have some rRNA. Typically, one oligo(dT) selection reduces the amount of rRNA to a level acceptable for any molecular biology procedures. For RNA-Seq, however, doubly oligo(dT)–selected poly(A)$^+$ RNA is recommended. Good results have been obtained with Ambion's RiboPure Yeast kit (Nagalakshmi et al., 2008). The total yield of poly(A)$^+$ RNA depends on cell type and their physiological state. Note that other methods of removing rRNA, e.g., Invitrogen's RiboMinus system, can also be used (unpub. observ.).

Availability of high-quality RNA virtually ensures a successful double-stranded cDNA synthesis because of the availability of high-quality commercial kits for cDNA synthesis. Difficulties, if any, normally occur during cDNA or RNA fragmentation, or during the ligation of adapters to the cDNA ends. Care should be taken to optimize the incubation time of the cDNA or RNA fragmentation step; the protocol presented in this unit is optimized for yeast. After incubation, the cDNA or RNA should consist of lengths that are normally distributed between about 150 bp and 350 bp. Agarose gel electrophoresis or a BioAnalyzer will provide the necessary information. Longer incubation of double-stranded cDNAs with DNase I may lead to cDNA fragments that are too small to be suitable for library preparation. Similarly, longer incubation of RNA with RNA fragmentation buffer may lead to RNA fragments that are too small.

A low concentration of the final cDNA library may be due to inefficient reverse transcriptase, deteriorated dNTPs (which are sensitive to freeze-thaw cycles), insufficient digestion of cDNA or fragmentation of RNA, and inefficient T4 DNA polymerase enzyme activity.

It is also very important to optimize adapter oligo concentrations during ligation to the cDNA ends. A molar ratio of adapter oligo to cDNA template in excess of 10:1 results in adapter-adapter dimerization. This may lead to preferential amplification of these dimers in the subsequent PCR step.

Anticipated Results

Typically, using 100 ng to 1 μg of poly(A)$^+$ RNA yields 40 to 350 ng of cDNA library product. If the amount of available poly(A)$^+$ RNA is limiting, the protocol can be scaled down. The quality of the cDNA library can be assessed after the PCR amplification step. During the final agarose gel electrophoresis stage, a normally distributed smear should be visible from 150 to 350 base pairs. After gel purification, the final concentration of the cDNA library should be determined by UV spectroscopy, preferably using NanoDrop technology, which requires only 1 µl of sample. Typically 18 to 24 pM adapter-ligated and amplified cDNA library is processed in one lane on the Illumina Genome Analyzer, yielding 15 to 25 million total reads

Time Considerations

In general, an experienced person can process eight samples for library preparation simultaneously. On day 1, it is possible to finish steps 1 to 13, i.e., first-strand cDNA synthesis, second-strand cDNA synthesis, and gel purification. The samples can then be stored at –20°C. On day 2, steps 14 to 24 can be completed, i.e., fragmentation of cDNA, end repair, and addition of an overhanging A base. On day 3, steps 25 to 34 can be completed, i.e., adapter ligation followed by PCR amplification and gel purification.

Literature Cited

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* 306:2242-2246.

Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., Roberstson, A.J., Perkins, A.C., Bruce, S.J., Lee, C.C., Ranade, S.S., Peckham, H.E., Manning, J.M., McKernan, K.J., and Grimmond, S.M. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5:585-587.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U.S.A.* 103:5320-5325.

Kent, W.J. 2002. BLAT- the BLAST-Like Alignment Tool. *Genome Res.* 12:656-664.

LaDeana, W.H., Reinke, V., Green, P., Hirst, M., Marra, M.A., and Waterston, R.H. 2009. Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Genome Res.* 19:657-666.

Li, R., Li, Y., Kristiansen, K., and Wang, J. 2008. SOAP: Short oligonucleotide alignment program. *Bioinformatics* 24:713-714.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Millar, A.H., and Ecker, J.R. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133:523-536

Mardis, E. 2008. The impact of next generation sequencing technology on genetics. *Trends Genet.* 24:133-141.

Marguerat, S., Wilhelm, T., and Bähler, J. 2008. Next-generation sequencing: Applications beyond genomes. *Biochem. Soc. Trans* 36:1091-1096.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18:1509-1517.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. 2008. Profiling the Hela S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* 45:81-94.

Morozova, O. and Marra, M.A. 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92:255-264.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5:621-628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. 2008. The transcriptional landscape of the yeast whole genome defined by RNA sequencing. *Science* 320:1344-1349

Shendure, J. 2008. The beginning of the end for microarrays? *Nat. Methods* 5:585-587

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., and Yaspo, M.L. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human genome. *Science* 321:956-960.

Tsuchihara, K., Suzuki, Y., Wakaguri, H., Irie, T., Tanimoto, K., Hashimoto, S.I., Matsushima, K., Sugano, J.M., Yamashita, R., Nakai, K., Bentley, D., Esumi, H., and Sugano, S. 2009. Massive transcriptional start site analysis of human genes in hypoxia cells. *Nucleic Acid Res.* 37:2249-2263

Wang, Z., Gerstein, M., and Snyder, M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10:57-63.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bahler, J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453:1239-1243.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., Pham, P., Cheuk, R., Karlin-Newmann, G., Liu, S.X., Lam, B., Sakano, H., Wu, T., Yu, G., Miranda, M., Quach, H.L., Tripp, M., Chang, C.H., Lee, J.M., Toriumi, M., Chan, M.M., Tang, C.C., Onodera, C.S., Deng, J.M., Akiyama, K., Ansari, Y., Arakawa, T., Banh, J., Banno, F., Bowser, L., Brooks, S., Carninci, P., Chao, Q., Choy, N., Enju, A., Goldsmith, A.D., Gurjal, M., Hansen, N.F., Hayashizaki, Y., Johnson-Hopson, C., Hsuan, V.W., Iida, K., Karnes, M., Khan, S., Koesema, E., Ishida, J., Jiang, P.X., Jones, T., Kawai, J., Kamiya, A., Meyers, C., Nakajima, M., Narusaka, M., Seki, M., Sakurai, T., Satou, M., Tamse, R., Vaysberg, M., Wallender, E.K., Wong, C., Yamamura, Y., Yuan, S., Shinozaki, K., Davis, R.W., Theologis, A., and Ecker, J.R. 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302:842-846.