Information Panel

Tips for Preparing mRNA-Seq Libraries from Poly(A)⁺ mRNA for Illumina Transcriptome High-Throughput Sequencing

Brenton R. Graveley

Many investigators who do high-throughput sequencing of mRNA (mRNA-Seq) use kits for library preparation purchased from Illumina. Recognizing that these kits are continually being updated and improved, we provide here some background information, tips, and troubleshooting advice for the kits available at the time of this writing.

OVERVIEW

Preparation of cDNA libraries from poly(A)⁺ mRNA for high-throughput sequencing (HTS) or mRNA-Seq is becoming a routine method for gene expression profiling (Mortazavi et al. 2008; Wang et al. 2009) and studies of alternative splicing patterns (Pan et al. 2008; Sultan et al. 2008; Wang et al. 2008). There are several different approaches to generating cDNA libraries compatible with short-read Illumina HTS (Mortazavi et al. 2008; www.illumina.com). One of the best approaches, in terms of more complete representation of all regions of a given transcript, is to fragment the mRNA using metal-ion-mediated cleavage before random-primed cDNA synthesis (Mortazavi et al. 2008; Wilhelm and Landry 2009; Nagalakshmi et al. 2010; Wilhelm et al. 2010). The procedure discussed here is adapted from the Illumina mRNA-Seq protocol (Mortazavi et al. 2008; www.illumina.com) to generate short cDNA libraries from fragmented poly(A)⁺ mRNA suitable for HTS on an Illumina Genome Analyzer. Bioinformatic analysis of the HTS data is a critical aspect of this procedure once the cDNA libraries generated by this protocol are performed (Pepke et al. 2009).

The RNA shatter method (Mortazavi et al. 2008; Wilhelm and Landry 2009) starts with ~10 µg of total cellular RNA and uses two cycles of oligo(dT) magnetic bead purification of poly(A)⁺ mRNA. The purified mRNA is then fragmented to ~200 nucleotides, and random-primed cDNA fragments are generated. These cDNA pools are made double-stranded by use of RNase H and *Escherichia coli* DNA polymerase. The cDNA ends are then repaired with T4 DNA polymerase, Klenow fragment DNA polymerase, and T4 polynucleotide kinase. Nontemplated A residues are added to the 3' ends of the double-stranded cDNA and then special Illumina adapters are ligated onto the cDNA pools, polymerase chain reaction (PCR)-amplified, and submitted to Illumina sequencing on the Illumina Genome Analyzer. (For information on exact protocols for library preparation and sequencing, see www.illumina.com.)

Note that the size of the fragments generated can be controlled by the length of the incubation. For instance, a 2-min incubation yields fragments of up to 700 nucleotides, a 3-min incubation gives fragments of ~100–500 nucleotides, and a 5-min incubation gives fragments of ~200 nucleotides. The fragmentation size should match the purpose of the experiment.

Adapted from *RNA: A Laboratory Manual*, by Donald C. Rio, Manuel Ares Jr, Gregory J. Hannon, and Timothy W. Nilsen. CSHL Press, Cold Spring Harbor, NY, USA, 2011.

© 2013 Cold Spring Harbor Laboratory Press

Cite this article as *Cold Spring Harb Protoc*; 2013; doi:10.1101/pdb.ip073908

NOTES AND CONSIDERATIONS BEFORE BEGINNING

For RNA-Seq, one must consider the purpose of the experiment when deciding how to make the libraries. For example, single reads and short reads (~36 nucleotides) are typically sufficient for performing gene expression analysis. In contrast, deeper coverage (in terms of total reads) and longer, paired-end reads significantly enhance the ability to detect alternative splicing events. Moreover, greater depth is necessary for discovery of rare RNA forms than for quantitation of well-expressed genes. It is also important to keep in mind that although paired-end reads definitely enhance the analysis of alternative splicing events, they complicate the downstream bioinformatics analysis because the mate pairs need to be tracked. In addition, the accuracy of using the insert size to infer splicing events between the two mate pairs depends entirely on how tightly selected the insert size is. There is also a balance between the cost of generating longer or paired-end sequences and whether there is likelihood that the read can be mapped. For example, with 50-nucleotide reads, fewer will map uniquely to the human genome than to the *Drosophila* genome due to the size of the fly genome.

As a guideline, when using a 76-nucleotide read length, between 10% and 13% of the reads span a splice junction. Furthermore, ~20 million uniquely mapped reads provide sufficiently robust coverage of splice junctions to obtain statistically relevant results for well-expressed genes when using cell lines. Obviously, for complex samples such as tissues or whole *Drosophila* or *Caenorhabditis elegans* animals, more reads are necessary and it is difficult to determine exactly the depth needed.

It is important to note that libraries prepared using paired-end linkers and PCR primers can be sequenced using either single or paired-end reads. As a result, paired-end libraries yield the versatility of performing either type of sequencing.

It is also advisable to use a set of spike-in RNAs (Mortazavi et al. 2008) that cover a broad range of concentrations. This can ensure the quality of the library and, in samples such as cell lines, the possibility of quantitating the number of molecules of each mRNA in a cell. These should be added to constitute between 0.1% and 1% of the library and can be added to the total RNA if they are polyadenylated or to the mRNA after the poly(A)⁺ selection.

Finally, it is advisable to generate independent libraries from biological replicate samples. After sequencing and determining the correlation between the two samples at the gene level, it is common practice to pool the reads from the replicates, provided that they have an $r^2 > 0.9$.

One can routinely begin with only 1 µg of total RNA and get great libraries. Starting with 10 µg is ideal if one has the luxury of generating copious quantities of RNA, but it is certainly possible to start with less.

TIPS AND TROUBLESHOOTING

General

1. Optimization of each step in this procedure is critical for full transcriptome representation. This has been done for genomic DNA sequencing at the Sanger Centre, Cambridge, UK and the reader is referred to this protocol for additional reagent troubleshooting of individual steps in library preparation (Quail et al. 2008, 2009).

Adapter Ligation

2. Residual ethanol in the samples is a problem at this step. If there is residual ethanol, the samples will float out of the gel and cause the loss of all of the DNA. To solve this problem, open the caps of the column tubes before the elution step and let them air-dry for ~5 min. Then, elute with 10 µL of QIAGEN Elution Buffer.

First Gel Purification

3. Add a large amount of loading dye to each sample to prevent them from floating away. Generally, 5 µL of 6× loading dye works well. Dilute the 6× loading dye with clear loading dye to decrease the amount of dye in the gel.

4. When loading the sample, leave a space, load the ladder, leave a space, etc., so that there is a ladder on each side of the sample.
5. For both gel purification steps, it is important to use as long a gel as possible and run the gel until the 100-bp band is at the bottom. The longer the gel, the tighter the insert size region, which is particularly important for paired-end applications. A tight insert size region will provide greater accuracy in inferring splicing of exons located between the mate pair reads, whereas a broad insert size will make this difficult.
6. Be sure to add isopropanol in equal volume to the gel slice to the QIAGEN QG solution (comes with QIAGEN columns) after the gel slice is dissolved (see QIAGEN protocol for small [<500 bp] fragments).

PCR Enrichment

7. Again, be careful with the residual ethanol (see above), which will cause the sample to float out of the well when the gel is loaded below.
8. It is common to skip the column purification step here and instead load the 50 μ L of PCRs directly onto agarose gels. It seems redundant to remove the polymerase and dNTPs before gel-purifying.

Second Gel Purification

9. The size of the PCR fragment is ~ 100 – 125 bp larger than the insert size of the library (depending on exactly which linkers and PCR primers are used). Thus, be sure to include the size of the linkers when determining the size fragment to cut out of the gel. It may be better to use MinElute columns and elute the library in 10 μ L of QIAGEN Elution buffer to give a higher library cDNA concentration, which generally gives better NanoDrop readings.
10. It is also possible to use the QIAGEN MinElute gel extraction kit. Regardless of using Qiaquick or MinElute columns, it is useful to elute with only 10 μ L of QIAGEN Elution Buffer. Again, add isopropanol to the QIAGEN QG solution for a small fragment size. The smaller the eluted volume, the more concentrated the sample, and it will be easier and more accurate to quantify using the NanoDrop spectrophotometer.
11. Typically, we subject the eluted library DNA to NanoDrop spectrophotometry. NanoDrop readings below ~ 10 – 20 ng/ μ L are very unreliable, and thus, libraries may need to be sequenced more than once.

REFERENCES

- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5: 621–628.
- Nagalakshmi U, Waern K, Snyder M. 2010. RNA-Seq: A method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol* Chapter 4: Unit 4.11. 1–13.
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40: 1413–1415.
- Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 6: S22–S32.
- Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5: 1005–1010.
- Quail MA, Swerdlow H, Turner DJ. 2009. Improved protocols for the Illumina Genome Analyzer sequencing system. *Curr Protoc Hum Genet* Chapter 18: Unit 18.2.
- Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhom-chuk D, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321: 956–960.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470–476.
- Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57–63.
- Wilhelm BT, Landry JR. 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 48: 249–257.
- Wilhelm BT, Marguerat S, Goodhead I, Bähler J. 2010. Defining transcribed regions using RNA-seq. *Nat Protoc* 5: 255–266.

Cold Spring Harbor Protocols

Tips for Preparing mRNA-Seq Libraries from Poly(A)⁺ mRNA for Illumina Transcriptome High-Throughput Sequencing

Brenton R. Graveley

Cold Spring Harb Protoc; doi: 10.1101/pdb.ip073908

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

[Analysis of Gene Expression](#) (197 articles)
[Analysis of Gene Expression, general](#) (129 articles)
[cDNA Libraries](#) (26 articles)
[High-Throughput Analysis, general](#) (155 articles)
[Libraries](#) (147 articles)
[mRNA](#) (132 articles)
[RNA](#) (317 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>