Science Education Collection

RNA-Seq

URL: https://www.jove.com/science-education/5548

Abstract

Among different methods to evaluate gene expression, the high-throughput sequencing of RNA, or RNA-seq. is particularly attractive, as it can be performed and analyzed without relying on prior available genomic information. During RNA-seq, RNA isolated from samples of interest is used to generate a DNA library, which is then amplified and sequenced. Ultimately, RNA-seq can determine which genes are expressed, the levels of their expression, and the presence of any previously unknown transcripts.

Here, JoVE presents the basic principles behind RNA-seq. We then discuss the experimental and analytical steps of a general RNA-seq protocol. Finally, we examine how researchers are currently using RNA-seq, for example, to compare gene expression between different biological samples, or to characterize protein-RNA interactions.

Transcript

RNA sequencing, or RNA-seq, is a technique that can provide information on the sequence and quantity of every RNA expressed, known as the “transcriptome,” in a cell population. Unlike other expression profiling methods such as microarrays, which involve probing for known RNA sequences, RNA-seq can profile gene expression from organisms with un-sequenced genomes. Additionally, RNA-seq can accurately measure a larger range of transcript expression levels than microarrays, especially at very low or very high levels.

This video will cover the principles of RNA-seq, a protocol for preparing an RNA-seq library and analyzing the data, and some applications of this technique.

First, let’s review some principles behind RNA-seq. Transcriptome sequencing requires isolating a population of transcripts whose levels are to be measured. Most RNA in cells is ribosomal RNA, or rRNA, the central component of the cell’s protein-production machinery. To facilitate recovery of other types of transcripts, rRNA is typically removed prior to sequencing by hybridizing the sample to complementary oligonucleotides attached to magnetic beads, and using a magnet to separate the rRNA from the rest of the sample.

Alternatively, a specific population of RNA can be selected for sequencing. For example, protein-coding messenger RNAs, or mRNAs, can be captured with “oligo-dT”—short stretches of deoxy-T nucleotides that bind to the sequence of A bases known as a poly-A tail at the end of these transcripts. The contaminating rRNA is then removed. MicroRNAs, which are 22-nucleotide regulatory RNAs, can be selectively isolated for sequencing based on their size. Because RNA is inherently prone to degradation, it is first reverse transcribed to double-stranded DNA.

Oligonucleotide sequences known as adaptors are then ligated onto the DNA fragments. The adaptors contain constant regions that serve as primer-binding sites for subsequent PCR amplification, and these are usually asymmetric so that the “strandedness” of the template is preserved. The adaptors also contain unique sequences, known as “barcodes,” that identify all fragments originating from a single sample. The library is then amplified by PCR.

A sequencing chip, on which there are oligonucleotides complementary to the adaptors, is used to immobilize the library sample, which is diluted such that the DNA molecules anneal onto the chip at low density. The DNA is amplified on the chip via a process called “bridge amplification” to form “clonal clusters.” Short fragments, each 30-150 bases in length, are then synthesized from one or both ends of these DNA templates, generating hundreds of millions of products known as sequencing reads.

The sequencing results are then analyzed for quality and the data are processed. Analysis of the sequences can reveal a wide variety of information, including differences in expression levels of RNAs between samples and previously unknown transcripts or forms of transcripts.

Now that we’ve seen how RNA-seq works, let’s go through a protocol for preparing an RNA-seq library and analyzing the sequence data. RNA is first obtained from the sample of interest, and its quality is checked by electrophoresis, for example by using a microfluidics device called a bioanalyzer. The RNA must be of high quality for accurate sequencing results. To ensure the absence of DNA contamination, RT-PCR for an expressed gene is conducted with or without reverse transcriptase. There should be no products in the absence of reverse transcriptase.

To select poly-A RNA, the samples are bound to oligo-dT probes attached to magnetic beads. The selected RNA is fragmented to 200-nucleotide pieces at high temperature in the presence of magnesium ions, reducing length-dependent biases in subsequent reactions and analyses. The fragments are then converted to double-stranded DNA, and adaptors are ligated. The library is amplified by PCR, and its quality is checked on a bioanalyzer and by performing qPCR. The bioanalyzer results should reveal a peak of products at the size expected based on the average fragment size and length of the adaptors.

Libraries from different samples, containing different barcoded adaptors, can be mixed together, along with a sample of reference DNA added at low concentration as a quality control for subsequent steps of the process, such as clonal cluster generation and the sequencing reactions. The mixture is added to a sequencing chip and loaded into the machine.

During the sequencing reaction the density of DNA clusters is monitored: it must not be too high, which can lead to cross-contamination, or too low, which can lead to insufficient data. The quality of the sequencing is given by the Q score, which indicates the likelihood of an incorrect base being identified. The Q scores for most bases should be greater than 30, which corresponds to a chance of less than 1 in 1000 for an incorrect read. Recovery of the reference DNA sequences at the expected rate indicates that all library sequences are evenly represented.

Reads generated by sequencing are then overlapped with each other to deduce the RNA that was sequenced. For organisms with genome information available, reads can be aligned to the reference genome. The number of reads per transcript is counted to measure the abundance of each RNA.

After seeing how RNA-seq works, let’s look at some ways it’s being used.

Transcriptome sequencing can identify genes that are differentially expressed under different conditions. For example, in this experiment, transcriptomes of mosquito larvae produced under different growth conditions were compared. Even though this particular species of disease-carrying mosquito does not have a sequenced genome, researchers were able to compare the obtained transcriptome information to other sequenced species, and identify genes with increased or decreased expression levels.

RNA-seq can also be used in “massively parallel reporter assays” to study gene regulatory mechanisms. This is done by transfecting mammalian cells with a library of thousands of plasmids, each containing a mutated variant of a gene regulatory site “driving” the transcription of a reporter sequence that is coupled to unique tags. Following RNA isolation and high-throughput sequencing, the levels of each tag are assessed to evaluate reporter expression from each construct, which gives insight into the functional importance of the nucleotides mutated in each regulatory site variant.

Finally, RNA sequencing can be adapted to study RNA-protein interactions, particularly to identify transcripts that a protein of interest binds to. The protein is immunoprecipitated with antibodies and the bound RNAs are defined by sequencing. If the RNA-protein complexes are crosslinked at the beginning, sequencing analysis can map the site of the crosslink and identify the protein-binding site on the RNA down to the nucleotide level.

You’ve just watched JoVE’s video on RNA-seq. In this video, we’ve seen how RNA samples are converted into libraries, sequenced, and the resulting data analyzed, as well as the types of information that sequencing analysis can provide. Thanks to its sensitivity, potential to be used in any organism, and the lowered cost of sequencing, RNA-seq is increasingly being used in multiple areas of genetics research, and will provide insight into many questions surrounding cell function and development. Thanks for watching!