Notes on RNA-seq

Logan Thrasher Collins

Experimental process of RNA-seq

- To perform RNA-seq, a sample's RNA fraction is isolated and converted to a cDNA library using poly-T primers (that bind to the poly-A tails of mRNAs) to initiate reverse transcriptase activity on the mRNAs within the sample. Finally, the cDNAs are amplified using PCR and sequenced via next-generation sequencing techniques.
- If the experiment requires profiling of other types of RNA besides mRNA, different primers can be used. The primers often are equipped with sequence extensions called adaptors that tag the cDNAs for compatibility with next-generation sequencing. Other primer variations are common as well.
- When using RNA-seq to profile relatively long RNA molecules (i.e. mRNAs), the RNAs require fragmentation in order to decrease their sizes and so make them compatible with next-generation sequencing platforms. This step is often not performed when profiling small RNAs such as miRNAs.
- To provide accurate results, RNA-seq necessitates performing many reads of the sample and averaging them. The number of reads is referred to as sequencing depth. Typically, RNA-seq experiments require tens of millions of reads in order to properly sequence all or nearly all of the expressed mRNAs in the sample. However, this can vary depending on the application.
- Using qPCR, RNA-seq can provide a quantitative measure of gene expression levels. Such experiments require controls called spike-in RNAs which are RNAs of known sequence and quantity that provide a point of comparison.

Experimental process of scRNA-seq

- For many applications, it is advantageous to sequence the transcriptomes of individual cells rather than bulk tissues. In this way, the resulting data represents the actual gene expression for a given cell rather than the average gene expression for the tissue sample.
- Single-cell RNA-seq (scRNA-seq) is performed by isolating individual cells using techniques like fluorescence activated cell sorting, micromanipulation, laser-capture microdissection, or microfluidics.
- Drop-seq and inDrop are especially powerful tools for isolating the RNA contents of individual cells. They enable efficient sequencing of many thousands of individual transcriptomes.
- Drop-seq encapsulates single cells into lipid droplets along with a microparticle carrying barcoded DNA primers. The cells are then lysed within the droplets, releasing their mRNAs. The mRNAs are captured by a poly-dT sequence at the end of the microparticle's barcoded primers. Next, the captured transcriptomes undergo PCR amplification, retaining the barcode associated with their cell of origin. When sequenced using next-generation sequencing, the transcriptomes can be traced back to their cells of origin via the barcodes.

- The inDrop technology encapsulates single cells into hydrogel droplets containing barcoded primers (each droplet possesses a unique barcode) with poly-dT sequences for capturing mRNAs, lyses the cells, and carries out reverse transcription on the captured transcriptomes of the lysed cells within the droplets. Next, the droplets are broken up and the cDNAs are sequenced using next-generation sequencing. Since each hydrogel droplet contained a unique barcode, the transcriptomes can be traced back to their cells of origin.
- For scRNA-seq, the necessary sequencing depth is lower than that of bulk RNA-seq experiments. Instead of tens of millions of reads, scRNA-seq can often achieve full coverage of a transcriptome using less than one million reads. Once again, this can vary depending on the application.

Analyzing RNA-seq and scRNA-seq data

- RNA-seq generates large amounts of data that require rigorous computational analysis. Many variations of the algorithms utilized for RNA-seq exist, but some general principles are presented here.
- Raw reads are examined for GC bias (GC-rich and GC-poor sequences are often underrepresented), duplicate reads, and other artifacts. Various algorithms correct for these sequencing errors.
- The first step in RNA-seq data analysis is read mapping (after preprocessing to correct for sequencing errors). Reads are matched to a reference genome or transcriptome by aligning the sequences.
- Next, gene expression levels are measured via counts computation. Most commonly, this involves counting the total number of reads overlapping the exons of each gene. However, reads sometimes map outside of known exons due to poor understanding of a given gene's structure. As such, an alternative computational strategy also includes reads from introns so as to ensure coverage of the entire gene. The latter technique also allows alternative splicing isoforms to be counted separately.
- Since different RNA-seq experiments may generate different numbers of reads for the same RNA expression levels, count data must also undergo normalization. This involves scaling by an estimated sequencing depth for the given experiment. But highly expressed genes often introduce a bias in the scaling factor since these genes "use up" the available reads and leave less for the rest of the genes. Algorithms that compensate for this bias are commonly used.
- Differential gene expression analysis is performed on normalized RNA-seq data. This involves using various statistical methods to find significant differences between the RNA expression of cells from distinct samples or under distinct conditions.

References

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Finotello, F., & Di Camillo, B. (2014). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, *14*(2), 130–142. https://doi.org/10.1093/bfgp/elu035

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, *9*(1), 75. https://doi.org/10.1186/s13073-017-0467-4

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., … Kirschner, M. W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, *161*(5), 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Shapiro, E., Biezuner, T., & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, *14*, 618. Retrieved from https://doi.org/10.1038/nrg3542

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*, 57. Retrieved from https://doi.org/10.1038/nrg2484