Primer on High-Throughput RNA Sequencing

RNA-Seq is used to measure gene expression of many genes simultaneously in a cell or population of cells. We are going to learn how to take raw RNA-Seq data and to transform it into gene level abundance measurements. This is called "low level analysis" of RNA-Seq or "pre-processing". Low level analysis is very important as it underpins everything we do after. This document gives some of the historical and technical background that led to RNA-Seq.

Genes in the DNA are transcribed into RNA, and RNA in turn serves a multitude of functions in the cell. Some RNAs are regulatory, some are structural, some carry information to build proteins, and new types and functions of RNA are being discovered on a regular basis. Every cell has the sequence of every RNA encoded in the DNA. However in any given cell, at any given time, there is only a subset of them floating around the nucleus and cytoplasm. The genes that are copied into RNA in a cell are called the "expressed" genes in that cell. To some extent which genes are expressed differentiates one type of cell from another. And for any given gene it is not just expressed or not, some genes are represented by thousands of RNA copies (high expressed), some genes are represented by just a few copies (low expressed) and some genes are represented by no copies (unexpressed). The situation is dynamic with many RNAs being transcribed at any given time and many others being degraded. So two cells might have the exact same genes expressed, however the levels at which the individual genes are expressed might be very different.

There is enormous information in RNA measurements, relating to the state of the cell's health and its normal function. Therefore since the discovery of RNA, biologists have desired to measure the abundances of the different RNA molecules at a given time in a cell or population of cells. And this has long been possible, at least for one or very few RNAs at one time. But there are typically tens of thousands of expressed genes in any given cell, and the ability to take a snapshot of a large part of the expressed transcriptome simultaneously has only recently been possible. Measuring RNA one gene at a time was has been possible since the 1980s using PCR technology. But PCR is slow and laborious, so it can only be done for a relatively few genes at a time. In the late 1990's microarrays were introduced as the first highly parallel method. There are many variations on microarrays, but essentially one creates spots on a glass slide each with many copies of a short probe - one spot for each gene of interest. One then fluorescently labels the RNA to be measured and hybridizes it to the slide. The higher the gene is expressed the brighter the spot glows under a laser. This is an oversimplification, but it is enough to give you an idea of the process.

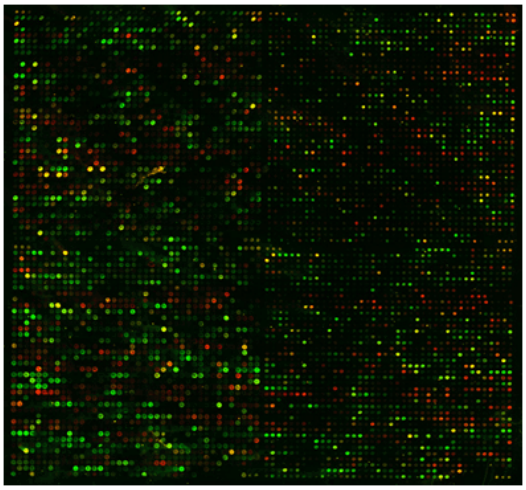

Figure 1. A picture of a microarray assay. Each spot corresponds to a gene. The brighter the spot the more highly expressed

Microarrays allow us to measure simultaneously the

expression level of tens of thousands of genes and as such microarrays revolutionized modern biology. However, they still have a great many limitations. For example, they can only measure genes that we know exist beforehand, because one needs to design a specific probe for each gene – an expensive and complex process. Thus extending to a new organism requires considerable upfront analysis and effort. Furthermore microarray probes are short, typically 25 to 60 bases. We may put a few probes for one gene, but most of the length of the transcript is not covered. Therefore, we do not get clean enough information to infer very much about things like "differential splicing" for example. Differential splicing is when one gene produces different RNA transcripts by preferentially leaving in or out certain exons[1]. Some genes have many exons and produce many different splice forms, and it is the differential behavior of the splice forms that are important to the biology and not the overall expression level of the gene as a whole. When we speak of "gene level" expression we mean the average expression averaged over all of its different splice forms.

Figure 2. Illumina Sequencing Machine

Microarrays have other shortcomings, such as a high level of background signal and a low signal saturation point. Both of these issues decrease the spectrum of detection. It is therefore impossible using microarrays to determine that a gene is not expressed, for example, if its signal is below the level of the background signal. One can only say *if* it is expressed then it is expressed at or below the level of the background signal. Also, different probes have different hybridization affinities. A probe with a higher hybridization affinity than another will have a brighter spot even if both genes are expressed at exactly the same level. Therefore, one cannot easily compare the expression level of two different genes in the same sample, simply by directly their signal intensities.

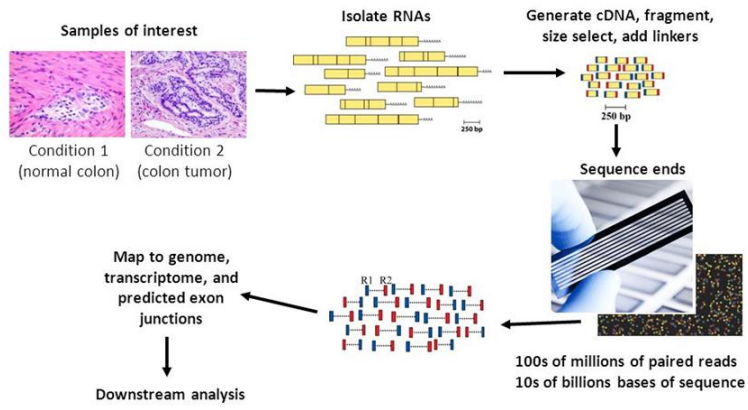

Figure 3. RNA-Seq flowchart

High-throughput sequencing appeared around 2009 and has largely replaced microarrays as the platform of choice for RNA abundance because of a number of advantages. However, with this greater power come a large number of technical issues that must be overcome in order to get useful information out of the data. The process of determining the best

[1] If you are not familiar with what an "exon" and an "intron" is you should read the Wikipedia page on RNA splicing: https://en.wikipedia.org/wiki/RNA_splicing

analysis methods is an ongoing and heated debate that will likely not settle down for a long time. There are hundreds of competing algorithms and there is more hype than agreement in the field. Performances of the available methods vary tremendously and one can waste a lot of time and end up with vastly inferior results by using the inferior methods. What is worse is that some of the most popular and most highly touted algorithms actually have some of the worst performance.

There have been a number of competing sequencing machines developed by various companies. But at this point Illumina has achieved near total market domination. As such, we will focus entirely on the Illumina platforms in what follows. But before we delve into the analysis issues, we must discuss a couple things about the nature of the data itself.

Ideally we would like to get from RNA-Seq the full length sequence of the RNA molecules in the cell. But in reality, the sequencing machines cannot generate contiguous sequences as long as most full length genes. Typically we can only get approximately 100-200 bases of contiguous sequence, while genes are typically thousands or tens of thousands of bases long. Therefore, to deal with this technical issue, the RNA is first fragmented into small pieces, usually between 100 and 500 bases. Then the 100 bases at one, or both, ends of the fragment are sequenced. The resulting data then consists of millions of short ends of these short fragments. The computational challenge then becomes determining which genes each fragment came from. See Figures 3 and 4. In Figure 4, the red bits show the parts of the fragments for which we get actual data.

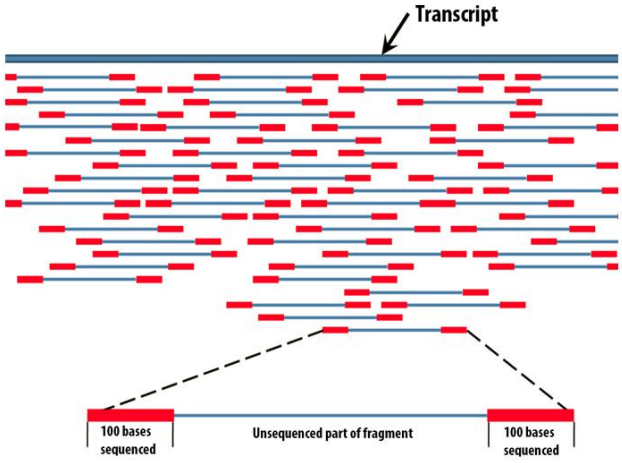

Figure 4. RNA-Seq fragments and paired-end reads.

Putting all of this fragmented information back together into information about the full length RNAs is the computational challenge. There are two standard approaches to do this. The most common is to align the short reads to a reference genome[2]. The alternative approach is to align the reads to each other to find overlapping reads and to assemble them into transcripts *de novo*. The second approach considered much harder and so if there is a reference genome available then it should be used. For human and model organisms, and many others, a high quality reference genome *is* available. Therefore, we will focus on the first kind of analysis.

[2] By "Reference Genome" we mean the full DNA sequence of the organism.

The pipeline for a genome guided analysis breaks down into the six steps as shown in Figure 5. We will be particularly interested in Alignment, Normalization and Quantification in our in-class work. We will just touch on Statistical Analysis which is a huge subject on its own. And we will not have time to talk at all about Data Management and Deployment, but that is as important as anything else. An analysis is only as good as it can be utilized by the biologists for whom the analysis is being done in the first place.

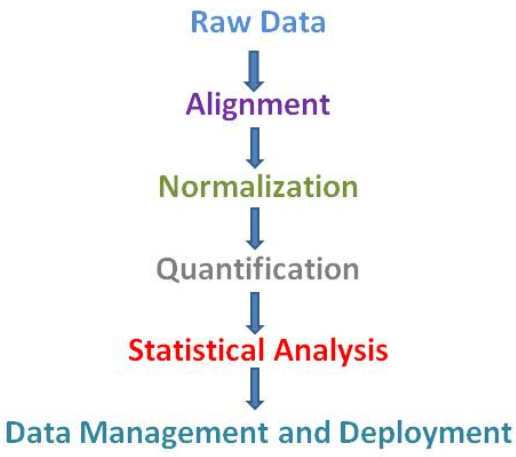

Figure 5. The RNA-Seq analysis pipeline

One of the most frustrating aspects of performing an RNA-Seq analysis is the sheer number of available applications that are available for each step of the pipeline. There are also many review articles giving conflicting guidance. One should not follow any review article's advice blindly. One should instead look for articles that publish unbiased benchmarking efforts. For example these papers:

- http://www.ncbi.nlm.nih.gov/pubmed/26338770
- http://www.ncbi.nlm.nih.gov/pubmed/24185836
- http://www.ncbi.nlm.nih.gov/pubmed/21775302

For those students interested in looking much deeper into this subject, we recommend this page called the "RNA-Seqlopedia" as a good starting point:

- http://rnaseq.uoregon.edu/

In class we will look at the standard push-button approach to RNA-Seq and we will also learn how to perform a much more sophisticated analysis. We will compare the effectiveness of the standard and the sophisticated approach.

As a final comment we should point out that if you ask ten RNA-Seq researchers what is the best method of analysis, you are likely to get ten different answers. Ultimately one has to make some judgement calls and even some arbitrary decisions in order to proceed. What we will show you in this class is just one way of doing things. It is the way we currently believe is best, but it will certainly be a long time before there is any kind of universal agreement in the field. Until then, proceed with caution.