11.1 Introduction to RNA-sequencing

11.1.1 Overview of the RNA-sequencing analysis pipeline

This week, we will be going over the steps involved in taking raw RNA-sequencing reads and converting them into a format that can be loaded into R for further analysis. Looking at the quantities of RNA in a given tissue can tell us a lot about its phenotype and can be used to compare samples from different conditions. In this class, we will be comparing RNA from a tumour tissue to that of a healthy tissue. This sort of analysis could provide clues as to which genes may be involved in cancer development or tumour suppression.

We will specifically be looking at RNA-seq data from three individuals who have human esophageal squamous cell carcinoma (ESCC). Each individual has had their tumour sequenced, as well as the healthy tissue adjacent to the tumour. entire dataset is six samples: three normal and three tumour. These are real data, so the sample IDs that you are using, like 16N and 16T, are the sample IDs that were used by the authors. The numbers correspond to the individuals, and the letters indicate whether the data are normal or tumour tissue.

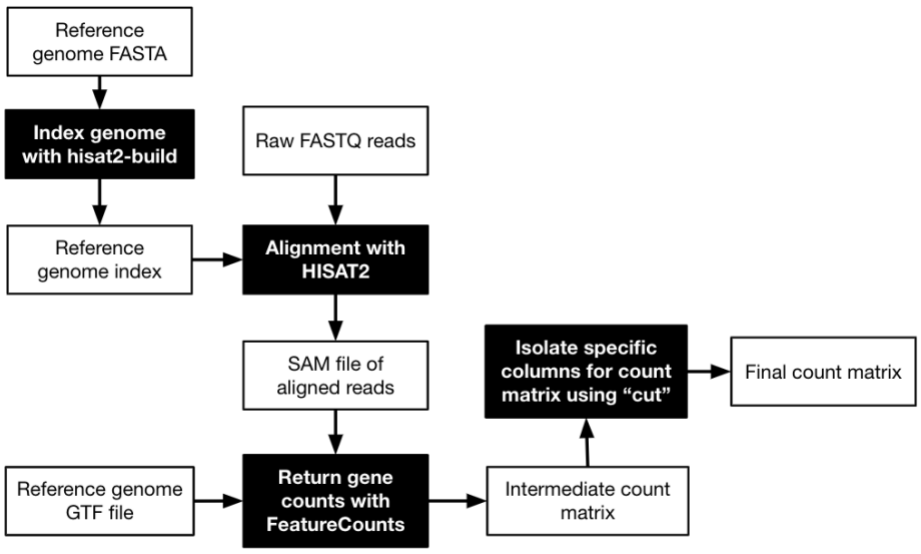
	INDIVIDUAL 1	**INDIVIDUAL 2**	**INDIVIDUAL 3**
NORMAL	16N	18N	19N
TUMOUR	16T	18T	19T

(Citation: Wei G, Luo H, Sun Y, Li J et al. Transcriptome profiling of esophageal squamous cell carcinoma reveals a long noncoding RNA acting as a tumor suppressor. Oncotarget 2015 Jul 10;6(19):17065-80. PMID: 26158411)

We'll start with an overview of the RNA-sequencing analysis pipeline before going over some of the individual steps in more detail. The steps are broadly as follows:

1. mRNA is isolated from samples and reverse transcribed to create the more stable complementary DNA (cDNA)
2. cDNA is then sequenced. This sequencing data is stored in FASTQ files
3. The cDNA "reads" are aligned to the reference genome to determine which gene likely encoded that transcript
4. The total number of alignments are tallied and converted into a "count matrix" that counts the total number of reads which align to a gene in that sample

A more detailed flowchart of the various steps is as follows. Not only does it include the major steps, but it also includes the program names and file types that we will be exploring this week. White boxes indicate files and black boxes indicate steps done with specific tools. Therefore, each white box that connects to a black box is either an input or output file for a certain tool.



All of the above steps are performed on the command line. After they're complete, the count matrices for different samples are combined and read into R to be analyzed. We will cover this analysis in R next week.

11.1.2 Raw RNA-sequencing reads

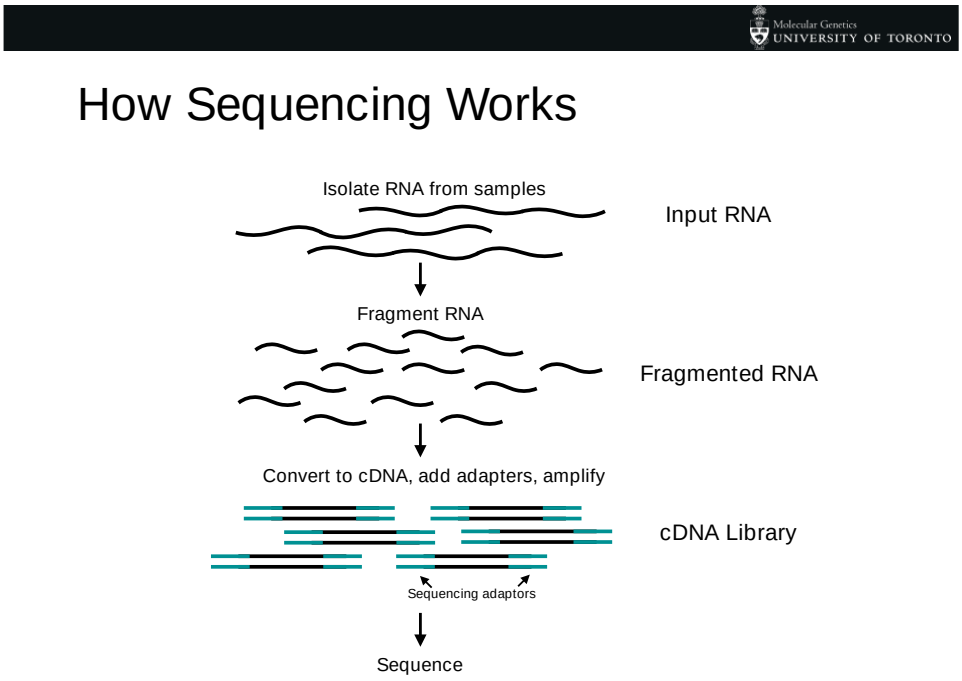
Now we're going to look closer at the first step: receiving the RNA-seq reads from the sequencing facility. It is always important to understand how the data that we use have been generated in an experiment. RNA-seq data can be generated by a variety of protocols, and this may change based on:

(1) What files you have

(2) What is in the files

(3) Biases in the data

The first step in the experiment is to fragment the RNA into pieces that are converted into cDNA by reverse transcriptase, and in the process, adaptors are ligated onto the ends of the cDNA molecules. Using primers that match the adaptor sequences, the cDNA is then amplified with PCR. Sequencing adaptors then bind to the adaptor sequences on the cDNA and the reads are sequenced. This can either happen from only one end of the cDNA molecule (single-end sequencing) or from both ends (paired-end sequencing).



(1) What files you have: Single-end sequencing and paired-end sequencing can create different files. Single-end sequencing produces one or multiple files of sequences in any order. Multiple files of sequencing data are only required if a very large number of reads are produced. On the other hand, paired-end sequencing produces two sets of files: one set of files contains one end of a single read, and the other set of files contains the other end of the read. The different sets of files have the same number of reads, are ordered analogously and are labeled as "1" and "2" respectively.

(2) What is in the files: Read lengths (i.e. the number of base pairs in a sequenced transcript) can change depending on the sequencing technology, and the lengths of the adaptors included in the file. As sequencing technology improves, read lengths tend to get longer, which helps them align to the genome more accurately. Reads are also often recorded with a quality score for each pair produced by the sequencing technology. If reads are recorded with quality scores, the files are FASTQ instead of FASTA files, which we will touch on more later.

(3) Biases in the data: Different RNA-seq protocols change how the reads are amplified and sequenced, which can impact which transcripts are more likely to be detected. For instance, different chemical properties of transcripts can make it so that some reads are sequenced while others are not, even if they are at similar abundance levels. Changes in sequencing depth can also introduce artificial zeros into the dataset (i.e. zeros that are due to low sequencing depth not picking up a transcript), even though the gene is actually transcribed in small amounts. Right now, it is not too important to know how to deal with these biases, but it is important to remember that changes in gene expression are not always due to true biological signal.

11.1.3 FASTQ files

Raw sequencing reads are stored in FASTQ files, which are an extension of the FASTA file. You have already learned about FASTA files, which contain a header followed by a nucleotide sequence.

In FASTQ files, the header begins with "@" and is followed by a unique descriptor for a particular read. This is almost the same as with FASTA files, but in FASTA files the header beings with ">". The headers are also used with paired-end RNA-seq data to help align the reads and make sure they are in the right order. Each read needs to have a "mate" and the header can help decipher any "missing mate" issues that may come up.

The length of your RNA sequencing reads is dependent on the technology used for the sequencing. Because the reads are cDNA, thymine is recorded instead of uracil. Adaptors are sometimes included in the sequences, in which case they would need to be removed ('adaptor trimming') before the reads are aligned to the genome, otherwise they may not align properly (although this is beyond the scope of this course).

FASTQ files are very similar to FASTA files, except each read has information that spans 4 lines instead of 2. The first line is still the header that begins with "@", instead of the ">" of the FASTA file. Line 2 is the actual sequence. Line 3 begins with "+" and is often a repeat of the header. Line 4 is the mapping quality score, with each character representing its corresponding base in line 2 (therefore line 2 is the same length as line 4). Symbols (e.g. !, ", #, $) are good scores, and letters (e.g. A, B, C) are worse scores. If you are interested in learning more about the specific symbols, resources such as the following exist: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm). These scores are important because it allows you to trim or filter your transcripts based on quality. For example, if the last 10 base pairs of most of your RNA-seq reads have a poor quality score, you may want to trim these bases off of all of your reads. Tools such as FastQC allow you to calculate such metrics across all of your reads, to get a good idea of the quality of the data you are working with.

Let's take a look at one of the FASTQ files from one of our tumour samples. The first 100 lines of sample 16T have been isolated and downloaded into your local directory. You should see it saved as 16T_100_lines.sra.fastq. Take a look at this file using `less`.

```
less 16T_100_lines.sra.fastq
```

Also note that there are Ts instead of Us even though these are RNA sequencing data. Again, this is because you're not looking directly at the sequence of the unstable mRNA, you're looking at the cDNA. If you're interested in learning more about this process, you can read a brief paragraph about it here: https://en.wikipedia.org/wiki/RNA-Seq#Complementary_DNA_sequencing_(cDNA-Seq).

11.1.4 Genome FASTA files and GTF files

The raw RNA-seq reads from the FASTQ files are aligned to a reference genome to determine the genes from which they were likely transcribed. This can therefore tell you which genes are being expressed. A genome file is a very large file containing the entire known sequence of a given genome Since a genome separates naturally into chromosomes, a perfectly sequenced genome would be bases segmented into the number of chromosomes that belong to your species of interest. However, because our genome sequences can be incomplete, the genome often exists in smaller segments or "contigs" that make up the reference genome FASTA file. These genome FASTA files often end in ".fna" which stands for "FASTA nucleic acid".

A GTF file keeps track of where things are in the FASTA file; it is essentially lines of genomic features (e.g. gene, mRNA, ncRNA, lncRNA, stop codon, start codon, exon, 3 prime UTR etc.), and their coordinates that map where these features are located in the FASTA. Two important components of the

GTF file are the contig names and their corresponding base pair number. Many of the features in this file are optional.

Here is an example line from the GTF file that we're using for the alignment. This line represents the gene *OR4F16*:

```
chr1    BestRefSeq%2CGnomon     gene    683910  720115  .       -
.       gene_id "OR4F16"; db_xref "GeneID:81399"; db_xref
"HGNC:HGNC:15079"; description "olfactory receptor family 4 subfamily
F member 16"; gbkey "Gene"; gene "OR4F16"; gene_biotype
"protein_coding"; gene_synonym "OR1-1"; gene_synonym "OR7-21";
```

The line above begins similarly to a TSV file, with specific items separated by a tab. The position of each item dictates what it represents. If the information that would fill a certain column is unknown, a period is used as a placeholder. After the first 8 mandatory tab-separated items, further pieces of information describing the feature are separated by semi-colons. In this example, the line can be read as follows:

- chr1 - the gene is located on chr1 (in the genome FASTA file, it is labeled as ">chr1")
- The second column is the "source" and would be filled with the software that found the gene
- gene - the type of feature (e.g. gene, mRNA, exon, etc.)
- 683910 - the starting base of the gene on the chr1 contig
- 720115 - the ending base of the gene on the chr1 contig
- The next column is the "score", an ill-defined optional feature that is often used to measure the confidence that the feature is correct. Here it is left blank.
- '–' - the direction of the feature on the strand, most often positive (+) or negative (–)
- The next column represents the "phase" which is similar but not the same as "frame". The value is usually either 0, 1, or 2, indicating how much the triplet shifts between a given coding region of a gene and the next coding region. This feature doesn't have phase information because it's describing an entire gene, not a single coding region.
- gene_id "OR4F16" - this is where the tab-separated columns end, and values are now separated by semi-colons. "gene_id" is a unique string used to describe the gene.
- The rest of the items describe the gene, often with optional categories. We will not go into any of them in detail.

As mentioned above, the features present in the file can also include mRNA, exons, CDS, and more. mRNA features are often more numerous than the gene features because a single gene may form multiple transcripts. The more you use GTF files, the more intuitive they will become to read. Knowing how to read these files can be very useful if you are looking for more information about a specific gene or chromosome, for instance if you are trying to figure out how long a gene is or where it is located on a contig.