Lecture 3: RNA-seq quantification

Sources: stat115 video and statquest: a gentle introduction to RNA-seq

- ▼ What is an RNA-seq library or what is library preparation for RNA-seq?
    - ▼ During sample preparation for RNA-seq, the RNA from biological samples is extracted and enriched, then fragmented and converted into stable cDNA molecules. The collection of cDNA molecules that is formed and can then be sequenced is called a library.
- ▼ How is RNA-seq different from DNA-seq?
    - ▼ RNA-seq is different in experimental preparation and data processing steps.
    - ▼ In experimental preparation, we have to isolate the RNA molecules that we are interested in sequencing from among all the other molecules in the cell, including other RNA types than the ones we are interested in.
    - ▼ In data processing, the alignment algorithm used has to be splice-aware and there has to be counting of how many fragments map to each gene and normalization of the counts to correct for differences in sample preparation and gene lengths.
- ▼ What is the difference between single and paired end RNA-seq experiment?
    - ▼ In a single-end RNA-seq experiment, any fragment is only sequenced from one end, i.e. an adapter is added to only one end of the fragment and therefore there is only one read per fragment.
    - ▼ In a paired-end RNA-seq experiment, adaptors are added to both ends of the fragment and it can be read from both ends. So there may be a maximum of two reads per fragment in the readout.
- ▼ What is sequencing depth in an RNA-seq experiment?
    - ▼ My understanding is that sequencing depth refers to the total number of fragments being sequenced from a given sample. The larger this number, the

greater the probability of capturing even lowly expressed transcripts. This number can range from 1M to 100M or more depending on the number of samples being run in a single lane of a flow cell.

- ▼ What are some applications of RNA-sequencing?
    - ▼ Comparing gene expression in different conditions like different developmental stages, different tissues, normal vs disease, drug treatment, gene perturbation etc.
    - ▼ Find novel genes or transcripts
    - ▼ Model alternative splicing that may not have been observed before.
    - ▼ Find gene mutations or gene fusions
    - ▼
- ▼ How many fragments can be sequenced simultaneously in a high throughput sequencing machine?
    - ▼ 100s of millions, upto a billion fragments can be sequenced simultaneously in a flow cell depending on the machine.

- ▼ Which end in an RNA molecule is more stable and which end degrades faster?
    - ▼ The degradation of RNA molecule usually happens from the 5′ end and the 3′ end is more stable.
- ▼ How can you check the quality of an RNA sample for degradation before running it in a sequencing machine?
    - ▼ We can run the sample to check the % of RNA molecules that are more than 200bp long. The metric is called DV200. It is recommended that DV200 should be greater than 30%.
- ▼ What are different types of RNA-seq experiments in terms of RNA enrichment steps?
    - ▼ Ribo-mibus - remove tRNA and rRNA and you still have RNA molecules that include introns as well.

▼ PolyA - uses PolyA to select for mature mRNA after splicing. It enriches for exons in addition to removing tRNA and rRNA and pre-mature RNA

▼ Strand-specific - it preserves directionality of RNA and is useful for novel long non-coding RNAs - lncRNA

▼ What does the quality score of a particular base in a particular fragment mean? Give some instances where the quality score of a base in a particular fragment can be low.

- The quality score indicates how confident the machine is that it correctly captured a base.
- If the probe is not bright enough to be captured by the machine.
- If there are lots of probes of same color in the same region, called low diversity, can blur the colors together, making it hard to identify individual sequences. This is especially the problem in the first few nucleotides because this is when machine determines where DNA fragments are located on the grid.

▼ What type of questions do you need to ask before you hand over your samples for RNA-sequencing to a core or another lab?

▼ Single end or Paired end

▼ Sequencing depth - ask how many fragments will we get per sample. Because you will get less per sample if many samples are run in the same lane.

▼ Sequencing length - What is the read length of the fragments being sequenced - 50bp or 150bp.

▼ What should be the minimum sequencing depth per sample for a differential expression analysis? What should it be for studying alternate splicing or transcript assembly?

▼ For differential expression analysis, the sequencing depth should be 20-50 Million and for transcript assembly or alternative splicing, it should be 100 million reads per sample.

▼ What is the difference between technical and biological replicates in an RNA-seq experiment?

    - ▼ Technical replicates - same RNA, library prep and sequencing separately but these are not done anymore.
    - ▼ Biological replicates come from different biological entities such as cell lines grown in different dishes, different animals that are similar in age and lifestyle and genetically identical or different human individuals.
- ▼ How many replicates should be included in a typical RNA-seq experiment?
    - ▼ At a minimum 3 biological replicates per condition, more for humans.
- ▼ Give some examples of splice aware aligners that can be used for RNA-seq and what algorithm do they use?
    - ▼ Tophat (Burrows-Wheeler (BW) algorithm)
    - ▼ HISAT and HISAT2 (BW algorithm)
    - ▼ STAR (Suffix Array)
- ▼ Which files do you need in addition to .fastq files for RNA-seq alignment?
    - ▼ Genome index file
    - ▼ Transcript annotation file (optional)
- ▼ What type of information does BAM file obtained after alignment of RNA-seq data?
    - ▼ Read ID, information whether the sequence is paired end (odd numbers)
    - ▼ Chromosome information - chromosome number and beginning location of read
    - ▼ cigar string: tells how well the sequence is a match. Example 27M1099N48M - 27 bp matched, spans a 1099bp intron, then maps 48bp on next exon
    - ▼ Same information for the other read if it’s a paired end.
- ▼ What is an algorithm that you can use to check quality of RNA-seq data?
    - ▼ RSeQC
- ▼ What type of quality metrics are used for RNA-seq data?

- ▼ Overall mappability of reads: You want atleast 50% reads mapped to the genome. The higher the mappability, the better it is.
- ▼ Quality of read relative to location/position of read. Typically quality is better in the beginning and drops later on.
- ▼ Nucleotide composition/frequency as a function of of location or position of read: Nucleotide frequency should be flat as a function of position of read.
- ▼ Look at insert size for paired-end read.? What is insert size?
- ▼ Look at Read distributions over known exons: How many reads map to coding exons.
- ▼ If a read spans a junction, look at how many are over known splice junctions. YOu want majority splice junctions to be known
- ▼ Transcript integrity number (TIN) on each transcript and then calculate median TIN score across all transcripts in a sample. You want the medTIN to be >50% for a sample.
- ▼ Gene body coverage: How many reads cover each location of the total gene length.

▼ What is the difference between a read, fragment and a transcript in RNA-Seq?

- ▼ A read is a spot on the flow cell whose sequence is captured in a sequencing machine.
- ▼ A fragment is a fragment of the cDNA library that is read in a sequencing machine. In a single-end sequencing experiment, one read will map to one fragment, but in paired-end sequencing experiment, a maximum of two reads can come from one fragment.
- ▼ A transcript is the original mRNA which was fragmented to create the cDNA library before sequencing. In RNA-seq data processing, our goal is to quantify the transcript abundance by mapping how many fragments/reads correspond to the original transcript.

▼ What is RPKM?

- ▼ Reads per Kilobase Million
- ▼ Step1: normalize data for sequencing depth - million part

  - ▼ Step2: normalize data for length of gene - kilobase part

- ▼ What is FPKM?
  - ▼ Fragments per kilobase million - same normalization as with RPKM

- ▼ How is FPKM different from RPKM?
  - ▼ RPKM is for single end RNA-seq and FPKM is for paired end RNA-seq
  - ▼ With single-end sequencing, there is always only one read per fragment, from 5′ end or 3′end.
  - ▼ With paired end sequencing, both ends can map, giving us two reads per fragment or one read for a fragment if one of the reads is low quality and will need to be discarded. So FPKM keeps track of fragments, so that one fragment with two reads mapped to it is not counted twice.

- ▼ What is TPM?
  - ▼ Transcripts per million
  - ▼ Step1: normalize for gene length - kilobase part
  - ▼ Step2: normalize for sequencing depth - million part

- ▼ How is TPM different from RPKM and FPKM? Why is TPM preferred in some cases?
  - ▼ TPM is like FPKM and RPKM, except order of operations is switched.
  - ▼ Sum of total normalized reads in each column is the same in TPM so that the value of a gene in a sample is proportional to the fraction of that gene in that sample. So if a gene has higher value in sample 1 versus sample 2, we know that of all the reads that mapped in sample 1, a larger proportion of them mapped to our gene compared to sample 2.
  - ▼ With TPM, every sample gets the same sized pie.

- ▼ What is CPM?
  - ▼ Counts per million - raw read counts in a sample normalized by total reads in the sample

- ▼ What is a tool called RSEM used for in RNA-seq data processing? What inputs does it take and what outputs does it produce?

    - RSEM is a tool that can be used to quantify transcript abundances from RNA-seq data.
    - It can take .fastq files and reference transcript annotation file as input or BAM files and generates read counts, TPM and FPKM values calculated based on effective transcript length (not the full gene length)
- What is a gene isoform?
    - Gene isoforms are mRNAs that are produced from the same locus in the DNA but differ in their transcription start site, protein coding DNA sequence and/or untranslated regions, thus potentially altering gene function.
- What is pseudoalignment in RNA-seq?
    - Instead of mapping to the whole genome, we map to known transcript annotation sites in the genome, which speeds up the mapping process. Therefore, pseudo aligners require a reference transcript annotation file.
    - used when we care about estimate of expression levels of transcript and not single mutations etc.
- What are some tools that can be used for pseudoalignment of RNA-seq data?
    - Kallisto
    - Salmon
- What factors does the count of a read depend on in RNA-seq?
    - expression of gene
    - length of gene
    - sequencing depth
    - library composition (e.g. abundance of other transcripts)
    - library preparation factors (e.g PCR)
    - in silico factors like alignment algorithm
- What type of distribution is used to model RNA-seq data?
    - Negative Bionomial distribution
- What is a QQ plot and what is it used for?

  - A QQ plot is used to test if data follow a normal distribution
  - x-axis is theoretical quantiles of data following a normal distribution
  - y-axis contains actual quantiles of the data
- What are the main steps in an RNA-seq experiment?
  - Prepare the RNA-seq library
  - High-throughput sequencing
  - Data processing
    - Quality control
    - Alignment
    - Transcript abundance quantification

Three main steps for RNA-seq:

1. Preparing an RNA-seq library
   a. Isolate the RNA from the cell
   b. Break the RNA into small fragments because mRNAs can be 1000s of bases long but the sequencing machine can only sequence short 50-300bp fragments
   c. Convert the RNA fragments into double stranded DNA
   d. Add sequencing adaptors to allow the seq. machine to recognize fragments, allow you to seq. diff sample at the same time since different samples can use different adaptors. Note: this step does not work 100% (not 100% of fragments get adaptors)
   e. PCR amplify the library. Only fragments with adaptors get amplified and enriched.

f. Quality control: verify the library concentration and library fragment lengths

2. Sequence

   a. Attach fluorescent nucleotides to the first base on each sequence (nearest to the flow cell surface)

   b. Image in a 4-channel fluorescent scanner

   c. Cleave the fluorescent probes

   d. Attach the next fluorescent probe, take picture, cleave and repeat.

3. Data processing

   a. Filter out garbage reads

      i. Reads with low quality base calls

      ii. reads that are clearly artifacts of the chemistry.

         1. Sometimes adaptors can bind to each other without a fragment in between.

   b. align high quality reads to a genome

   c. Count the number of reads per gene

   d. Normalize the data

      i. one sample might have more low quality reads or another sample may have a higher library concentration