Selected slides on RNA-seq intro

- Taken from https://github.com/hbctraining/Intro-to-rnaseq-hpc-orchestra/blob/master/lectures/

Mariam Quiñones, PhD

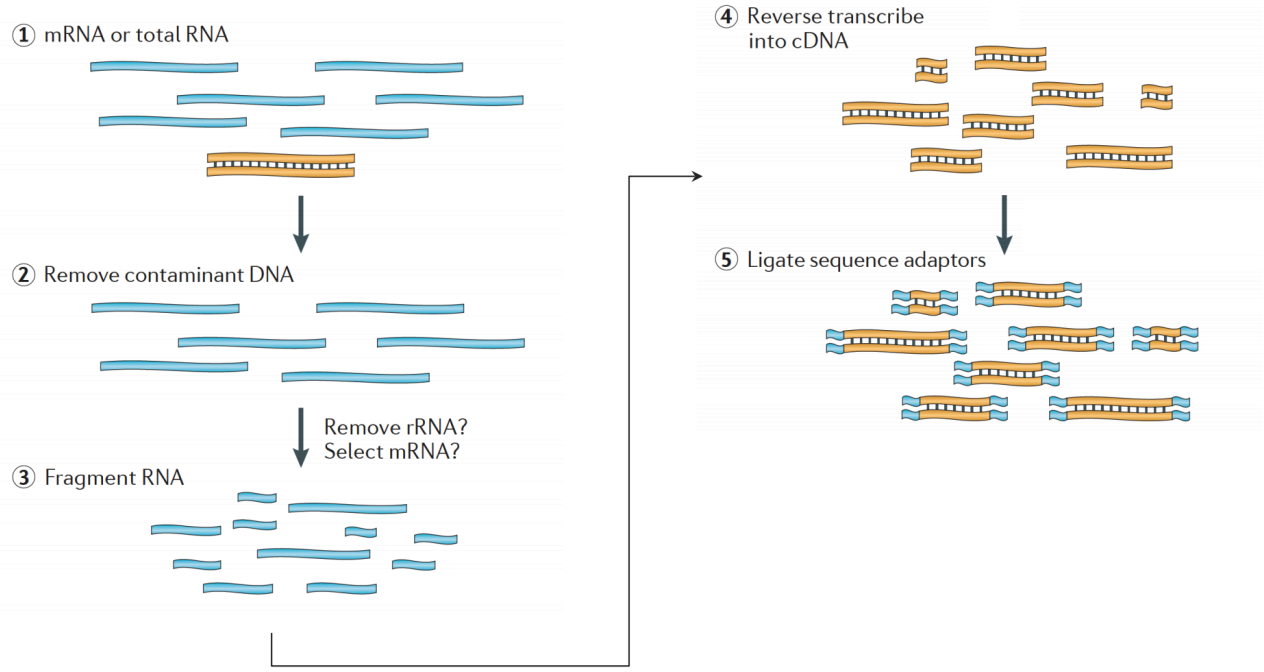
May 19, 2020

Transcriptomics (RNA-Seq)

- The process of sequencing the “transcriptome”
- Uses include –
  - Differential Gene Expression

    Quantitative evaluation and comparison of transcript levels
  - Transcriptome assembly

    Building the profile of transcribed regions of the genome, a qualitative evaluation.
  - Can be used to help build better gene models, and verify them using the assembly
  - Metatranscriptomics or community transcriptome analysis



RNA-Seq library prep

http://rnaseq.uoregon.edu/#rna-prep

Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Experimental and Practical considerations

1. Experimental Design
2. Poly(A) enrichment or ribosomal RNA depletion?
3. Single-end or Paired-end data?
4. Stranded libraries?
5. How much sequencing data to collect?
6. Multiplexing

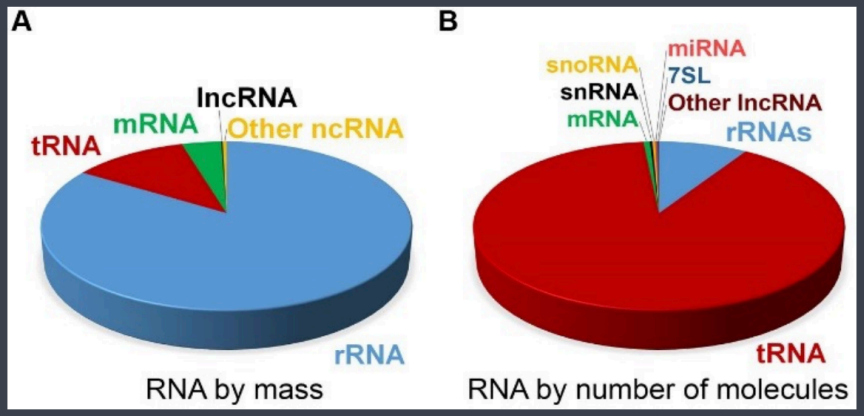
Experimental and Practical considerations

1. Experimental design

- Technical replicates: Illumina has low technical variation unlike microarrays, hence technical replicates are unnecessary.
- Biological replicates, are absolutely essential. Have at least 3!
- Batch effects are still a problem. Be consistent!
- For differential gene expression, pooling RNA from multiple biological replicates can be tricky; do so only if you have multiple pools from each experimental condition.

Experimental and Practical considerations

2. Poly(A) enrichment or ribosomal RNA depletion?



Depends on which RNA entities you are interested in…

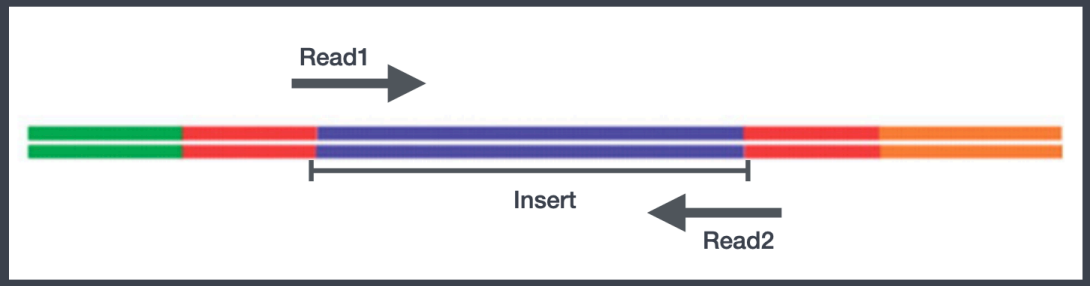
- For differential gene expression, it is best to enrich for Poly(A)+
  - EXCEPTION – If you are aiming to obtain information about long non-coding RNAs, then do a ribosomal RNA depletion.

Experimental and Practical considerations

3. Single-end or Paired-end data?

Depends on your goals, paired-end reads are better for reads that map to multiple locations, for assemblies, and for splice isoform differentiation.

- For differential gene expression, which one you pick depends on-
  - If you are specifically interested in isoform-level differences
  - The abundance of paralogous genes in your system of interest
  - Your budget, paired-end data is usually 2x more expensive



- ✓ SE - Single end dataset => Only Read1
- ✓ PE - Paired-end dataset => Read1 + Read2
  - can be 2 separate FASTQ files or just one with interleaved pairs
- ✓ Fragment length: ~300-500bp
- ✓ Read length: 50bp - 250bp, depends on the sequencer (HiSeq2500, MiSeq, NextSeq)

Options for sequencing

http://tucf-genomics.tufts.edu/home/faq

Experimental and Practical considerations

4. Stranded libraries?

Stranded libraries are now standard with Illumina's TruSeq stranded RNA-Seq kits. This means that with a great amount of certainty you can identify which strand of DNA the RNA was transcribed from.

3 types of libraries –

- Reverse (firststrand)– reads resemble the complementary sequence (TruSeq)
- Unstranded
- Forward (secondstrand) – reads resemble the gene sequence

Experimental and Practical considerations

5. How much sequencing data to collect?

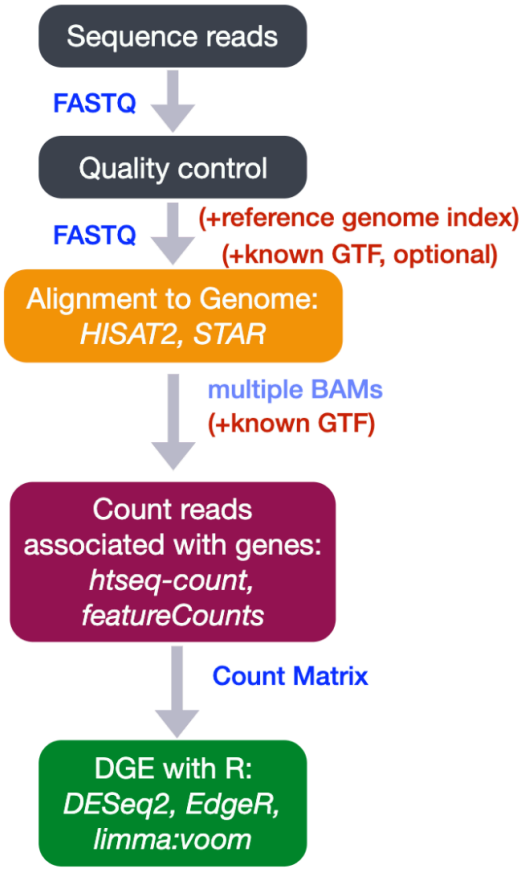
- Only ~2% of the human genome transcribes protein-coding RNA
- Some mRNAs will be much more abundant than others
- Some genes are much longer than others

Recommendations:

- For human samples ~30-50 million reads/sample (ENCODE guidelines)
- Modify that number based on the size of your transcriptome (crude estimate)
- If working with a tight budget:

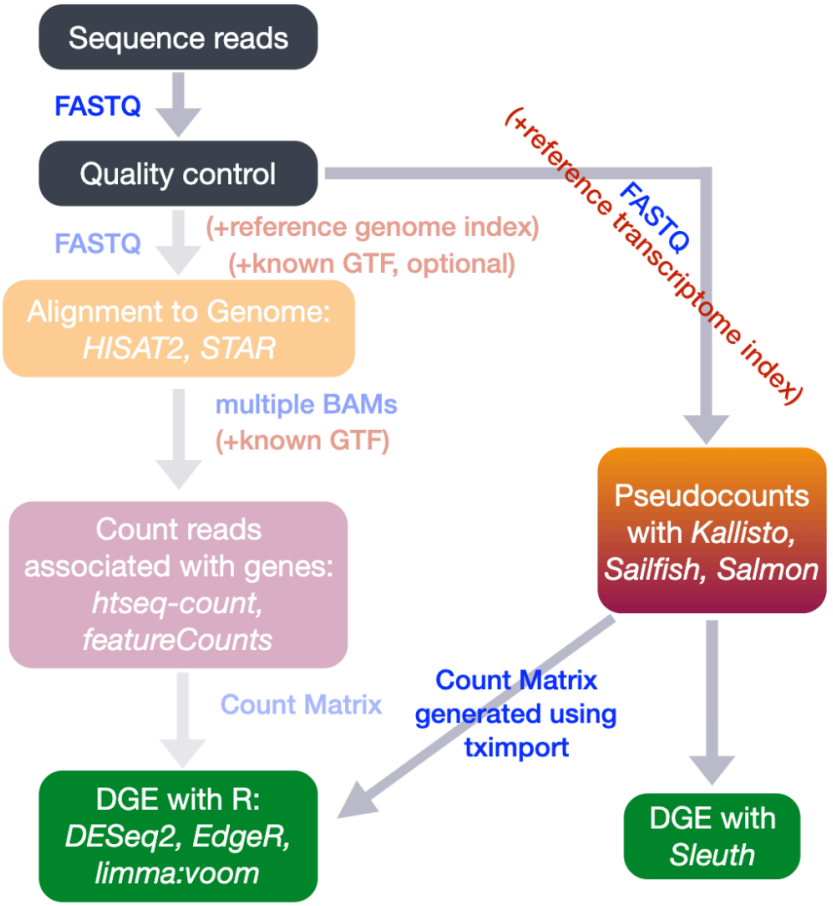
  More replicates >> More reads (for standard differential expression analysis)

- ✓ Genome
- ✓ Transcriptome
- ✓ Gene-level quantification



Differential Expression Analysis Workflow

- *Genome?*
- ✓ Transcriptome
- ✓ Isoform-level and/or gene-level quantification



Differential Expression Analysis Workflow