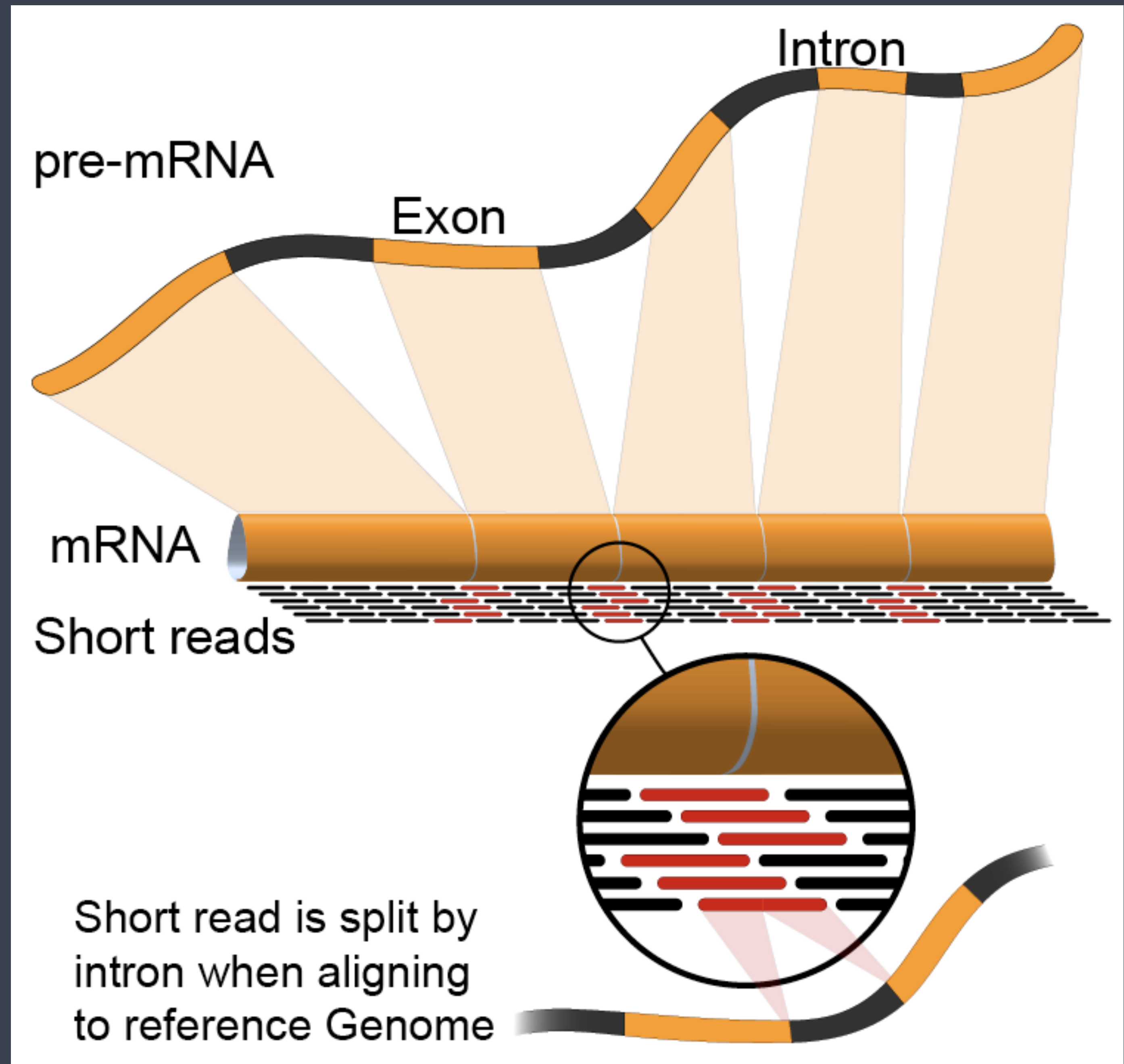
RNA-seq workflow



http://upload.wikimedia.org/wikipedia/commons/0/01/RNA-Seq-alignment.png

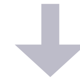
Transcriptomics (RNA-Seq)

- 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

Biological samples/Library preparation



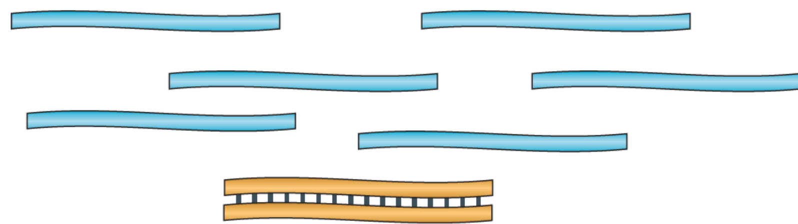
Sequence reads

Analysis Workflow

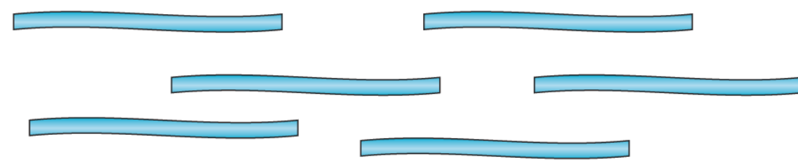
Outline

- Library preparation and sequencing with Illumina
- Experimental and Practical Considerations
- Analysis workflow

1. mRNA or total RNA

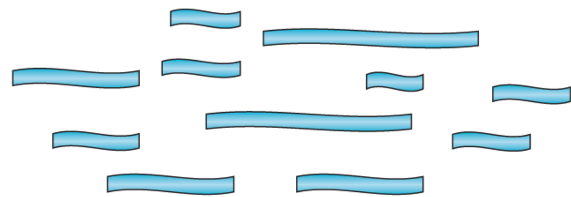


2. Remove contaminant DNA

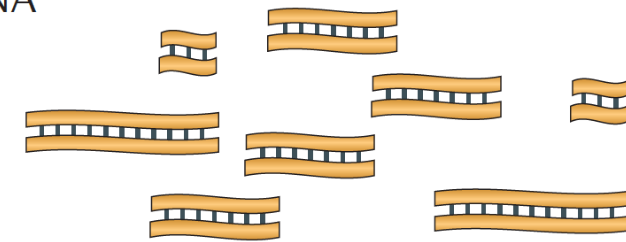


Remove rRNA?
Select mRNA?

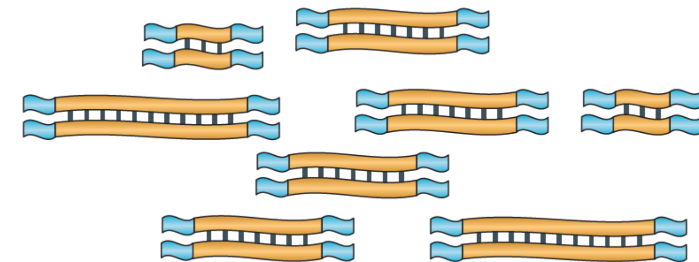
3. Fragment RNA



4. Reverse transcribe into cDNA



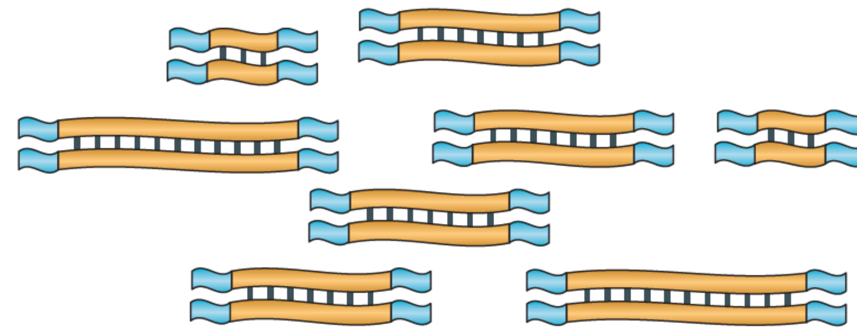
5. Ligate sequence adaptors

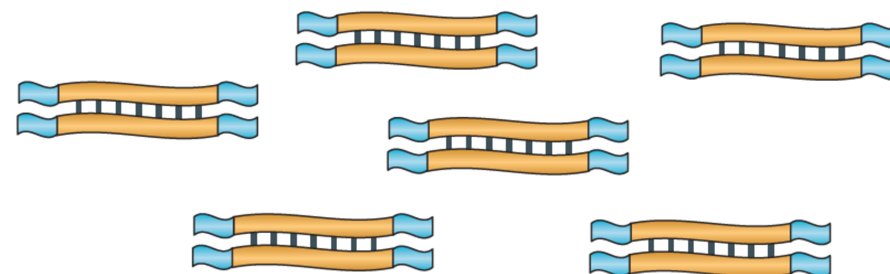


RNA-Seq library prep

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

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

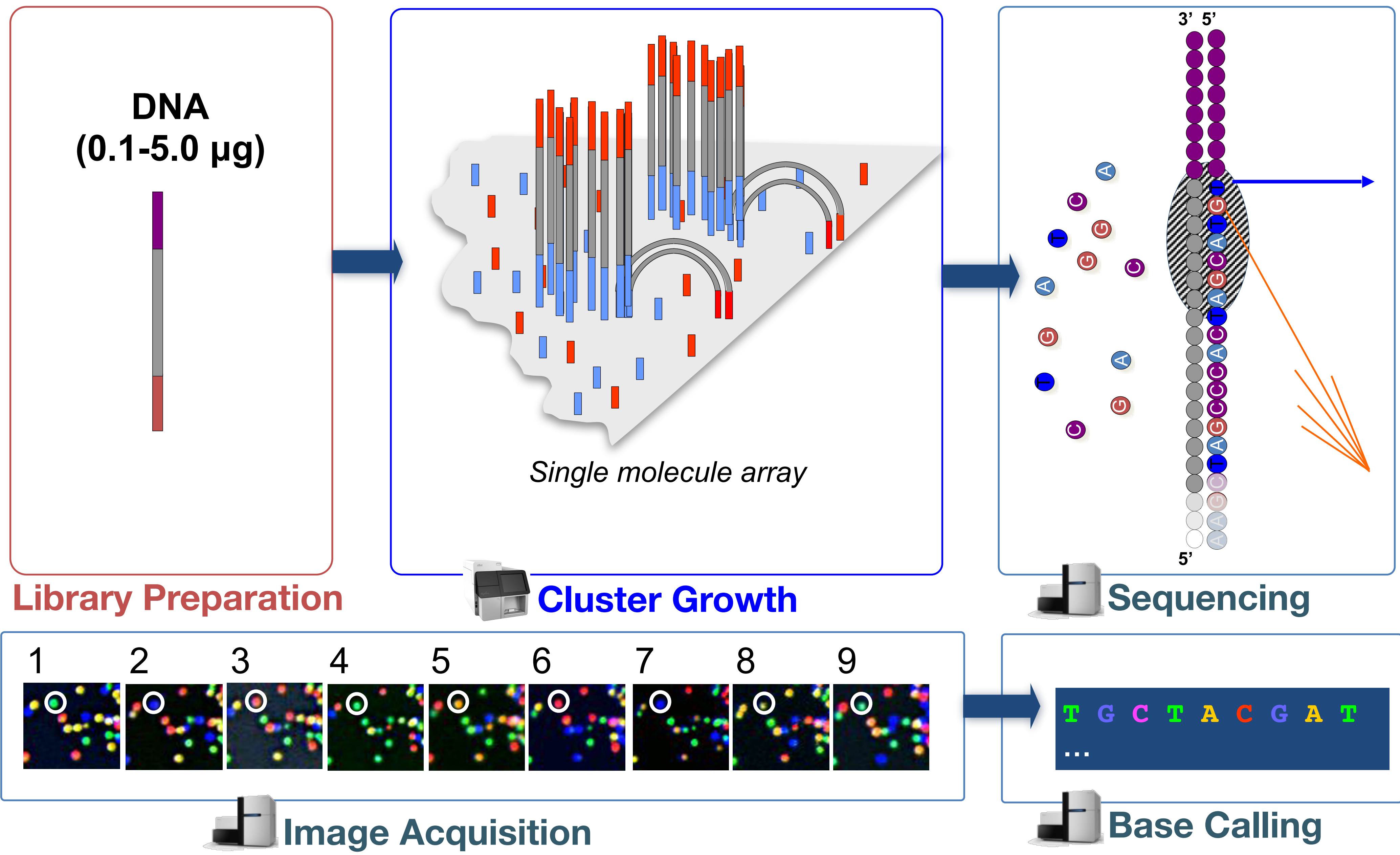




RNA-Seq library prep

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

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



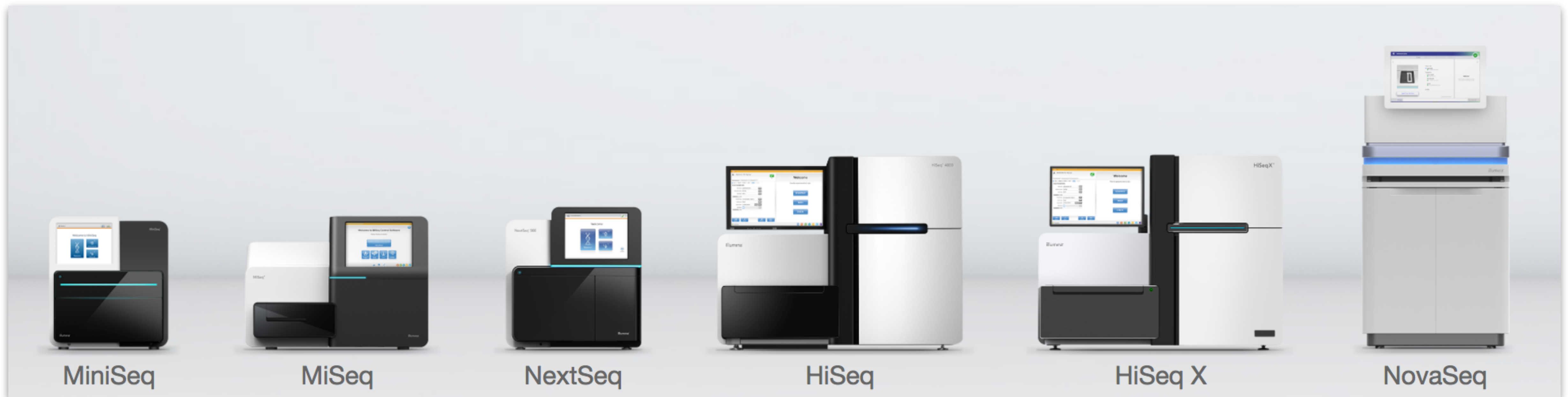
https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=3s

Illumina: Sequencing by Synthesis

Number of clusters ~= Number of reads

Number of sequencing cycles ~= Length of reads

Illumina: Sequencing by Synthesis



https://www.illumina.com/systems/sequencing-platforms.html

Illumina: Sequencing Platforms

Oxford Nanopore (MinION): https://nanoporetech.com/

Pacific Biosciences: http://www.pacb.com/

Other Sequencing Platforms

Outline

- Library preparation and sequencing with Illumina
- Experimental and Practical Considerations
- Analysis workflow

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

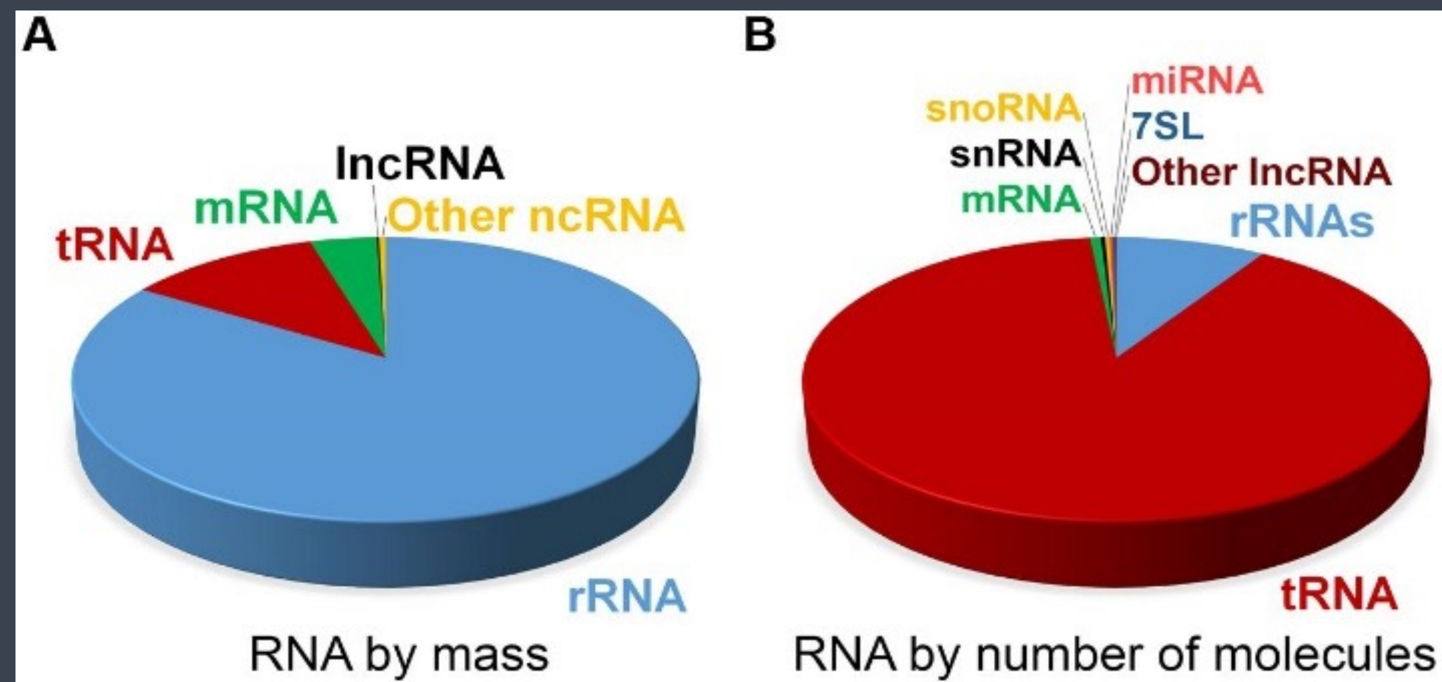
Experimental and Practical considerations

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?



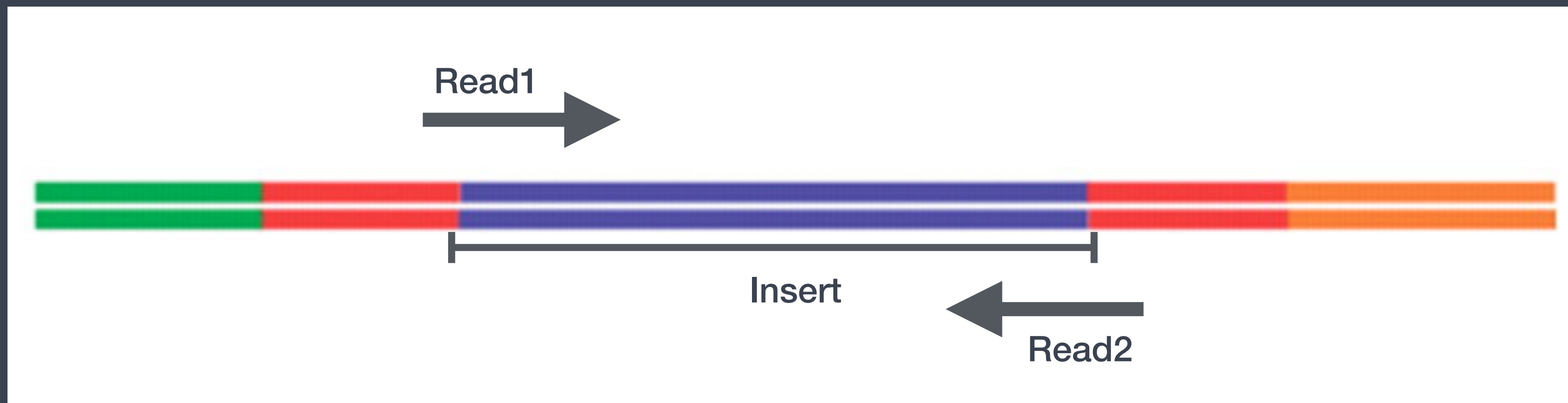
Depends on which RNA entities you are interested in…

- 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?

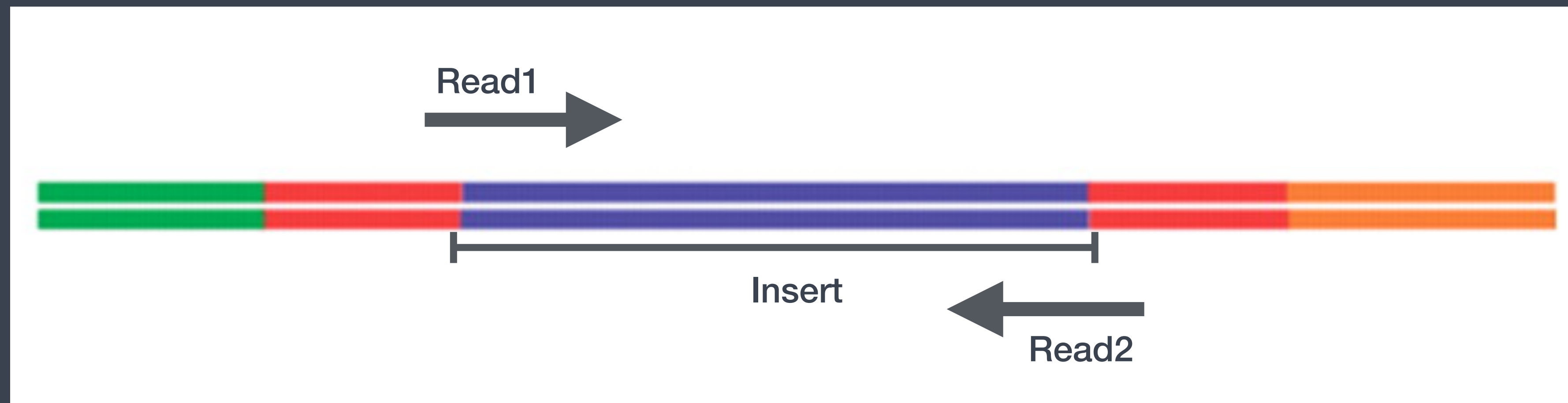
Depends on your goals, paired-end reads are better for reads that map to multiple locations, for assemblies and for splice isoform differentiation.



- ✓ SE - Single end dataset => Only Read1
- ✓ PE - Paired-end dataset => Read1 + Read2
    - can be 2 separate FASTQ files or just one with interleaved pairs

Options for sequencing

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



- ✓ 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 - 300bp, depends on the sequencer (HiSeq2500, MiSeq, NextSeq)

Options for sequencing

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

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

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:

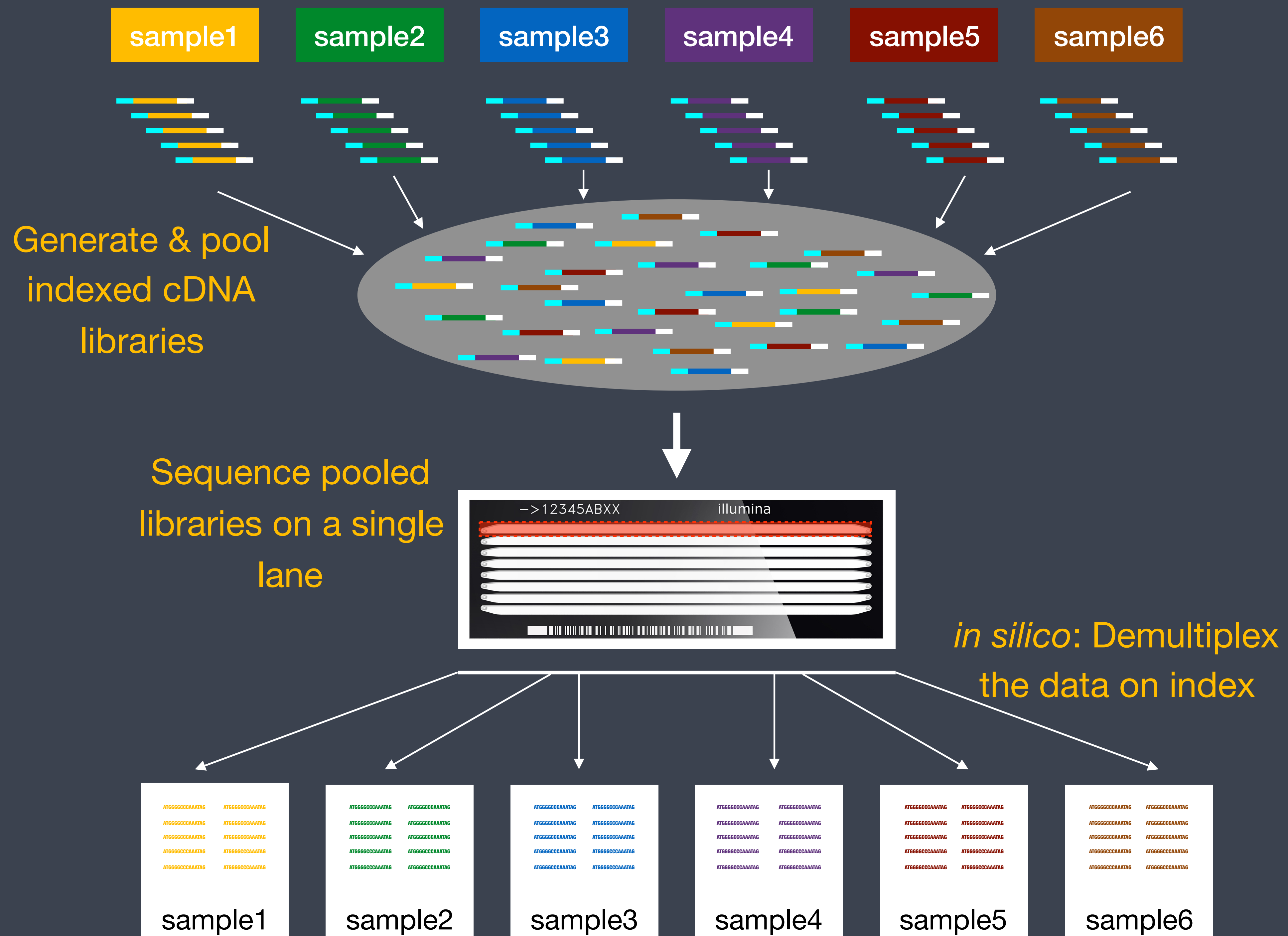
  More replicates >> More reads (for standard differential expression analysis)

Experimental and Practical considerations

6. Multiplexing (with barcodes and indices)

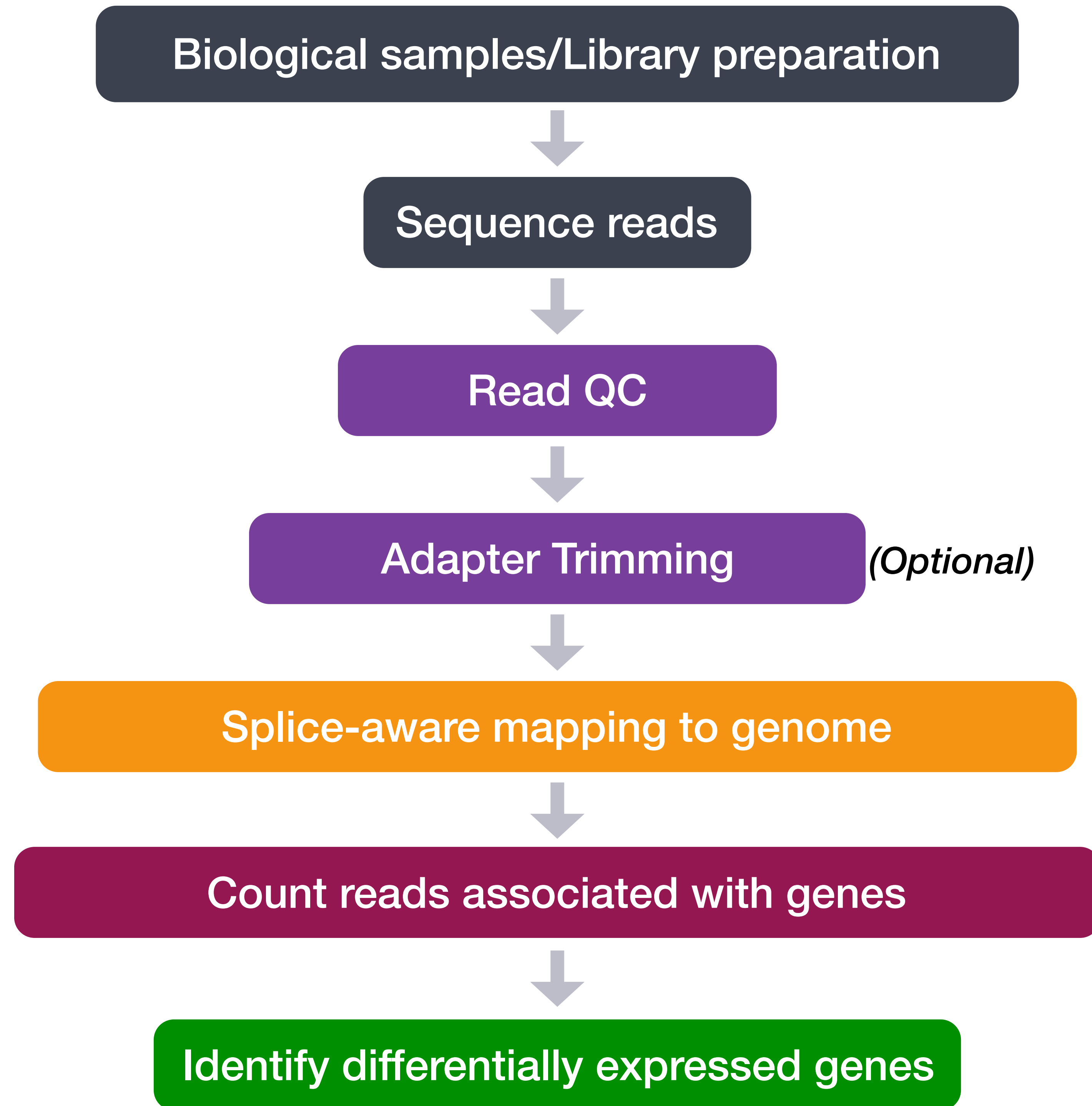
- Charges for sequencing are usually per lane of the flow cell
- Each lane generates ~150 million reads
- For RNA-Seq, the required data per sample is much lower than that
- Sequencing of multiple samples per lane possible with addition of indices (within the Illumina adapter) or special barcodes (outside the Illumina adapter).

6. Multiplexing (with barcodes and indices)

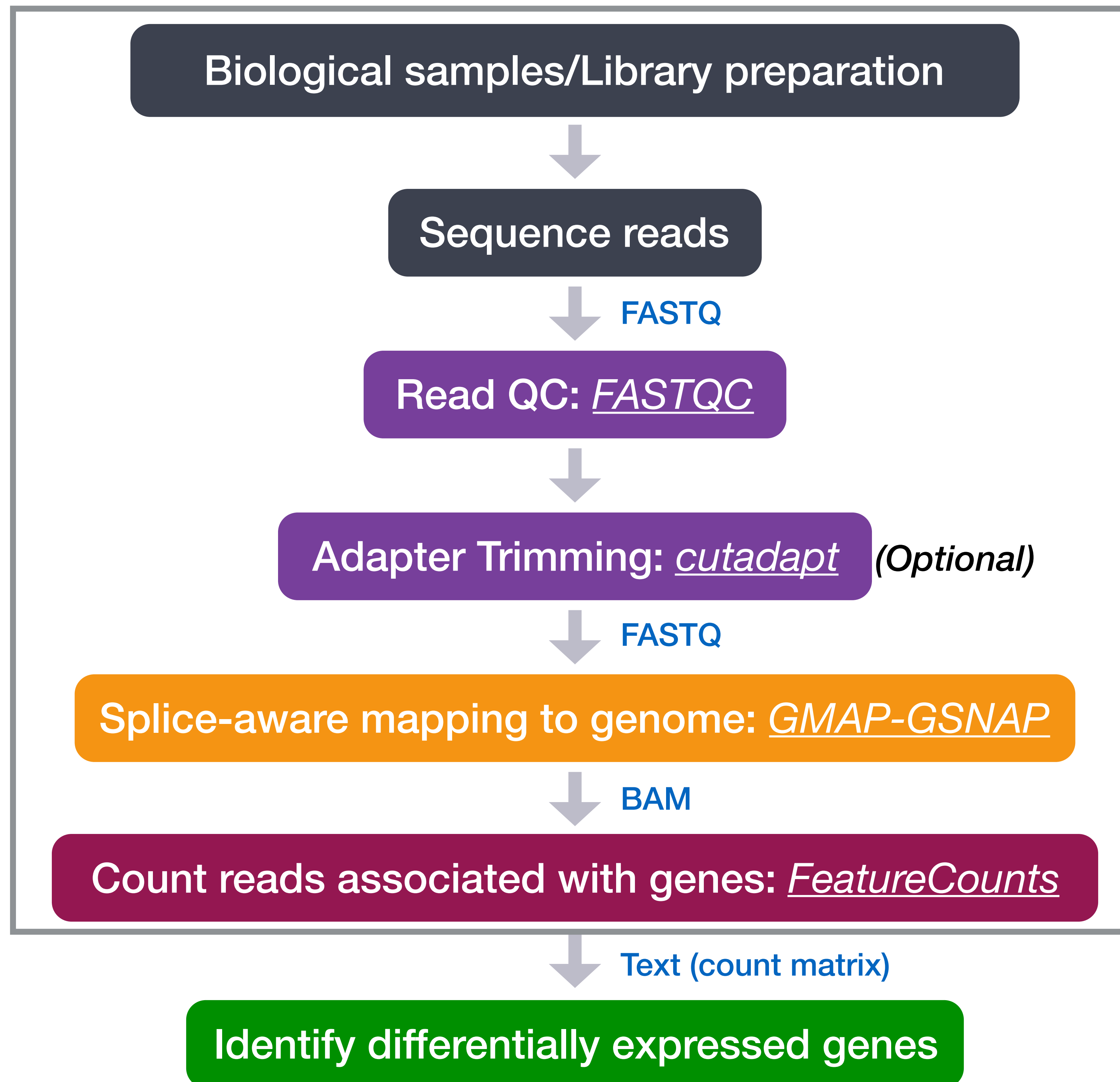


Outline

- Library preparation and sequencing with Illumina
- Experimental and Practical Considerations
- Analysis workflow



Analysis Workflow



Analysis Workflow

These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.