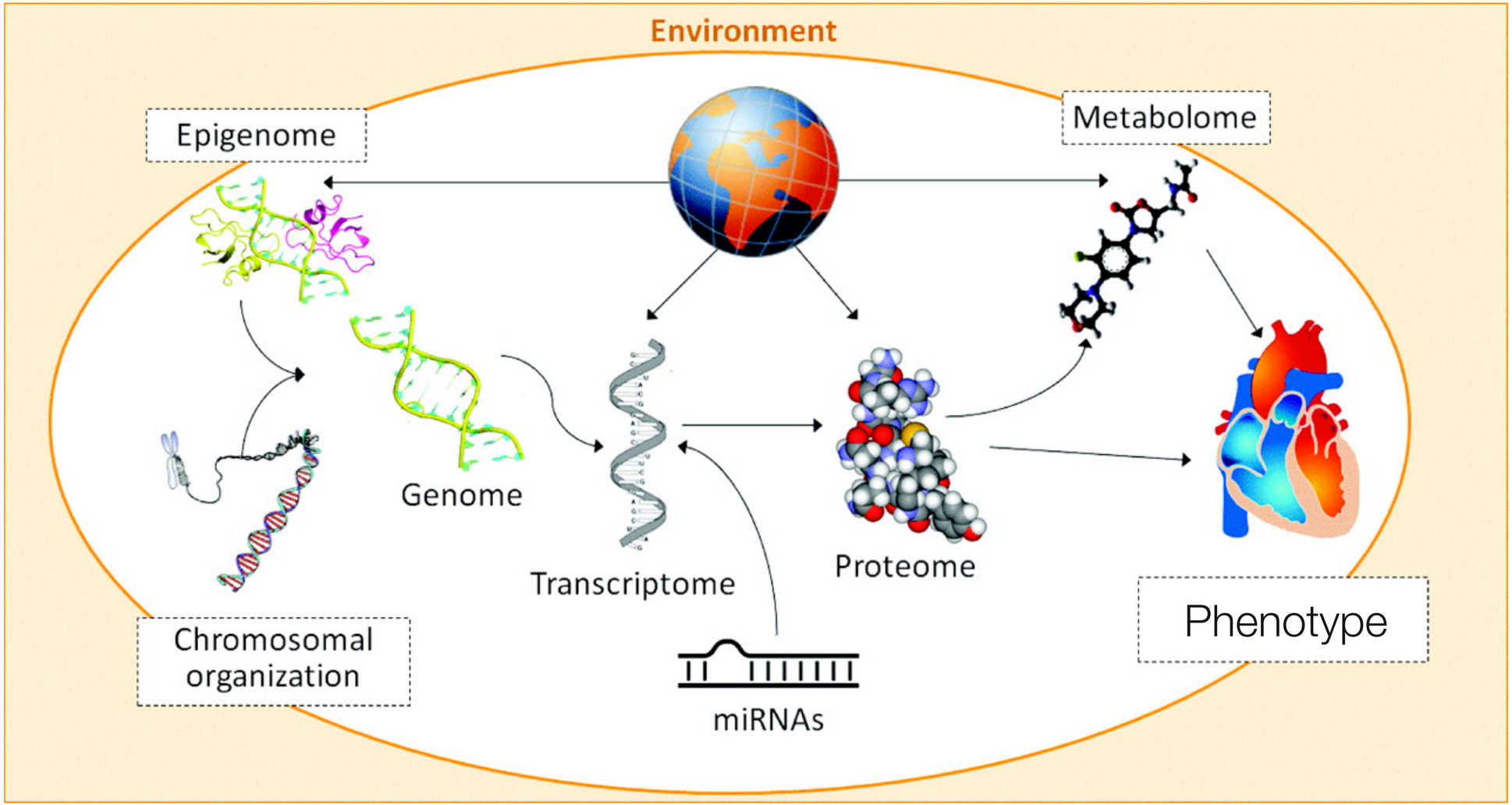
RNA-Seq Session Organization

- Introduction to RNA-Seq Experiment and Design
- RNA-Seq Alignment and Normalization
- RNA-Seq Differential Expression and Functional Analysis
- [Hands-On] RNA-Seq Differential Expression Analysis in R walkthrough

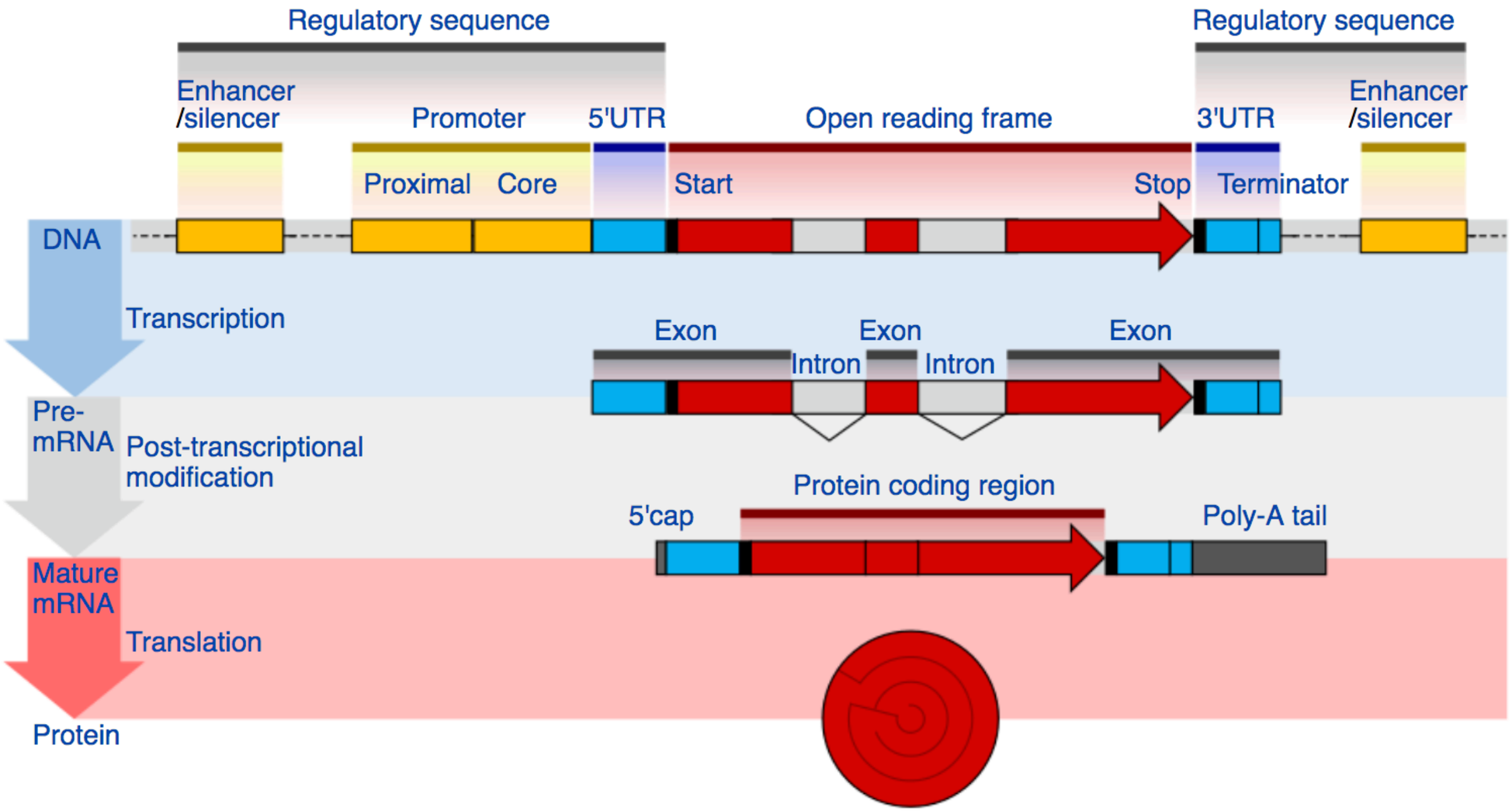
Introduction to RNA-Seq Experiment and Design

Eukaryotic Gene Regulation In Global Context



http://clinchem.aaccjnls.org/content/58/1/113

Eukaryotic Gene and mRNA Transcript Structure



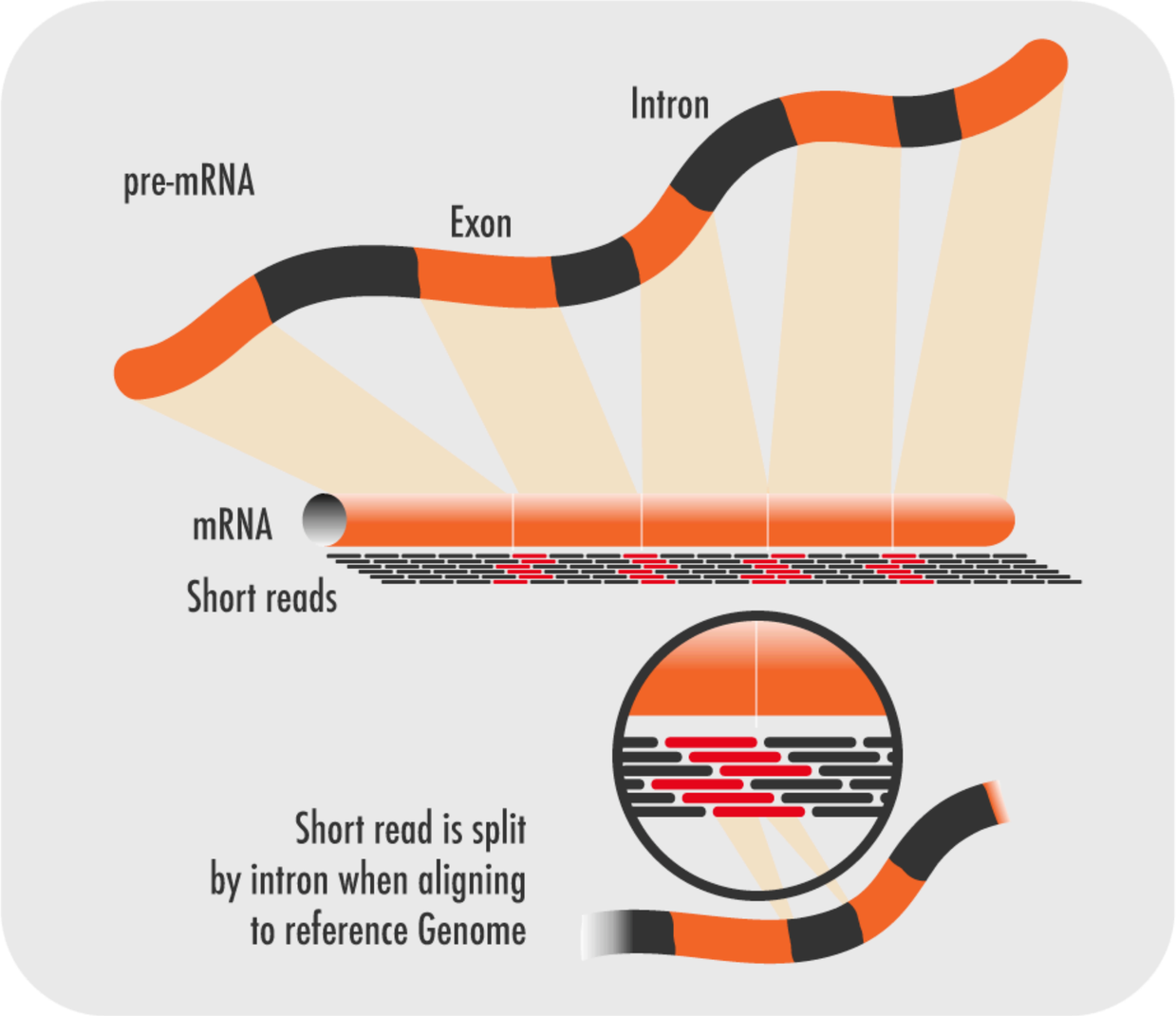
Shafee and Lowe, WikiJ of Med. (2017)

Transcriptome Profiling

- Transcriptome profiling represents a static gene expression state of a biological sample across the genome
- Allows for direct genomic comparisons with multiple samples to determine genes that exhibit *differential expression* in different state (i.e. normal vs. tumor)
- Allows for *hypothesis generation* on molecular abnormalities and mechanisms that may contribute to the tumor phenotype
- Provides information on molecular subtypes, the development of prognostic and predictive molecular signatures

What is RNA-Seq

- Using NGS technology to sequence RNA transcripts
- Typically refers to the sequencing of mRNA
- Different RNA species (i.e. miRNA, snoRNA, tRNA) require different preparation protocol
- Any type of RNA from any sample sources, such as cell, body fluid, stool, water, etc. can be the sequenced
- Sample from different sample sources, such as cell, body fluid, stool, water, etc, require different extraction method
-

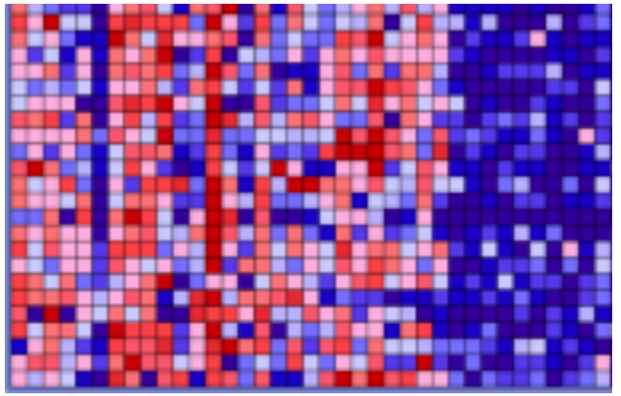


MacKenzie, 2018, Tech. Network.com

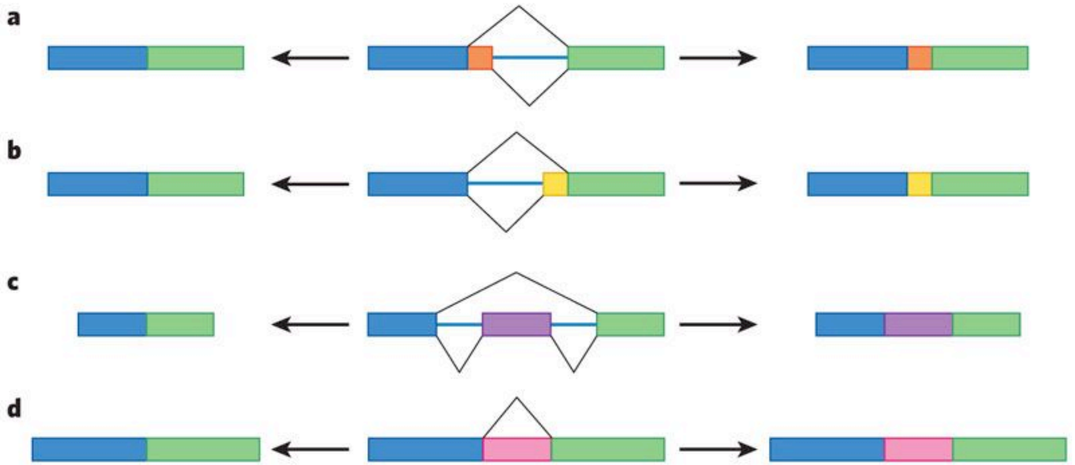
Why do RNA-Seq?

- Which genes are differentially expressed in different conditions?
- Are genes being transcribed in alternatively spliced transcript isoforms?
- Are there mutations being transcribed such as insertions, deletions, or novel isoforms?

Transcriptomic Profiling

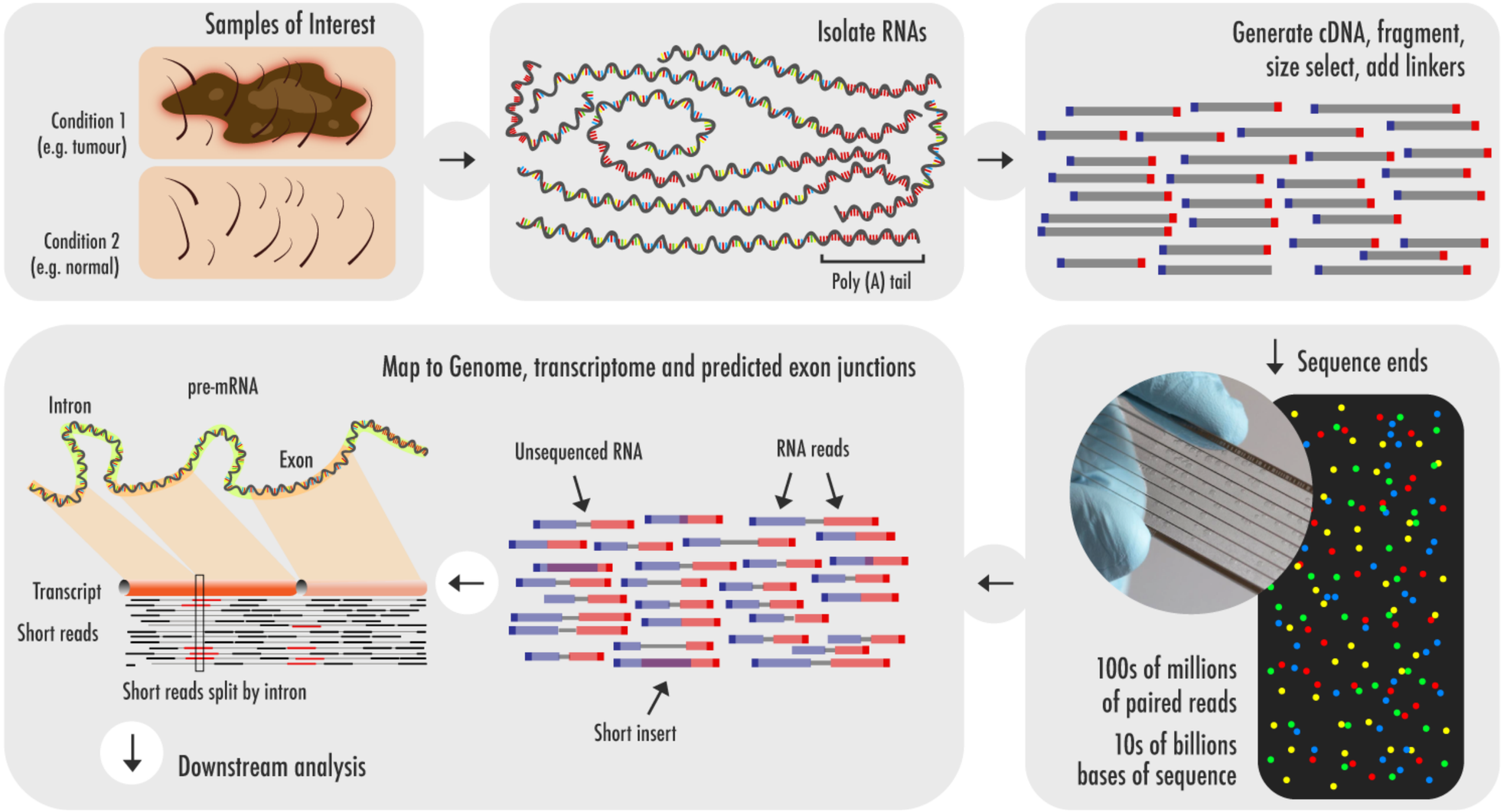


Basic types of alternative splicing



Nielsen and Graveley, 2010, Nature

RNA-Seq Experiment Workflow



MacKenzie, 2018, Tech. Network.com

Sample Acquisition

Fresh Frozen Tissues

- Sample tissues freeze to -80C or immerse in liquid nitrogen shortly after sample extraction
- All RNA is intact in natural form but with slow degradation process
- Produce highest quality data
- Expensive to keep and rare to acquire

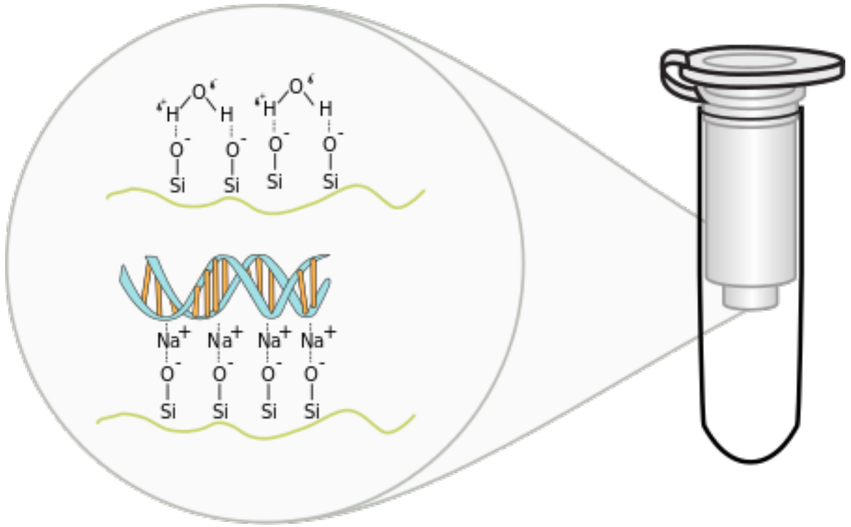
Formalin Fixed Paraffin Embedded (FFPE) Samples

- Fix sample tissues in paraffin wax immediately after extraction
- All RNA are immediately sheared into fragments
- Most common sample available from clinic
- Very cheap to store

RNA Extraction - Column-based Extraction

Protocol:

- Lyse cells by lysis buffer
- Bind RNA to silica membrane
- Filter through a column in a centrifuge
- Series of washes
- Elute in sterile H20



Advantages

Commercially available from vendors (i.e. Qiagen)

Fast and easy

Can isolate total and small RNA simultaneously or separately

Many custom kits for FFPE RNA, blood, saliva, etc

Disadvantages

Can be expensive

Can lose small RNA (<100bp) if not careful

May not work for low RNA quantity samples

RNA Quality Control - RNA Integrity Number (RIN)

- Developed by Agilent on Agilent BioAnalyzer System
- Based on Bayesian based model trained over a set of RNA integrity features from large number of RNA samples
- Samples assign to 10 different categories ranging from 1 (worst) to 10 (best),
- Most important feature is 18S/28S rRNA concentration ratio
- RIN > 6 is typically considered adequate for RNA-Seq

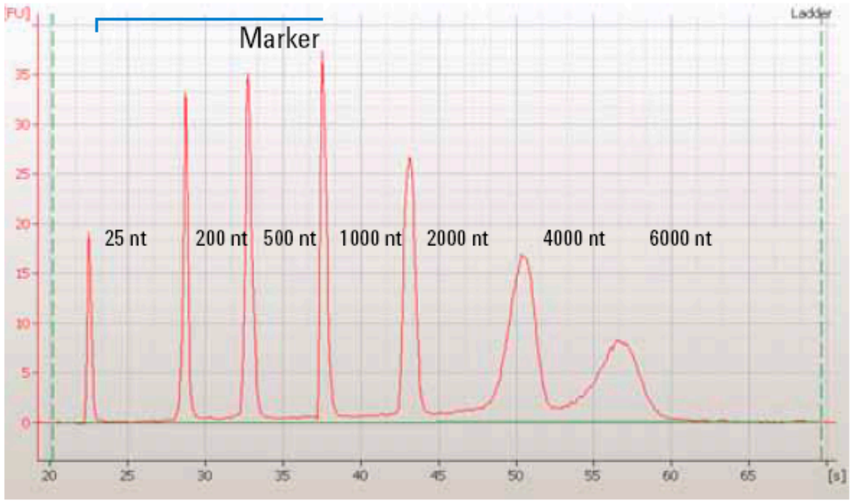
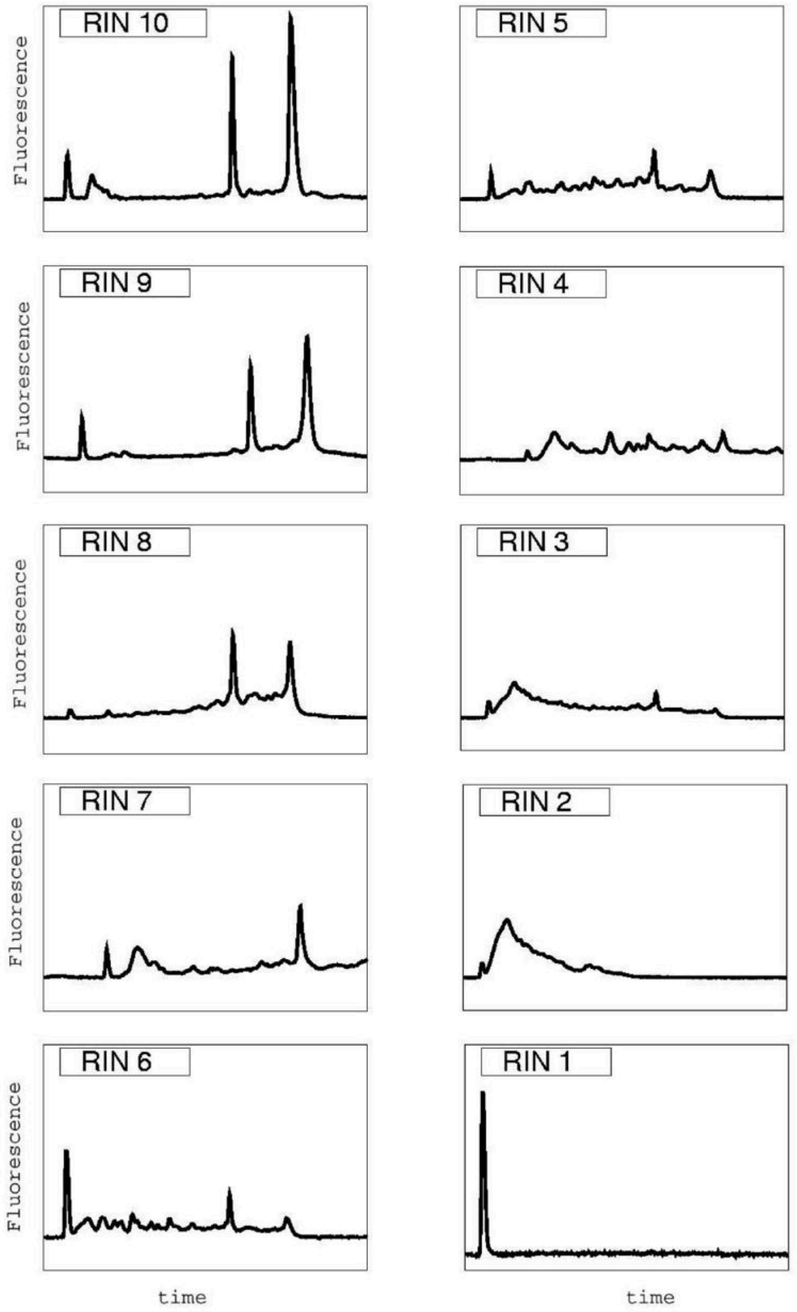


Figure 1 RNA 6000 Nano ladder

Schroeder, et al, 2006, BMC Mol. Bio

RNA Quality Control - RNA Integrity Number (RIN)

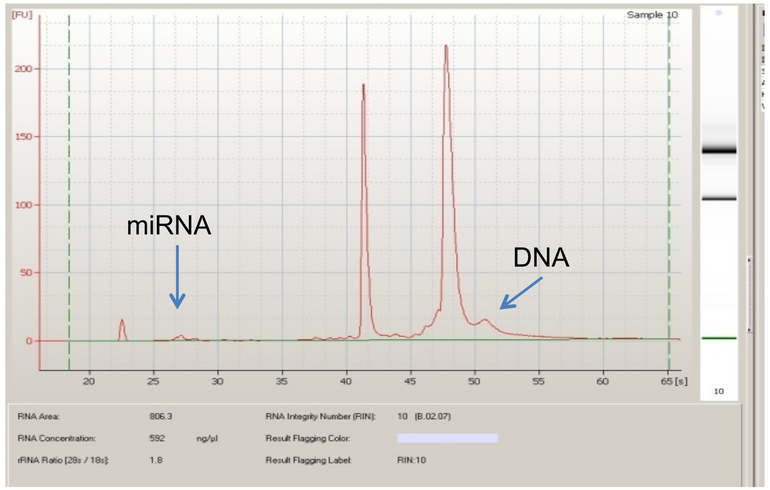
- Developed by Agilent on Agilent BioAnalyzer System
- Based on Bayesian based model trained over a set of RNA integrity features from large number of RNA samples
- Samples assign to 10 different categories ranging from 1 (worst) to 10 (best),
- Most important feature is 18S/28S rRNA concentration ratio
- RIN > 6 is typically considered adequate for RNA-Seq



Schroeder, et al, 2006, BMC Mol. Bio

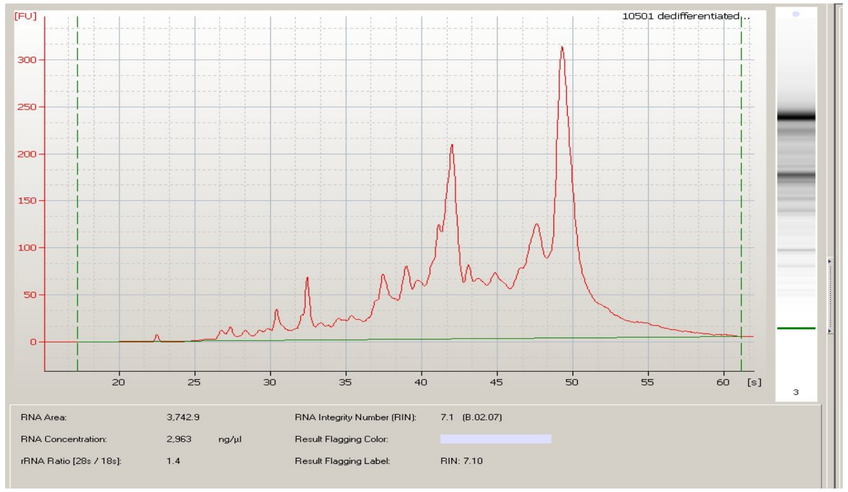
RNA Quality Control - RNA Integrity Number (RIN)

Excellent



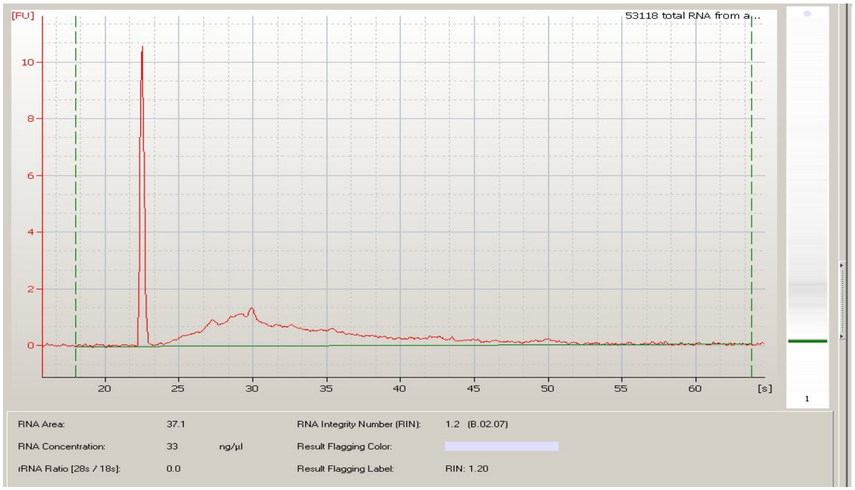
RIN > 8
Typically from fresh sample

Adequate



RIN~6
Typically from fresh frozen samples

Poor



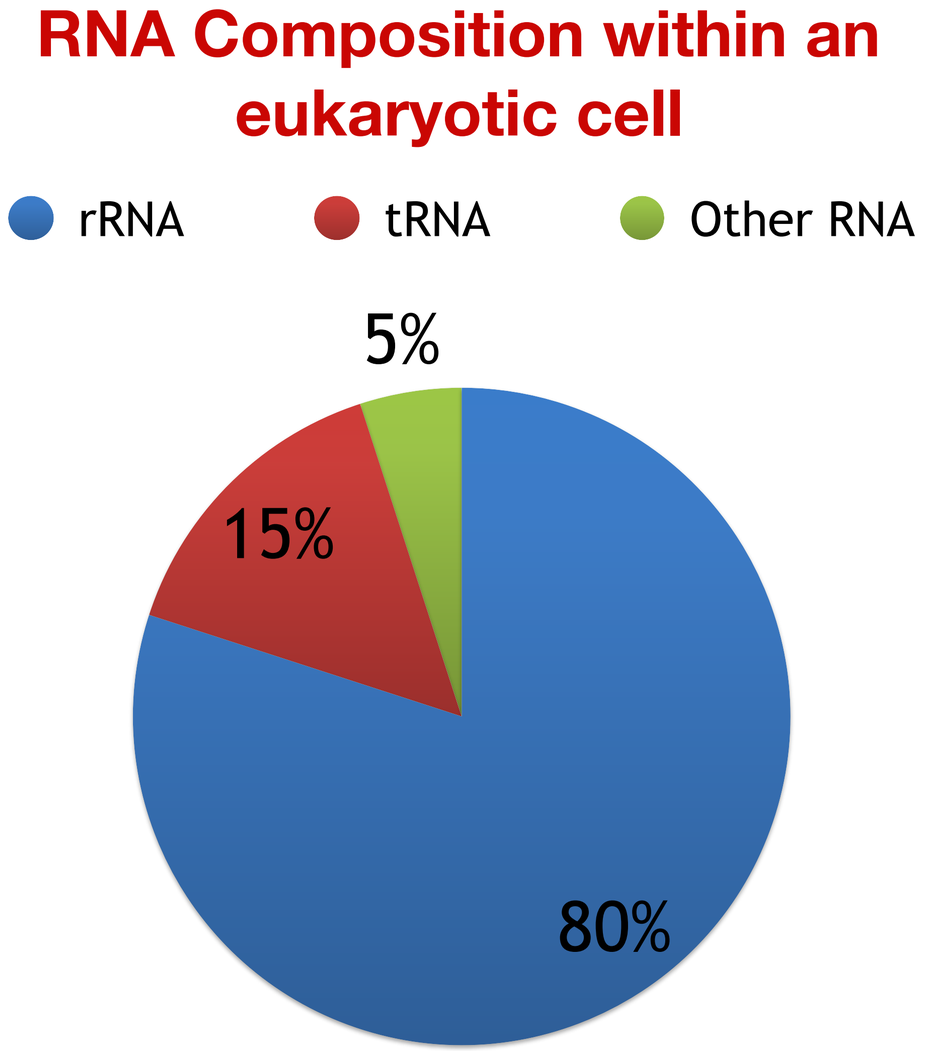
RIN<3
Highly degraded sample such as FFPE sample

Sequencing Library Generation

RNA-Seq sequencing library generation focuses on selecting the proper RNA species to sequence

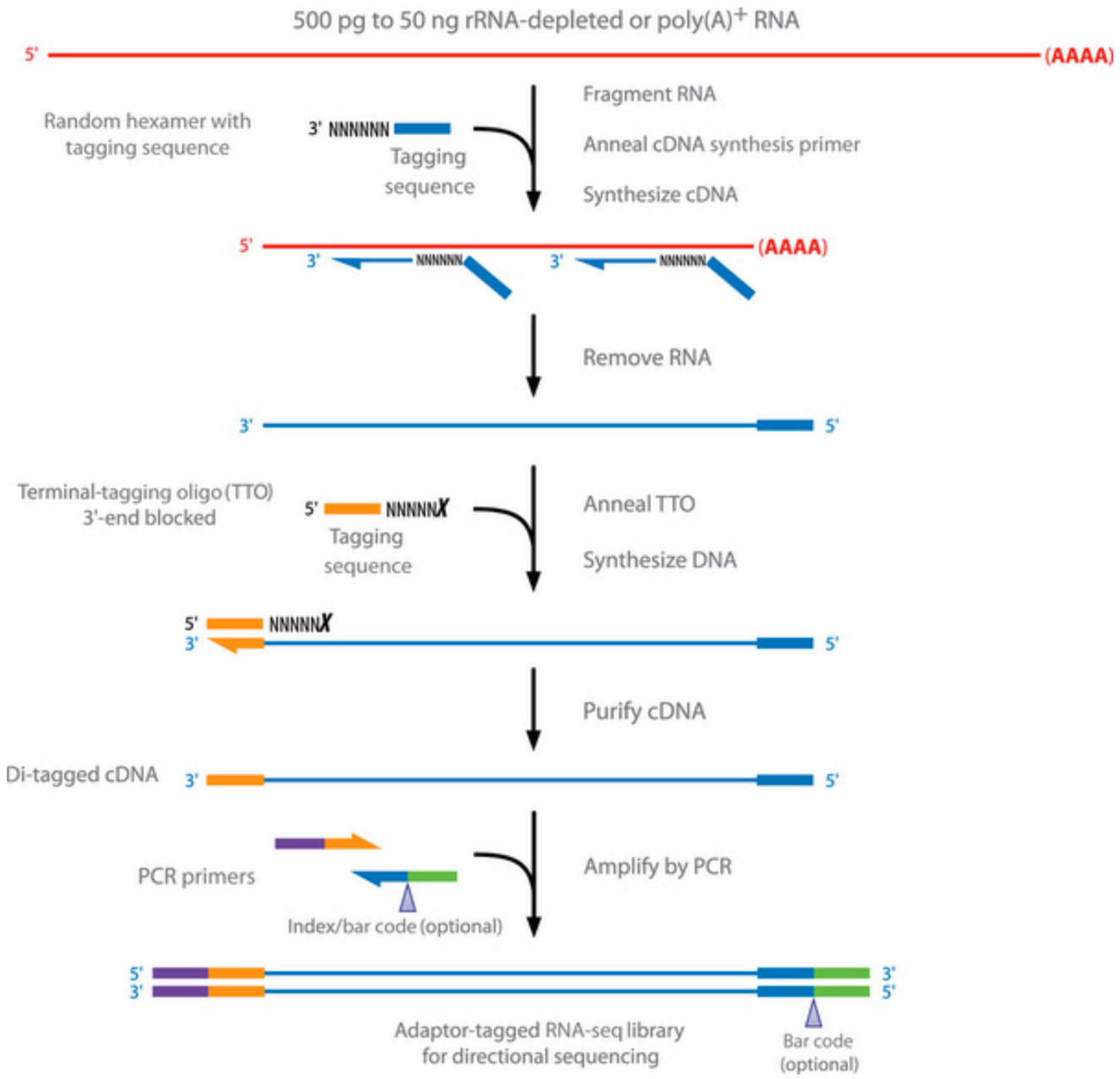
RNA Species Selection

- Pre-mRNA and mature mRNA composed of very small portion of total RNA
- MicroRNA, ncRNA, and others composed of even smaller number
- Library generation starts with **RNA species selection**



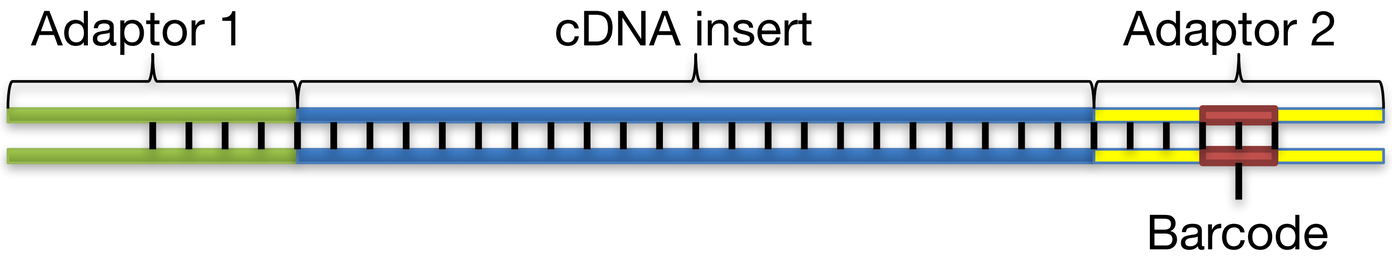
RNA-Seq Library Generation

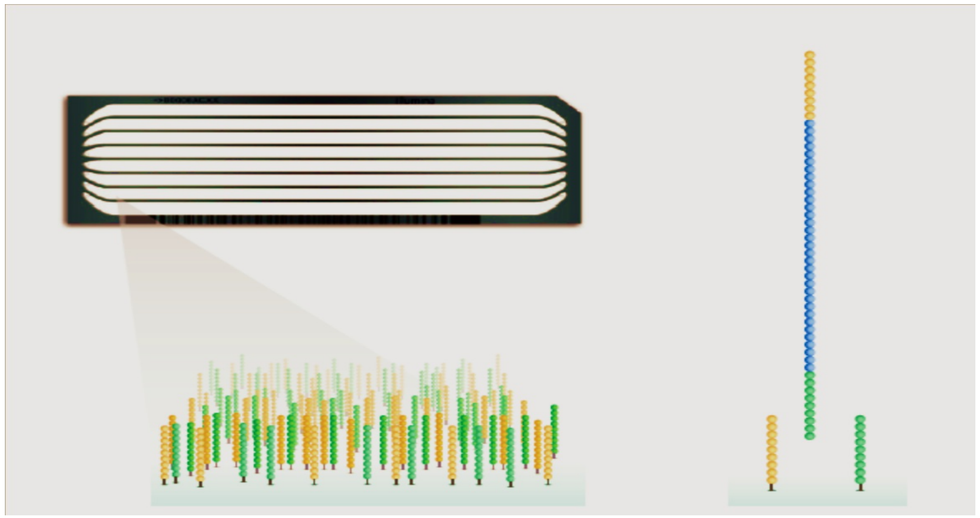
1. RNA selection by poly-A purification
2. Fragment RNA and random priming to generate cDNA
3. Clean up RNA
4. Ligate adapter sequences on both ends
5. Purify cDNA library by size selection
6. PCR amplify the library



Pease, Nature Methods 9 (2012)

Sequencing Library Structure



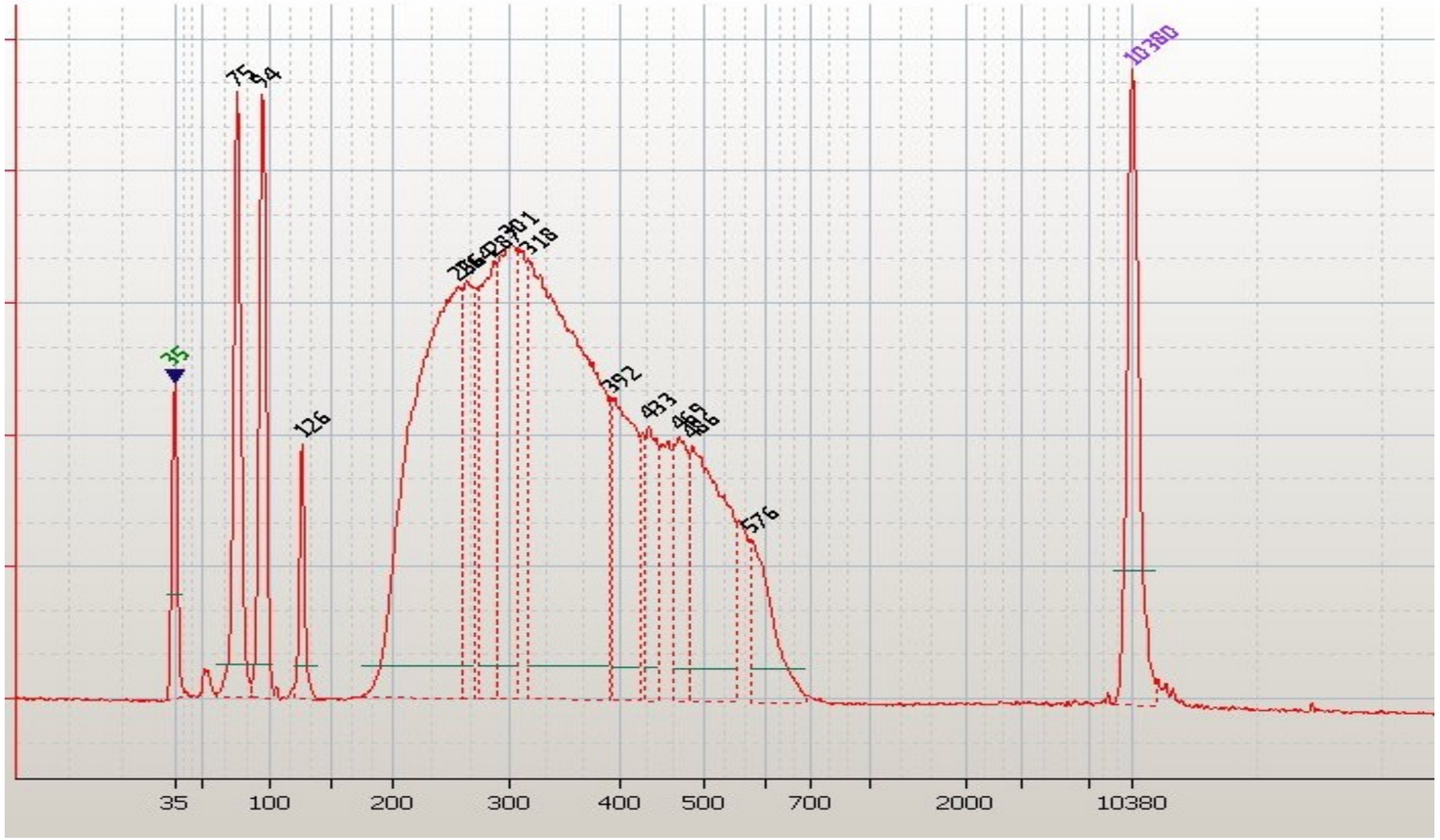


Adaptor – 58 bp nucleotide sequence to fix sequence library onto flow cell

Barcode – optional index sequence for sample multiplexing

cDNA insert – fragmented cDNA sequence generated from mRNA of interest. The insert typically range between 300-500bp for mRNA

Library Clean up and Size Selection



Sequence libraries contain sub-optimal fragments and adaptor dimers

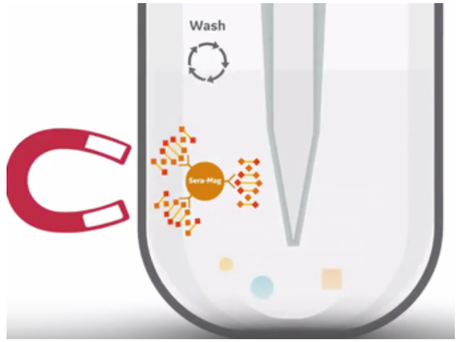
Library Cleanup and Size Selection

Magnetic Bead-based Approach

Magnetic Particles are added to sample and bind to target molecule

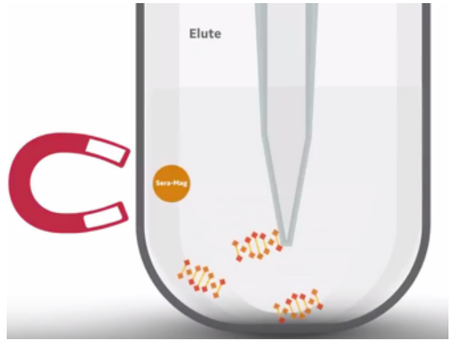
Bind

Magnetic Particles are captured and remainder of sample is washed away



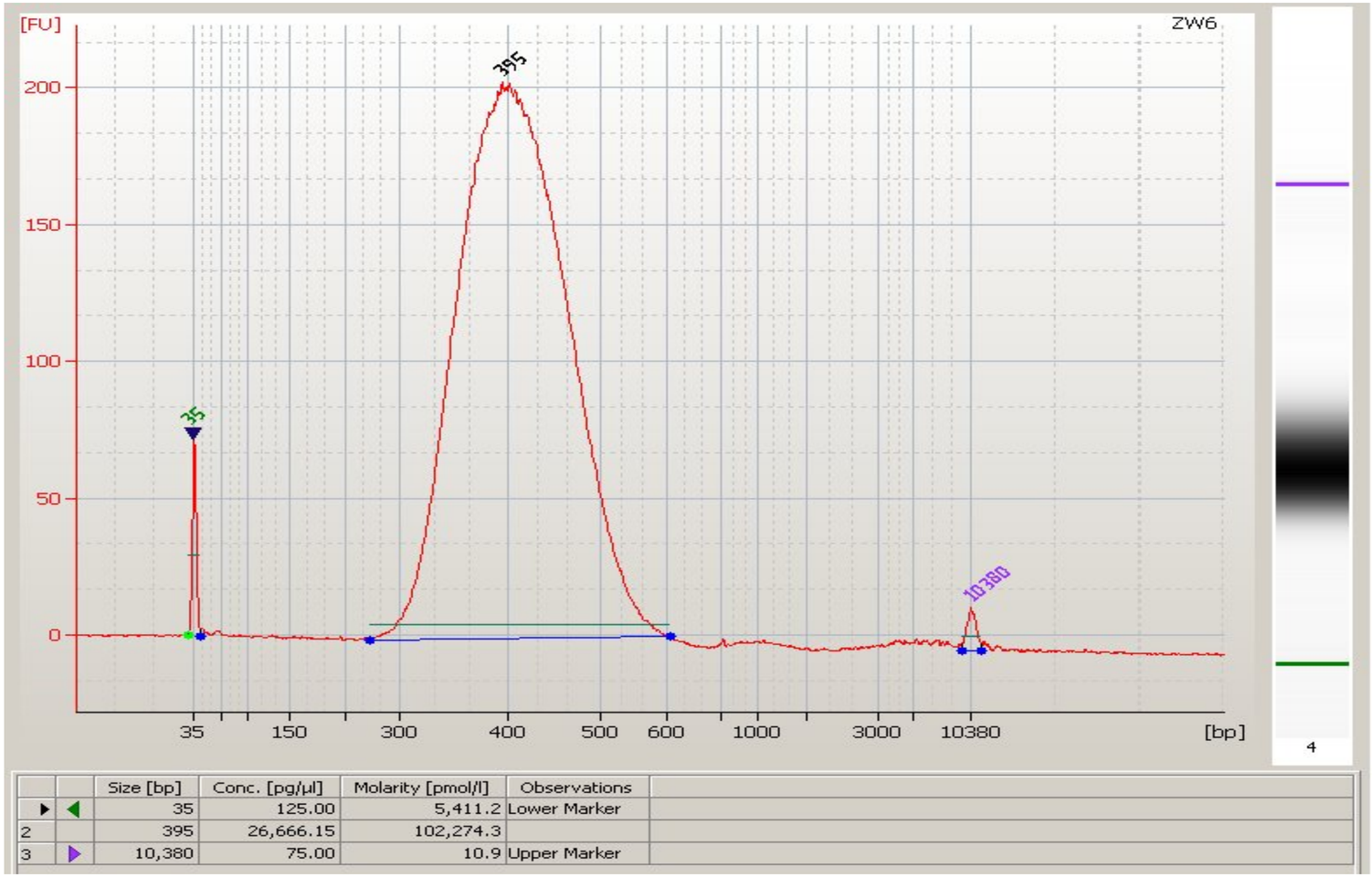
Capture

Target molecule is released from magnetic particles for further analysis



Release

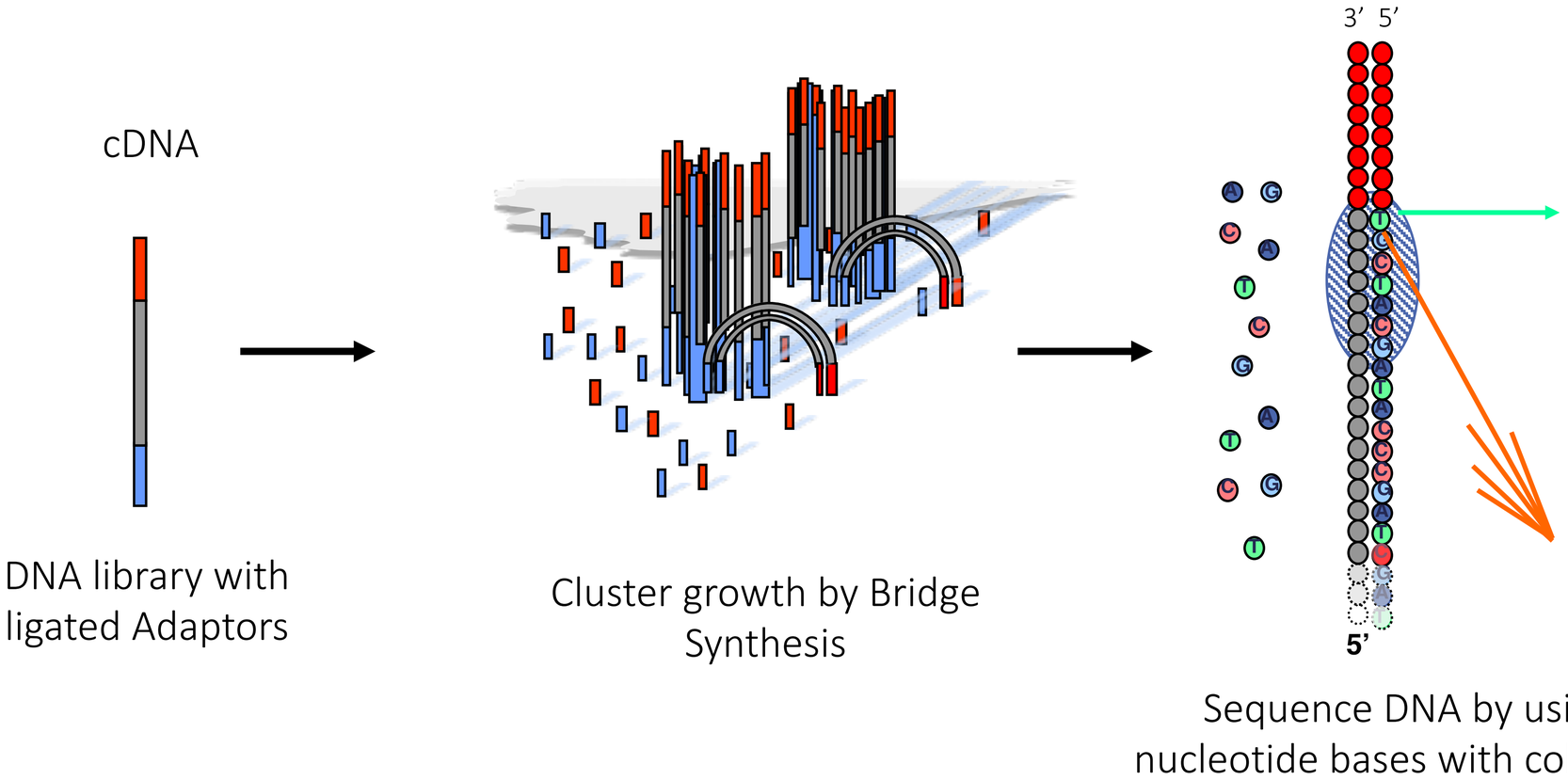
Library Clean up and Size Selection



		Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
▶	◀	35	125.00	5,411.2	Lower Marker
2		395	26,666.15	102,274.3	
3	▶	10,380	75.00	10.9	Upper Marker

Library size centers at 395bp

Sequence by Synthesis (SBS)



DNA library with ligated Adaptors

Cluster growth by Bridge Synthesis

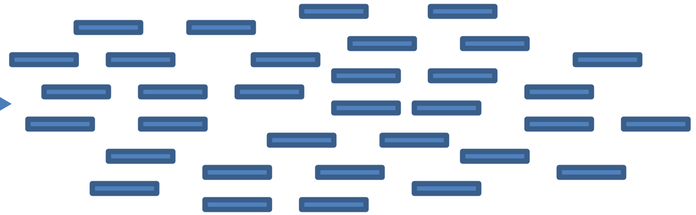
Sequence DNA by using nucleotide bases with color dye

Image acquisition

Base calling

Reads are ready.

bcl2fastq



Big Fastq files (2-30Gb)

- Reads represent real biology.
- More reads corresponding to a transcript indicate higher abundance of that transcript.
- Reads may represent novel transcripts or novel arrangements of exons that are not present in any known reference genome.

Typical biases of Illumina sequencing

- sequencing errors
- miscalled bases
- **PCR artifacts (library preparation)**
- duplicates (due to low amounts of starting material)
- length bias
- GC bias

Sample-specific problems!

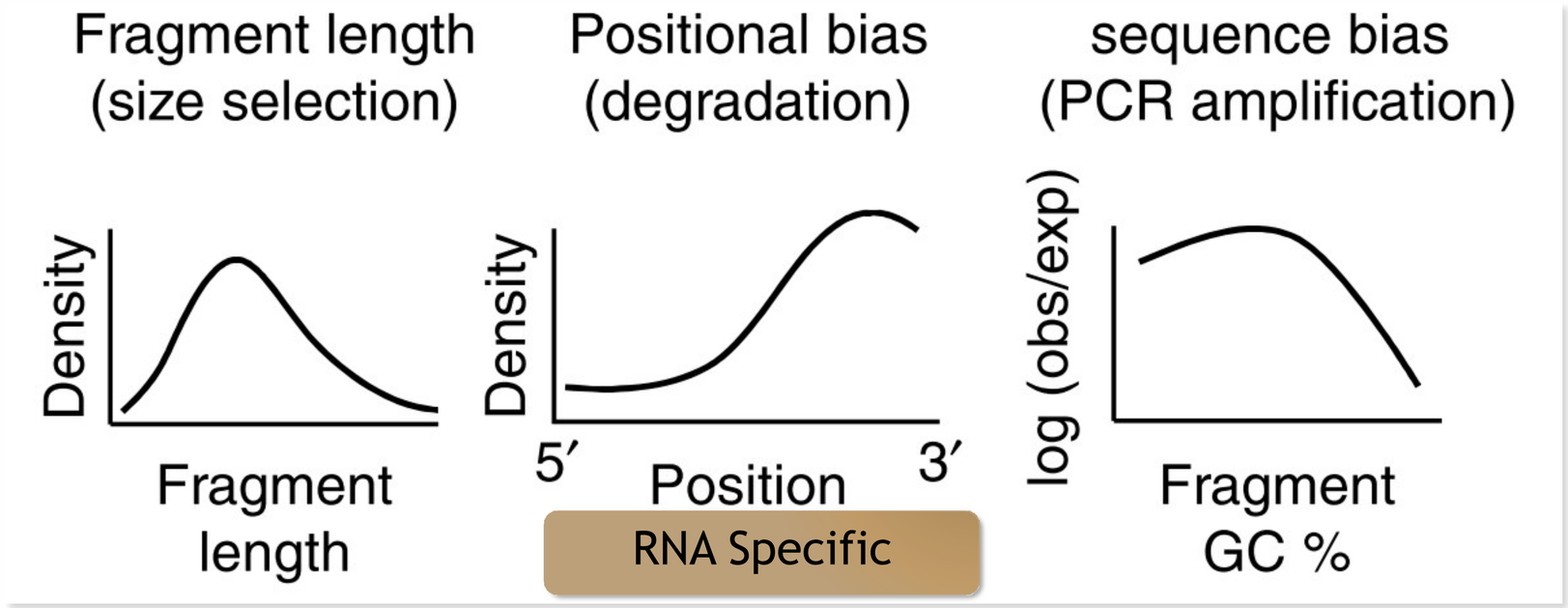
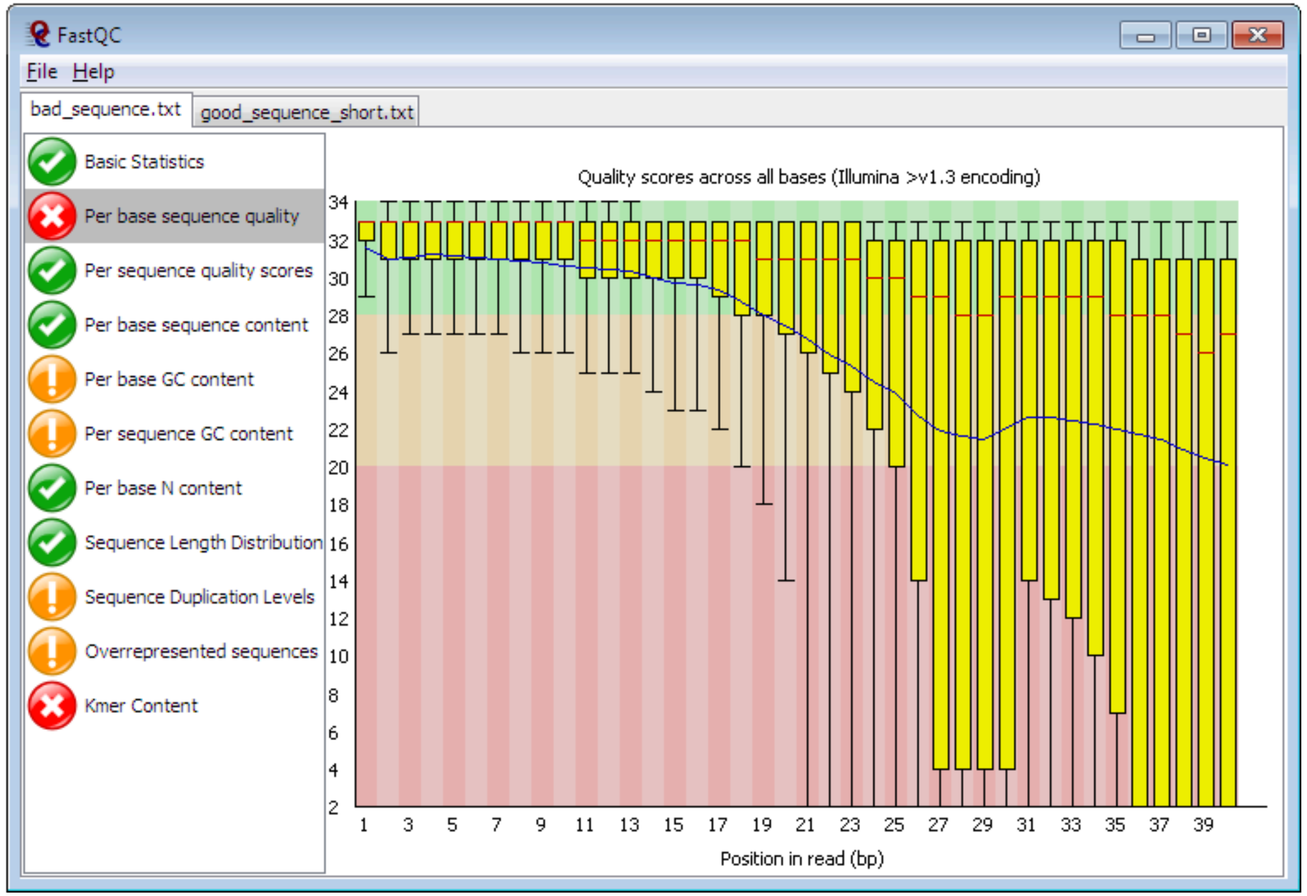


Figure from Love et al. (2016). Nat Biotech, 34(12). More details & refs in course notes (esp. Table 6).

Quality control of raw reads: FastQC

FastQC aims to provide a simple way to do some quality control checks on raw sequence data The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application



RNA-Seq Experiment Design

I want to do RNASeq on my samples, what do I need to consider?

Experimental Questions

What are my goals?

- Differential expression analysis of genes?
- Differential expression analysis of transcripts?
- Identify rare transcript isoforms?
- Identify transcript polymorphism?
- Identify non-coding RNA populations such as miRNA, lincRNA?

What are the characteristics of systems?

- Large, complex genome ? (ie. Human)
- Highly heterogeneous sample population ? (i.e. breast tumor)
- No reference genome or transcriptome ?
- High degree of alternative splicing?

What are the characteristics of available samples?

- Fresh and controlled samples (cell line, mouse, model organisms)
- Fresh/archived frozen clinical samples?
- New/Aged FFPE
- Biofluid?
- Extracellular vesicles?
- Low quantity rare samples?

How much money do I have for this project?

Experimental Design

Technical replicates

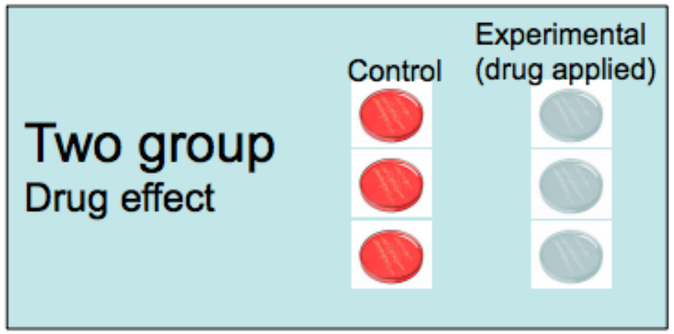
- Not necessary: RNASeq have low technical variation
  - Minimize batch effects

Biological replicates

- Not needed for novel transcript identification and transcriptome assembly
- Essential for differential expression analysis
- Difficult to estimate the minimum number
  - 3+ for cell lines
  - 5+ for inbred lines (i.e. mouse, model organisms)
  - 20+ for human samples (usually unachievable)
- Must have 3+ to perform statistical analysis

Biological System in Question

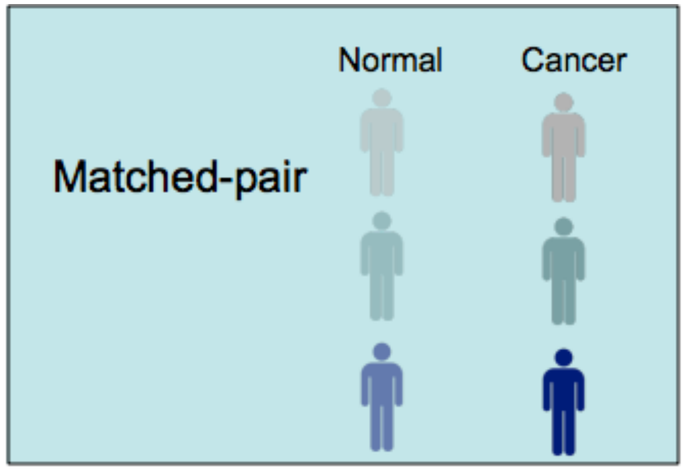
Simple Question



Examples:

- Cell line groups treated with different conditions
- Patient groups with the same disease treated with different treatment

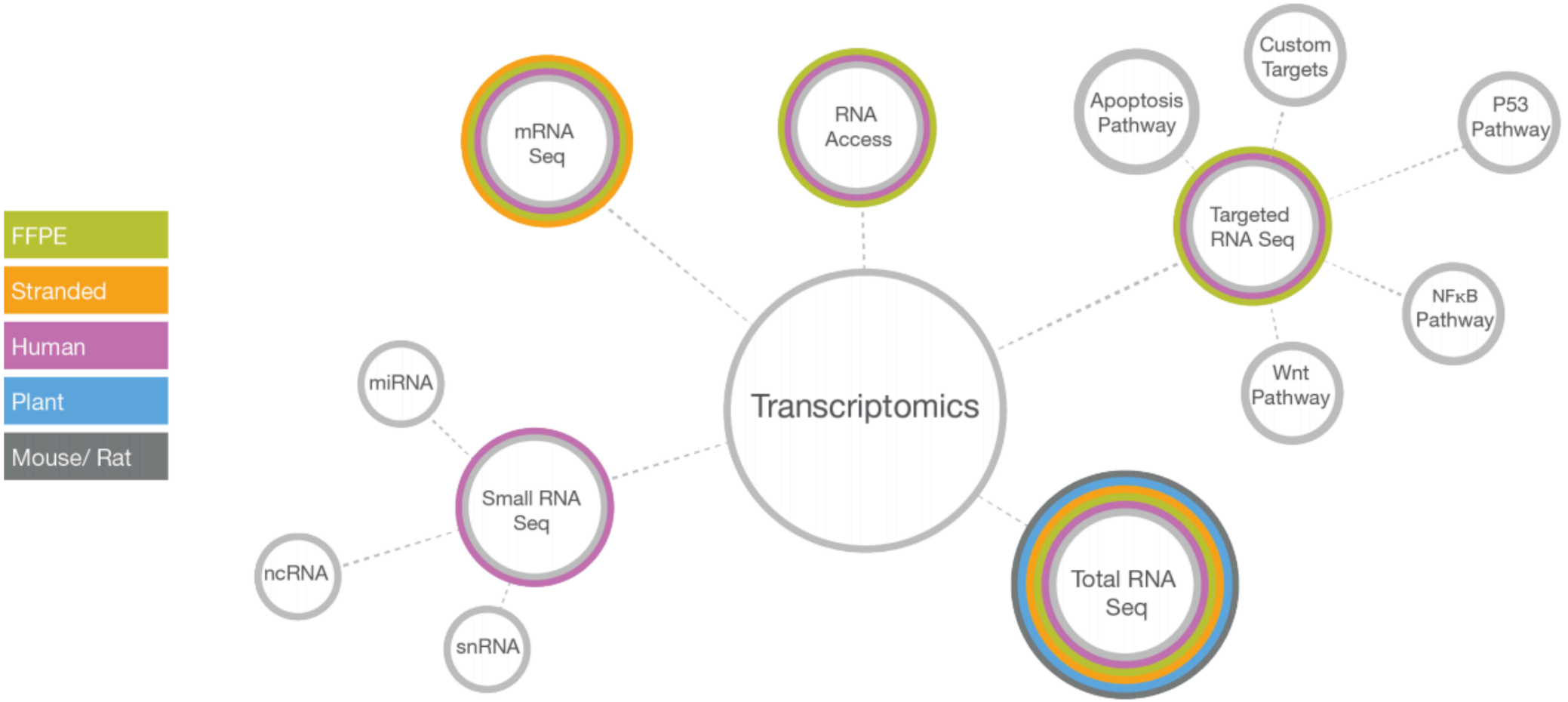
Complex Question



Examples:

- Matched patient samples from both normal and diseased tissues
- Normal and cancer samples obtained from genotypically diverse population

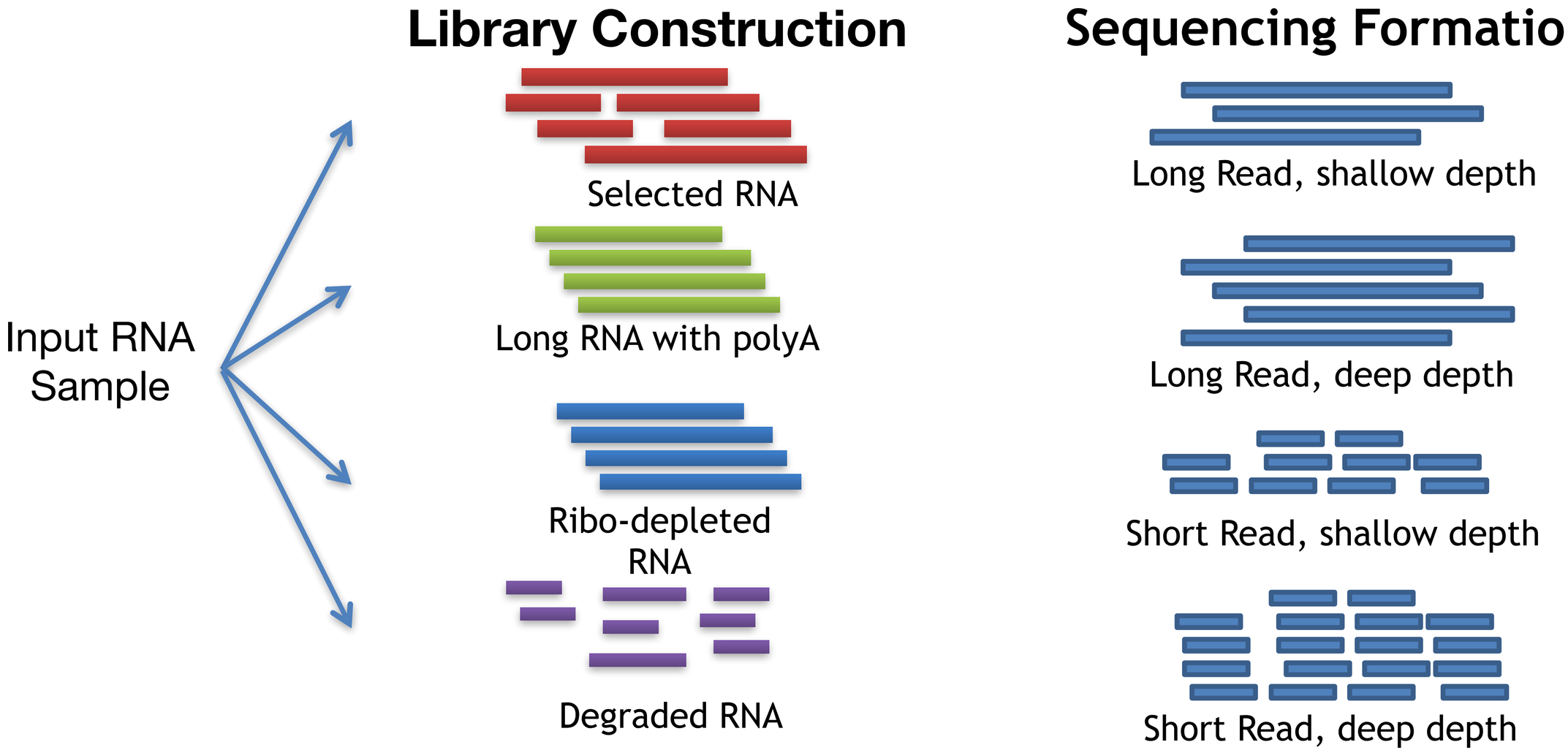
Library Type Selection



Available RNA-Seq Library preparation methods through Illumina

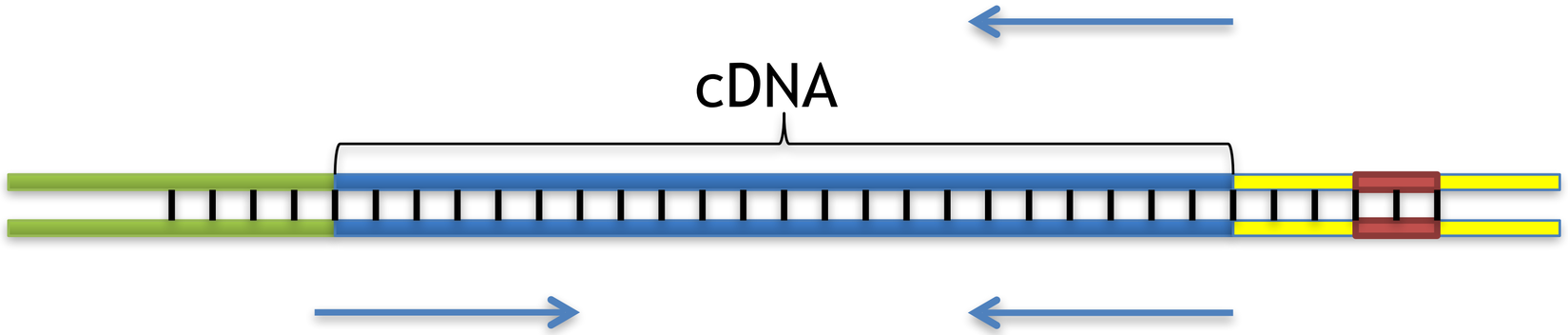
https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Experimental Questions



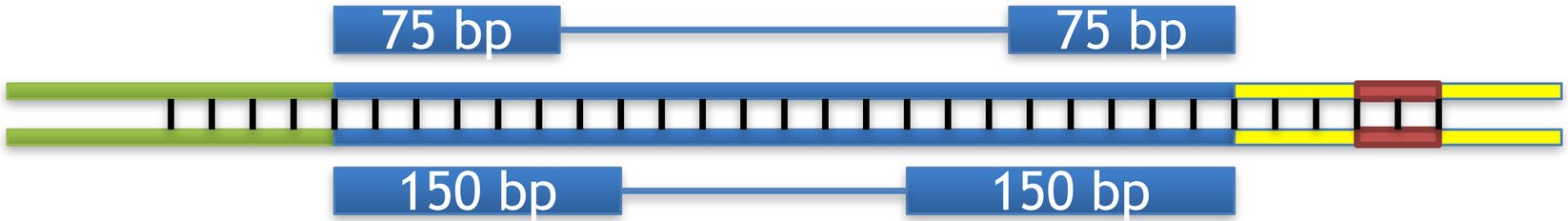
Sequencing Type

Single Read (SR) : only one end from each cDNA fragment is sequenced to generate one read per fragment. Use for gene quantification



Paired End (PE) : the cDNA fragment is sequenced from both ends to generate two reads per fragment from two directions. Use for transcript isoform quantification or identification

Sequencing Length



Longer read length provides (ie. 150bp vs 75bp):

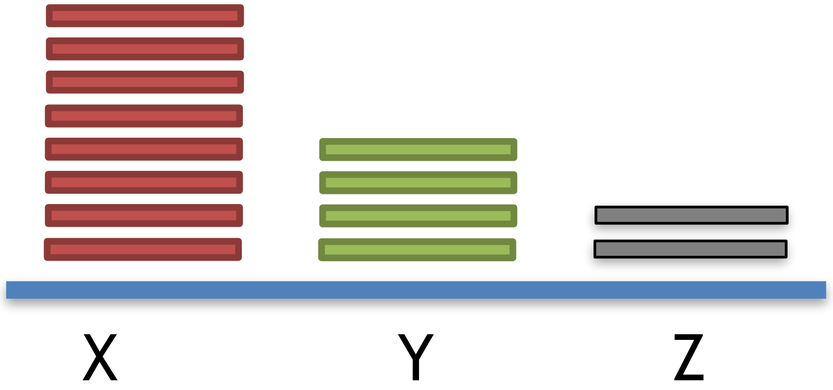
- better ability to assemble unknown transcripts
- Higher accuracy to map reads to complex regions (i.e. repeats, high polymorphic regions)
- Splice junction detection is most affected by read length
- Does not need high depth for splice junction detection

Is long read length (ie. 150bp vs 75 bp) always give better?

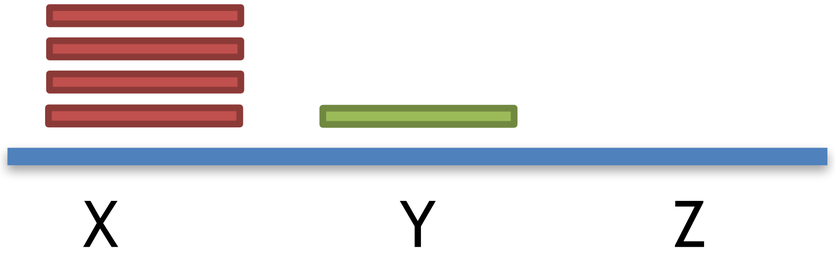
- Not necessarily
- Long reads convey minimal to no advantage for differential gene expression analysis

Impacts of Sequencing Depth

RNASeq 1, 30 million reads

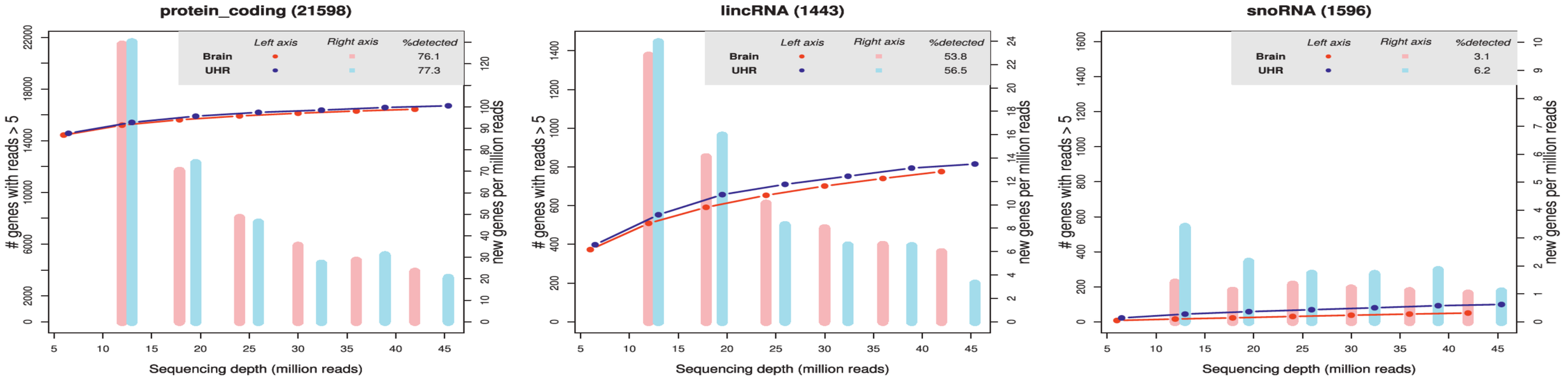


RNASeq 2, 10 million reads



- Quick means to detect more genes and transcript variants with low expression (*the more reads you sequence, the more genes you find*)
- Require logarithmic increase in depth for linear increase in gene detected

Impacts of Sequencing Depth



- Different RNA sequencing require different number of reads
- More genes are detected with higher sequencing depth
- However, the increase of detected genes reduces substantially
- Understand your sequencing system before deciding on depth
- Can always increase depth by additional sequencing on the same library

Tarazona, Gen Res.,2011.

Invest More in Replicates than Depth!

30 million reads

1.5 Fold Change

Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

The most effective way to improve detection of differential expression in low expression genes is to add more biological replicates, rather than adding more reads

Conesa et al, Genom. Bio, 2016

Major factors determine RNA-Seq experiment

Library preparation methods used

- Different RNA species selection process resulted in different noise level

Sequence length and depth

- Increasing depth improves detection but with diminishing return
- Longer read length is usually good, but may not improve results with additional expense

Experimental variability

- Technical and biological variability
- Three replicates being the ***absolute*** minimum for any inference analysis
- Invest in replica rather than depth