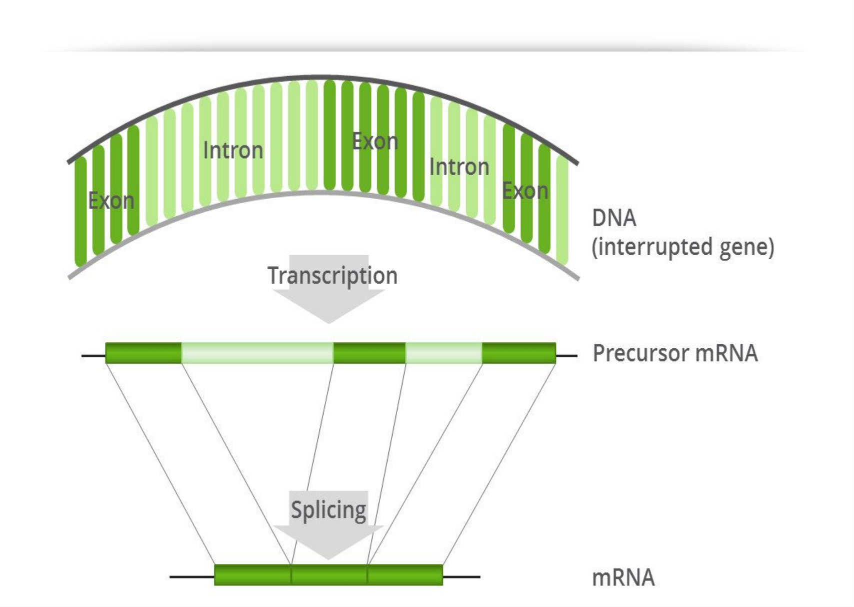
Sequence Analysis - RNA-Seq 1

The Transcriptome

- Complete set of RNA transcripts in the cell
- Many types of RNA, e.g.:
  - mRNA
  - lincRNA
  - Anti-sense
  - rRNA
  - Small molecules:
    - tRNA, snoRNA, miRNA, piRNA, etc



Transcriptome Studies

From abundance

- Gene expression
- Transcriptional regulatory networks
- Biological pathway discovery

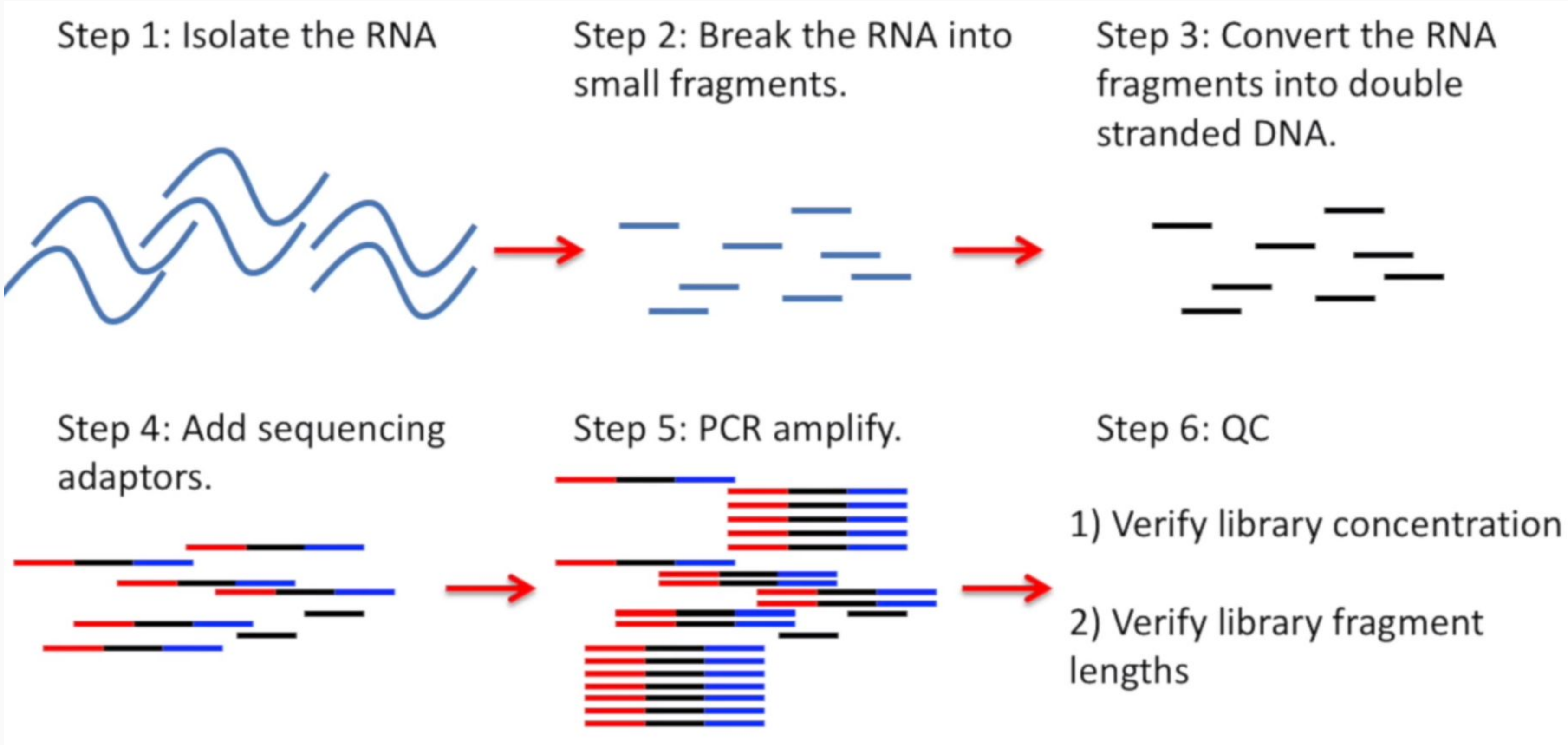
From sequence

- Alternative splicing
- Amino acid sequence
- Fusion transcripts
- RNA editing
- Gene discovery
- Coding variants

RNA-Seq

- Identify sequence of RNA molecules
- “Unbiased” - possible to sequence any molecule in sample
- Molecules sequenced in proportion to relative abundance in sample
- Most often used for gene abundance estimation

RNA-Seq Library Construction



Step 1: Isolate the RNA

Step 2: Break the RNA into small fragments.

Step 3: Convert the RNA fragments into double stranded DNA.

Step 4: Add sequencing adaptors.

Step 5: PCR amplify.

Step 6: QC

1) Verify library concentration

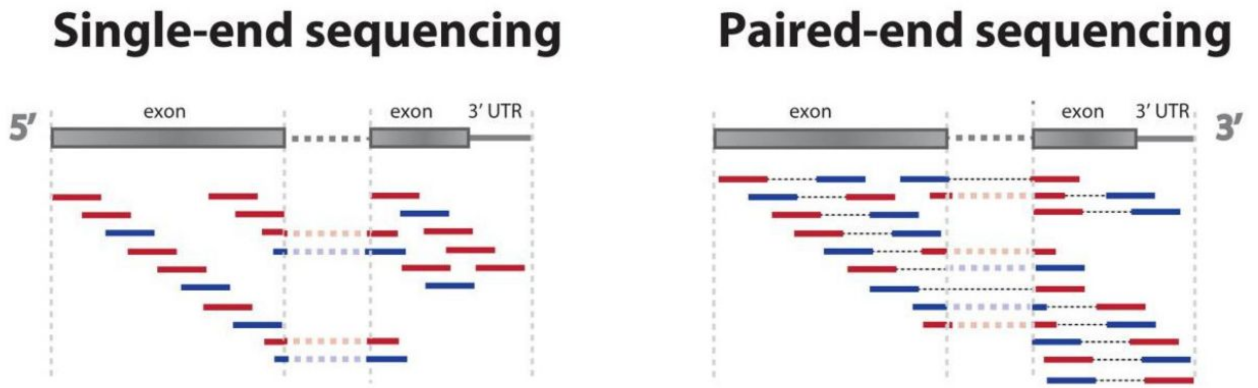
2) Verify library fragment lengths

Design Choices & Considerations

- Single vs paired end
- Read length
- Ribosomal Depletion Strategy
- Fragment length
- RNA Integrity
- Stranded vs Unstranded
- Library size
- Multiplexing

Design Choice: Single vs Paired End

- Single end vs paired end
  - 2x more distinct molecules sequenced
  - Harder to find reads spanning splice junctions
- For RNA-Seq, use paired end



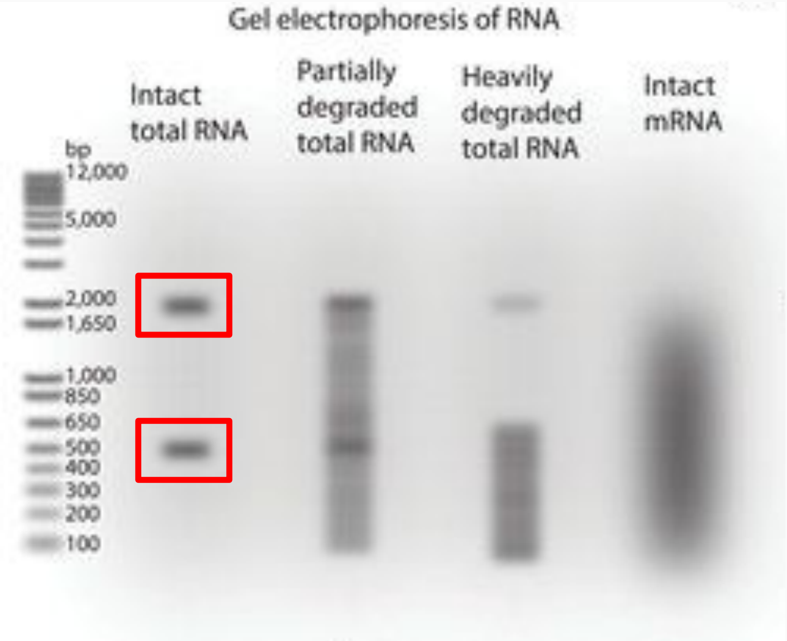
adapted from: Zhernakova et al., PLoS Genet. 2013 Jun; 9(6)

Design Choice: Read Length

- Read length determines mappability
- Longer reads:
    - more unique sequence → more uniquely mappable
    - more likely to span splice junction
- Shorter paired reads better than longer single end (why?)
- 2 x 75bp enough for hg, 2 x 150bp overkill

Design Choice: poly-A or Ribo-depletion

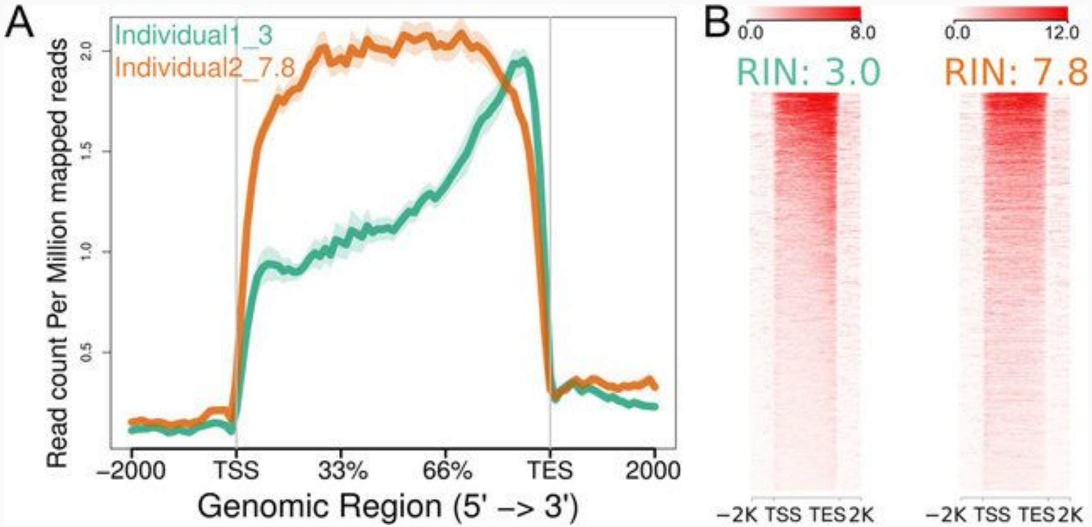
- ~95% of RNA in cell is ribosomal RNA
- 5S, 5.8S, 18S, 28S in humans
- Two removal strategies:
  - poly-A selection (positive)
    - poly-A capture
    - Only poly-A transcripts (mRNA)
  - Ribo-depletion (negative)
    - Probe-based rRNA capture
    - Leaves all other RNA sequence


Gel electrophoresis of RNA
Intact total RNA
Partially degraded total RNA
Heavily degraded total RNA
Intact mRNA
bp
12,000
5,000
2,000
1,650
1,000
850
650
500
400
300
200
100

Removing rRNA: poly-A

poly-A (mRNA-Seq)

- Enriched for mRNA (protein coding)
- Little pre-mRNA/ lincRNA/etc
- 👍Splicing analysis
- 👎Sensitive to low RIN
- 👎3’ degradation bias

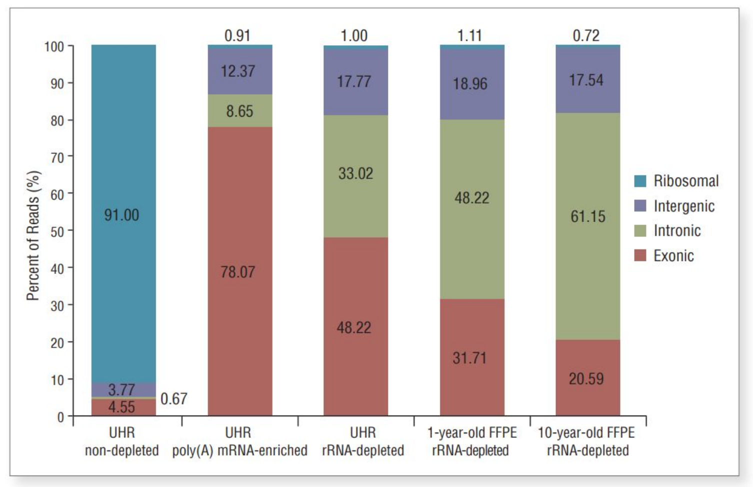


https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-284

Removing rRNA: Ribo-depletion

Ribo-depletion (RNA-Seq)

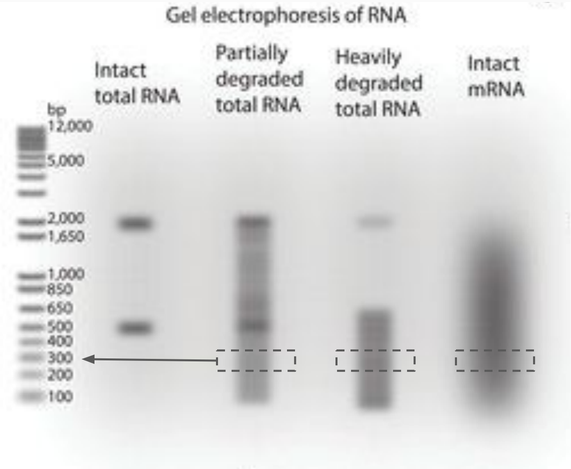
- Removes rRNA with probes
- 👍Diverse RNA sequences
- Relatively less protein coding
- Little to no 3’ bias
- Fewer spliced/exonic reads
- 👍Effective for degraded RNA
- 👎Harder to interpret protein effects

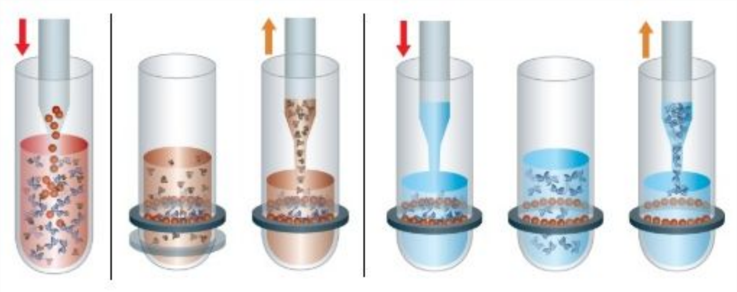


https://www.neb.com/-/media/catalog/application-notes/selective-depletion-of-abundant-rnas-to-enable-transcriptome-e6310.pdf?rev=214e1d46d2834c12876fa0867ea5197d

Large RNA Size Selection

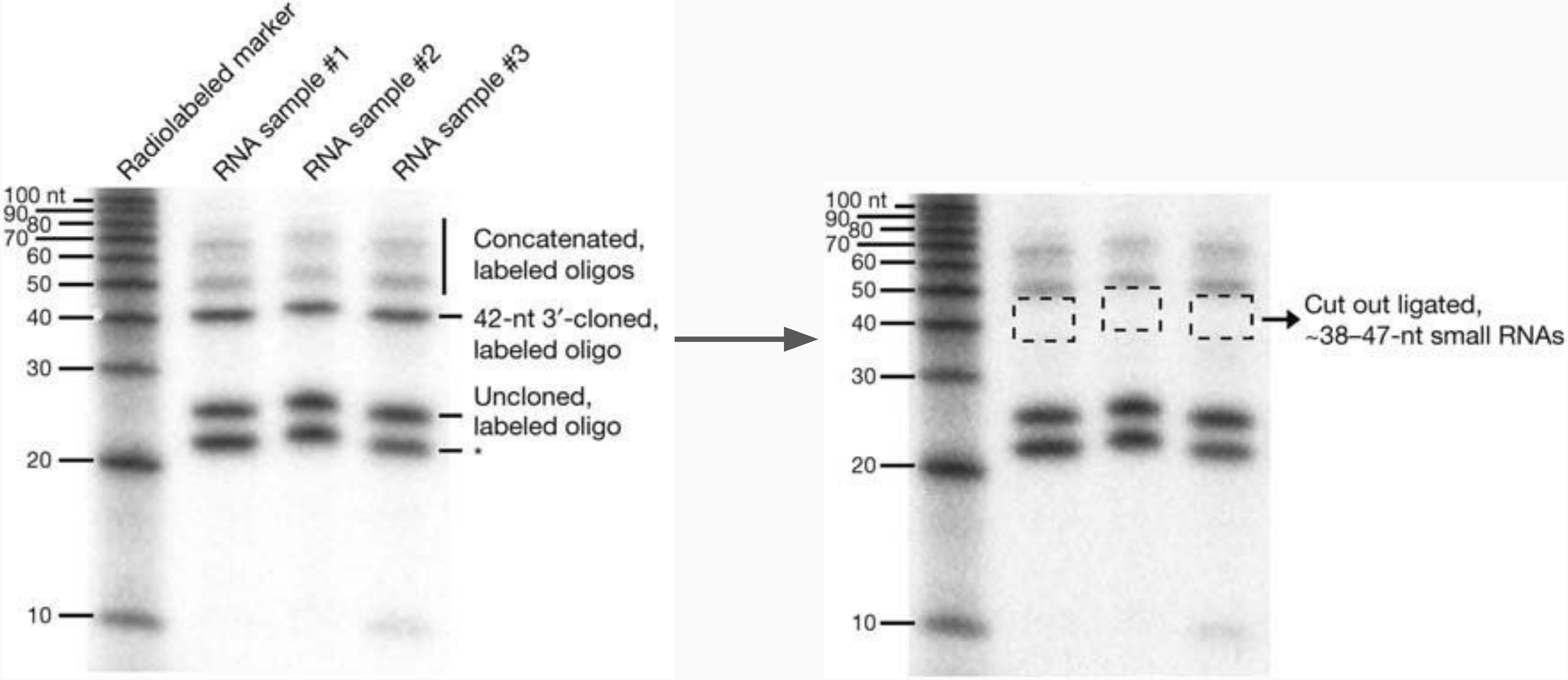
- Gel cut (old method):
  - Size select with gel electrophoresis
  - Fragment size distribution may indicate RNA quality
  - Select ~300nt fragments by gel cut
- SPRI Beads (current method)


Gel electrophoresis of RNA
Intact total RNA
Partially degraded total RNA
Heavily degraded total RNA
Intact mRNA
bp
12,000
5,000
2,000
1,650
1,000
850
650
500
400
300
200
100



https://www.researchgate.net/publication/280870031_Informatics_for_RNA_Sequencing_A_Web_Resource_for_Analysis_on_the_Cloud/figures?lo=1

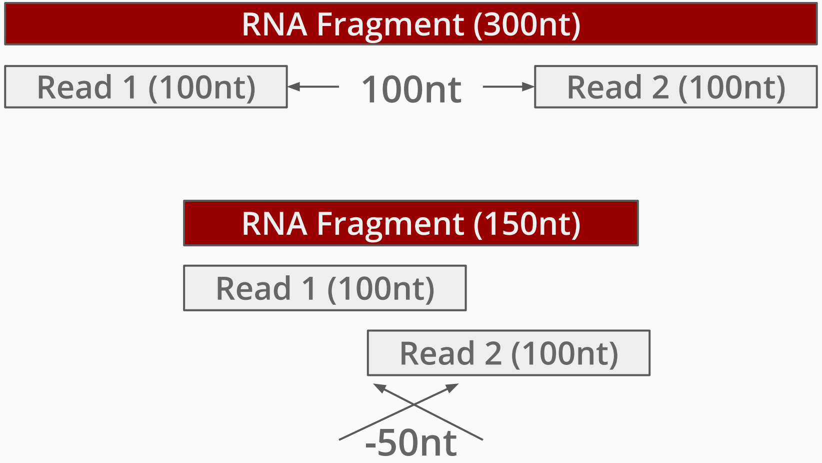
Small RNA Size Selection



https://www.researchgate.net/publication/231742684_Preparation_of_Small_RNA_Libraries_for_High-Throughput_Sequencing/figures

Batch effect: Fragment Length Distribution

- Inner mate distance: unsequenced length between read pair



RNA Fragment (300nt)

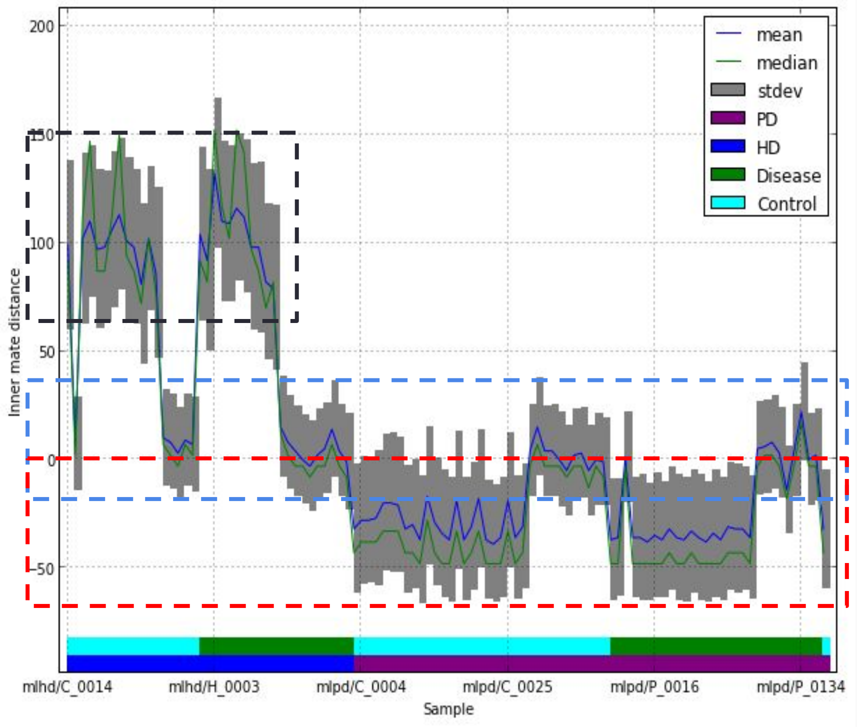
Read 1 (100nt) ← 100nt → Read 2 (100nt)

RNA Fragment (150nt)

Read 1 (100nt)

Read 2 (100nt)

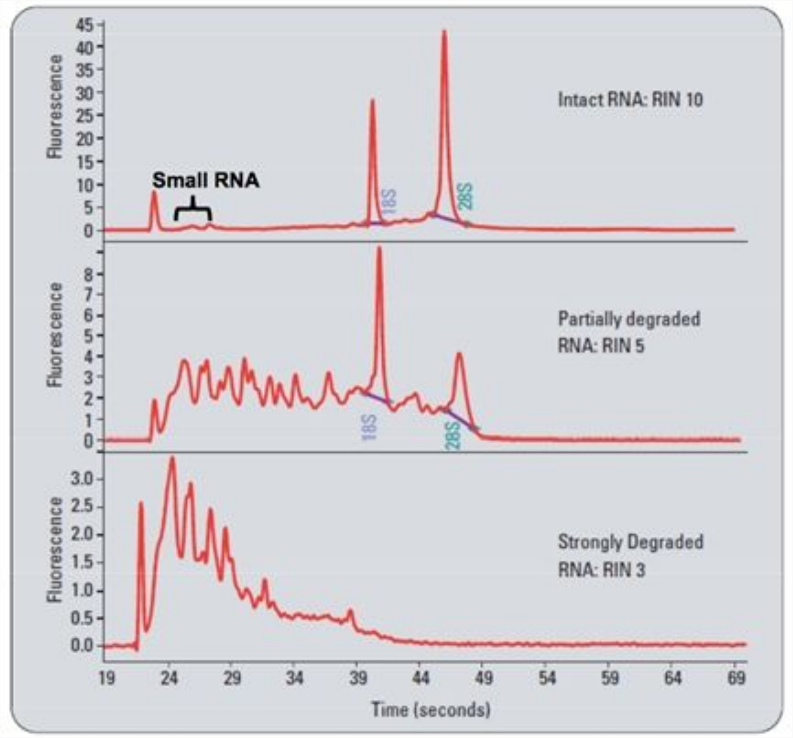
-50nt



Inner Mate Distance

Design Consideration: RIN

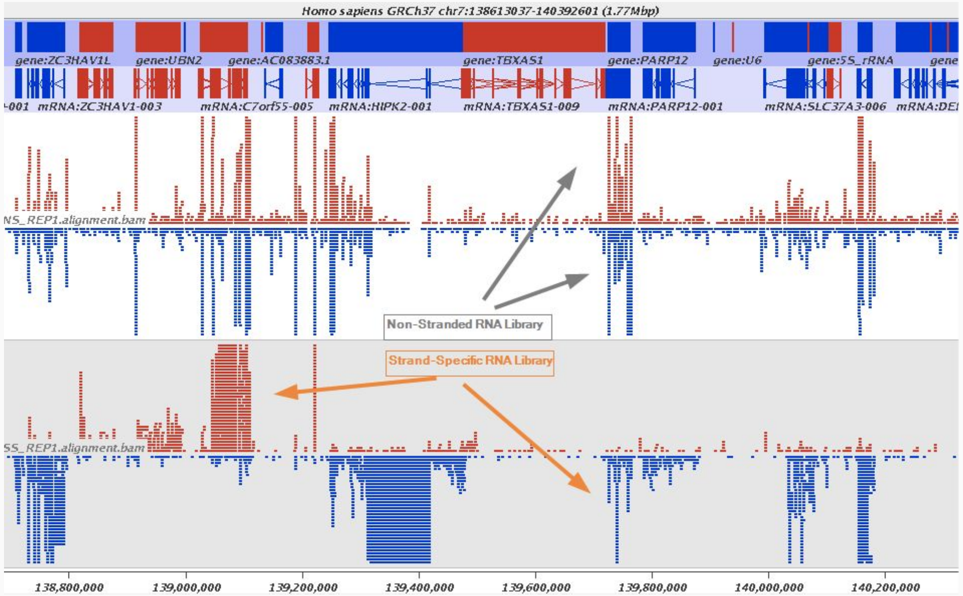
- RNA Integrity Number
- Measurement of RNA quality
- 10 - best, 0 - worst
- Transcripts
    - degrade 5’ → 3’
    - At different rates!
- Rules of thumb
    - >8 👍
    - 6-8 is ok if necessary
    - 3-6 is ok only if very necessary
    - <3 👎



https://infravec2.eu/rna_seq/

Design Choice: Stranded vs Unstranded

- **Stranded** libraries maintain strand of molecule in reads
- **Unstranded** do not
- Important to resolve:
  - ○ Bi-directional transcription
  - ○ Anti-sense transcripts
  - ○ Overlapping genes
- Modern RNA-Seq library prep kits are stranded

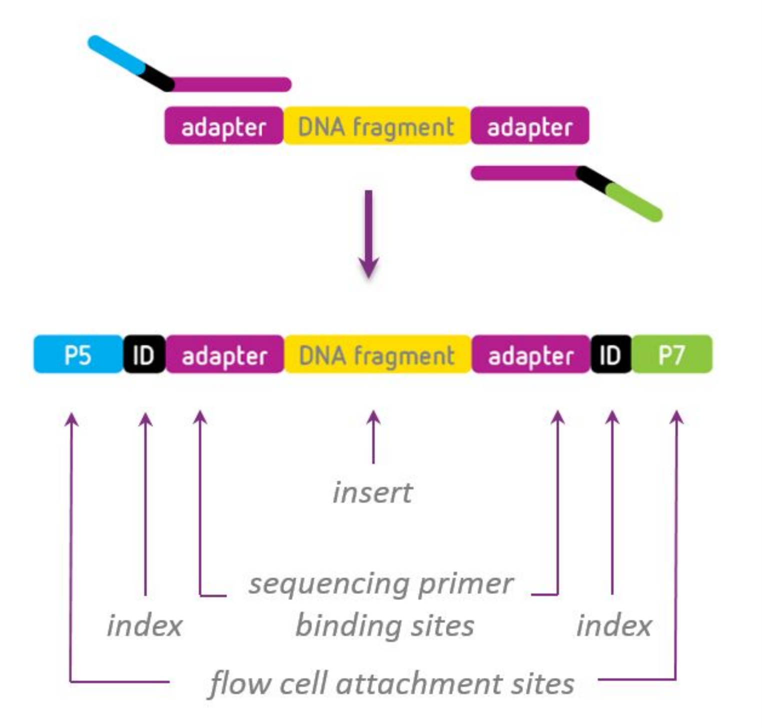


Design Choice: Library Size & Multiplexing

- **Library size**: # of reads per sample
- Depending on who you ask, a read is:
  - A RNA fragment (same for single/paired end)
  - One FASTQ record (not same for single/paired end)
- Library size is *target*, # reads will vary
- Rules of thumb for human transcriptome:
  - poly-A: 30M for expression, 80M alternative splicing
  - ribo: 50M for expression, 100M alternative splicing

Design Choice: Multiplexing

- Add unique barcode (index) to each sample library
- Multiplexed samples pooled and sequenced together → avoid lane batch effects
- Data will usually be demultiplexed for you



Design Choices & Recommendations

- Fragment length: ~300nt (large RNA)
- RNA Integrity: >8👍, >6 ok, >3 if need be
- Ribosomal Depletion Strategy: depends
- Single vs paired end: paired
- Read length: 2 x 75bp
- Stranded vs Unstranded: stranded
- Library size: poly-A 30-80M, ribo 50-100M
- Multiplexing