RNA-seq Introduction

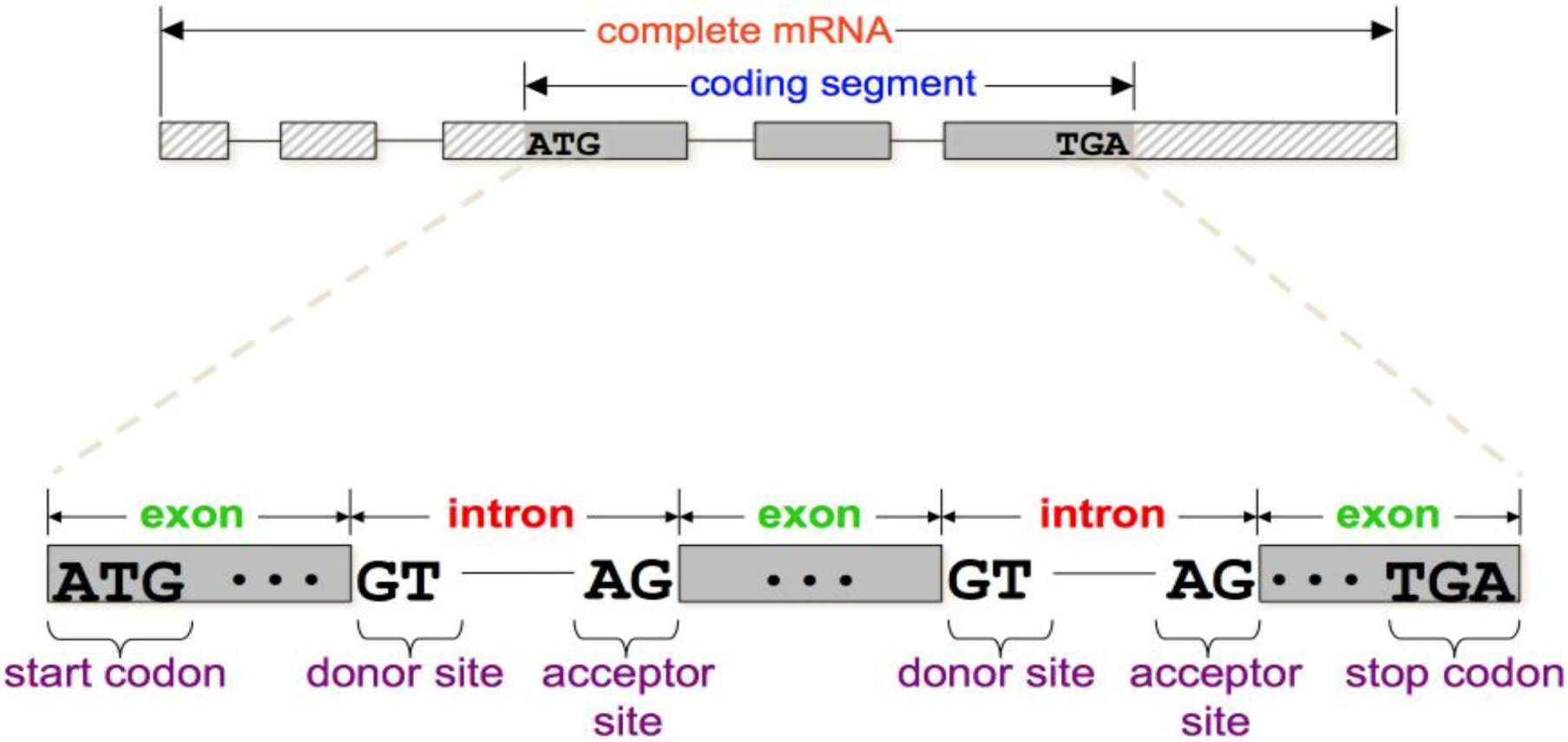
Mikhail Dozmorov

2021-03-03

Section 1

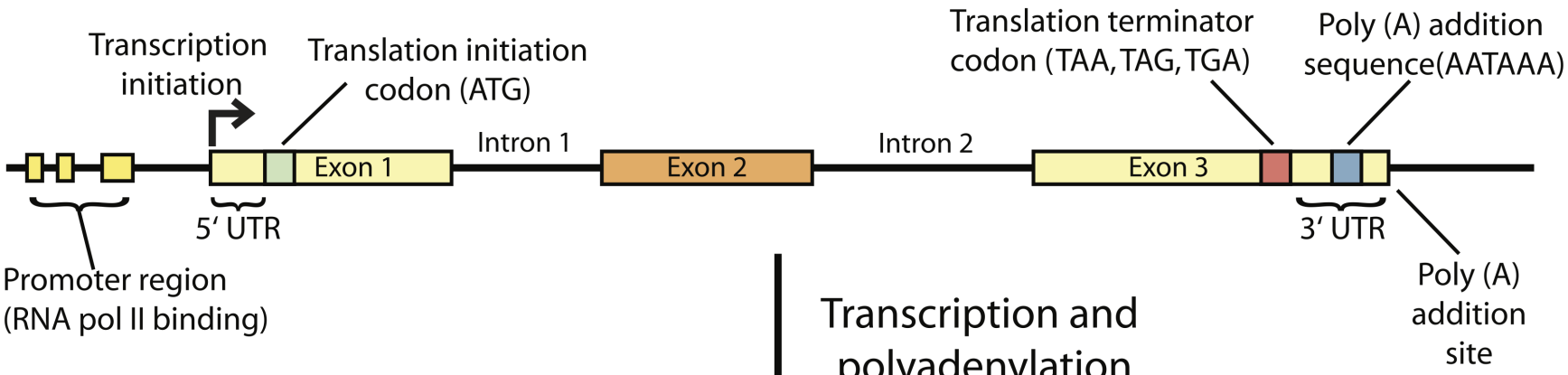
Introduction

Eukaryotic gene structure



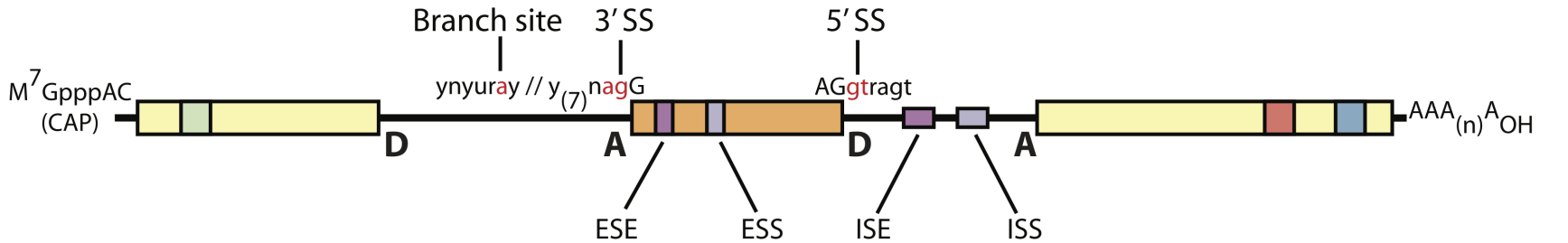
Gene expression

Double-stranded genomic DNA template



Transcription and polyadenylation

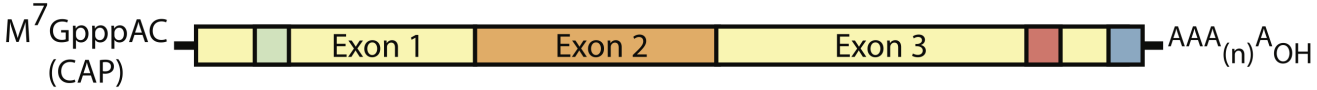
Single-stranded pre-mRNA (nuclear RNA)



RNA processing

Gene expression

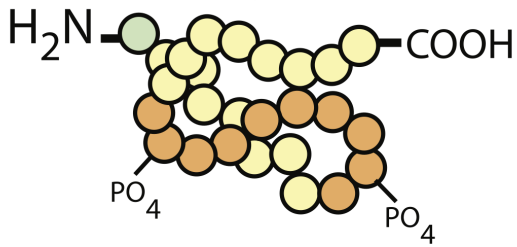
Mature mRNA



Export to cytoplasm and translation

Protein (amino acid sequence)

Folding, posttranslational modification, subcellular localization, etc.



What is RNA sequencing?

- Massive parallel sequencing to **characterize and quantify transcriptomes** (all actively transcribed genes)
- Detection of **differential gene expression**
- **Transcriptome reconstruction**, identification of **new transcripts**
- Detection of **alternative splicing events**
- Detection of **structural variants**, e.g., fusion transcripts
- **Allele-specific** gene expression measurements
- **Mutation analysis** – presence of genomic mutations and their effect on gene expression

RNA-seq analysis techniques

Sequencing technologies

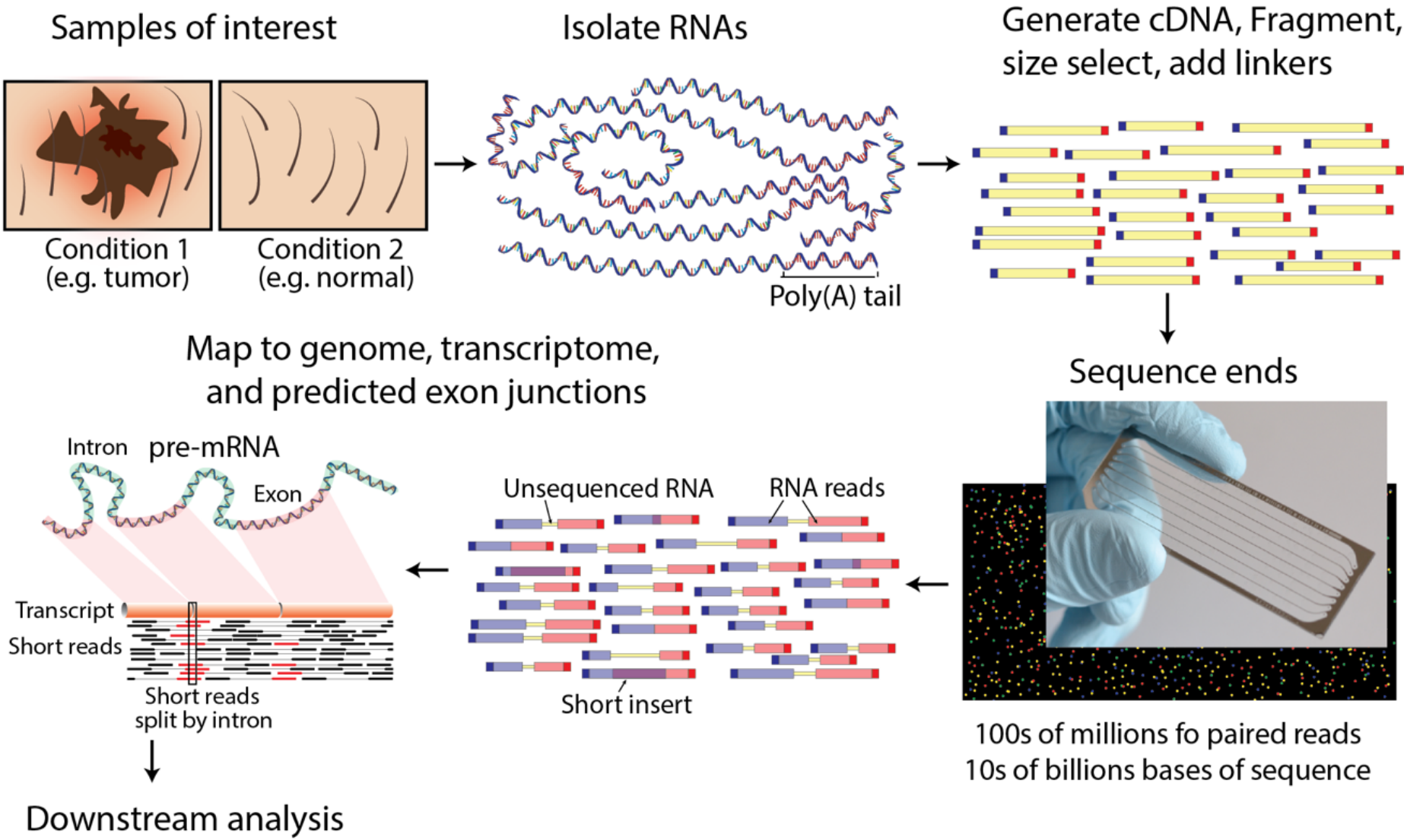
Commercially available

- **Illumina/Solexa** - short reads, sequencing-by-synthesis
- **Life Technologies Ion Torrent/Proton** - short reads, Ion Semiconductor sequencing
- **Pacific Biosciences** - long reads, Single Molecule Real Time sequencing

Experimental

- **Nanopore sequencing** - continuous sequencing (very long reads), fluctuations of the ionic current from nucleotides passing through the nanopore

Overview of RNA sequencing technology



Source: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

RNA-Seq Limitations

Quantitation influenced by many confounding factors

- "Sequenceability" - varying across genomic regions, local GC content and structure-related
- Varying length of gene transcripts and exons
- Bias in read ends due to reverse transcription, subtle but consistent
- Varying extent of PCR amplification artifacts
- Effect of RNA degradation in the real world
- Computational bias in aligning reads to genome due to aligners

RNA-Seq Limitations

SNP discovery in RNA-seq is more challenging than in DNA

- Varying levels of coverage depth
- False discovery around splicing junctions due to incorrect mapping

De novo assembly of transcripts without genome sequence:
computationally intensive but possible, technical improvements will help

- Longer read length
- Lower error rate
- More uniform nucleotide coverage of transcripts - more equalized transcript abundance

Section 2

Library preparation

Library preparation steps

- **RNA isolation and QC**, to extract RNA relevant to the experimental question
- **Fragmentation**, to recover short reads across full length of long genes
- **Size selection**, suitable for RNA sequencing. 300-500bp - mRNA, 20-150bp - small/miRNA
- **Amplification**, typically by PCR. Up to $0.5 - 10ng$ of RNA
- **Library normalization/Exome capture**
- **Barcoding** and **multiplexing**
- Optionally, add **External RNA Control Consortium (ERCC) spike-in controls**
- **Single** or **paired end** sequencing. The latter is preferrable for the *de novo* transcript discovery or isoform expression analysis

Sample prepatation and library construction strategies:
http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s005

RNA isolation

- **Ribosomal RNA (rRNA) depletion**
  - $0.1 - 1\mu g$ original total RNA (One cell contains ~10 picogram of total RNA)
  - rRNAs constitute over 90 % of total RNA in the cell, leaving the 1–2 % comprising messenger RNA (mRNA) that we are normally interested in (One cell contains ~0.1 picogram mRNA)
  - Enriches for mRNA + long noncoding RNA.
  - Hybridization to bead-bound rRNA probes

RNA isolation

- **Poly(A) selection (for eukaryotes only)**
  - Enrich for mRNA.
  - Hybridization to oligo-dT beads
- **Small RNA extraction**
  - Specific kits required to retain small RNAs
  - Optionally, size-selection by gel

Description of RNA-seq library enrichment strategies:
http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s006

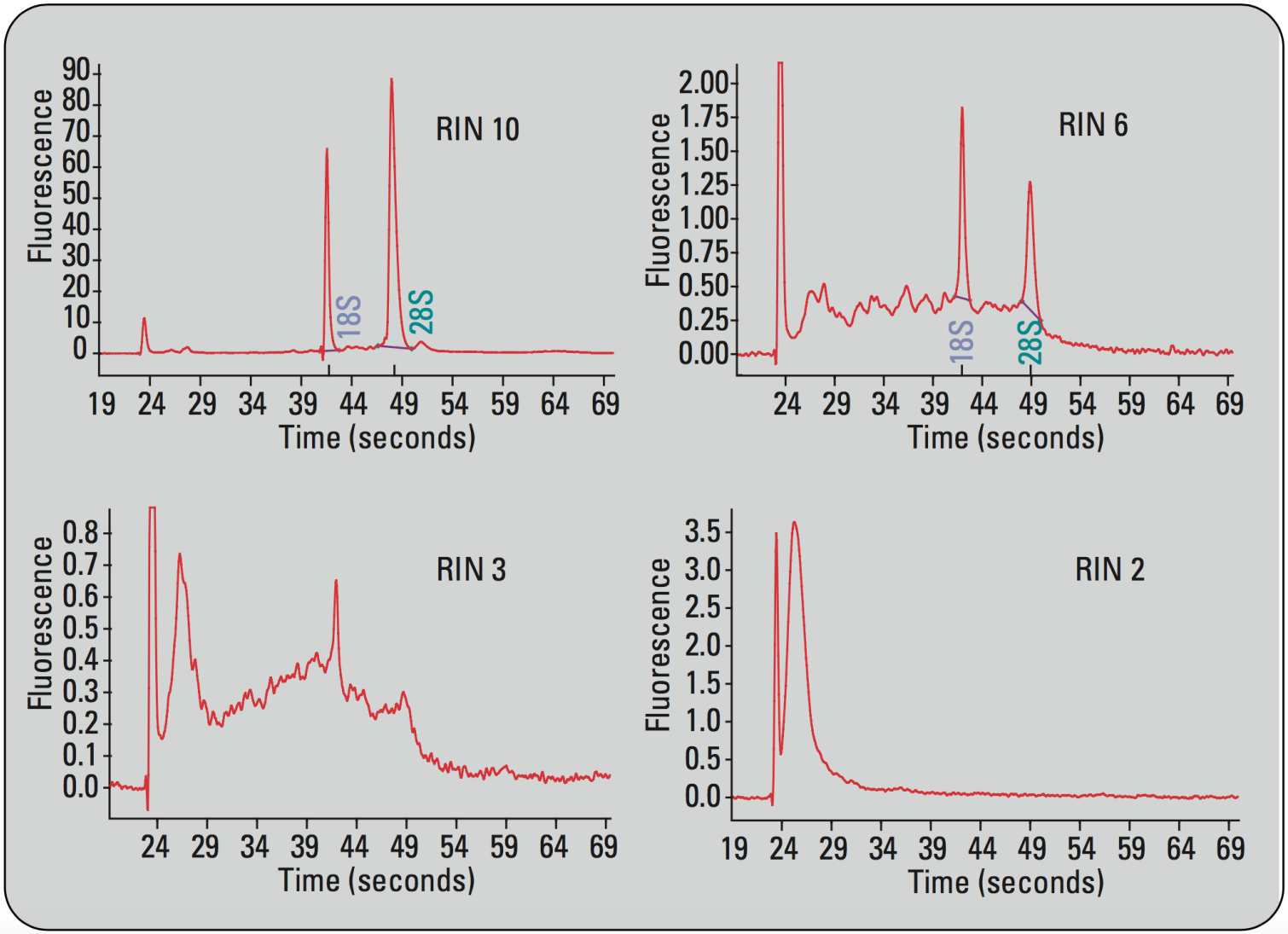
Poly-A selection or ribosome depletion protocol?

- Poly-A excels at gene quantification for classification/prediction purposes, better represents total RNA content
- Ribosomal depletion - more noncoding RNAs, better alignment of reads, more gene fusion events
- Overall, comparable performance

Detailed comparison of RNA-seq library construction protocols:
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4039-1

RNA quality

Agilent 2100 bioanalyzer. RIN - RNA integrity number (should be >7)



Unstranded vs. Strand-specific library

- **Unstranded**: Random hexamer priming to reverse-transcribe mRNA
- **Stranded**: dUTP method - incorporating UTP nucleotides during the second cDNA synthesis, followed by digestion of the strand containing dUTP

Strand-related settings for RNA-seq tools:
http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s007}

Unstranded vs. Strand-specific library