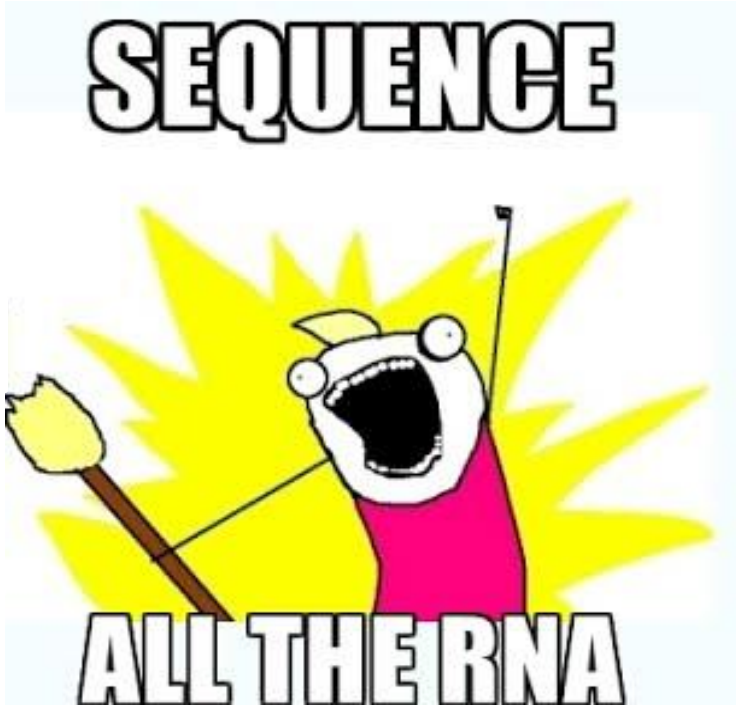
An Introduction to RNA-Seq

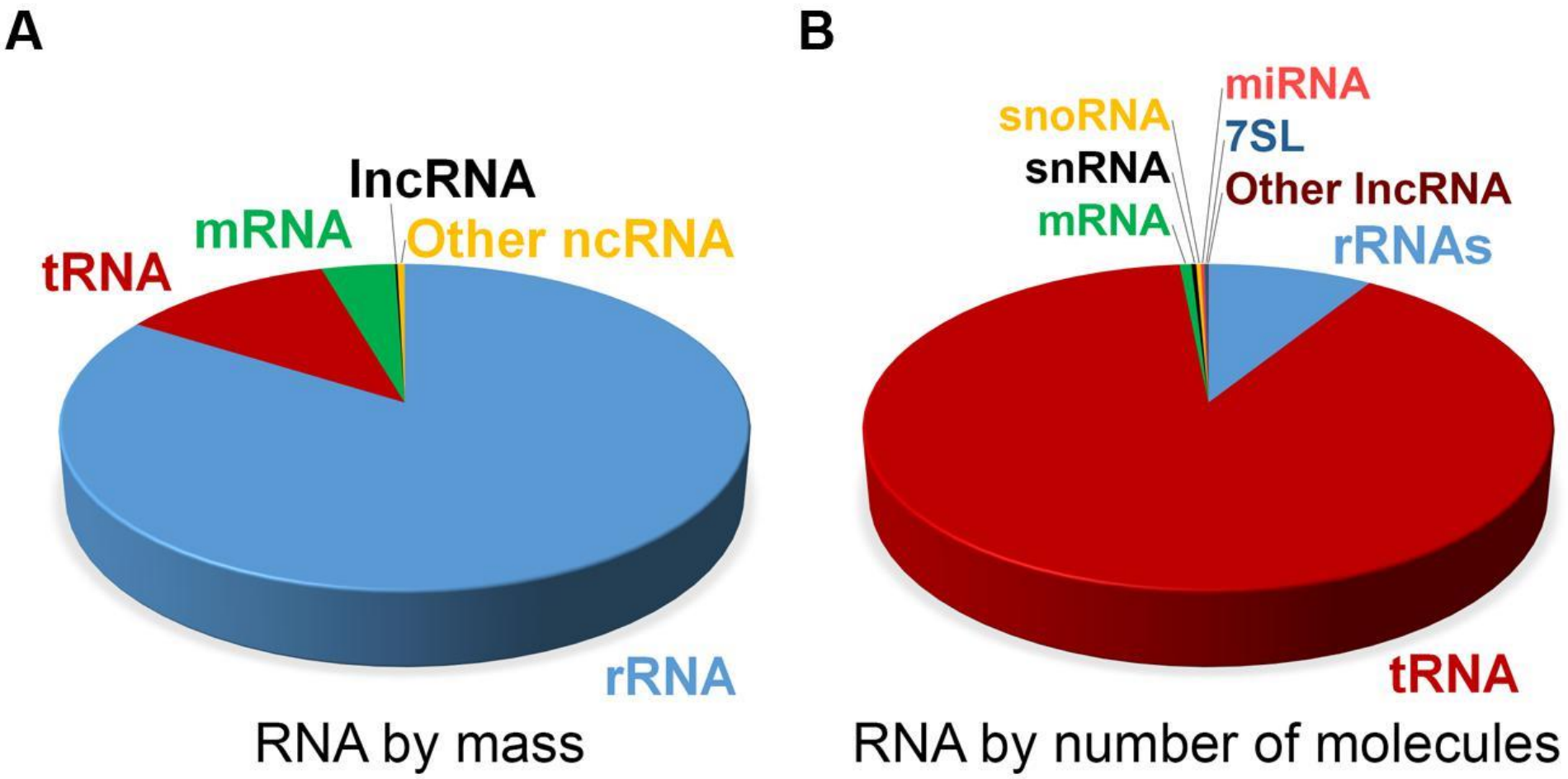
Daniel P Depledge, Ph.D



Overview

- **The (very) basics of RNA biology**
- **The many applications of RNA-Seq**
- **A (long!) list of considerations when undertaking RNA-seq**
- **Strategies for aligning RNA-Seq data**
- **Counting genes and transcripts**

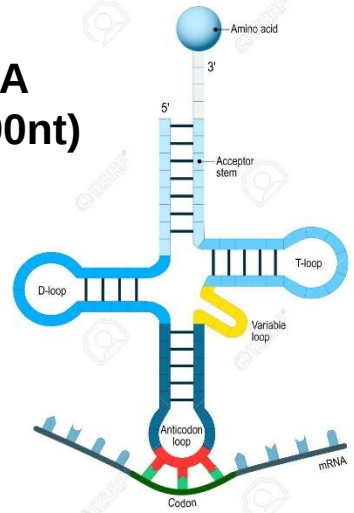
The makeup of RNA

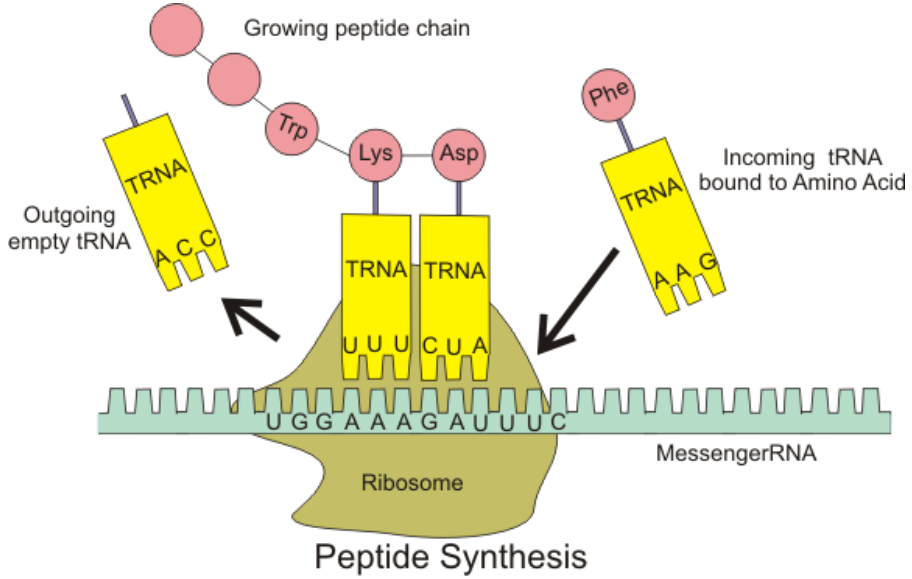


Palazzo & Lee, 2015

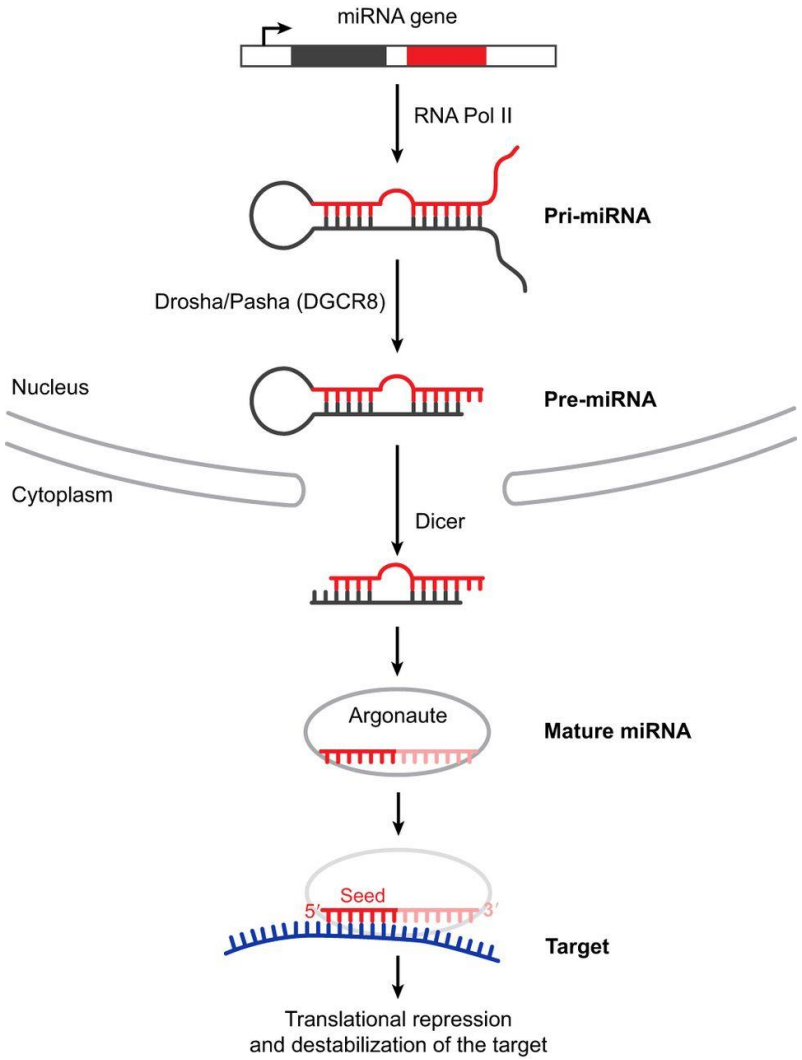
tRNAs and miRNAs

tRNA
(76 – 90nt)



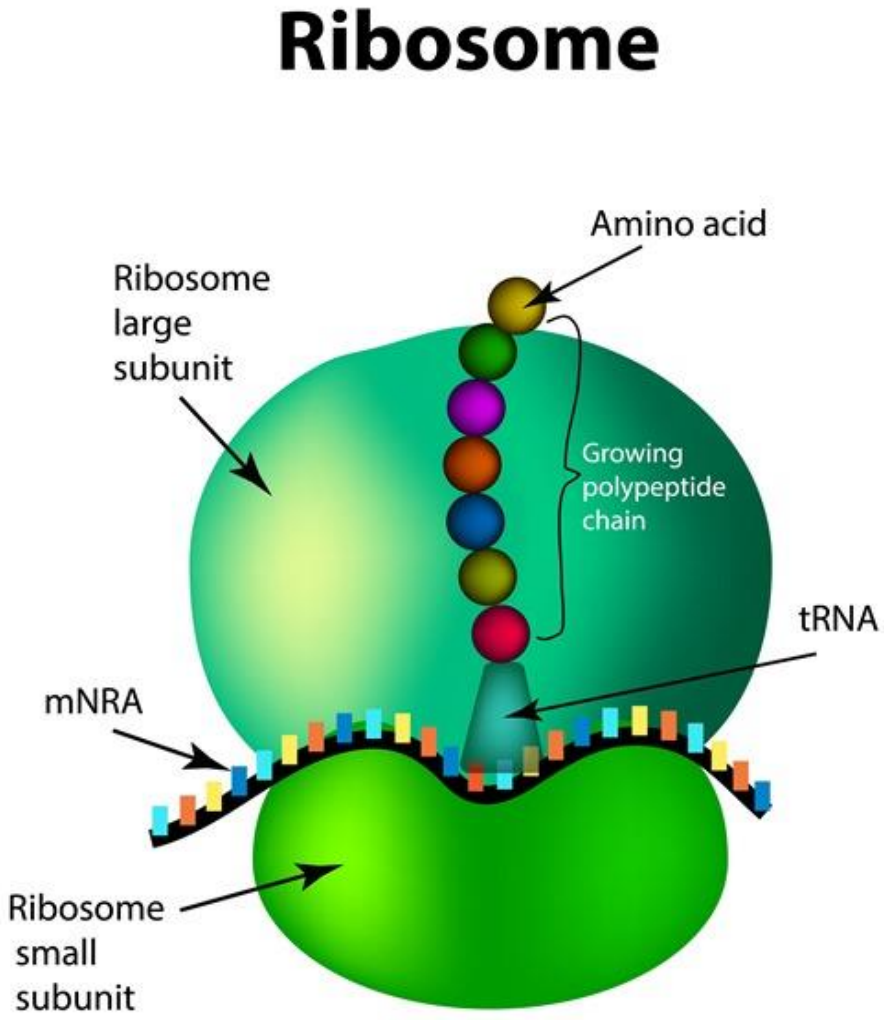


miRNA
(19 – 24nt)



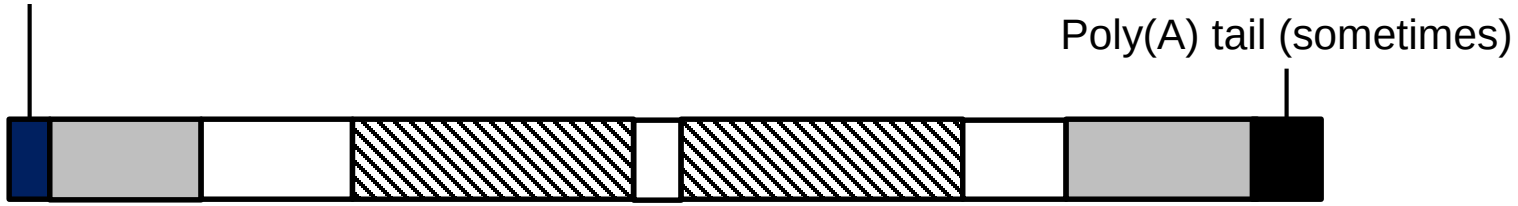
Ribosomal rRNA

- Ribosomes translate mRNAs into proteins
- Consists of protein (40% mass) and two ribosomal RNAs (60% mass
  - Small rRNA subunit
  - Large rRNA subunit
- rRNAs are found in all walks of life
  - Hugely important in studies of evolutionary biology
- 16s rRNA sequencing often used for bacterial metagenomics



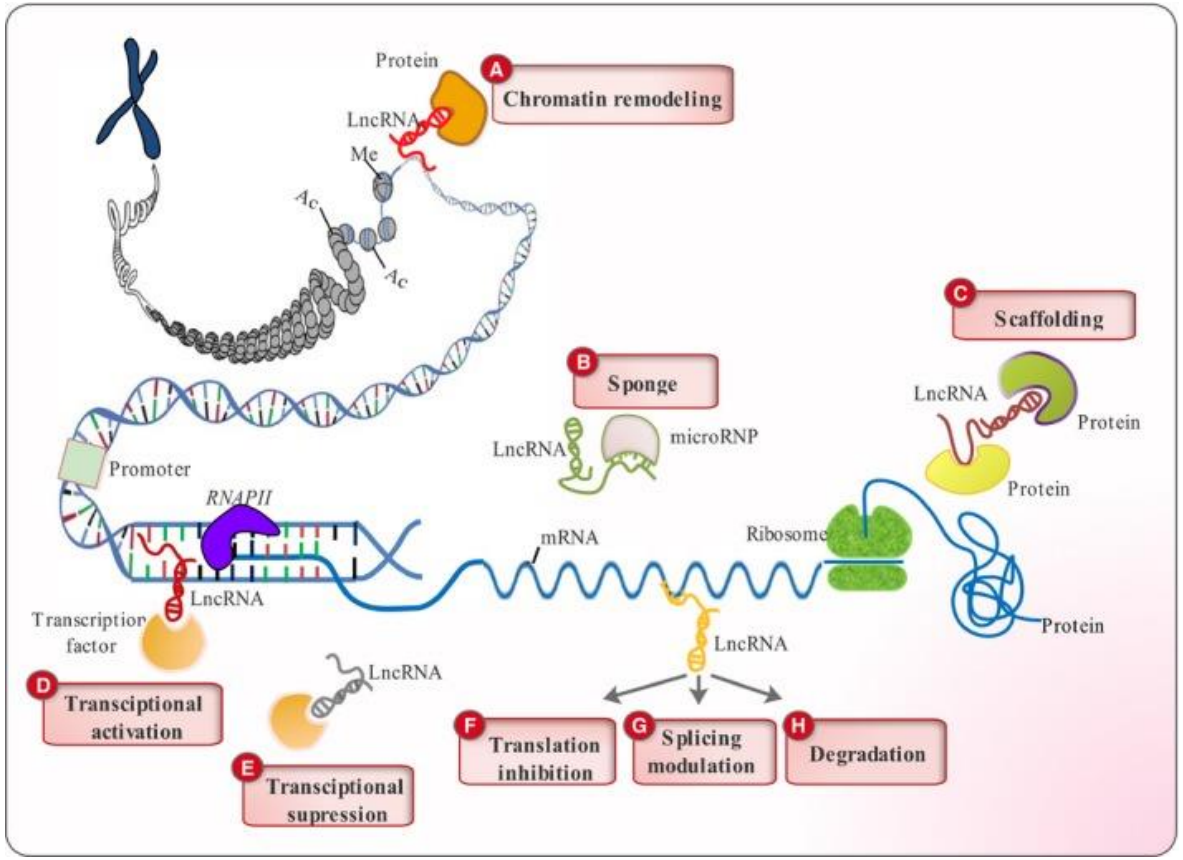
lncRNAs

Cap - Protects against degradation



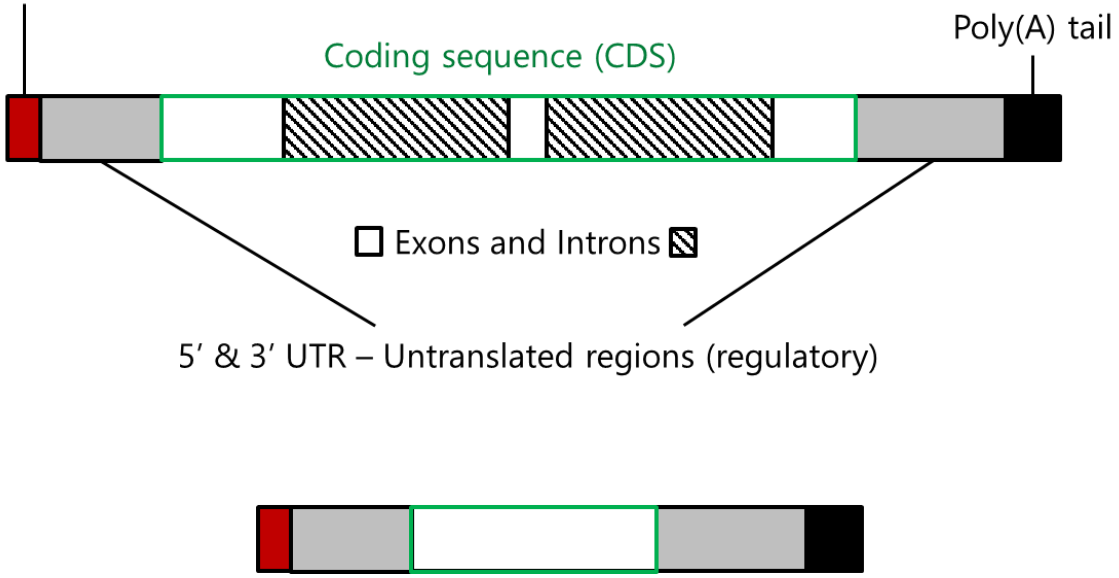
Exons and Introns

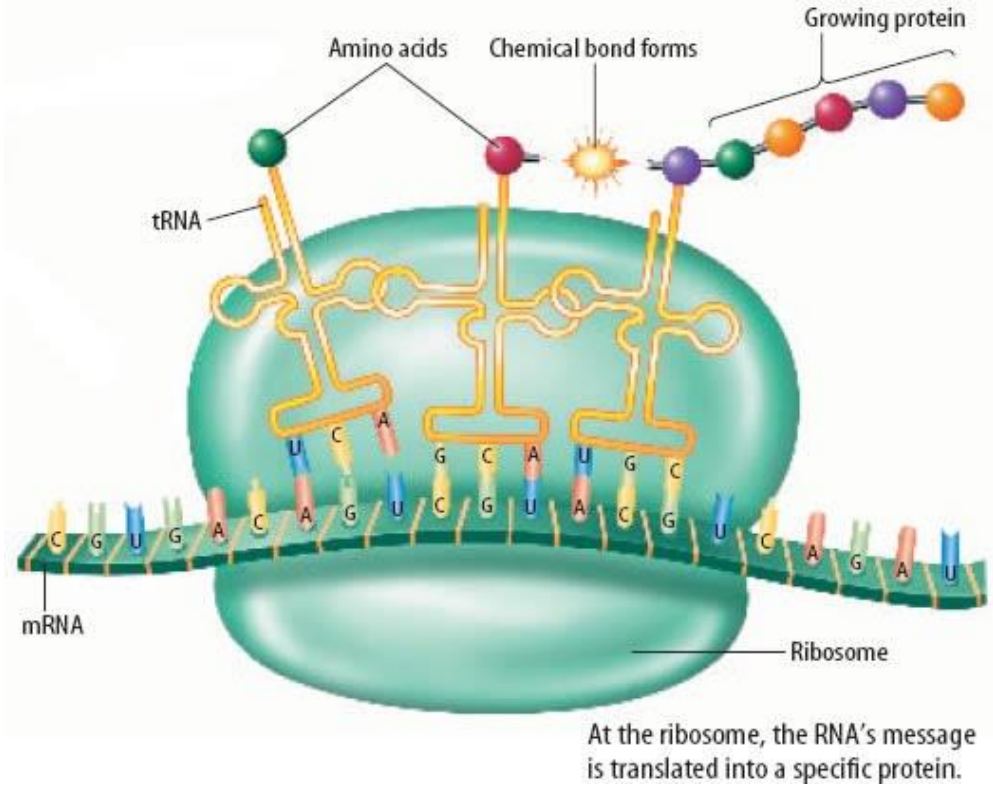
- lncRNAs regulate many aspects of transcription and translation
- Chromatin remodeling
- miRNA sponges
- Scaffolding
- linear or circular / lariats



mRNAs

Cap - Protects against degradation
+ required for translation





At the ribosome, the RNA's message is translated into a specific protein.

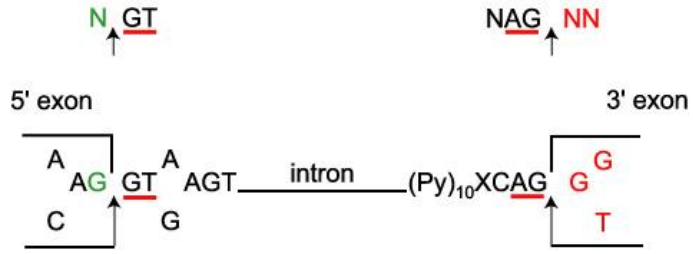
Splicing

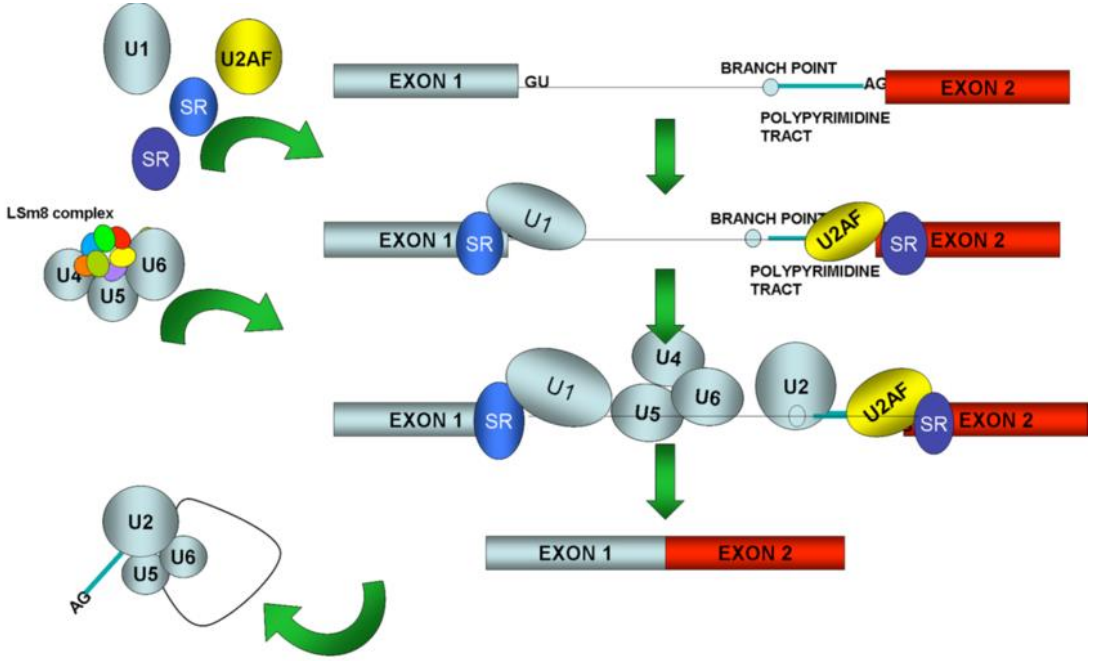
- A co/post-transcriptional process that generates mature RNAs from immature pre-RNAs

- Functionally removes and degrades (usually) intron sequences
  - Introns usually made into rapidly-degraded **lariats**
  - May also become persistent **circular RNAs**

- Catalyzed by spliceosome complexes
  - Major spliceosome targets GU | AG pairs (common introns)
  - Minor spliceosome targets other pairs (rare introns)

- Disruption of splicing (e.g. by viruses) is a popular strategy to disable host cell defenses





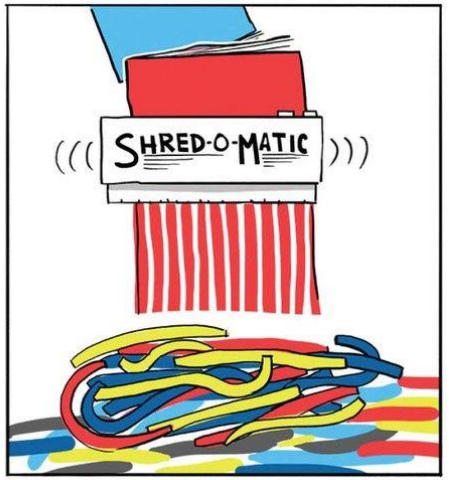
Why perform RNA-Seq?

Utilities of RNA sequencing

- RNA expression analysis
- Transcript discovery
- Pathogen discovery
- Epitranscriptomics
- RNA:RNA interactions
- RNA:DNA interactions
- RNA:Protein interactions

Specific considerations for RNA sequencing

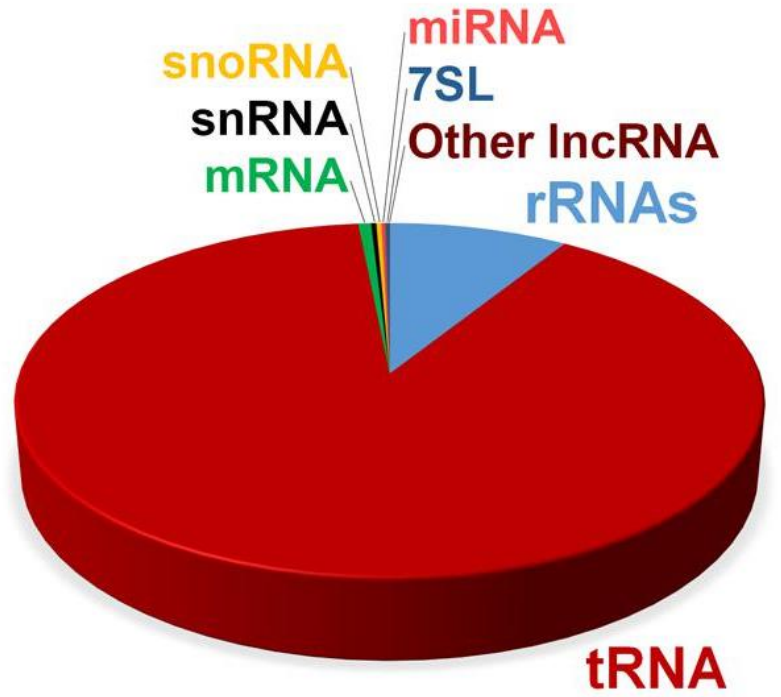
- Strandedness
- Size of transcripts
- Polyadenylation state
- Secondary structure
- Splicing



Picking an approach

- **Total RNA sequencing**
  - Data will be dominated by rRNAs
- **Poly(A) enrichment**
  - Captures mRNAs and polyadenylated ncRNAs > 150 nt
  - Not suited for degraded samples (RIN < 7.0)
- **rRNA depletion**
  - Captures mRNAs and non-adenylated RNAs > 150 nt
  - Best option for degraded samples (RIN < 7.0)
  - Requires deeper sequencing than poly(A) approaches
- **Small RNA analysis**
  - Only profiles small RNAs (< 200nt) and miRNAs
  - Alternative strategies can be used to profile (only) tRNAs
- **Targeted enrichment**
  - Compatible with both rRNA depletion and poly(A) selection approaches
  - Not generally suited for gene expression analyses
  - Can also target non-adenylated, non-ribosomal RNAs

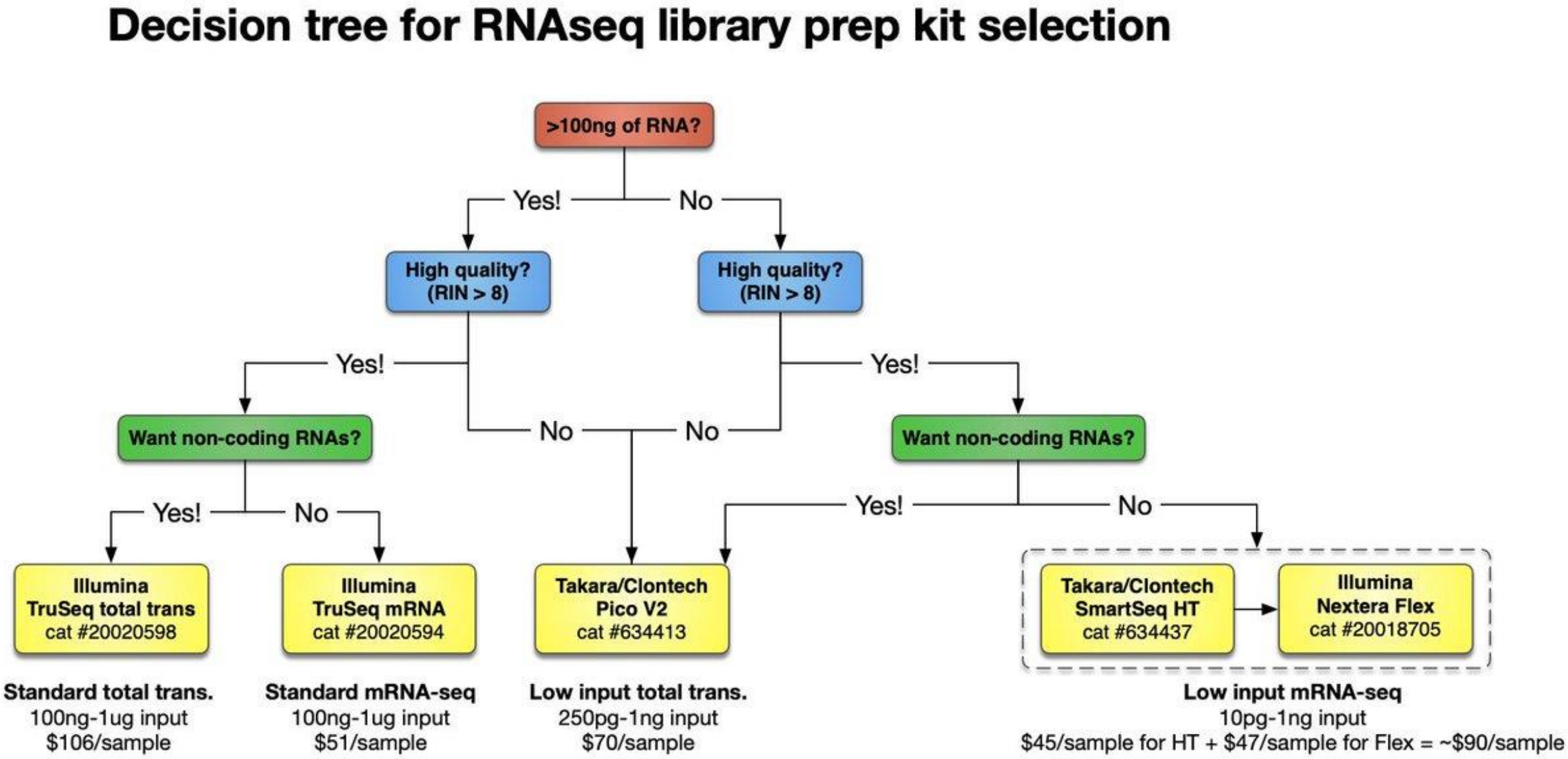
Note: tRNAs and small RNAs are generally excluded by standard library approaches (they are too small)



RNA by number of molecules

rRNAs dominate ‘library prep-able’ RNA

Picking an approach



Considerations: read length

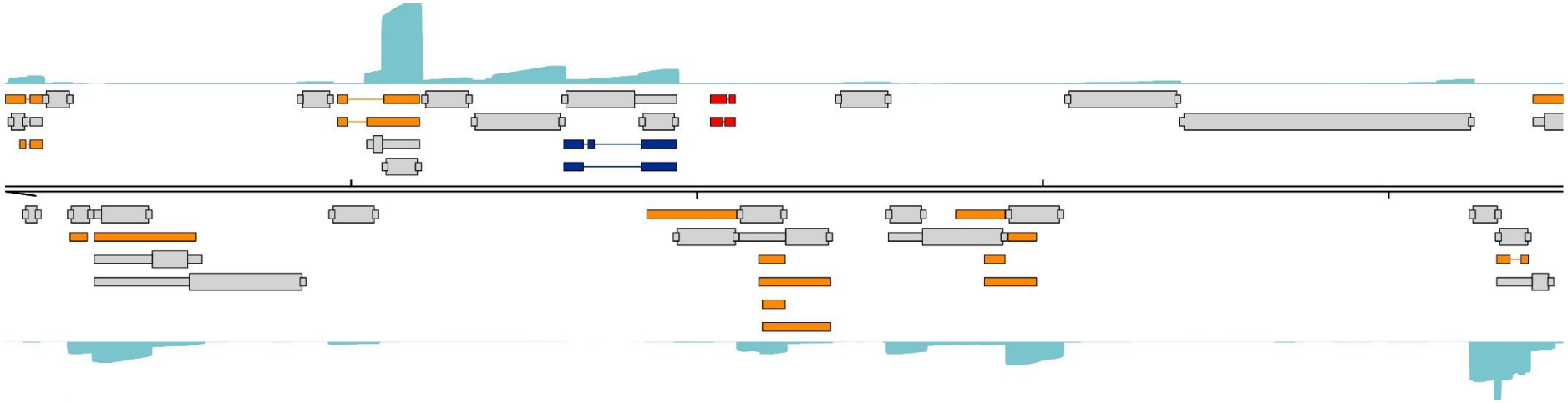
- Illumina approaches are standard for gene expression analysis in eukaryotes
  - PE 50 fine for (simple) gene expression analysis (gene profiling)
  - PE 100 / 150 for complex gene expression analysis (transcript profiling & splicing analyses)
  - Recoding and amplification bias are typical

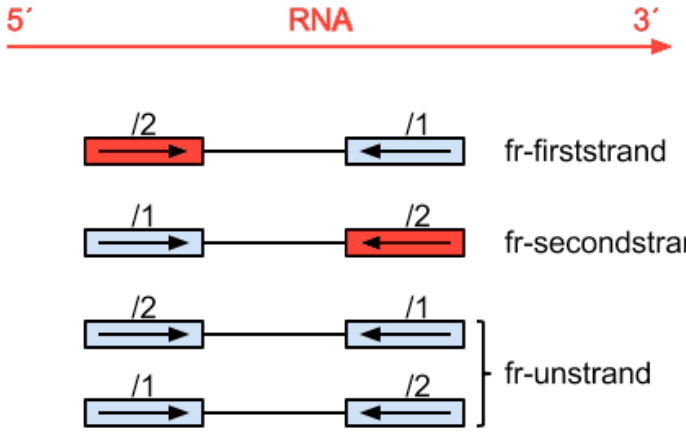
  - 25-30 million PE reads for poly(A) expression profiling
  - 30-40 million PE reads for rRNA-depletion expression profiling
  - 5-10 million SE reads for small RNA sequencing

  } Expression profiling of *H. sapiens*

- Long read approaches (i.e. nanopore) are becoming competitive
  - Similar read depths achievable
  - Read lengths allow unbiased assigning of isoforms
  - cDNA (recoding bias) and native RNA sequencing protocols available (minimal bias)
  - Sequencing of native RNA allows RNA modification detection and poly(A) tail length estimates
  - Input requirements (micrograms of total RNA) remain very high and not suitable for all situations (unlike Illumina)

Considerations: strandedness





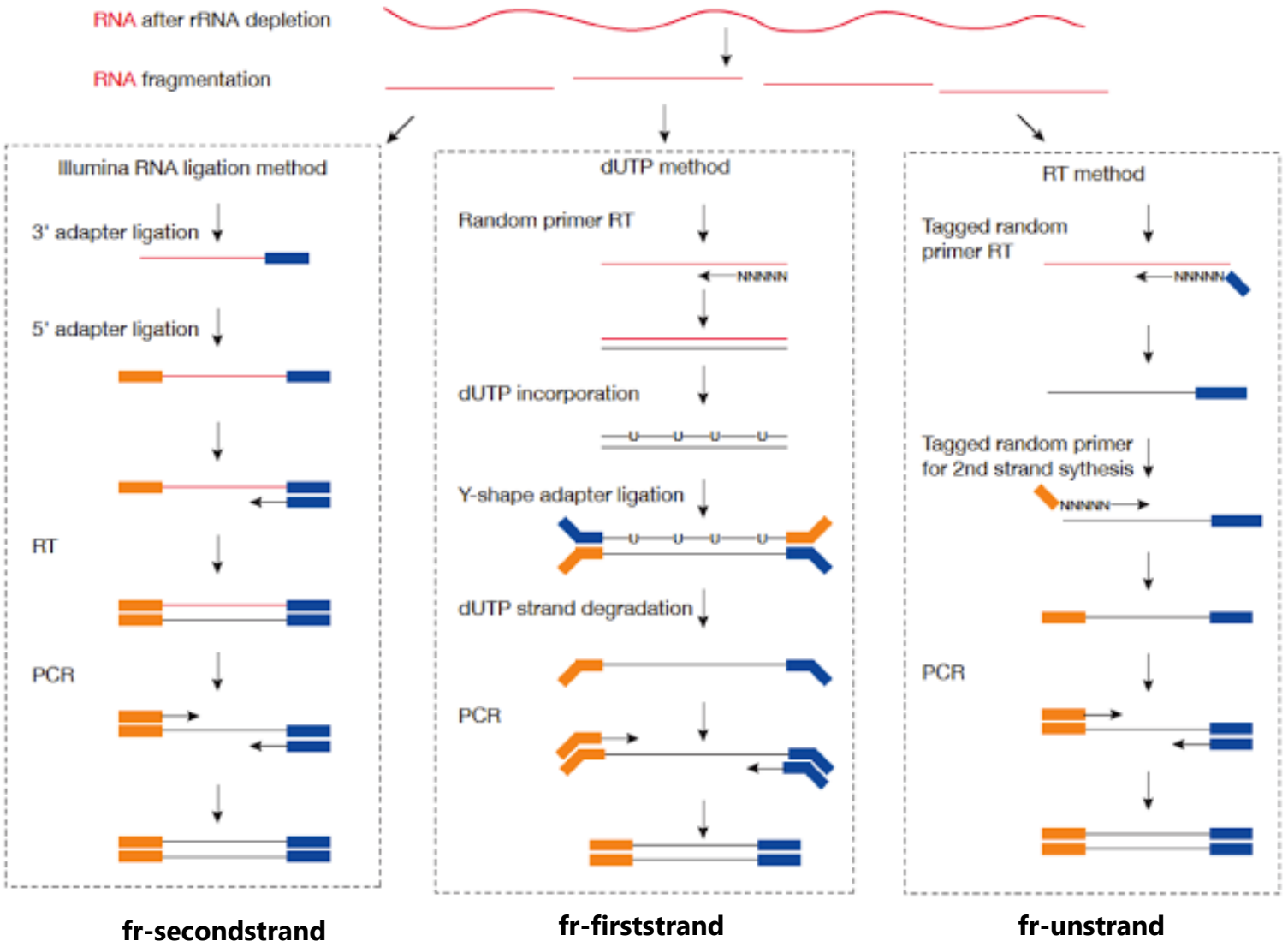
The second read (read 2) is from the original RNA strand/template, first read (read 1) is from the opposite strand. The information of the strand is preserved as the original RNA strand is degradated due to the dUTPs incorporated in the second synthesis step.

The first read (read 1) is from the original RNA strand/template, second read (read 2) is from the opposite strand. The directionality is preserved, as different adapters are ligated to different ends of the fragment.

Information regarding the strand is not conserved (it is lost during the amplification of the mRNA fragments).

https://chipster.csc.fi/manual/library-type-summary.html

Considerations: strandedness

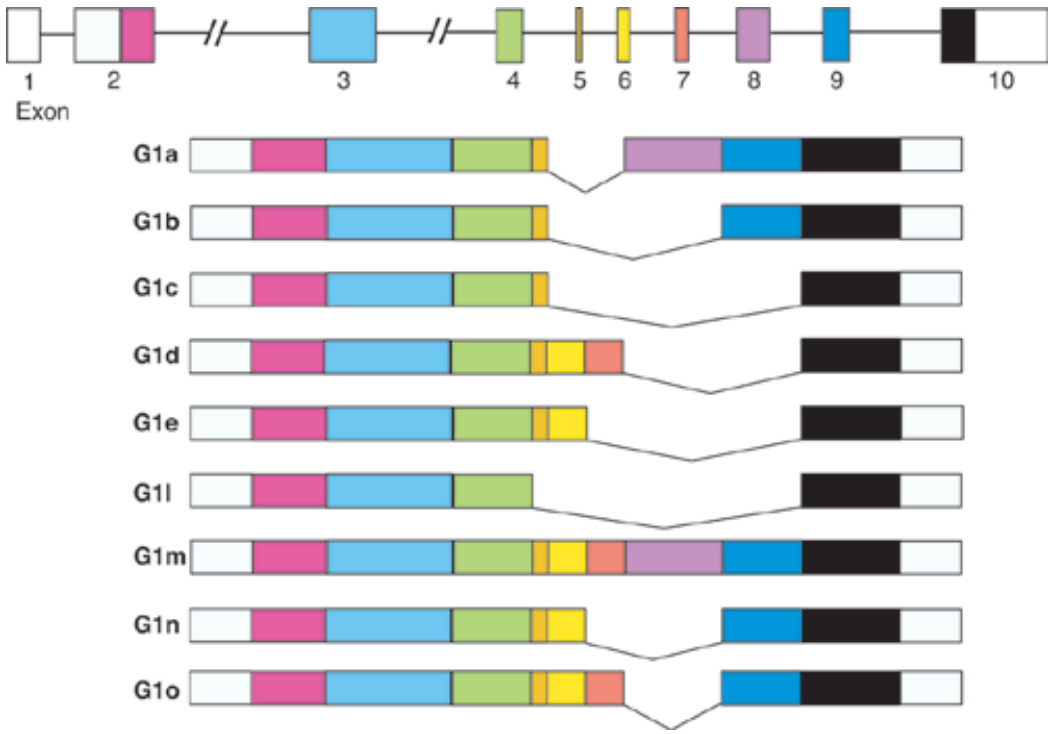


fr-secondstrand **fr-firststrand** **fr-unstrand**

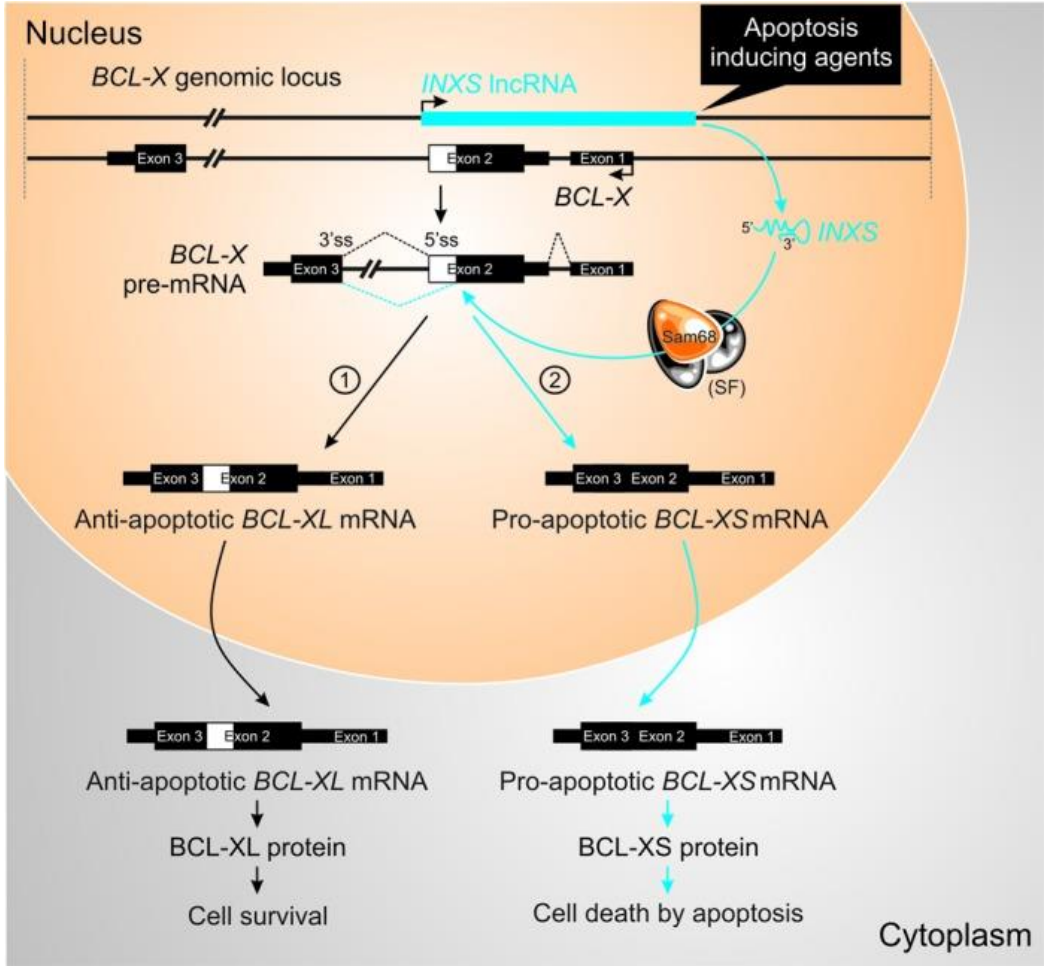
https://chipster.csc.fi/manual/library-type-summary.html

Considerations: annotation

- A crucial aspect of RNA-Seq analysis
- Most genes encode multiple distinct RNAs yet most expression studies refer to gene expression and do not distinguish between RNA isoforms



- This is a problem particular where individual isoforms may serve contrasting functions



Considerations: annotation

Two major file formats – GTF & GFF

GFF3 format is better described and allows for a richer annotation

- But… flexible annotation reduces cross-compatibility

Chromosome	Annotator	feature	start	stop	n/a	strand	flag	feature-relations
HSV1-Patton-Us11gfp	Dan	gene	2129	5706	.	+	.	ID=RL2a; Name=RL2a
HSV1-Patton-Us11gfp	Dan	mRNA	2129	5706	.	+	.	ID=mRNA_RL2a; Parent=RL2a
HSV1-Patton-Us11gfp	Dan	exon	2129	2318	.	+	.	ID=exon1_RL2a; Parent=mRNA_RL2a; Name=RL2a
HSV1-Patton-Us11gfp	Dan	exon	3084	3750	.	+	.	ID=exon2_RL2a; Parent=mRNA_RL2a; Name=RL2a
HSV1-Patton-Us11gfp	Dan	exon	3887	5706	.	+	.	ID=exon3_RL2a; Parent=mRNA_RL2a; Name=RL2a
HSV1-Patton-Us11gfp	Dan	CDS	2262	2318	.	+	0	ID=cds1_RL2a; Parent=mRNA_RL2a; Name=ICP0
HSV1-Patton-Us11gfp	Dan	CDS	3084	3750	.	+	0	ID=cds1_RL2a; Parent=mRNA_RL2a; Name=ICP0
HSV1-Patton-Us11gfp	Dan	CDS	3887	5490	.	+	0	ID=cds1_RL2a; Parent=mRNA_RL2a; Name=ICP0

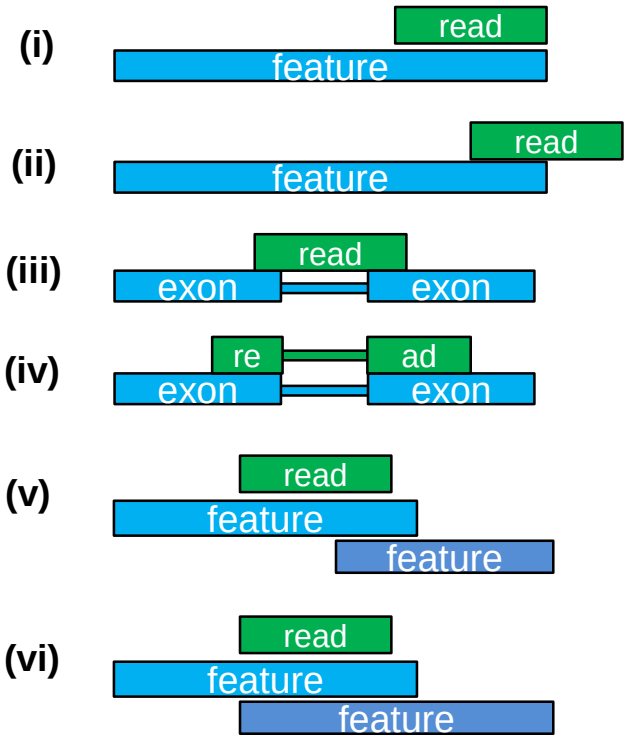
HSV1-st17	Cufflinks	transcript	513	1540	0	+	0	gene_id "RL1A.gene"; transcript_id "RL1A.gene"
HSV1-st17	Cufflinks	exon	513	1540	0	+	0	gene_id "RL1A.gene"; transcript_id "RL1A.gene"; exon_number "1"
HSV1-st17	Cufflinks	transcript	2087	5699	0	+	0	gene_id "RL2A.gene"; transcript_id "RL2A.gene"
HSV1-st17	Cufflinks	exon	2114	2318	0	+	0	gene_id "RL2A.gene"; transcript_id "RL2A.gene"; exon_number "1"
HSV1-st17	Cufflinks	exon	3084	3750	0	+	0	gene_id "RL2A.gene"; transcript_id "RL2A.gene"; exon_number "2"
HSV1-st17	Cufflinks	exon	3887	5699	0	+	0	gene_id "RL2A.gene"; transcript_id "RL2A.gene"; exon_number "3"
HSV1-st17	Cufflinks	transcript	9338	10949	0	+	0	gene_id "UL1"; transcript_id "UL1.gene"
HSV1-st17	Cufflinks	exon	9338	10949	0	+	0	gene_id "UL1.gene"; transcript_id "UL1.gene"; exon_number "1"
HSV1-st17	Cufflinks	transcript	9885	10949	0	+	0	gene_id "UL2.gene"; transcript_id "UL2.gene"
HSV1-st17	Cufflinks	exon	9885	10949	0	+	0	gene_id "UL2.gene"; transcript_id "UL2.gene"; exon_number "1"

Considerations: read quantification

- **Aim:** count number of reads overlapping with user-detailed genomic features

- **Considerations:**
  - overlap size (full read vs. partial overlap)
  - multi-mapping reads
  - reads overlapping multiple genomic features of the same kind
  - reads overlapping introns
  - ***Annotation quality is everything!!!***

Counting protocol defined by user!!!



Read counted as part of feature (independent of strictness)

Read may be counted as part of feature (dependent on strictness)

Read likely originates from pre-mRNA (should we count this?)

Read spliced correctly, counted as part of feature

Read partially overlaps second feature (count only as feature #1?)

Read overlaps two features (count as fraction for each feature?)

Considerations: transcript quantification vs. gene quantification

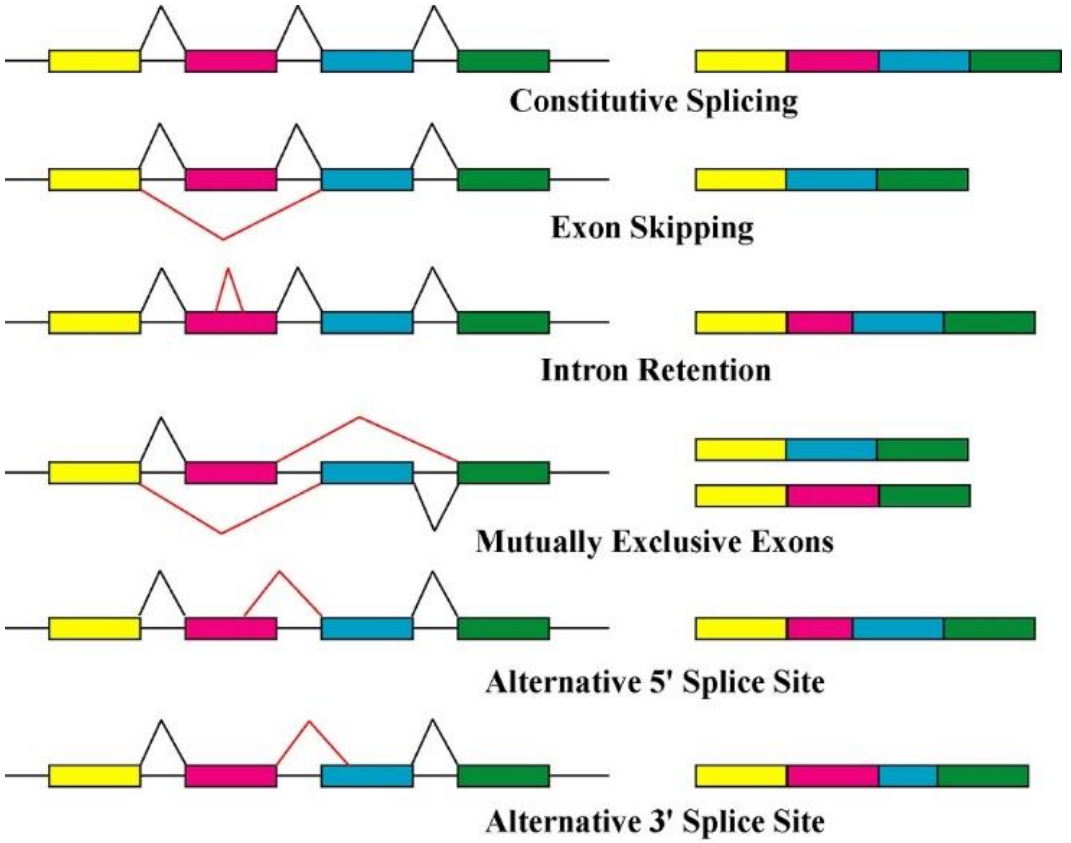
- Gene quantification is ‘standard’ and is *assumed* to work optimally

- Transcript quantification is faster and provides (theoretically) higher resolution but has several caveats
  - inconsistent annotation of transcripts
  - multiple isoforms of widely differing lengths
  - anti-sense/overlapping transcripts of different genes
  - does not detect novel isoforms

- An increasingly popular strategy is to count transcript isoforms and then merge transcripts from the same gene to generate gene counts

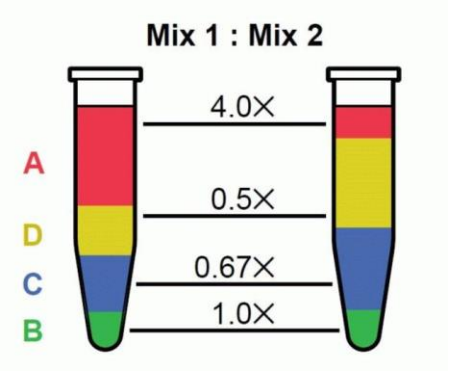
Considerations: detecting splice variants

- Alternative/differential splicing is a highly regulated process that greatly increases protein-coding diversity of an organism
- Mediated by the binding of splicing activators and splicing suppressors
- Detection relies on identification of splice junctions
- These are disconnected in short-read sequencing
  - Junction usage can be detected but is very hard to quantify
  - Numerous tools available
- Splice junctions are connected in long-read sequencing
  - Estimates of isoform abundances can be readily generated
  - Mapping of exact splice junction locations still problematic due to error rate
  - No specialized tools available



Considerations: spiking with defined RNAs

- Variation in RNA expression data is impacted by
  - Quality of the starting material
  - Level of cellularity
  - RNA extraction yield
  - Sequencing platform employed
  - Human error



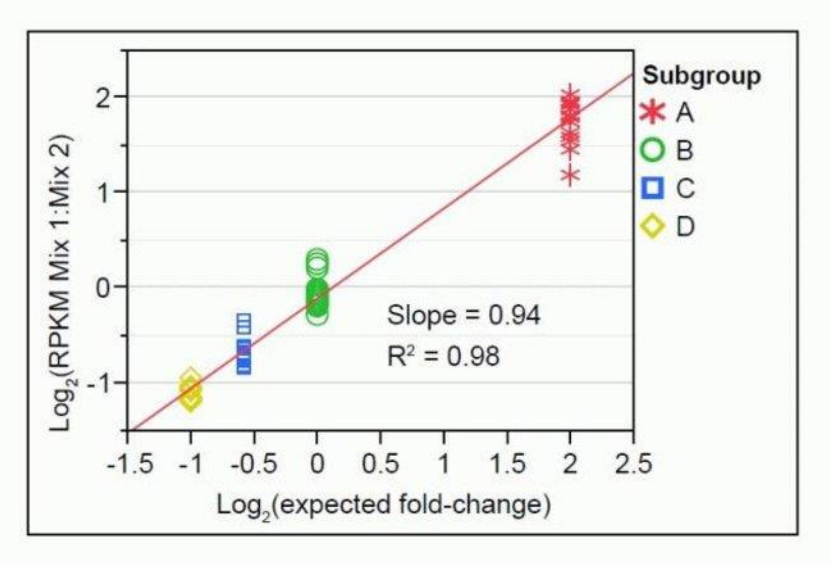
- Spike-in controls (a set of unlabeled, polyadenylated transcripts) are added to an RNA analysis experiment after sample isolation and allow measuring data against defined performance criteria

- The transcripts are designed to be 250 to 2,000 nt in length, which mimic natural eukaryotic mRNAs.

➢ Achieve a standard measure for data comparison across gene expression experiments

➢ Measure sensitivity (lower limit of detection) and dynamic range of an experiment

➢ Quantitate differential gene expression



Considerations: experimental design

- Gene expression analyses aims to identify genes whose expression varies between two or more experimental settings (conditions)

- During analysis, every single gene is tested to determine if it's expression changes between conditions

Our goal is to observe a reproducible effect that can be due only to the treatment (avoiding confounding and bias) while simultaneously measuring the variability required to estimate how much we expect the effect to differ if the measurements are repeated with similar but not identical samples (replicates)
-- Altman and Krzywinski, 2014

- Desired outcome is a list of genes that differ by a user-defined fold change in a statistically robust manner

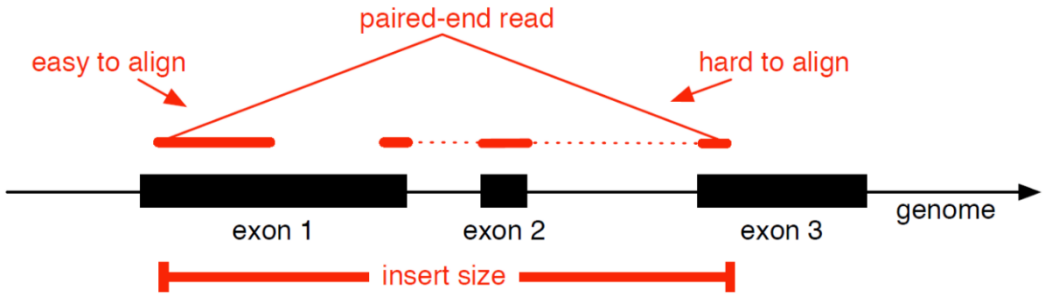
- Each condition should ideally have three or more biological replicates associated with it
  - number of replicates defines sensitivity (low fold changes)
  - more replicates = more sensitive = more money

Considerations: defining biological and technical replicates

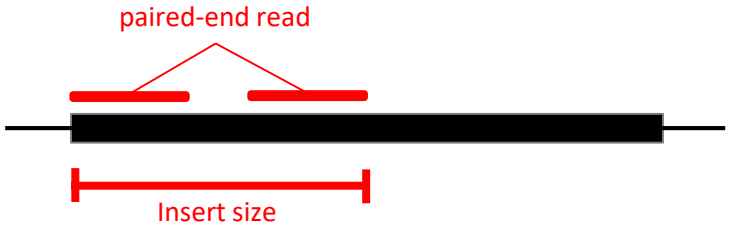
	Replicate type	Category
Subjects	Colonies	Biological
	Strains	Biological
	Cohoused groups	Biological
	Gender	Biological
	Individuals	Biological
Sample preparation	Organs from sacrificed animals	Biological
	Methods for dissociating cells from tissue	Technical
	Dissociation runs from given tissue sample	Technical
	Individual cells	Biological
	RNA-seq library construction	Technical
Sequencing	Runs from the library of a given cell	Technical
	Reads from different transcript molecules	Variable
	Reads with unique molecular identifier from a given transcript molecule	Technical

So you've got some RNA-Seq data… Now what?

- Strategy #1 – align to a reference genome



- Strategy #2 – align to a reference transcriptome

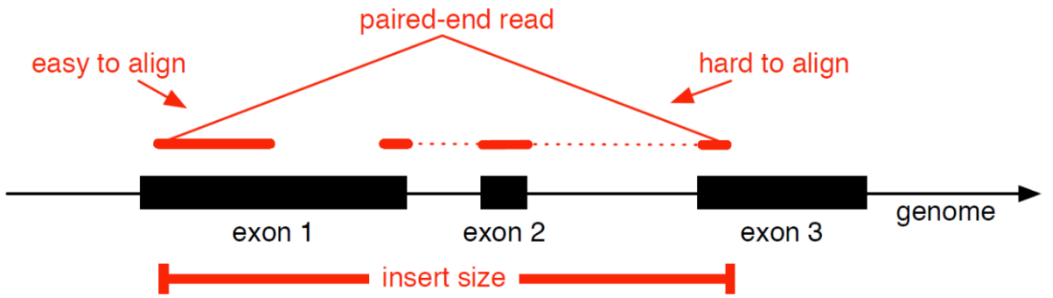


- *[Strategy #3 – de novo transcript assembly]*

So you've got some RNA-Seq data… now what?

Strategy #1 – align to a reference genome

- Gene expression (low-resolution)
- Novel transcription
- Transcriptional abnormalities (e.g. read-through)
- Slow, heavy computational load
- Most reads have unique mappings



Aligning RNA-Seq data against a reference genome

Short-read data – global aligners

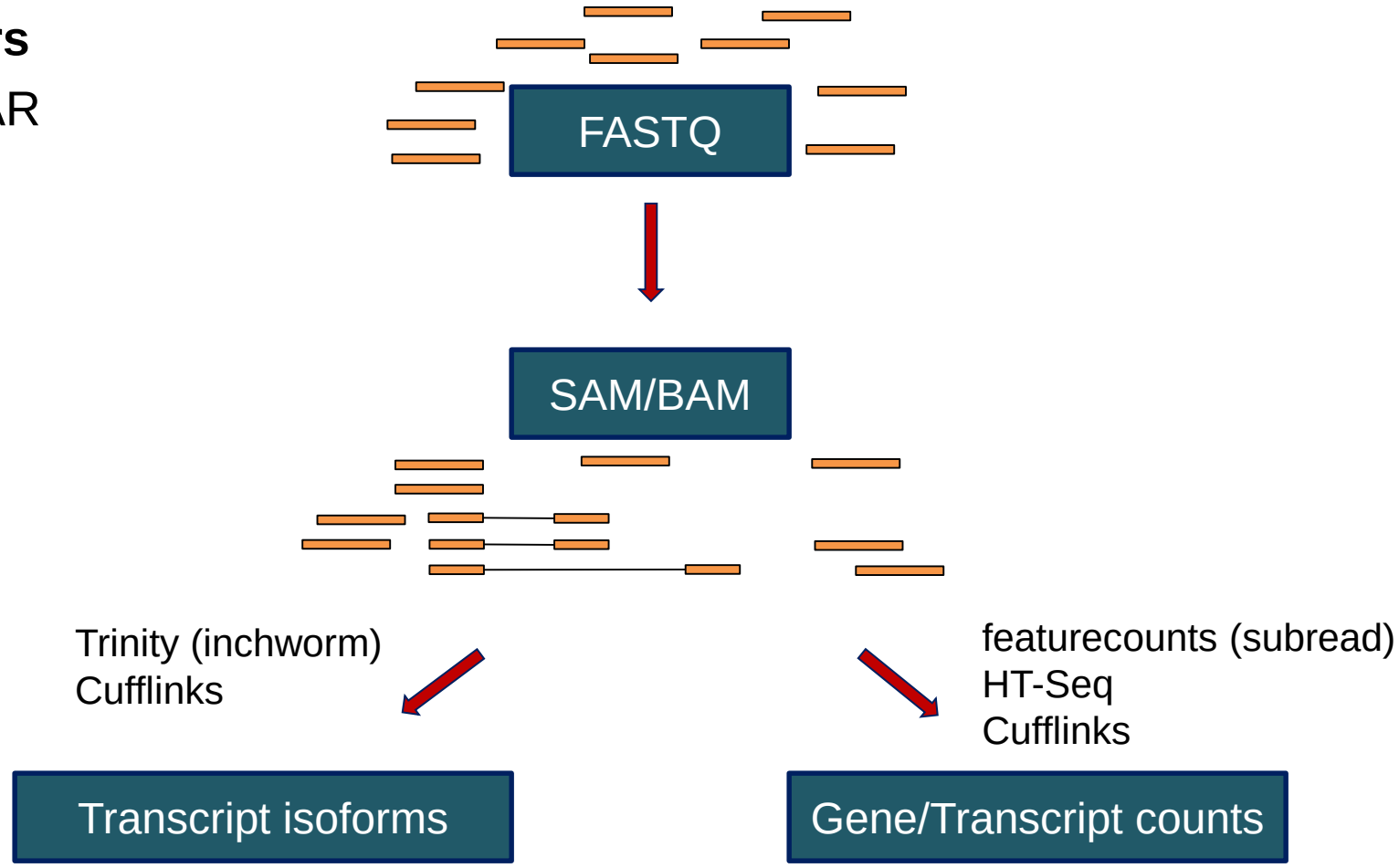
- Tophat2, Bowtie2, bbmap, STAR

Parameters to optimized

- Strandedness
- Intron lengths
- Splice junctions

Gene/Transcripts counts

Transcript isoforms



Generating count data

- A counts file contains the number of sequence reads mapped to each gene

- Standardized input for gene expression analyses packages such as DeSeq2 and edgeR

- Note: Counts files should always contain **raw** (i.e. non-normalized) counts

- Usually combine multiple samples and experimental conditions

gene	ctrl_1	ctrl_2	exp_1	exp_1
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0

- FeatureCounts (part of the subread package) and HTSeq are the primary softwares used to generate count files from BAM files
  - http://subread.sourceforge.net/
  - https://htseq.readthedocs.io/en/release_0.11.1/

- Important parameters to consider are strandedness, and how to deal with overlapping genes / transcripts

Example of featureCounts command

- A counts file contains the number of sequence reads mapped to each gene

featureCounts -M -s 1 -p -B -a annotation.gtf -o outfile firstbamfile secondbamfile thirdbamfile …

-M # *counts multi-mapping reads (be careful to ensure multi-mapping reads are randomly allocated!)*

-s # *perform strand-specific read counting. 0 (unstranded), 1 (stranded) and 2 (reversely stranded)*

-p # *count fragments rather than reads (e.g. two reads = one fragment for PE sequencing*

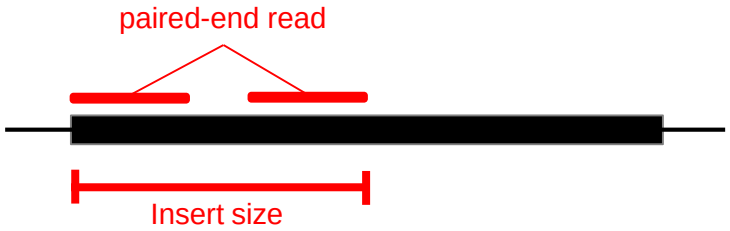
-B # *required read pairs are both mapped*

-L # *long reads (nanopore / PacBio)*

So you've got some RNA-Seq data… Now what?

Strategy #2 – (pseudo)align to a reference transcriptome

- Transcript expression (high-resolution)
- Rapid, low computational load
- Requires a (very) high quality annotation
- Not suitable for transcript discovery or any other 'non expression-related' analyses

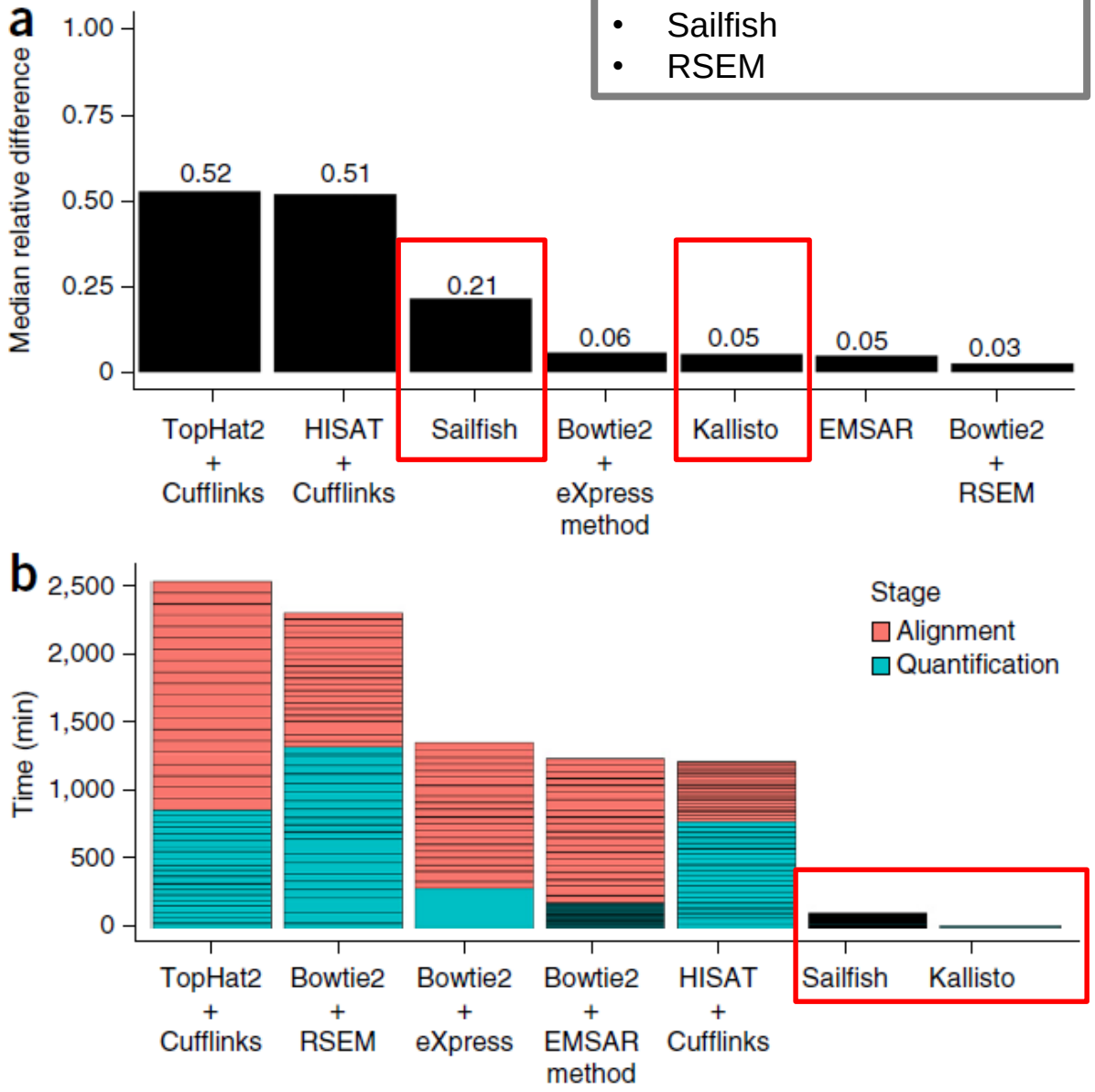


Pseudoalignment

- Pseudoaligned data is highly accurate and highly reproducible

- Psuedoalignment is incredibly rapid

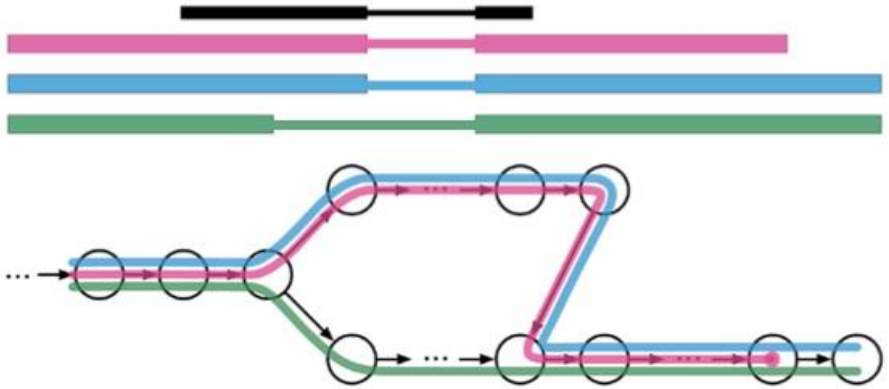
Other pseudoaligners
- Salmon
- Sailfish
- RSEM



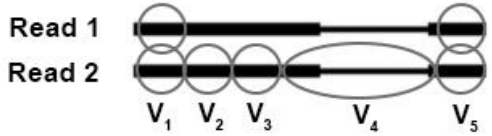
Pseudoalignment with Kallisto

…determine, for each read, not where in each transcript it aligns, but rather which transcripts it is compatible with…

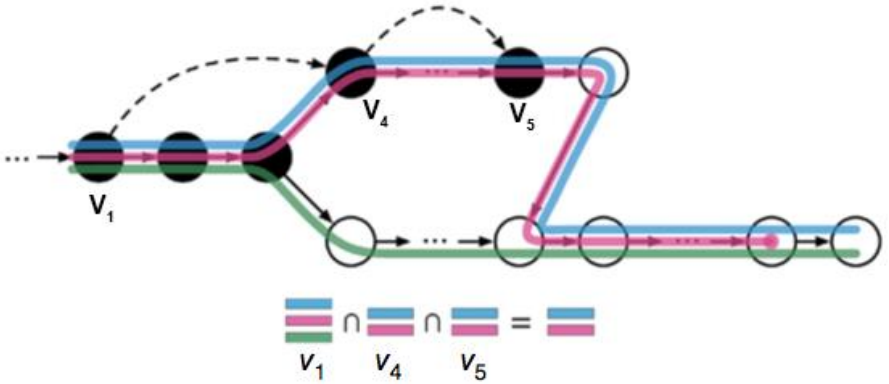
- Incredible fast method that discards 40-60% of reads



1. All transcripts are deconstructed into short k-mers



2. Each read is split into short k-mers
3. Find equivalent classes (i.e. group reads)



4. Find matches for first & last k-mer
5. Find matches for internal k-mers (if no unique match)
6. Take union to identify transcript
7. Determine probability of assigning correct transcript

http://nextgenseek.com/2015/05/kallisto-paper-summary-near-optimal-rna-seq-quantification/

Example mapping parameters (kallisto)

 # Step 1 – generate an index file using a fasta-formatted transcriptome file

kallisto index -i HomoSapiens Homo_sapiens.GRCh38.all.cds.ncrna.fa

Useful flags
-i desired index name

 # Step 2 – generate pseudoalignment (untrimmed fastq files)

kallisto quant -i HomoSapiens -o outputfile infile_R1.fastq.gz infile_R2.fastq.gz

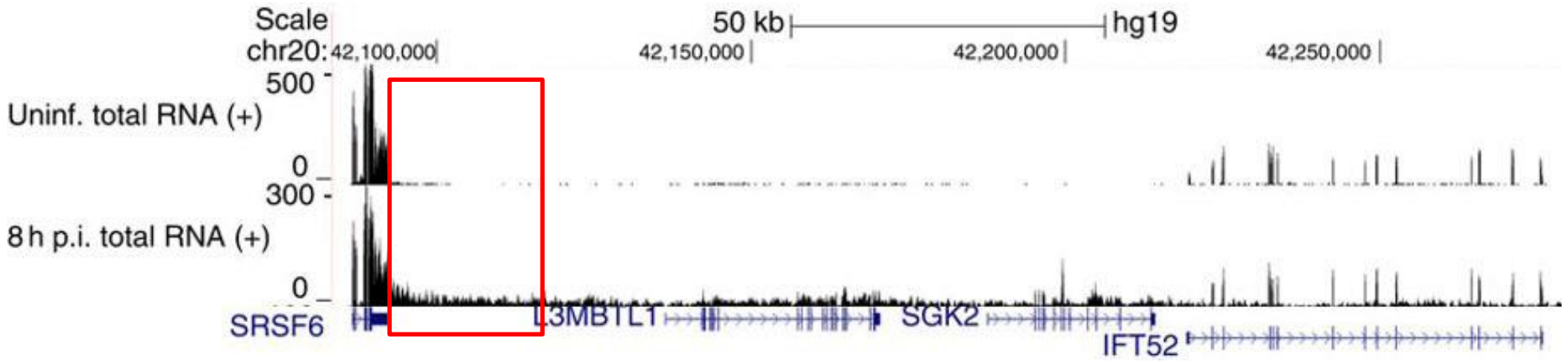
Useful flags
-b *bootstrap (100)*
--bias *performs sequence based bias correction*
--fr-stranded *or* --rf-stranded *strandedness*
--single *single-end mode (default paired-end)*

Output files

- abundance.h5 (HDF5 binary file - run info, abundance estimates, bootstrap estimates, and transcript length information length) → Sleuth + Shiny
- abundance.tsv (transcript counts – text based) → DeSeq2 / edgeR
- run_info.json (assembly metrics)

A warning: complications are everywhere, so be aware!

- Disruption of transcription termination – stressed / infected cells
  - RNA transcription does not terminate correctly and is not adenylated
  - rRNA-depletion approaches required to detect this BUT resulting data is <u>not</u> suitable for gene expression analysis (as we are only interested in mRNAs



- Host shut-off
  - Specific viral proteins 'clip' mRNAs to remove cap and promote degradation before translation
  - Impacts on gene expression estimates as a proxy for translation (protein abundance

Assignment #2 – Investigating the cellular response to exogenous DNA (cont…)

Introduction

Double-stranded DNA (dsDNA) in the cytosol of human cells stimulates the type 1 interferon (IFN) response, a component of innate immunity that is active against invading pathogens and many cancers. Over the course of Assignment #1 and Assignment #2, we will examine the host genes that are transcriptionally regulated upon detection of invading dsDNA.

For assignment #2, you will generate count data using two different alignment strategies, and subsequently perform differential gene expression (DGE) analysis on each with the final aim of comparing the two alignment strategies. Here, we will use all three bioreps for each of our two conditions in the analysis.

Assignment #2

1. **Use featureCounts to generate gene counts from the alignments generated in Assignment #1.**
   Import into DeSeq2
   Perform basic data exploration and DGE analysis

2. **Use Kallisto or Salmon to pseudoalign raw fastq files to generate alternative count datasets.**
   Convert transcript counts to gene counts
   Import into DeSeq2
   Perform basic data exploration and DGE analysis

3. **Compare outputs from the two strategies – what have you learned?**

- **Several R scripts are provided**
  - conversion of transcript counts to gene counts
  - basic data exploration (i.e. similarity of biological replicates)
  - DGE analysis

Preparing for Thursdays practical

- Ensure that you have R & Rstudio installed locally along with the following packages
  - DESeq2
  - pheatmap
  - RColorBrewer
  - vsn
  - AnnotationDbi
  - org.Hs.eg.db
  - genefilter
  - biomaRt
  - IHW

- Consider also installing R markdown

Thank you for your attention

Questions?