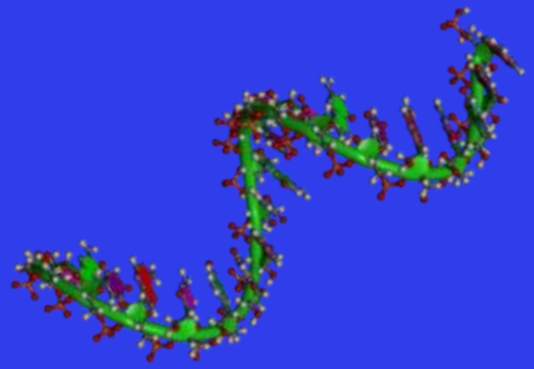
RNA-seq

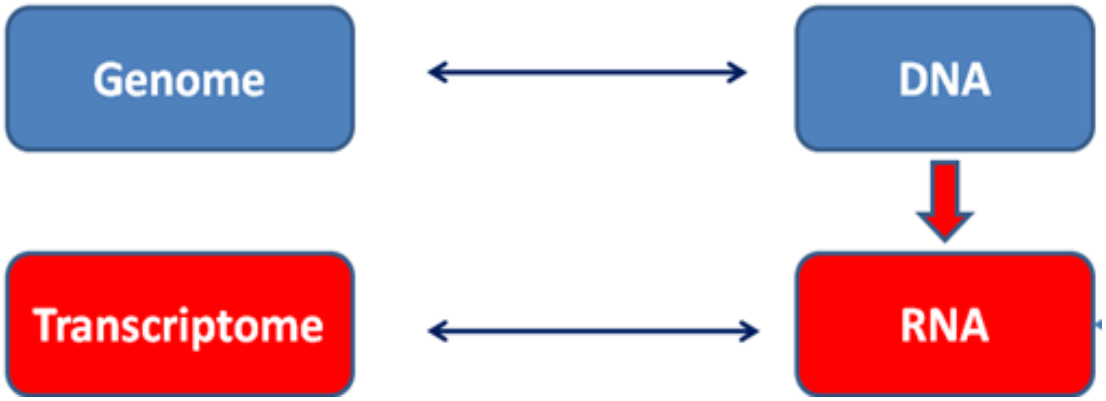
Next Generation Sequencing Analysis, 2021

Francesca Bertolini
DTU Aqua
franb@aqua.dtu.dk



Transcriptome

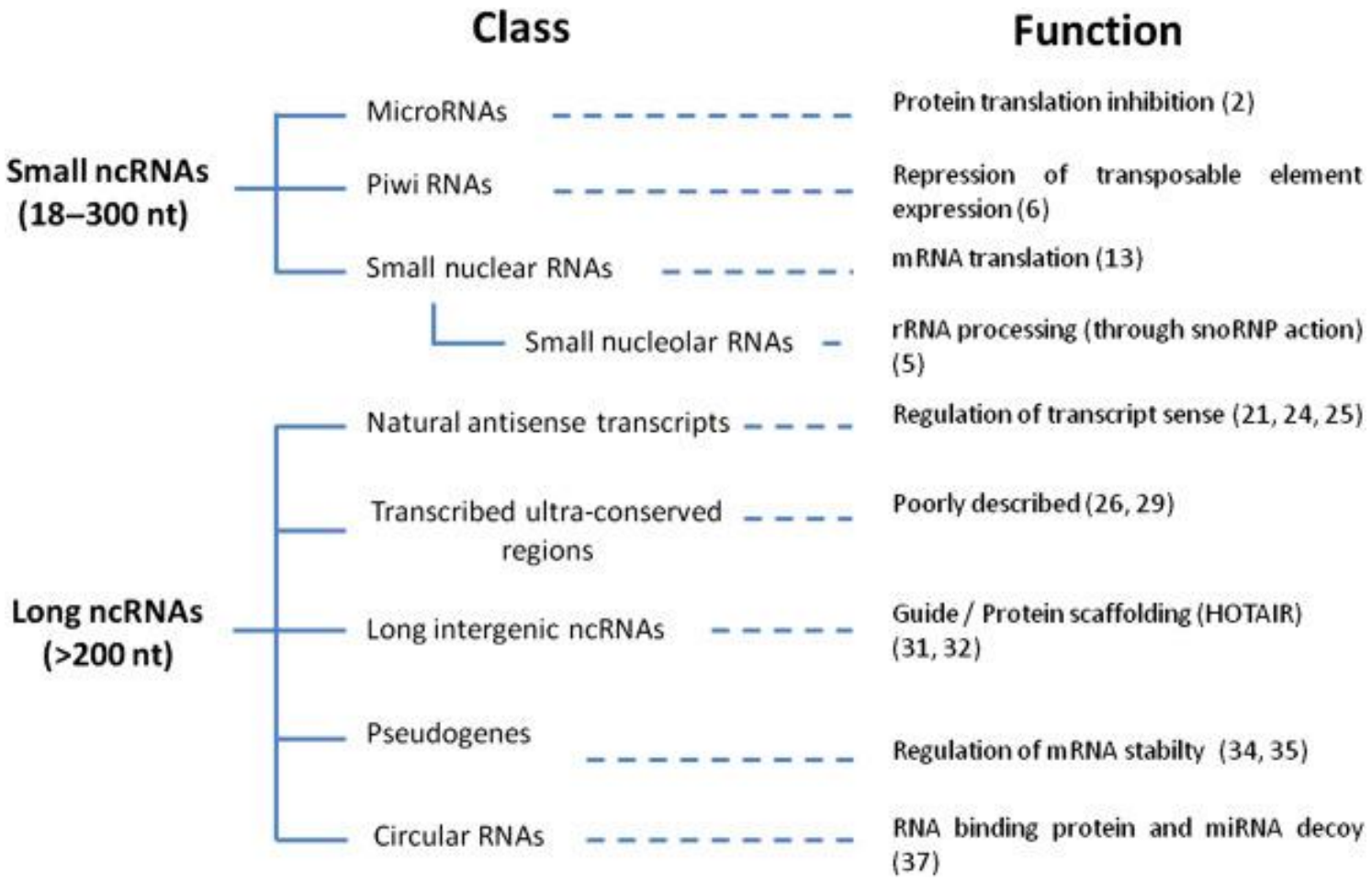
- The transcriptome is defined as the complete set of transcripts (RNA) in a cell, and their quantity, for a specific developmental stage or physiological condition (Wang et al. 2009).
- Transcriptome is therefore dynamic and a good representative of the cellular and tissue state (Srivastava et al 2019).



RNA classification

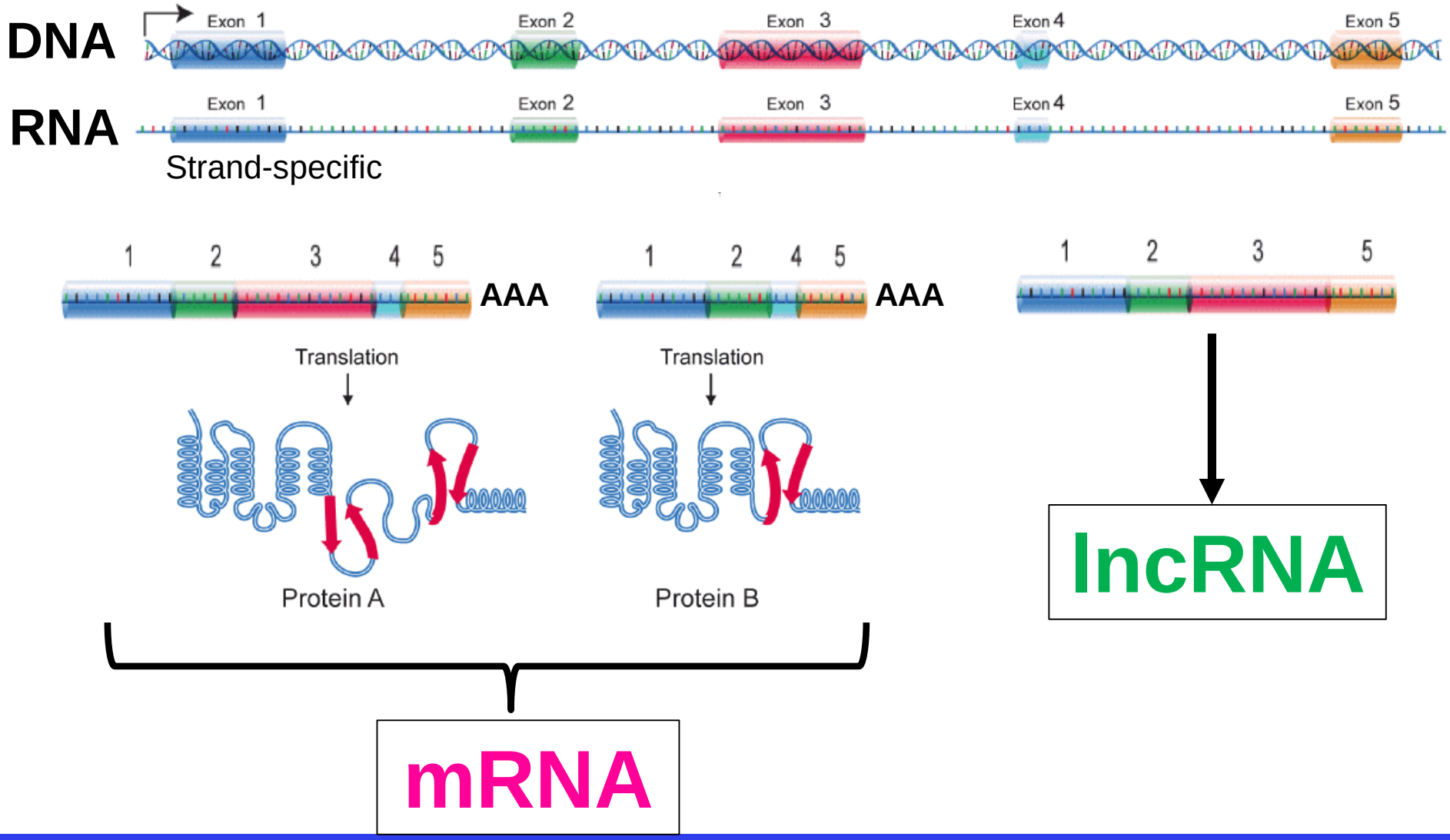
- **Ribosomal** RNA (rRNA): catalytic component of ribosomes (about 80-85%)
- **Transfer RNA** (tRNA): transfers amino acids to polypeptide chain at the ribosomal site of protein synthesis (about 15%)
- **Coding RNA(mRNA):** carries information about a protein sequence to the ribosomes
- **Other Non coding regulatory RNAs**

Other non coding regulatory RNAs



Delpu et al. 2016. *Drug Discovery in Cancer Epigenetics*

Long RNAs: splicing



RNA-seq

High-throughput sequencing technology used for probing the transcriptome of a sample

The types of information that can be gained from RNA-seq can be divided into two broad categories: **qualitative** and **quantitative**.

•**Qualitative** data includes identifying expressed transcripts, and identifying exon/intron boundaries, transcriptional start sites (TSS), and poly-A sites.

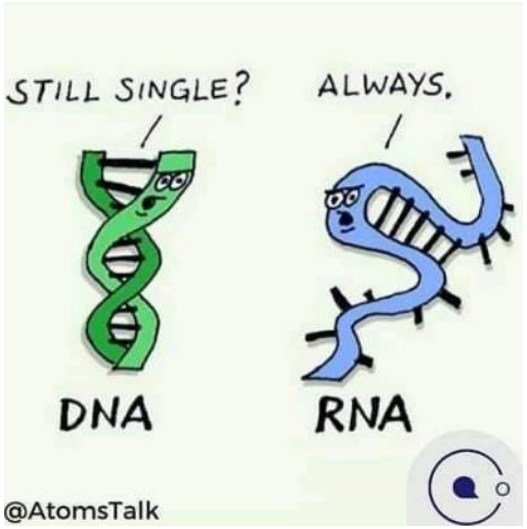
•**Quantitative** data includes measuring differences in expression, alternative splicing, alternative TSS, and alternative polyadenylation between two or more treatments or groups.

Before RNA extraction

RNA is more unstable than DNA, therefore higher precautions are needed to avoid degradation

TISSUE COLLECTION:

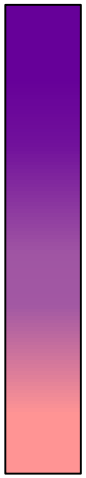
- **Liquid nitrogen**
- **RNA later (for solid tissues)**
- **Tempus/Pax tubes (for liquid tissue)**



After RNA extraction

RIN (RNA integrity number)**:** algorithm for assigning integrity values to RNA measurements.

Integrity

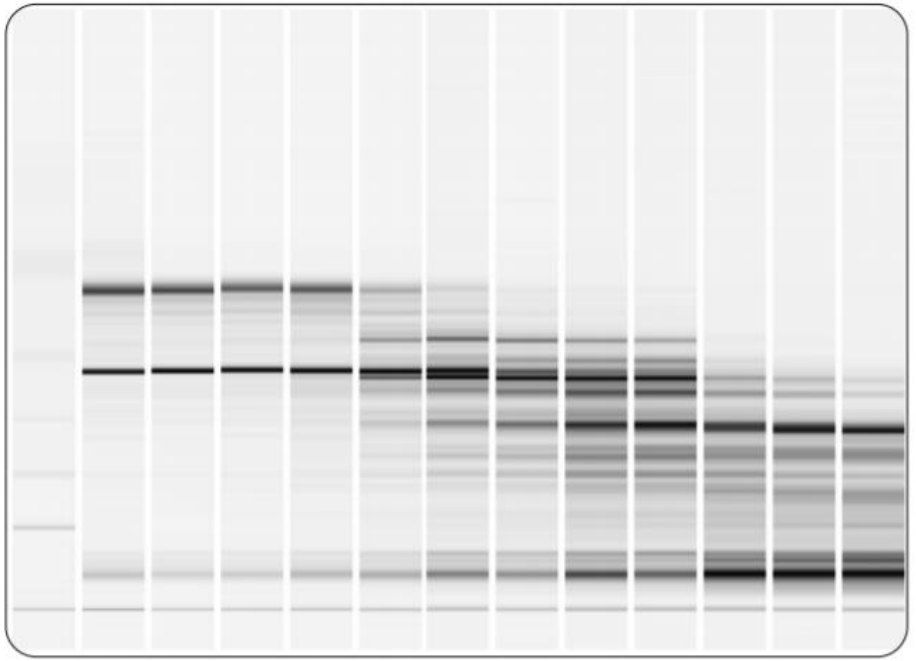


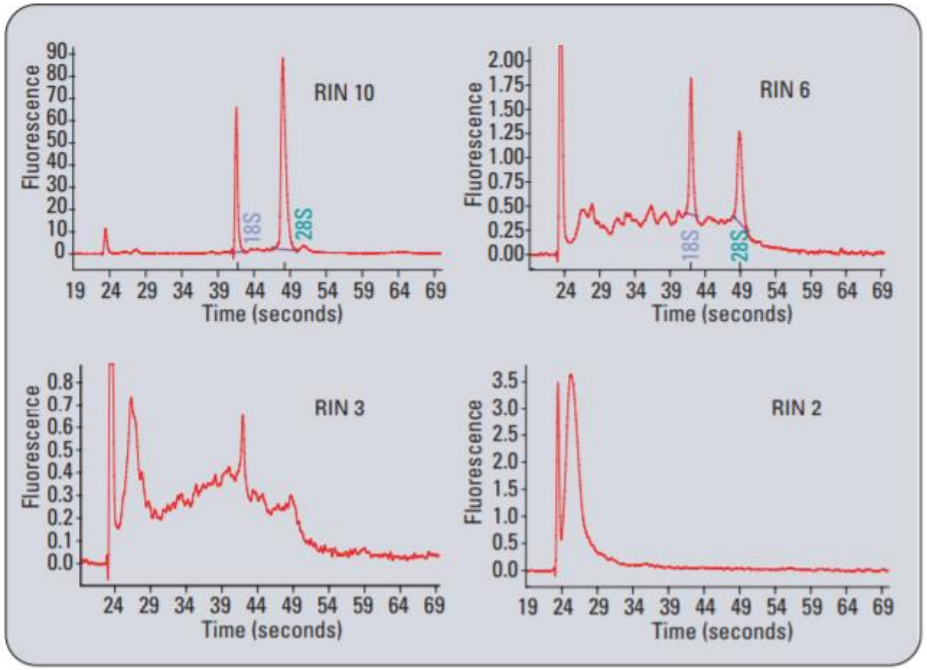
10: maximum

0: minimum

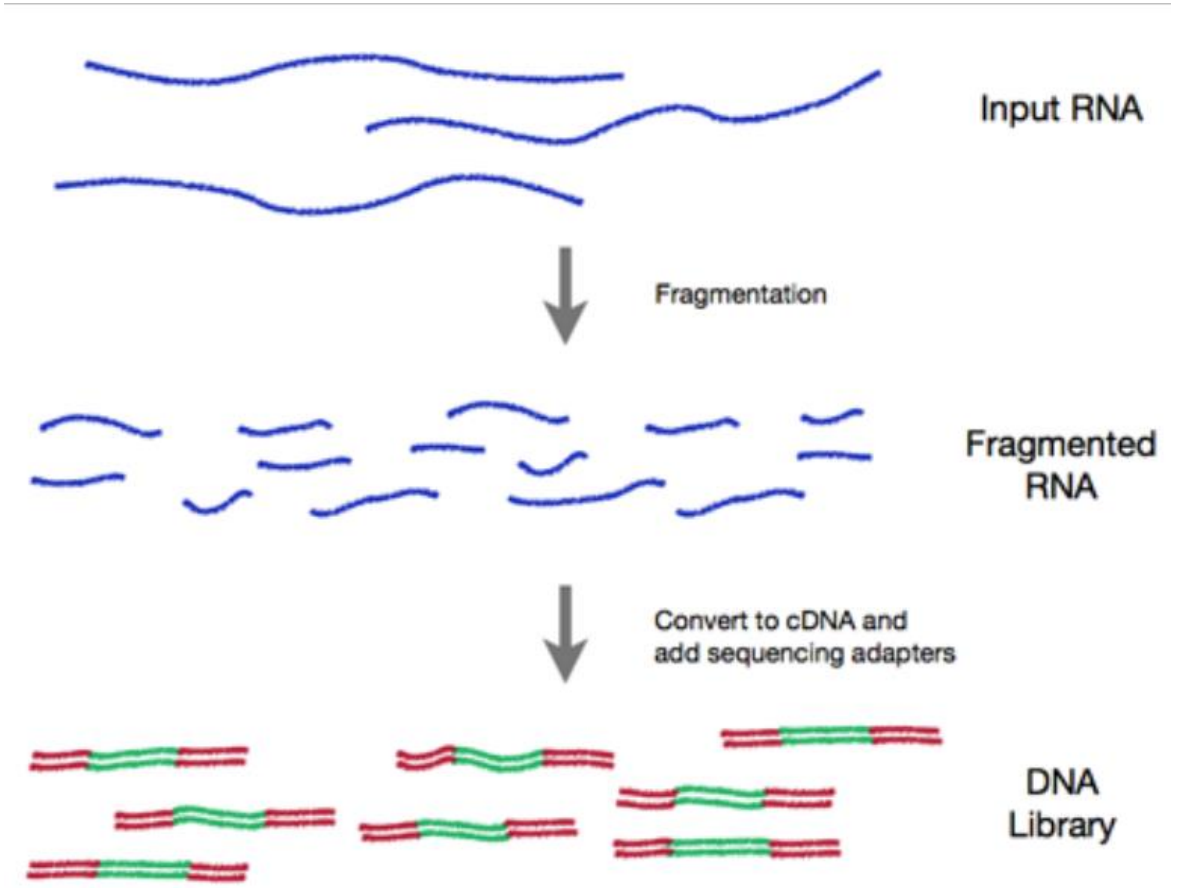
RIN>7 ok

RNA quality (RIN) and quantification: Bioanalyzer





A typical RNA-seq experiment on a 2nd generation seq platform



Different steps for different RNAs

- **Total RNA seq**

1. DNase treatment
2. Ribosomal depletion
3. library preparation

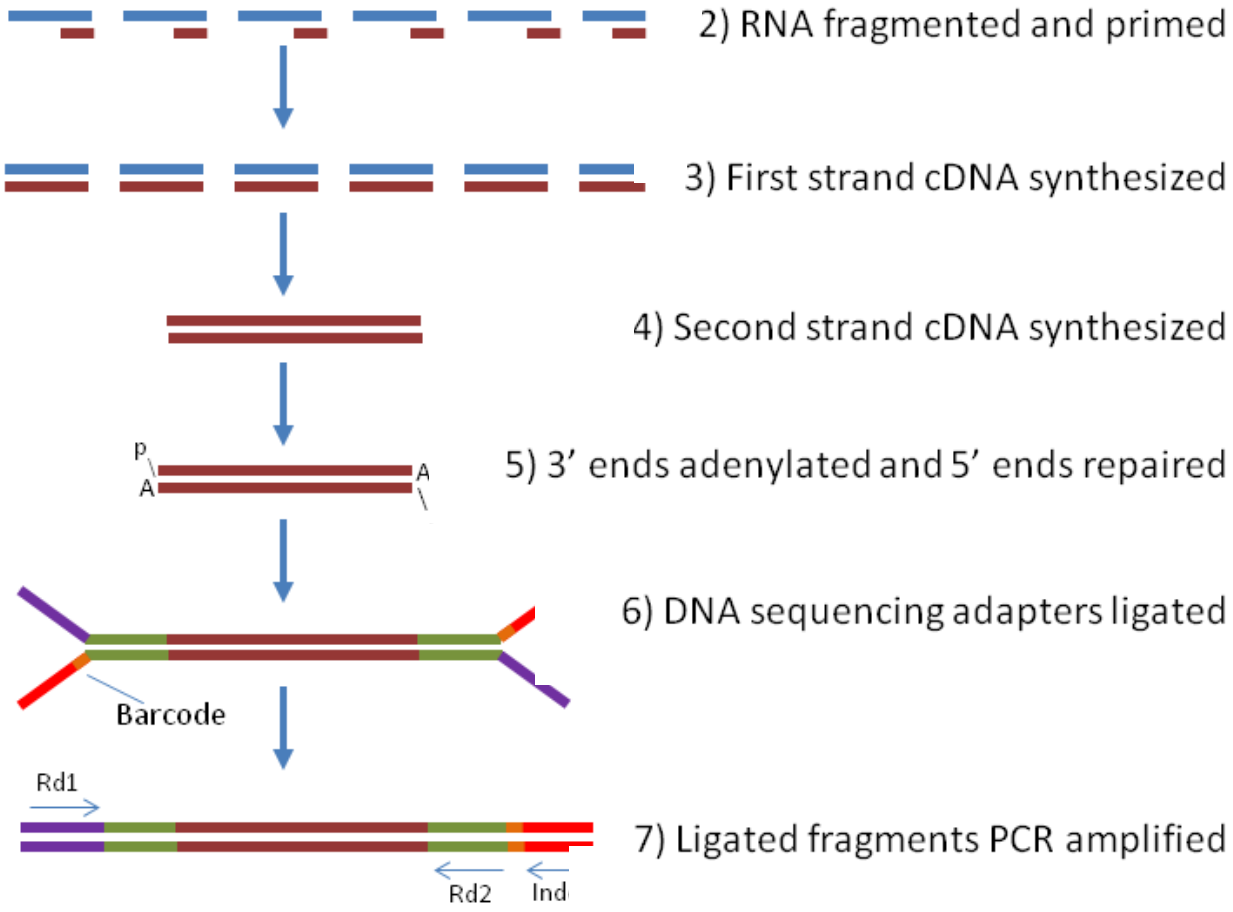
- **mRNA+lnc (polyA+) RNA seq**

1. DNase treatment
2. polyA enrichment (oligo-dT)
3. library preparation

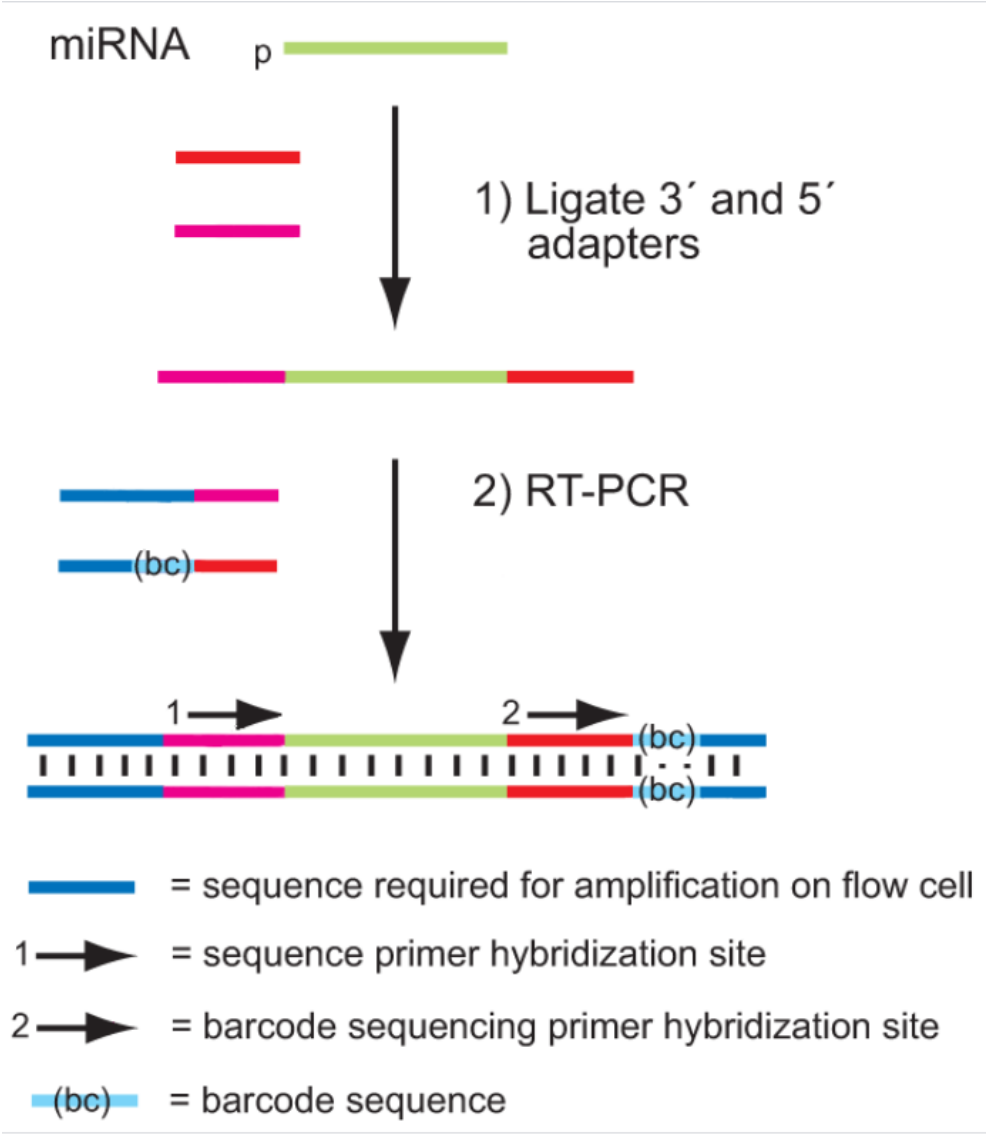
- **miRNA seq**

1. DNase treatment
2. Size selection
3. library preparation

Library preparation: mRNA-seq



Library preparation: miRNA

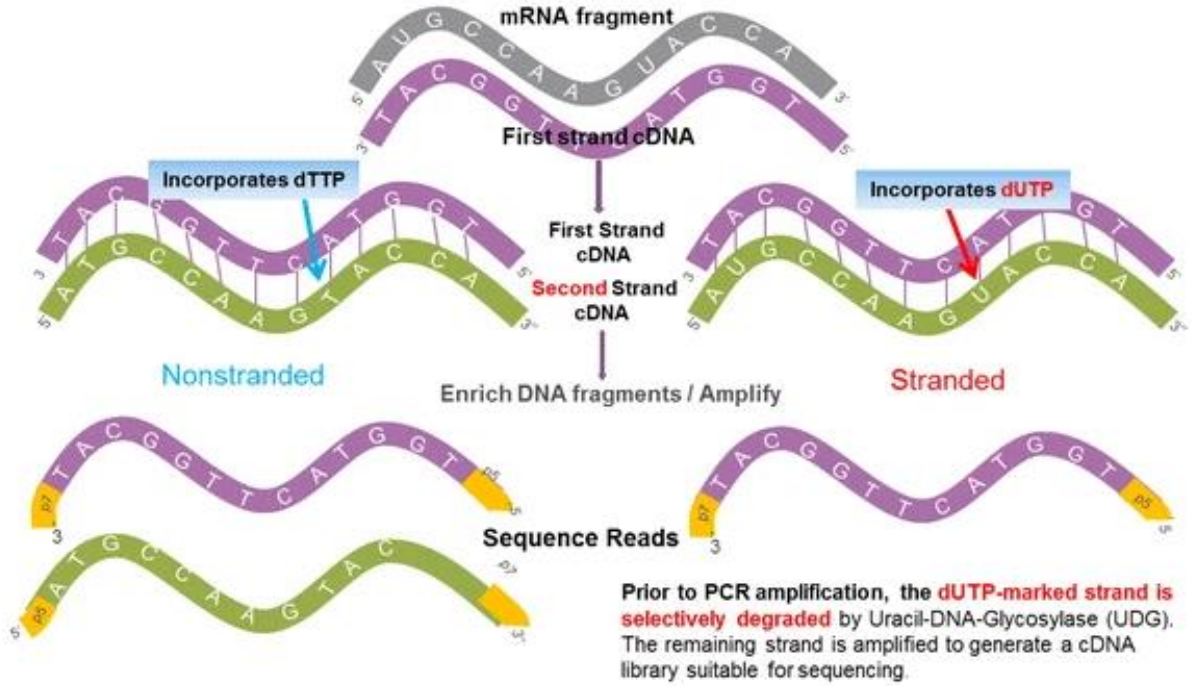


Stranded vs non-stranded

The stranded protocol differs from the non-stranded protocol in two ways.

1) During cDNA synthesis, the second-strand synthesis continues as normal except the nucleotide mix includes dUTPs instead of dTTPs.

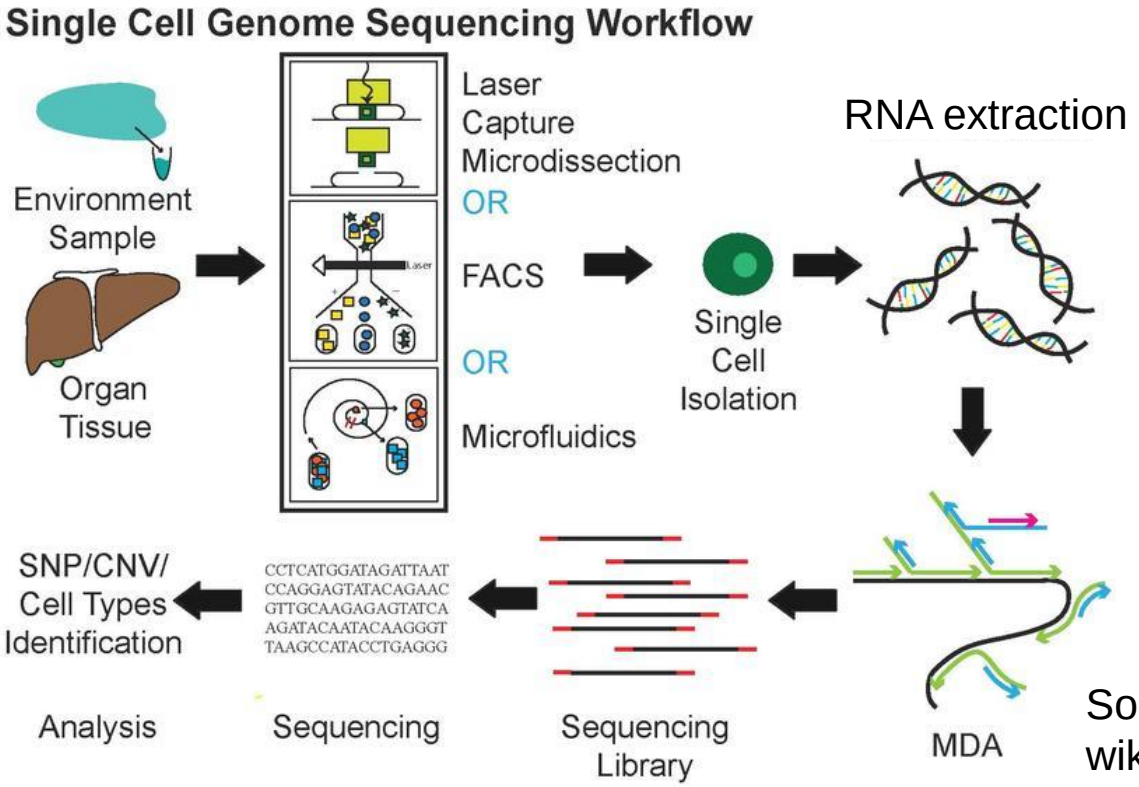
2) After library preparation, a second-strand digestion step is added. This step ensures that only the first strand survives the subsequent PCR amplification step and hence the strand information of the libraries



Zhao et al. 2015, BMC genomics

Single cell RNA-seq (scRNA-seq)

The first, and most important, step in conducting scRNA-seq has been the effective isolation of viable, single cells from the tissue of interest.

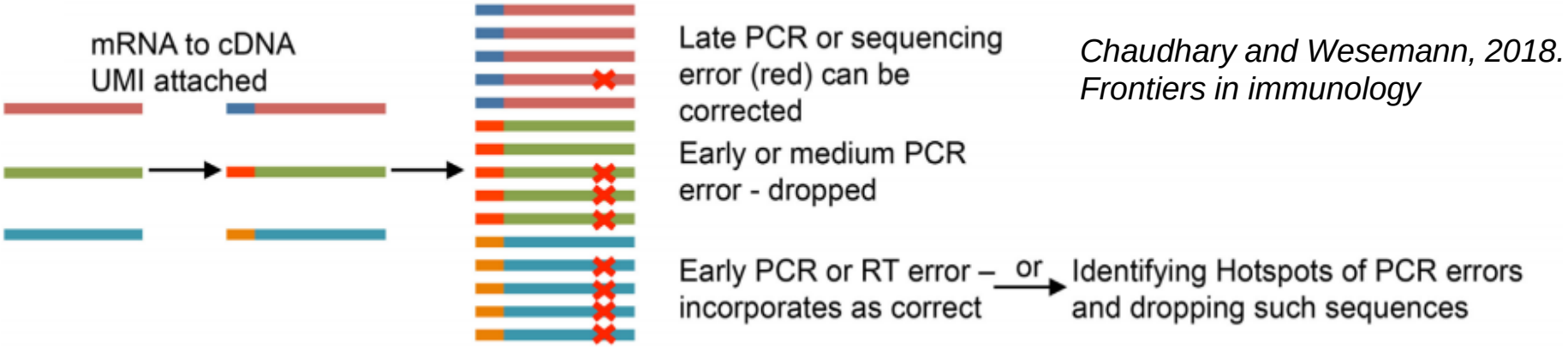


Source figure: wikipedia

Due to the efficiency of reverse transcription and other noise introduced in the experiments, more cells are required for accurate expression analyses and cell type identification

Unique mocecular identifiers (UMI)

Unique molecular identifiers (**UMIs**), or **molecular barcodes** (**MBC**) are short sequences or molecular "tags" added to DNA fragments in some next generation sequencing library preparation protocols to identify the input DNA or RNA molecules. These tags are added before PCR amplification, and can be used to reduce errors and quantitative bias introduced by the amplification.



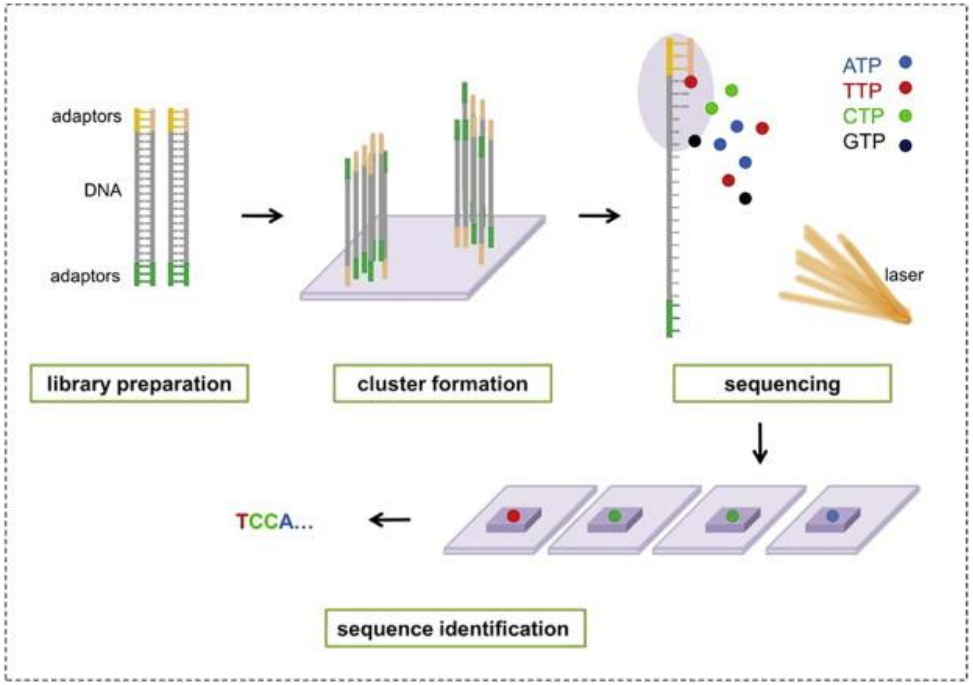
Chaudhary and Wesemann, 2018. Frontiers in immunology

Applications include variant calling and gene expression in RNA-seq

2nd Generation seq

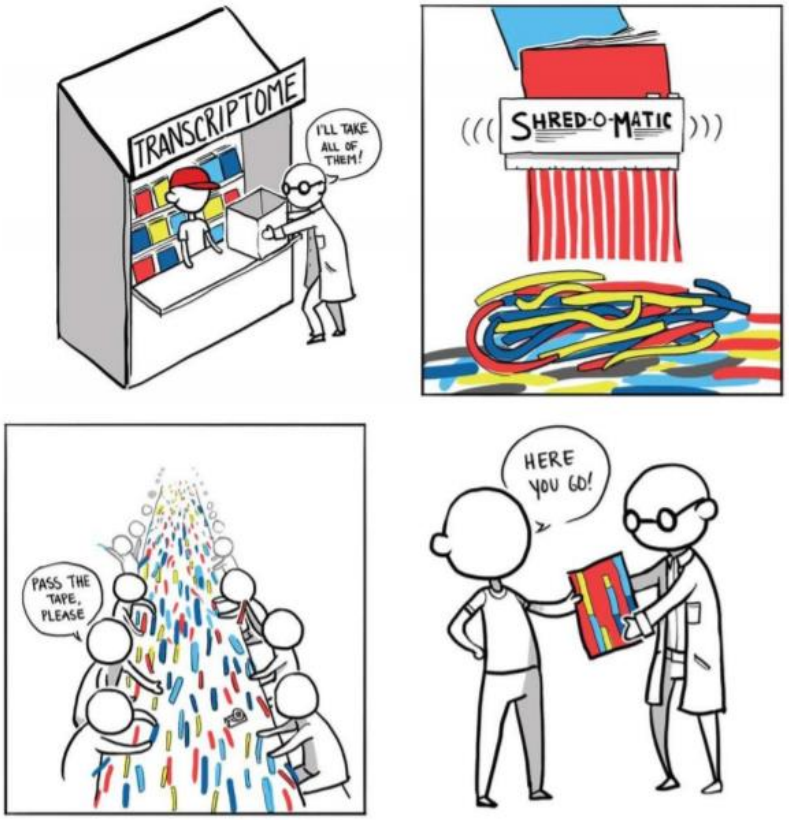
After library prep, the workflow is the same as the DNA-seq

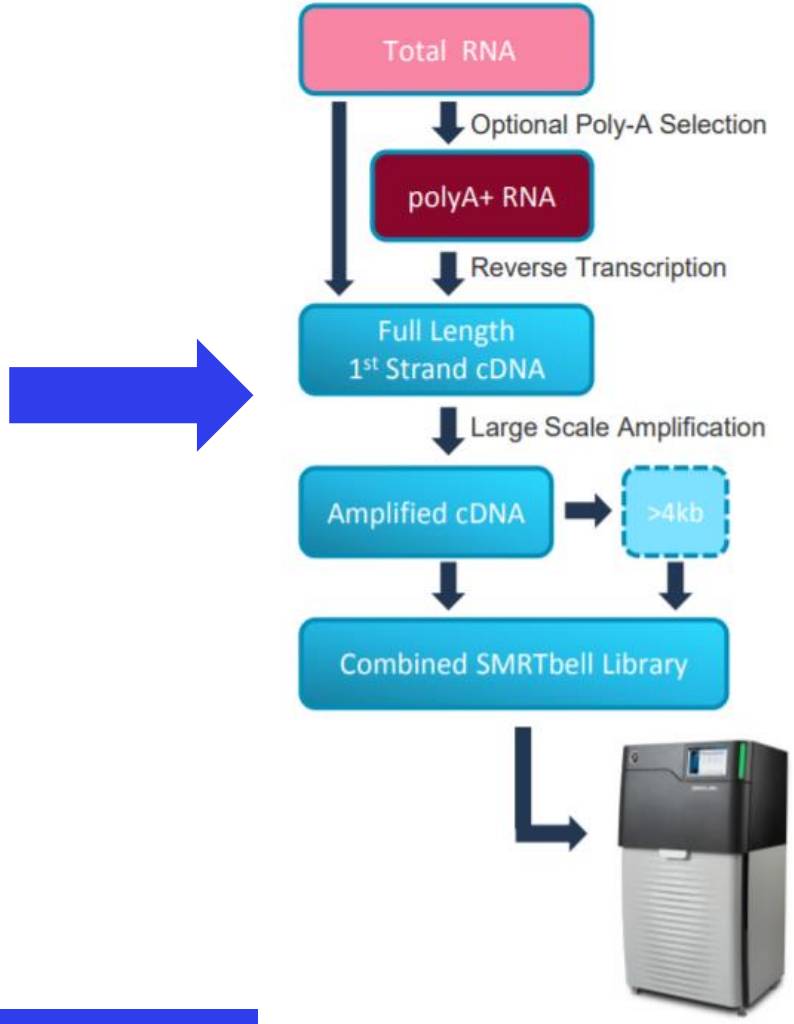
E.g.



Sequenced reads in Fastq format
(FastQC, trimming,…..)

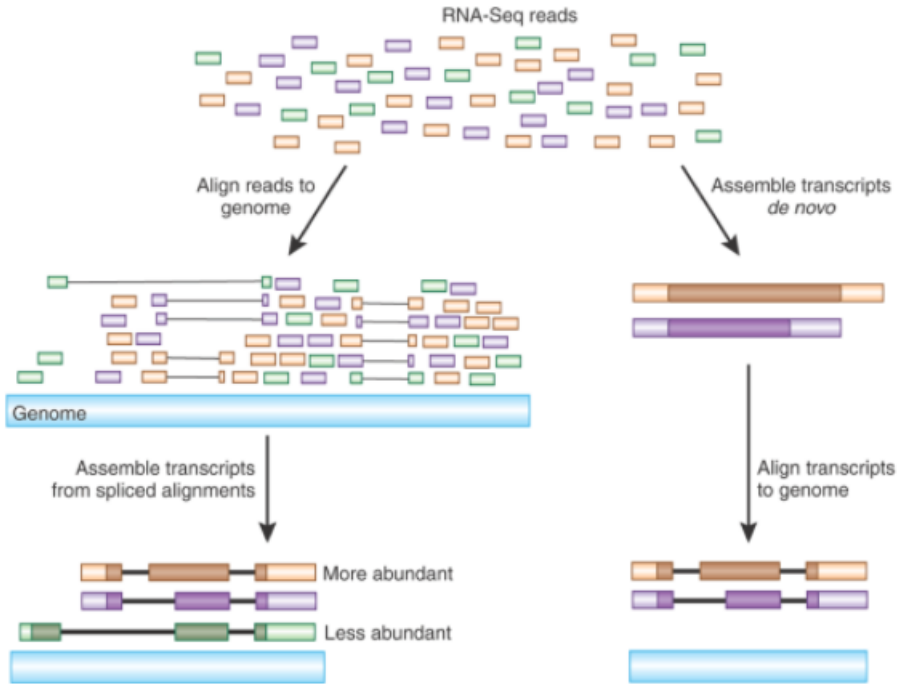
An eye on the 3rd generation seq





Read mapping strategies

- *De novo assembly*
- Reference-based
- Combined



Hass and Zody, Advancing RNA-Seq analysis, Nature Biotechnology 28:421-423

De novo assembly: Most common tools

- **Velvet**
  - ✓ Genomics and transcriptomics
- **Trinity**
  - ✓ Transcriptomics

Directional seq is a plus

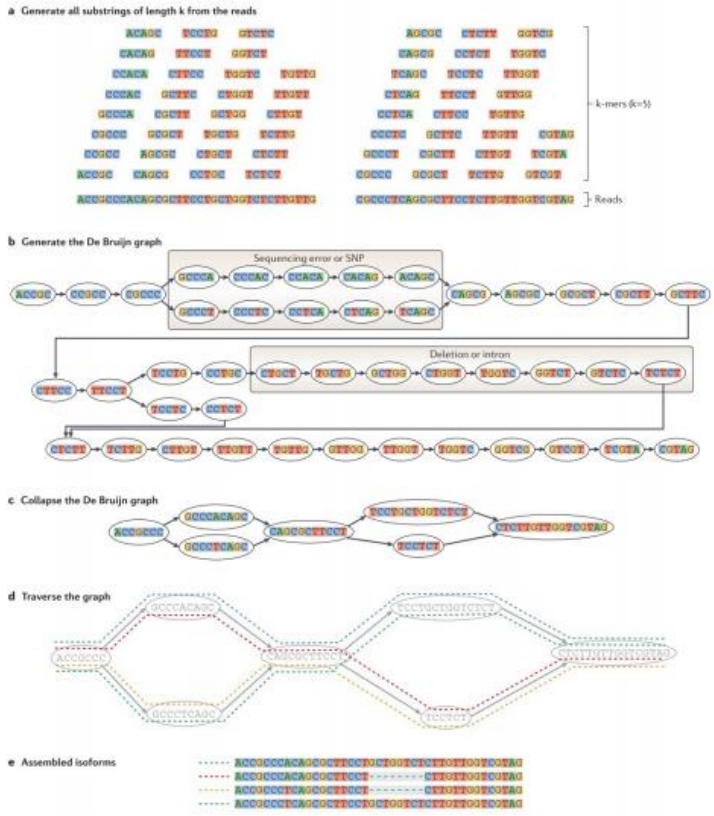
Kmers

De Bruijn graph

Collapse

Traverse graph

Assemble



Do you want to have a complete annotation map? Then different tissues, different developmental stages, different conditions, different sexes,…

Reference-based: Most common tools

- **Unspliced read aligner**

  ✓BWA
  ✓Bowtie2
  ✓Novoalign
  ✓Kallisto

  - Splice-junction not considered
  - Ideal for mapping against cDNA databases (also from de-novo outputs)

- **Spliced read aligner**

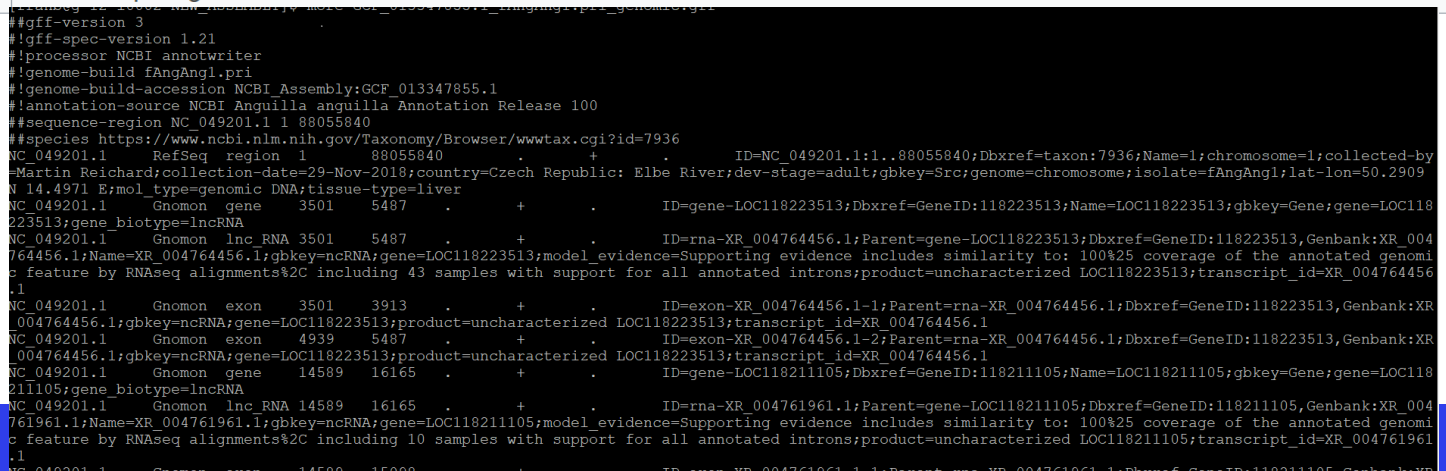
  ✓Tophat2
  ✓STAR
  ✓Hisat2

  - Novel splice-junction detected
  - Better performance for polymorphic regions and pseudogenes

ANNOTATION FILES

General GFF structure

Position index	Position name	Description
1	sequence	The name of the sequence where the feature is located.
2	source	Keyword identifying the source of the feature, like a program (e.g. Augustus or RepeatMasker) or an organization (like TAIR).
3	feature	The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow their parents in a single block (so all exons of a transcript are put after their parent "transcript" feature line and before any other parent transcript line). In GFF3, all features and their relationships should be compatible with the standards released by the Sequence Ontology Project.
4	start	Genomic start of the feature, with a **1-base offset**. This is in contrast with other 0-offset half-open sequence formats, like BED.
5	end	Genomic end of the feature, with a **1-base offset**. This is the same end coordinate as it is in 0-offset half-open sequence formats, like BED. [*citation needed*]
6	score	Numeric value that generally indicates the confidence of the source in the annotated feature. A value of "." (a dot) is used to define a null value.
7	strand	Single character that indicates the strand of the feature; it can assume the values of "+" (positive, or 5'->3'), "-", (negative, or 3'->5'), "." (undetermined).
8	phase	phase of CDS features; it can be either one of 0, 1, 2 (for CDS features) or "." (for everything else). See the section below for a detailed explanation.
9	attributes	All the other information pertaining to this feature. The format, structure and content of this field is the one which varies the most between the three competing file formats.



```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build fAngAng1.pri
#!genome-build-accession NCBI_Assembly:GCF_013347855.1
#!annotation-source NCBI Anguilla anguilla Annotation Release 100
##sequence-region NC_049201.1 1 88055840
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=7936
NC_049201.1     RefSeq  region  1       88055840        .       +       .       ID=NC_049201.1:1..88055840;Dbxref=taxon:7936;Name=1;chromosome=1;collected-by
=Martin Reichard;collection-date=29-Nov-2018;country=Czech Republic: Elbe River;dev-stage=adult;gbkey=Src;genome=chromosome;isolate=fAngAng1;lat-lon=50.2909
N 14.4971 E;mol_type=genomic DNA;tissue-type=liver
NC_049201.1     Gnomon  gene    3501    5487    .       +       .       ID=gene-LOC118223513;Dbxref=GeneID:118223513;Name=LOC118223513;gbkey=Gene;gene=LOC118
223513;gene_biotype=lncRNA
NC_049201.1     Gnomon  lnc_RNA 3501    5487    .       +       .       ID=rna-XR_004764456.1;Parent=gene-LOC118223513;Dbxref=GeneID:118223513,Genbank:XR_004
764456.1;Name=XR_004764456.1;gbkey=ncRNA;gene=LOC118223513;model_evidence=Supporting evidence includes similarity to: 100%25 coverage of the annotated genomi
c feature by RNAseq alignments%2C including 43 samples with support for all annotated introns;product=uncharacterized LOC118223513;transcript_id=XR_004764456
.1
NC_049201.1     Gnomon  exon    3501    3913    .       +       .       ID=exon-XR_004764456.1-1;Parent=rna-XR_004764456.1;Dbxref=GeneID:118223513,Genbank:XR
_004764456.1;gbkey=ncRNA;gene=LOC118223513;product=uncharacterized LOC118223513;transcript_id=XR_004764456.1
NC_049201.1     Gnomon  exon    4939    5487    .       +       .       ID=exon-XR_004764456.1-2;Parent=rna-XR_004764456.1;Dbxref=GeneID:118223513,Genbank:XR
_004764456.1;gbkey=ncRNA;gene=LOC118223513;product=uncharacterized LOC118223513;transcript_id=XR_004764456.1
NC_049201.1     Gnomon  gene    14589   16165   .       +       .       ID=gene-LOC118211105;Dbxref=GeneID:118211105;Name=LOC118211105;gbkey=Gene;gene=LOC118
211105;gene_biotype=lncRNA
NC_049201.1     Gnomon  lnc_RNA 14589   16165   .       +       .       ID=rna-XR_004761961.1;Parent=gene-LOC118211105;Dbxref=GeneID:118211105,Genbank:XR_004
761961.1;Name=XR_004761961.1;gbkey=ncRNA;gene=LOC118211105;model_evidence=Supporting evidence includes similarity to: 100%25 coverage of the annotated genomi
c feature by RNAseq alignments%2C including 10 samples with support for all annotated introns;product=uncharacterized LOC118211105;transcript_id=XR_004761961
.1
```

TOPHAT2

Kim et al. 2013

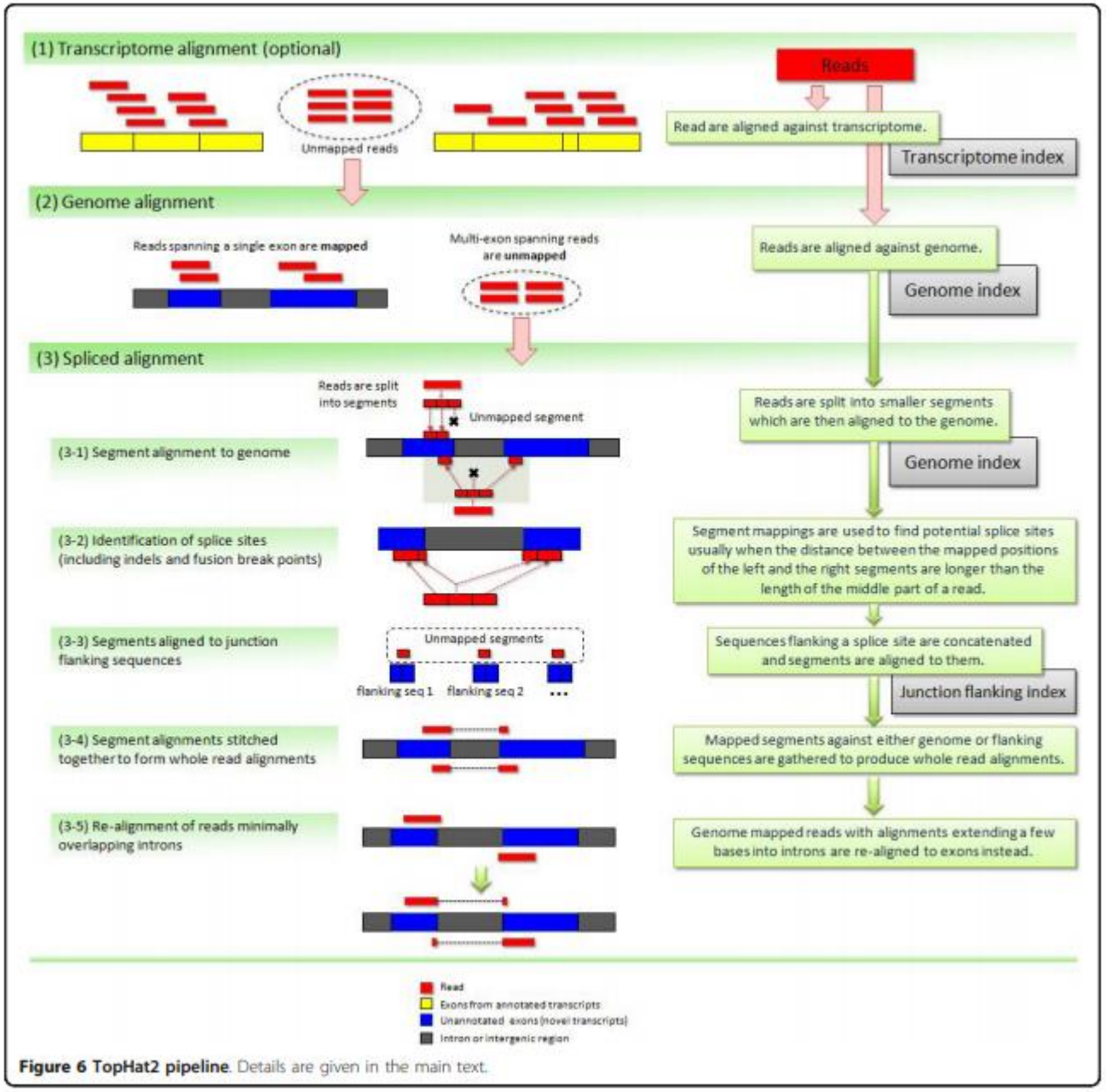
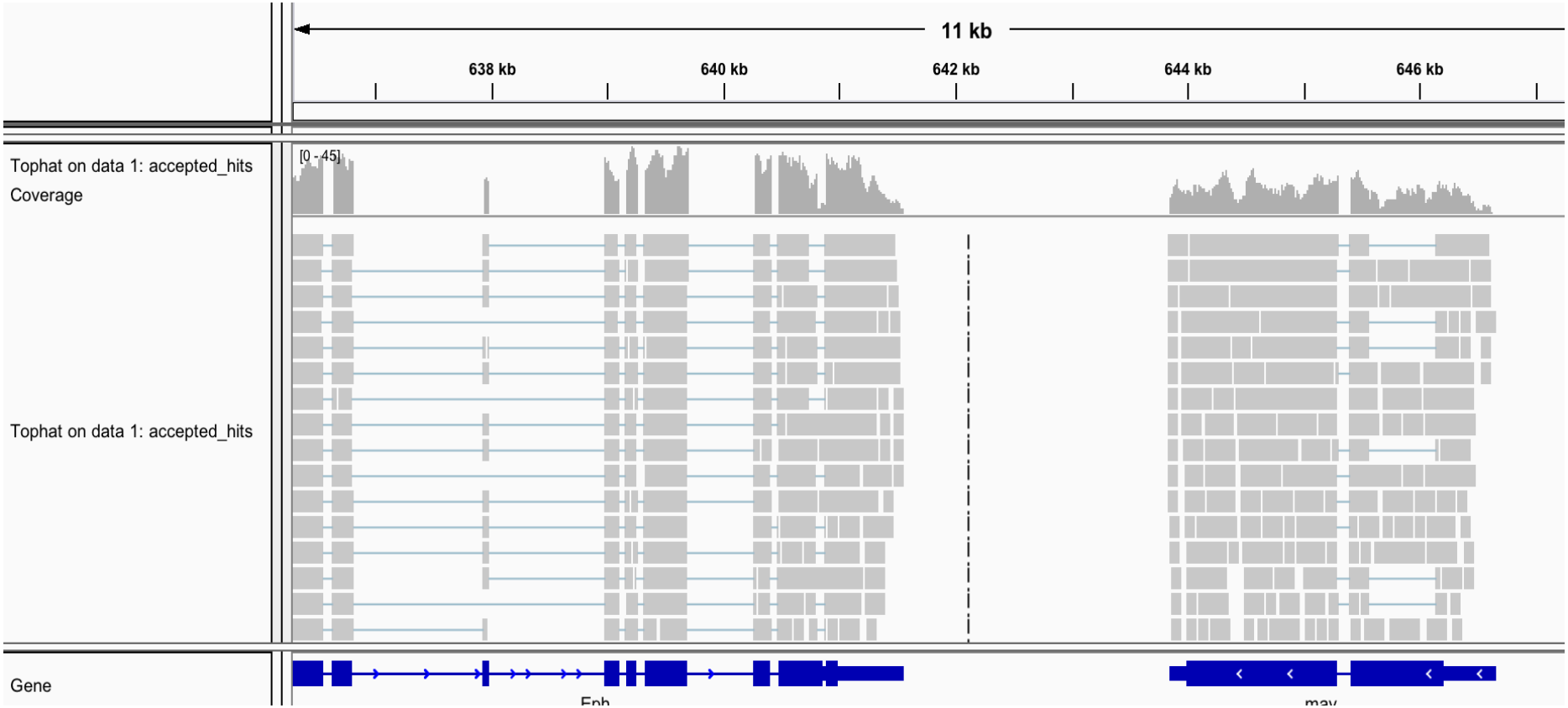


Figure 6 TopHat2 pipeline. Details are given in the main text.

Splice junctions view through IGV (Integrative Genomics Viewer)



REFERENCE-GUIDED ASSEMBLY

- **(Cufflinks/StringTie)**

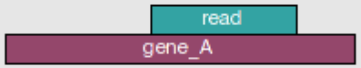
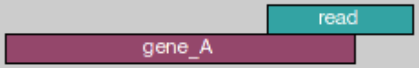
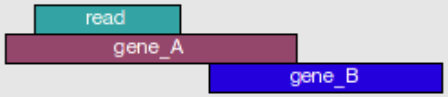
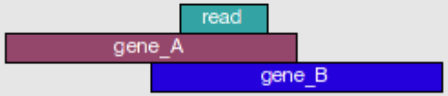
1) First you map all the reads from your experiment to the reference sequence.
2) Then you run another step where you use the mapped reads to assemble potential transcripts and identify the genomic locations of introns and exons.

The output is a de novo annotation file in gff/gtf format that can be used for read count

READ COUNT

Count the number of reads aligned to each known transcripts/isoform

E.g **HTSeq-count**
-It needs a gtf/gff file

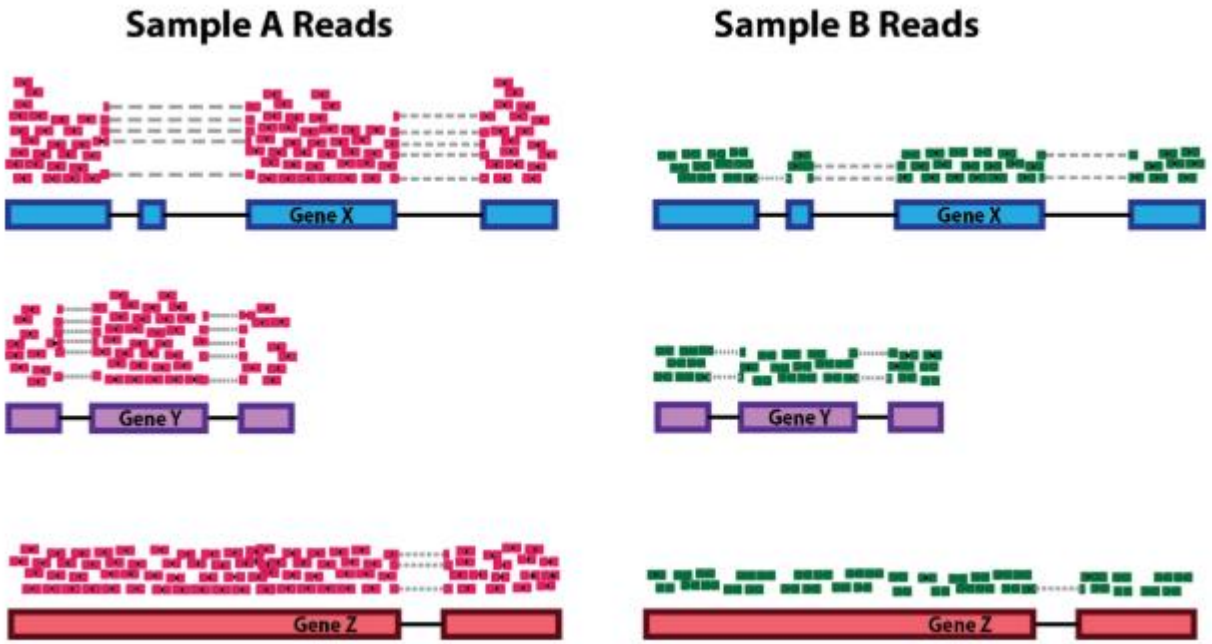
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

NORMALIZATION

- Longer genes will have more reads mapping to them (within samples)
- Sequencing run with more depth will have more reads mapping on each gene (between samples)

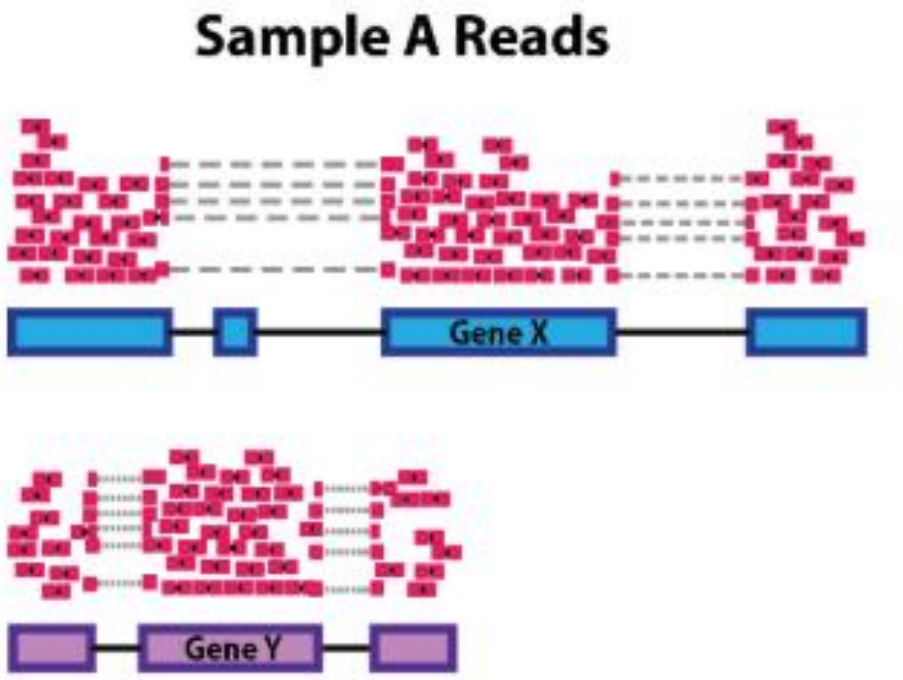
Main factors during normalization

Sequencing depth



Main factors during normalization

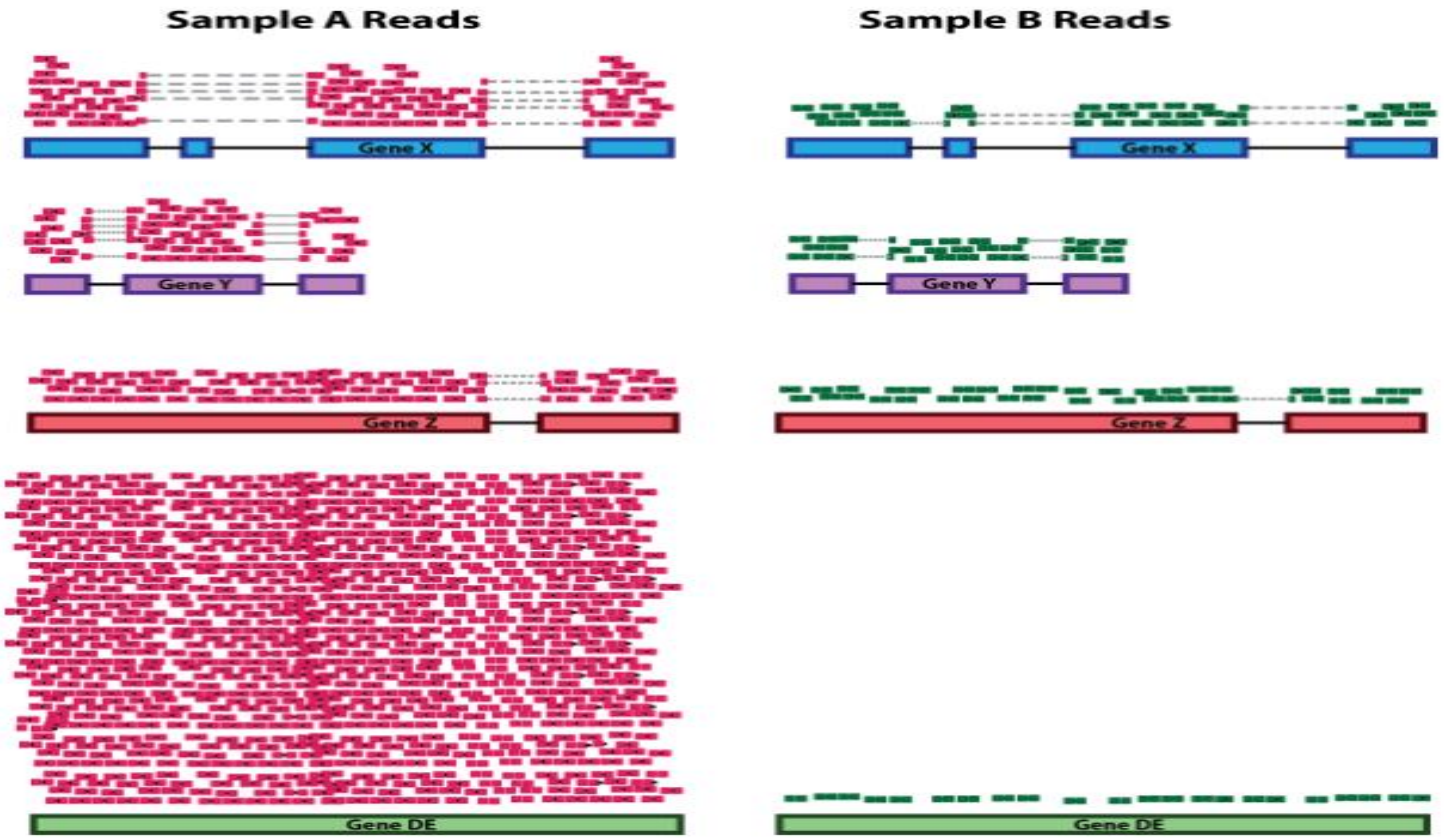
Gene length



Main factors during normalization

RNA composition

Anders and Huber , 2010 Genome Biol.



NORMALIZATION

Common normalization methods

Normalization method	Description	Accounted factors	Recommendations for use
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM(reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's median of ratios	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for DE analysis; **NOT for within sample comparisons**

Differential expression

DeSeq2

Differential gene expression analysis based on the negative binomial distribution

- **Input**: Read count tables (HTSeq)
- **Output**: Table containing statistics for whether a gene is differential expressed between two conditions

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 6 rows and 6 columns
##                 baseMean log2FoldChange     lfcSE      stat    pvalue      padj
##                <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## FBgn0039155          453          -3.71     0.160     -23.2 4.01e-119 3.11e-115
## FBgn0029167         2165          -2.08     0.104     -20.1  6.68e-90  2.59e-86
## FBgn0035085          367          -2.23     0.137     -16.3  1.89e-59  4.87e-56
## FBgn0029896          258          -2.21     0.159     -13.9  5.85e-44  1.13e-40
## FBgn0034736          118          -2.57     0.185     -13.9  8.07e-44  1.25e-40
## FBgn0040091          611          -1.43     0.120     -11.9  1.11e-32  1.44e-29
```

Gene id

Mean read count

Log2 fold change and standard error

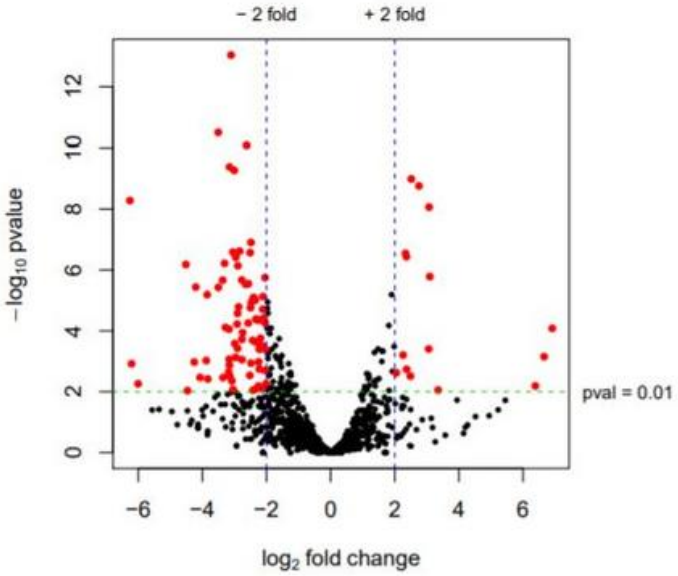
Test statistic

P-value and adjusted p-value

Why Log2fold change?

Fold change is a measure describing how much a quantity changes going from an initial to a final value. For example, an initial value of 30 and a final value of 60 corresponds to a fold change of 2 (or equivalently, a change to 2 times), or in common terms, a one-fold increase. **Fold change is calculated simply as the ratio of the difference between final value and the initial value over the original value.**

A disadvantage is that it is biased and may miss deferentially expressed genes with large differences (B-A) but small ratios (A/B), leading to a high miss rate at high intensities.



Let's say there are 50 read counts in control and 100 read counts in treatment for gene A. This means gene A is expressing twice in treatment as compared to control (100 divided by 50 =2) or fold change is 2. This works well for over expressed genes as the number directly corresponds to how many times a gene is over-expressed. But when it is other way round (i.e, treatment 50, control 100), the value of fold change will be 0.5 (all under expressed genes will have values between 0 to 1, while over expressed genes will have values from 1 to infinity). To make this leveled, we use **log2** for expressing the fold change. I.e, log2 of 2 is 1 and log2 of 0.5 is -1.

Functional enrichment analysis

Identification of classes of genes that are over-represented among the differentially expressed genes, and may have an association with the disease/phenotype investigated

Gene Ontology project provides an ontology of **defined terms** representing gene product properties. The ontology covers three domains:

- **Molecular function:** molecular activities of gene products
- **Cellular component:** where gene products are active
- **Biological process:** pathways and larger processes made up of the activities of multiple gene products.

Biological databases available

- Gene Ontology (GO)
- KEGG (Kyoto Encyclopedia of Genes and Genomes)
- Reactome
- Ingenuity Pathway Analysis (IPA)
- MSigDB (Molecular Signatures Database)
- DAVID (Database for Annotation, Visualization and Integrated Discovery)
- Panther
- Gorilla

Some GO and pathway analyses websites

Enrichr

http://amp.pharm.mssm.edu/Enrichr/

http://cbl-gorilla.cs.technion.ac.il/

https://david.ncifcrf.gov/

IPA®

https://cytoscape.org/

ARE YOU LOOKING FOR THESIS/PROJECT?

You can learn more about RNA-seq and its application in fish:

- ecology
- health
- aquaculture

You can learn more about NGS and its application in fish with the course **25334 Genomic methods in breeding and management of aquatic living resources** (fall 2021)