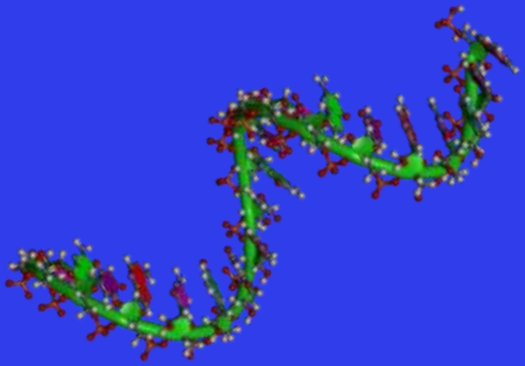
RNA-seq

Next Generation Sequencing Analysis, 2022



Francesca Bertolini
DTU Aqua
franb@aqua.dtu.dk

Lecture program

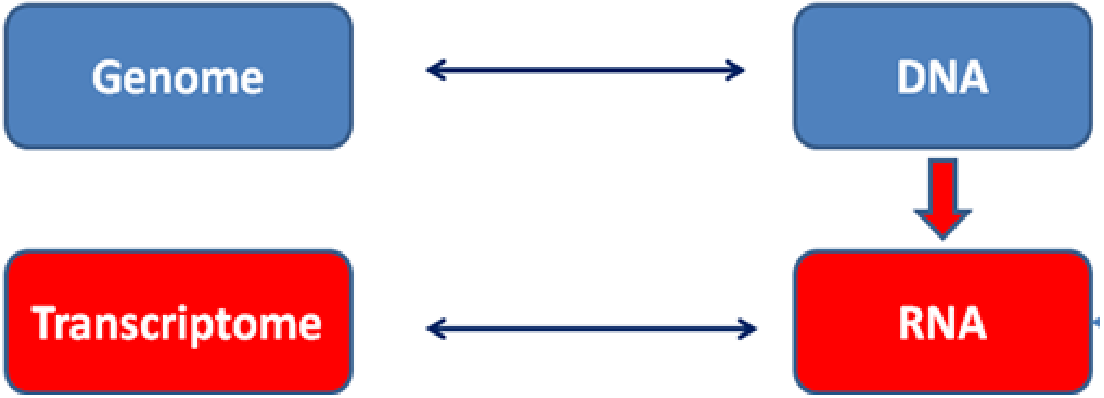
1. Types of RNAs and RNA quality
2. RNA-seq: the basics
3. RNA-seq: some variations
4. Data analyses
   - Alignment
   - Read count
   - Differential expression
   - Gene enrichment

Lecture program

1. **Types of RNAs and RNA quality**
2. RNA-seq: the basics
3. RNA-seq: some variations
4. Data analyses
   - ❖ Alignment
   - ❖ Read count
   - ❖ Differential expression
   - ❖ Gene enrichment

Transcriptome

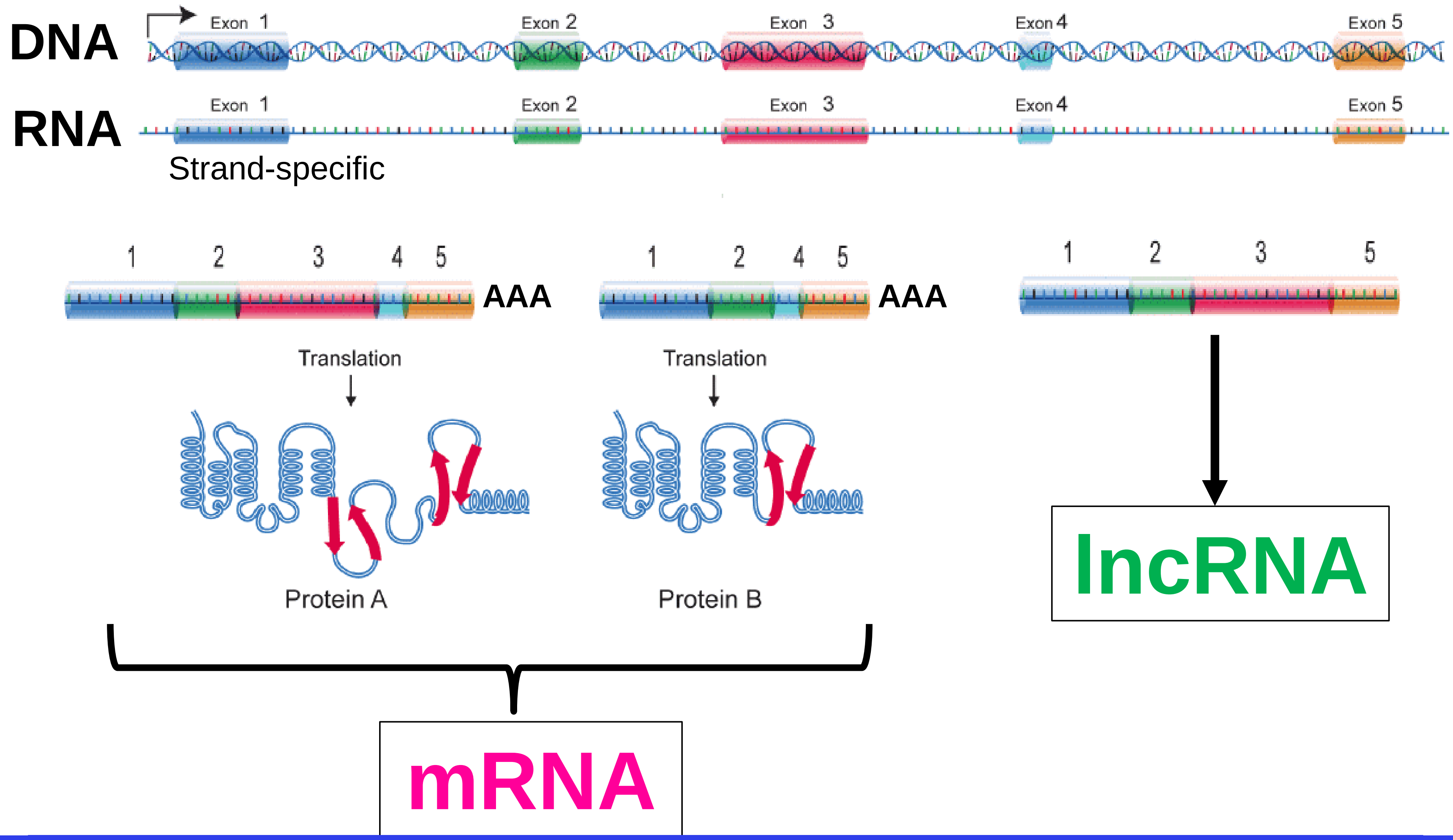
- The transcriptome is defined as the complete set of transcripts (RNA) in a cell, and their quantity, for a specific developmental stage or physiological condition (Wang et al. 2009).
- Transcriptome is therefore dynamic and a good representative of the cellular and tissue state (Srivastava et al 2019).



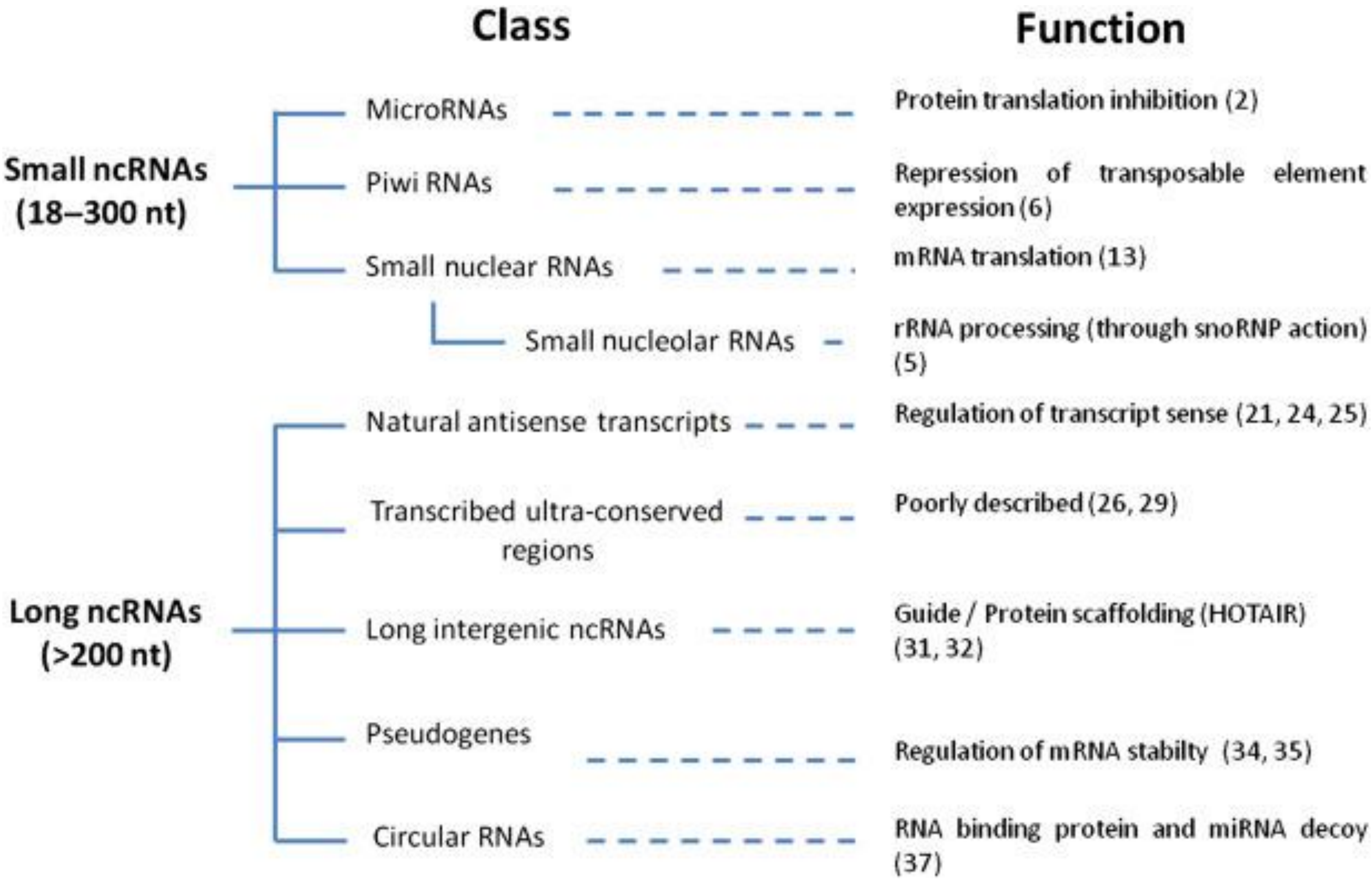
RNA classification

- **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**

mRNAs: splicing



Non coding regulatory RNAs



	Class	Function
Small ncRNAs (18–300 nt)	MicroRNAs	Protein translation inhibition (2)
	Piwi RNAs	Repression of transposable element expression (6)
	Small nuclear RNAs	mRNA translation (13)
	Small nucleolar RNAs	rRNA processing (through snoRNP action) (5)
Long ncRNAs (>200 nt)	Natural antisense transcripts	Regulation of transcript sense (21, 24, 25)
	Transcribed ultra-conserved regions	Poorly described (26, 29)
	Long intergenic ncRNAs	Guide / Protein scaffolding (HOTAIR) (31, 32)
	Pseudogenes	Regulation of mRNA stabilty (34, 35)
	Circular RNAs	RNA binding protein and miRNA decoy (37)

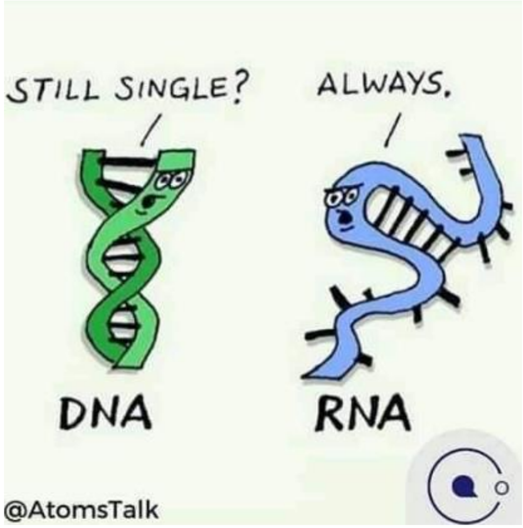
Delpu et al. 2016. *Drug Discovery in Cancer Epigenetics*

Before RNA extraction

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

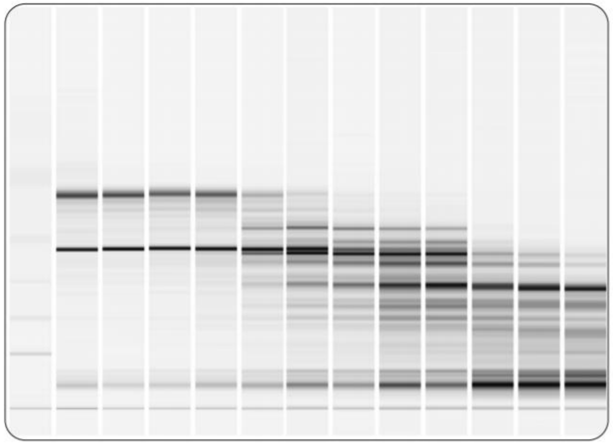
TISSUE COLLECTION:

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

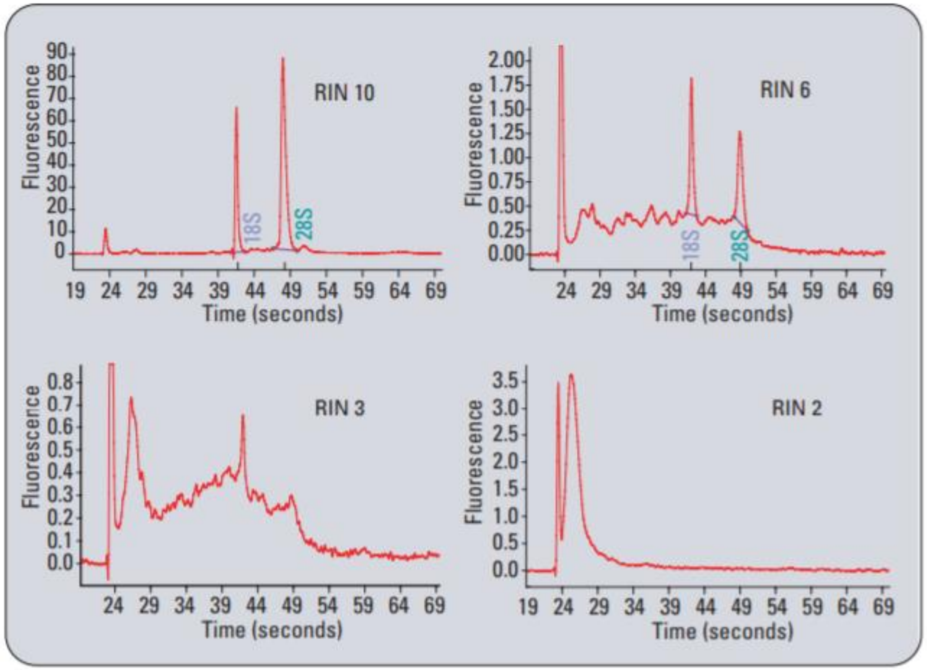


After RNA extraction

RIN (RNA integrity number)**:** algorithm for assigning integrity values to RNA measurements. ***RIN >7 is ok***



Bioanalyzer



Lecture program

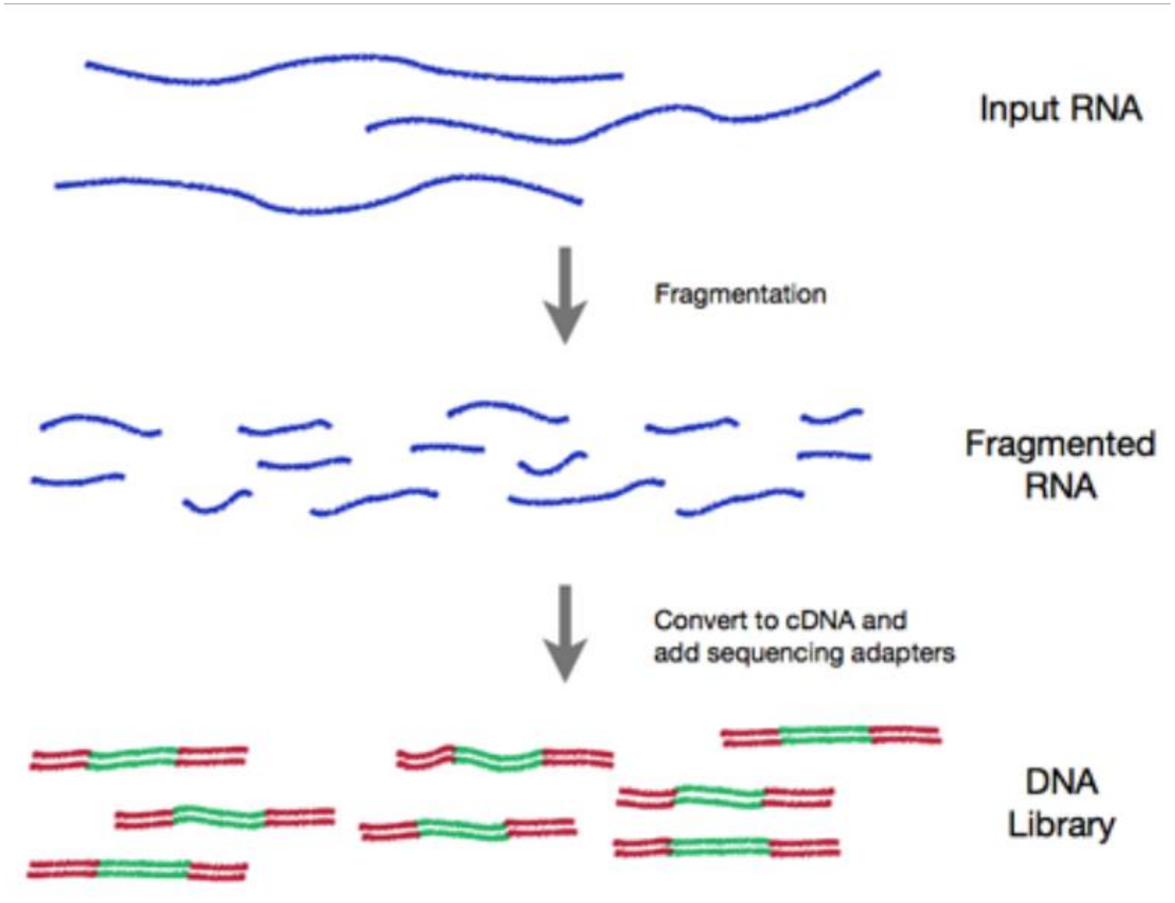
1. Types of RNAs and RNA quality
2. **RNA-seq: the basics**
3. RNA-seq: some variations
4. Data analyses
   - ❖ Alignment
   - ❖ Read count
   - ❖ Differential expression
   - ❖ Gene enrichment

RNA-seq

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

- Alternative splicing
- RNA editing
- Novel transcripts
- Allele specific expression
- Fusion transcripts
- **Abundance estimation/differential expression**

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. Fragmentation
4. library preparation

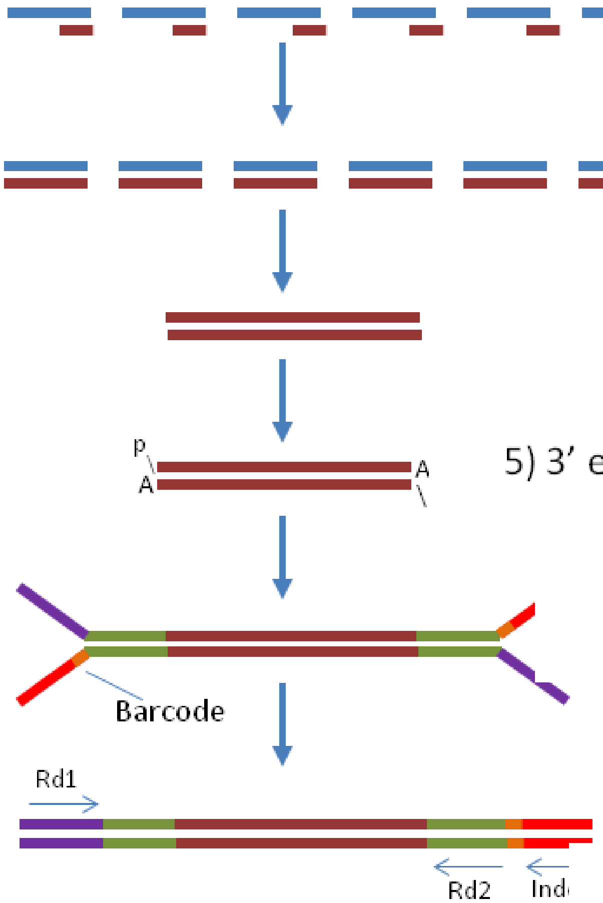
- **mRNA seq**

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

- **miRNA seq**

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

Library preparation: mRNA-seq



2) RNA fragmented and primed

3) First strand cDNA synthesized

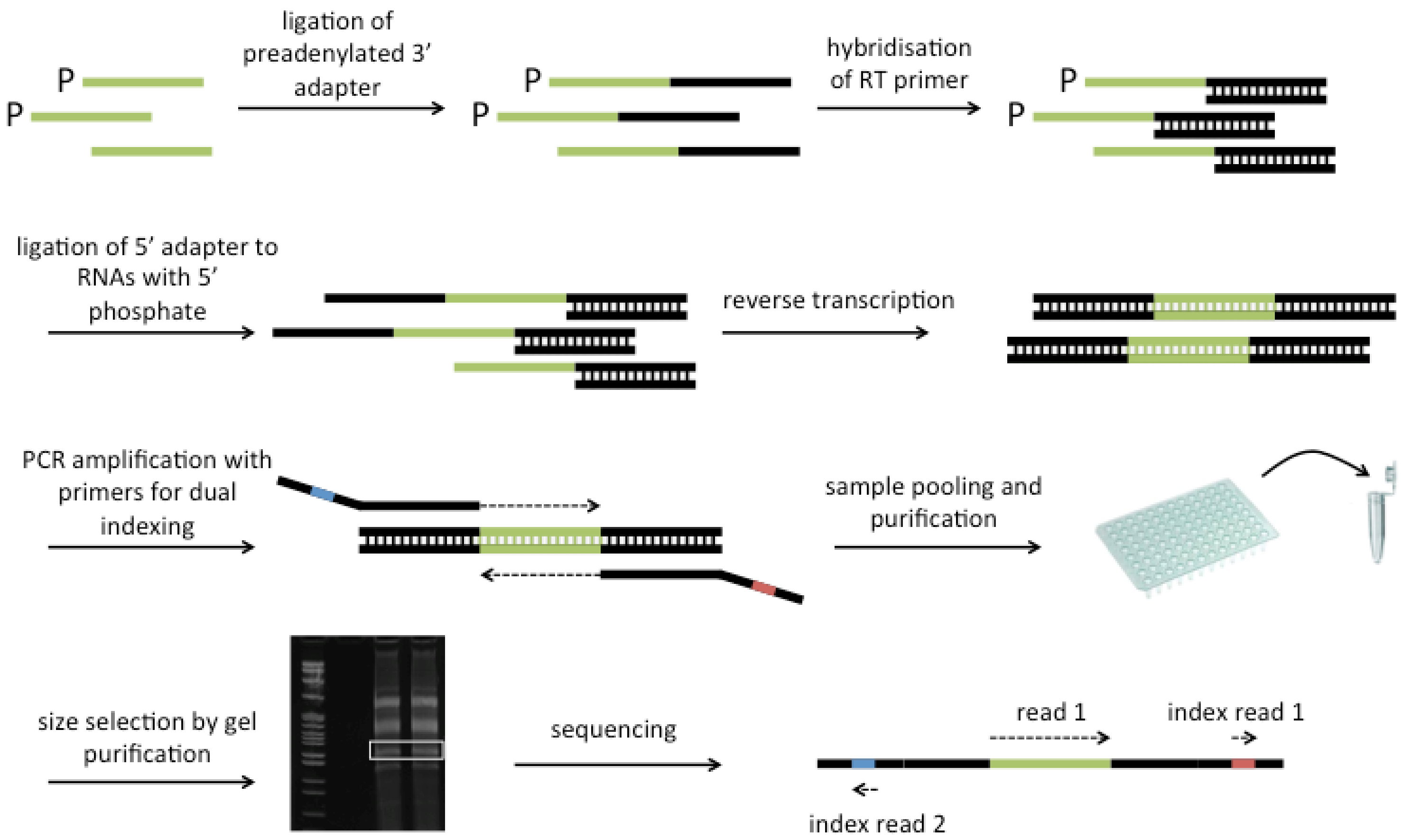
4) Second strand cDNA synthesized

5) 3’ ends adenylated and 5’ ends repaired

6) DNA sequencing adapters ligated

7) Ligated fragments PCR amplified

Library preparation: miRNA-seq

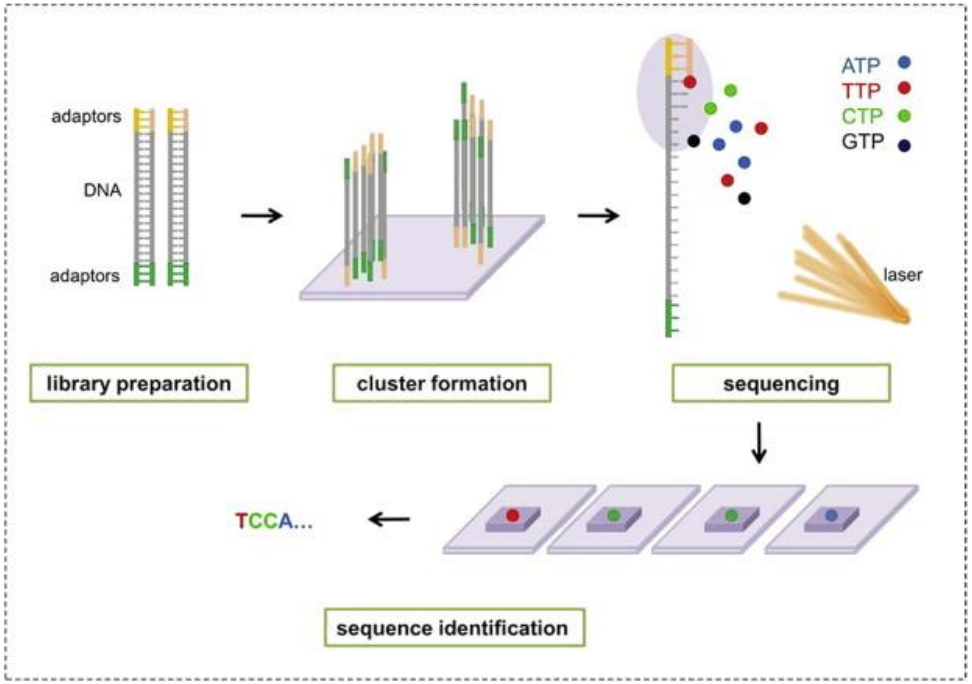

ligation of preadenylated 3' adapter
hybridisation of RT primer
P
P
P
P
P
P
ligation of 5' adapter to RNAs with 5' phosphate
reverse transcription
PCR amplification with primers for dual indexing
sample pooling and purification
size selection by gel purification
sequencing
read 1
index read 1
index read 2

https://www.rna-seqblog.com/preparation-of-highly-multiplexed-small-rna-sequencing-libraries/

2nd Generation seq

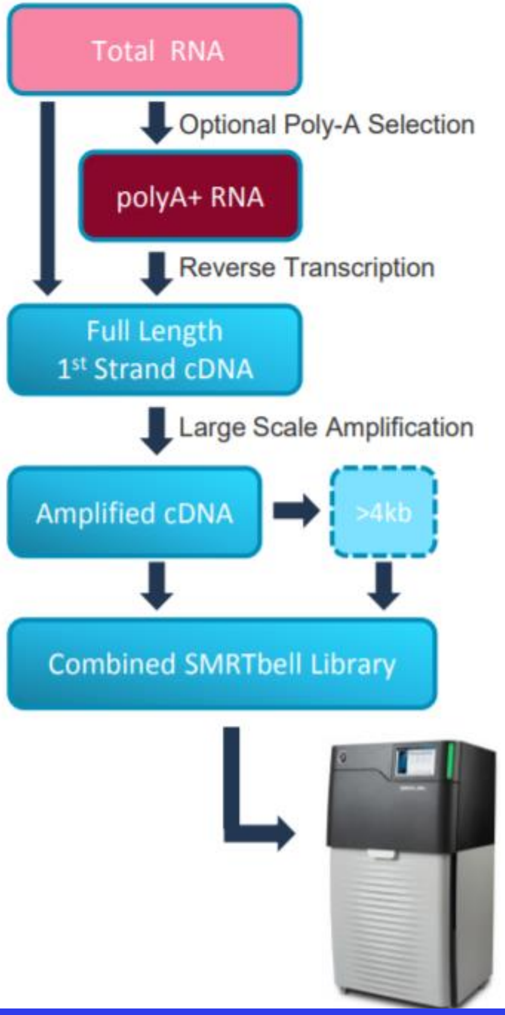
After library prep, the workflow is the same as the DNA-seq

E.g.



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

An eye on the 3rd generation seq



Total RNA

Optional Poly-A Selection

polyA+ RNA

Reverse Transcription

Full Length 1st Strand cDNA

Large Scale Amplification

Amplified cDNA

>4kb

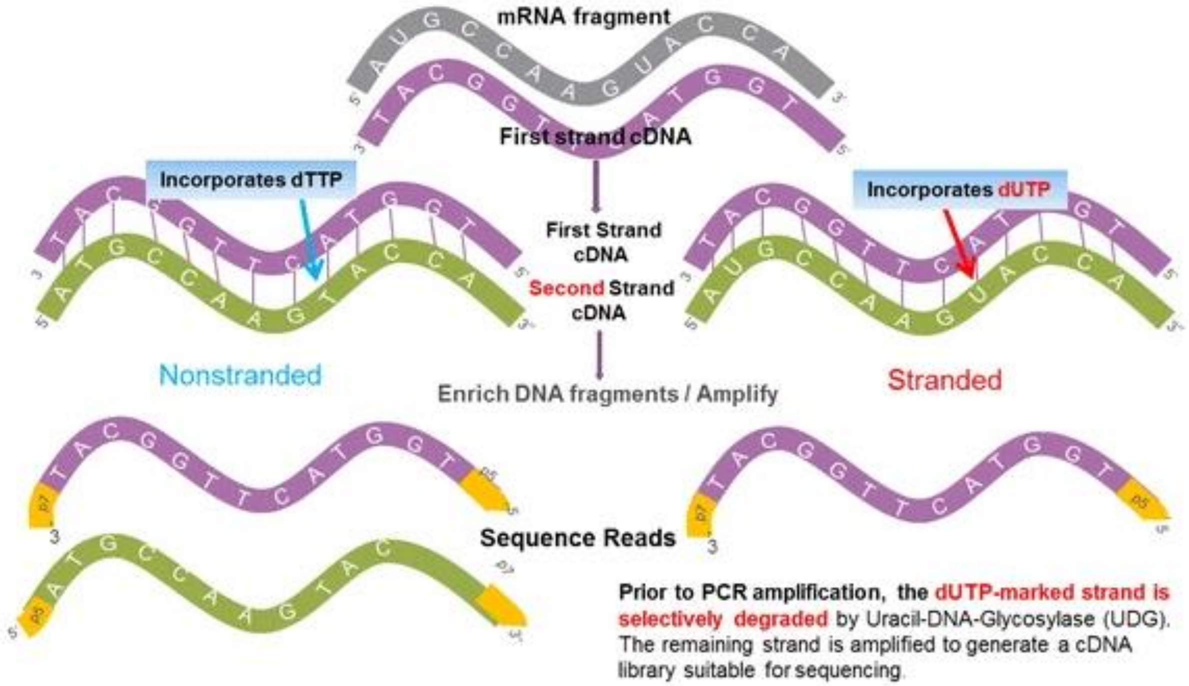
Combined SMRTbell Library

Lecture program

1. Types of RNAs and RNA quality
2. RNA-seq: the basics
3. **RNA-seq: some variations**
4. Data analyses
   - ❖ Alignment
   - ❖ Read count
   - ❖ Differential expression
   - ❖ Gene enrichment

Stranded sequencing

Standard RNA-seq protocol does not retain the strand specificity of origin for each transcript. Without strand information it is difficult and sometimes impossible to accurately quantify gene expression levels for genes with overlapping genomic loci that are transcribed from opposite strands.

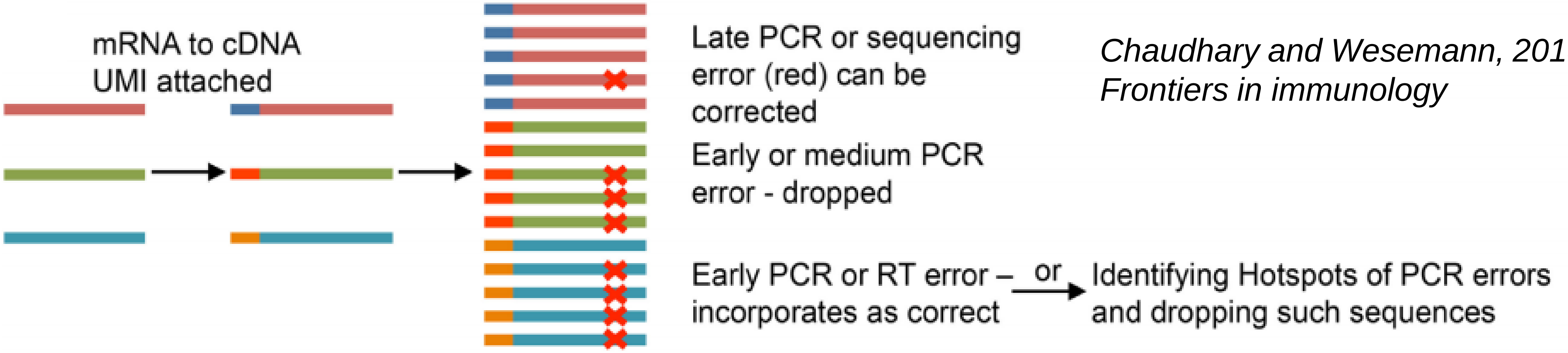


Prior to PCR amplification, the dUTP-marked strand is selectively degraded by Uracil-DNA-Glycosylase (UDG). The remaining strand is amplified to generate a cDNA library suitable for sequencing.

Zhao et al. 2015, BMC genomics

Unique molecular identifiers (UMI)

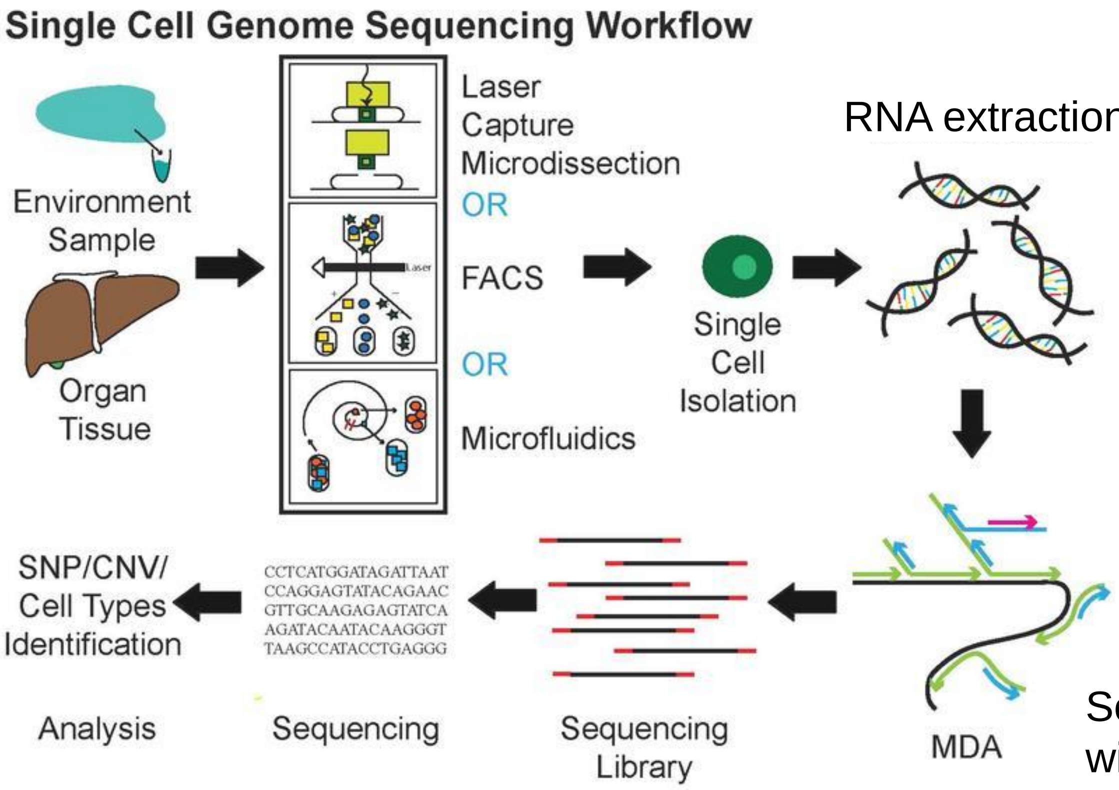
Unique molecular identifiers (**UMIs**), or **molecular barcodes** (**MBC**) are short sequences or molecular "tags" added to DNA fragments in some next generation sequencing library preparation protocols to identify the input DNA or RNA molecules. These tags are added before PCR amplification, and can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


mRNA to cDNA
UMI attached
Late PCR or sequencing error (red) can be corrected
Early or medium PCR error - dropped
Early PCR or RT error – incorporates as correct
or
Identifying Hotspots of PCR errors and dropping such sequences

Chaudhary and Wesemann, 2018. Frontiers in immunology

Single cell RNA-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.



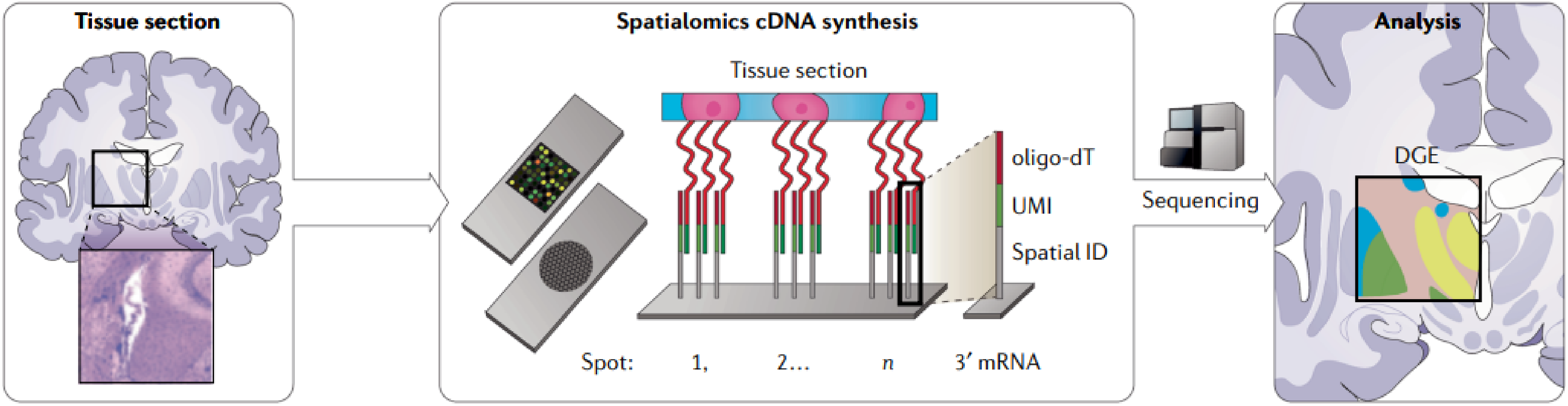
RNA extraction

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

Spatialomics

Current bulk and scRNA-seq methods provide users with highly detailed data regarding tissues or cell populations but do not capture spatial information, which reduces the ability to determine how cellular context relates to gene expression.

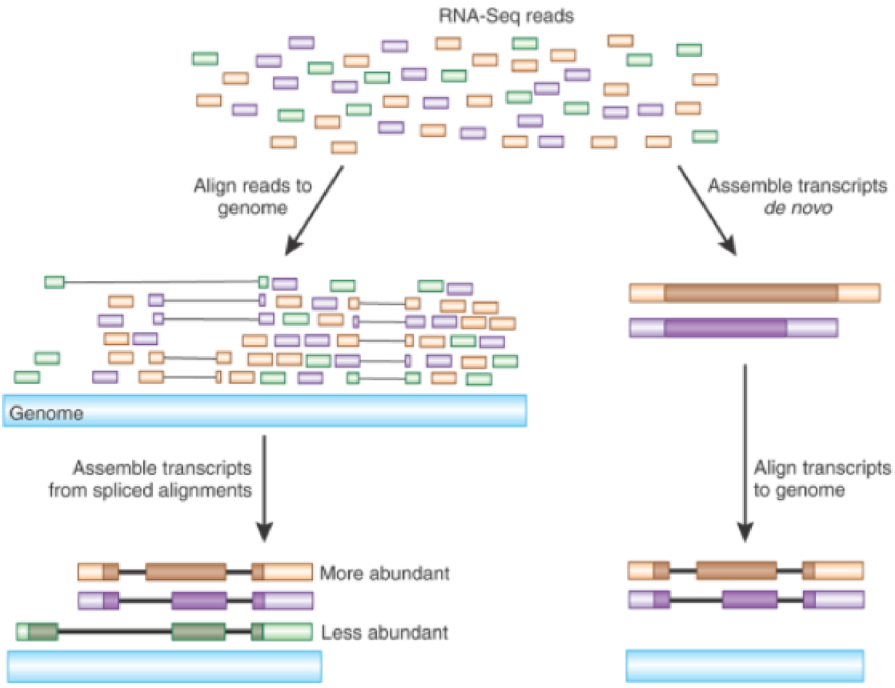


Lecture program

1. Types of RNAs and RNA quality
2. RNA-seq: the basics
3. RNA-seq: what more
4. Data analyses
    - ❖ Alignment
    - ❖ Read count
    - ❖ Differential expression
    - ❖ Gene enrichment

Read mapping strategies

- ❑ **Reference genome-based** - an assembled genome exists for a species for which an RNAseq experiment is performed. It allows reads to be aligned against the reference genome and significantly improves our ability to reconstruct transcripts.
- ❑ **Reference genome-free** - no genome assembly for the species of interest is available. In this case one would need to assemble the reads into transcripts using *de novo* approaches.


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

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

De novo assembly: Most common tools

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

Stranded seq is a plus

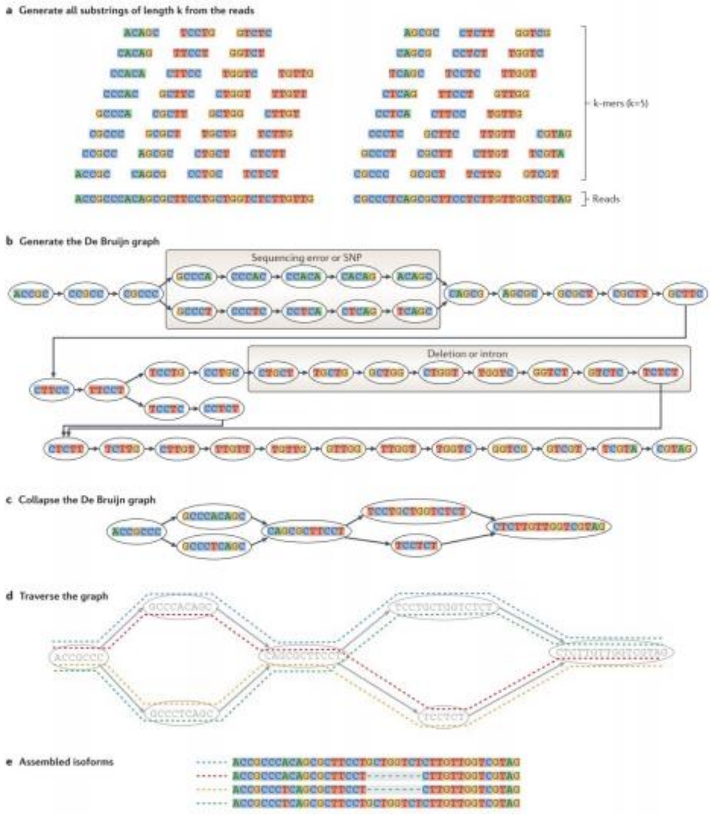
Kmers

De Bruijn graph

Collapse

Traverse graph

Assemble


a Generate all substrings of length k from the reads
k-mers (k=5)
Reads
b Generate the De Bruijn graph
Sequencing error or SNP
Deletion or intron
c Collapse the De Bruijn graph
d Traverse the graph
e Assembled isoforms

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)

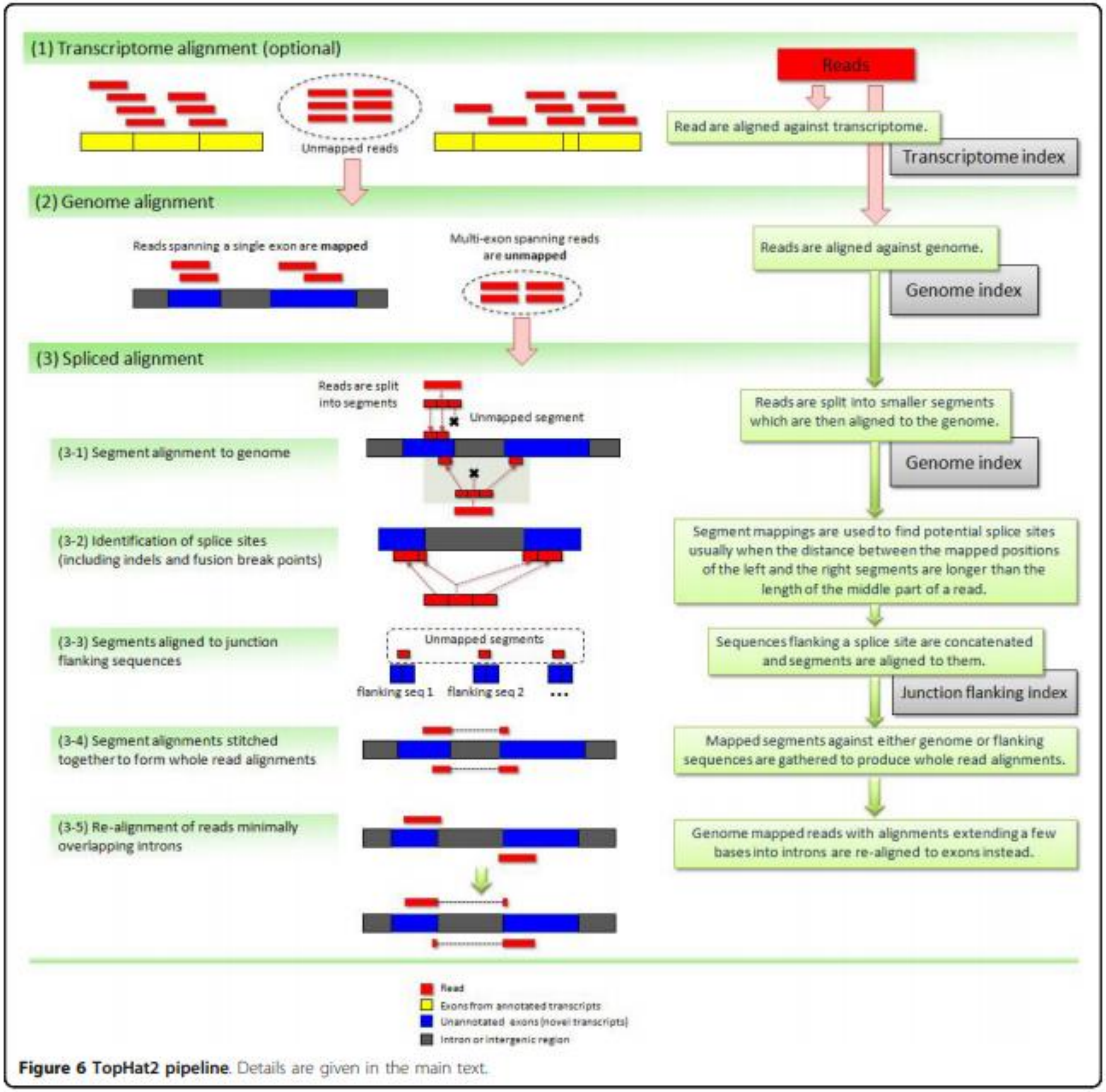
- **Spliced read aligner**

✓**Tophat2**
✓STAR
✓Hisat2

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

TOPHAT2

Kim et al. 2013


(1) Transcriptome alignment (optional)
Reads
Unmapped reads
Transcriptome index
(2) Genome alignment
Reads spanning a single exon are mapped
Multi-exon spanning reads are unmapped
Genome index
(3) Spliced alignment
Reads are split into segments
Unmapped segment
(3-1) Segment alignment to genome
Genome index
(3-2) Identification of splice sites (including indels and fusion break points)
(3-3) Segments aligned to junction flanking sequences
Unmapped segments
flanking seq 1
flanking seq 2
...
Junction flanking index
(3-4) Segment alignments stitched together to form whole read alignments
(3-5) Re-alignment of reads minimally overlapping introns
🟥 Read
🟨 Exons from annotated transcripts
🟦 Unannotated exons (novel transcripts)
⬛ Intron or intergenic region

Read are aligned against transcriptome.

Reads are aligned against genome.

Reads are split into smaller segments which are then aligned to the genome.

Segment mappings are used to find potential splice sites usually when the distance between the mapped positions of the left and the right segments are longer than the length of the middle part of a read.

Sequences flanking a splice site are concatenated and segments are aligned to them.

Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.

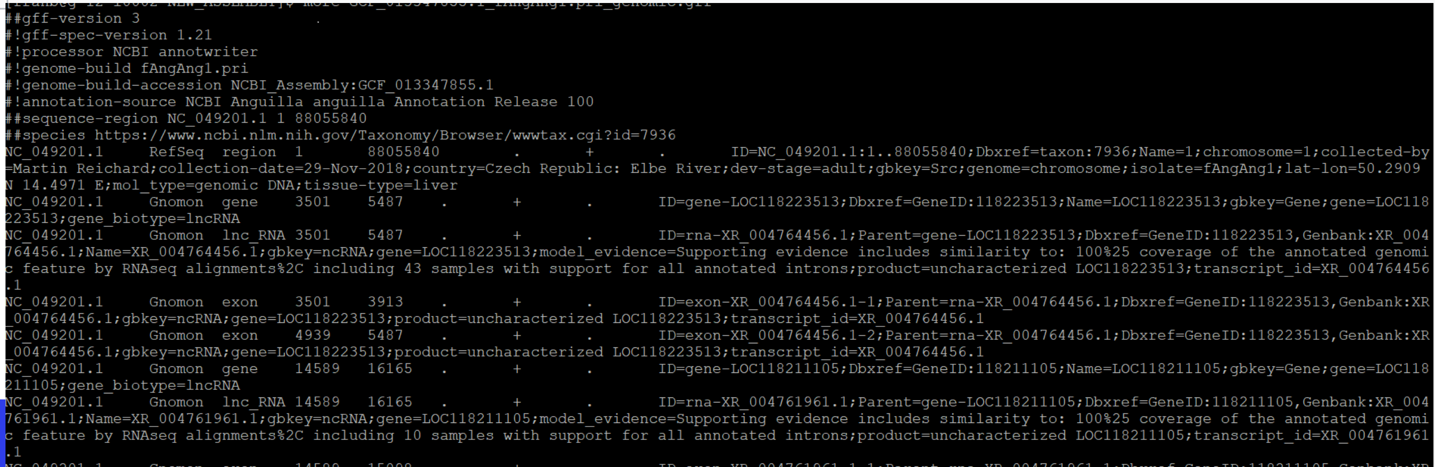
Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.

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

Annotation file

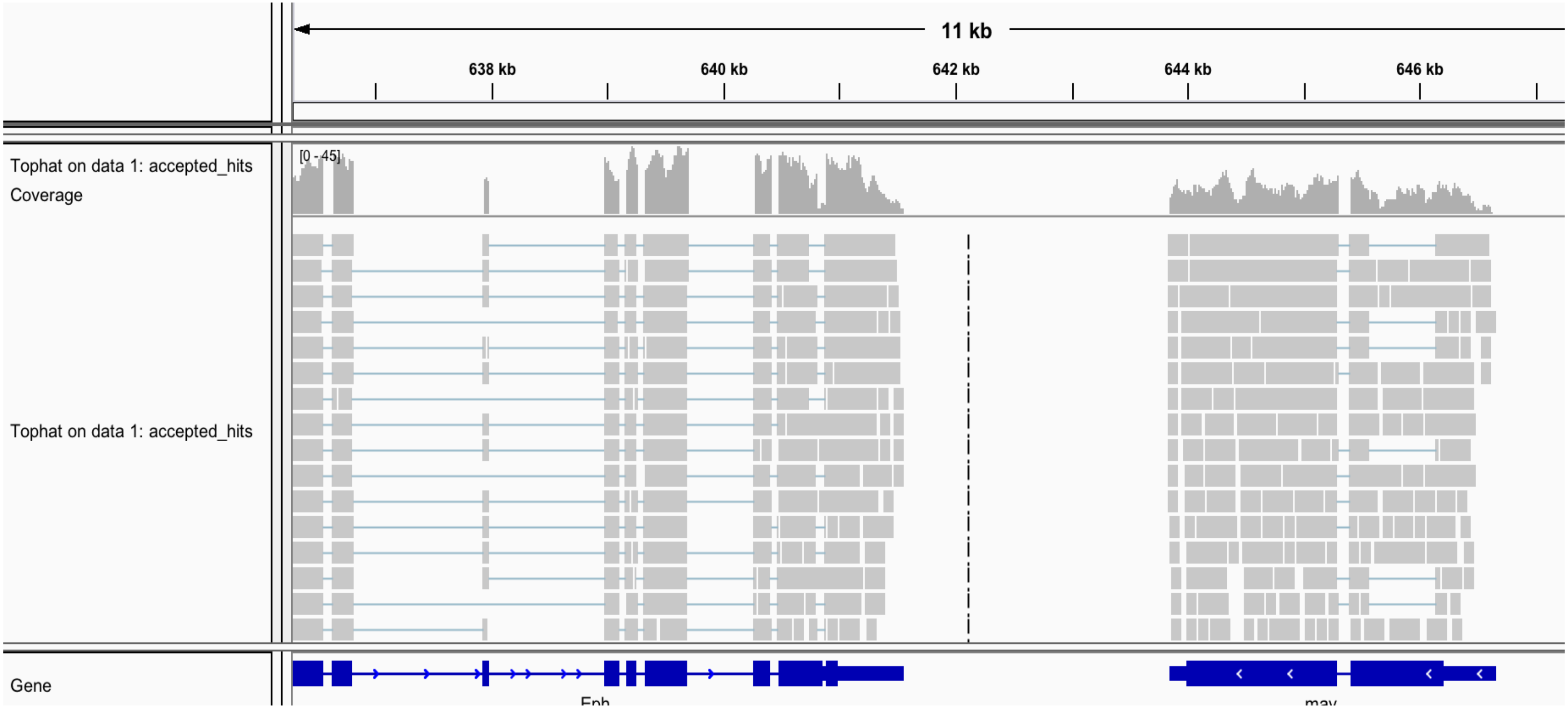
General GFF structure

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.



```
[framb@g-12-10002 NEW_ASSEMBLY]$ more GCF_013347855.1_fAngAng1.pri_genomic.gff
##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
NC_049201.1     Gnomon  exon    14589   15009   .       +       .       ID=exon-XR_004761961.1-1;Parent=rna-XR_004761961.1;Dbxref=GeneID:118211105,Genbank:XR
```

Splice junctions view along the genome


11 kb
638 kb
640 kb
642 kb
644 kb
646 kb
Tophat on data 1: accepted_hits
Coverage
[0 - 45]
Tophat on data 1: accepted_hits
Gene
Eph
may

miRNA-seq

The short sequence length makes small RNA difficult to map in large and complex reference genome. Common aligner for long RNA are therefore not accurate for short RNA mapping

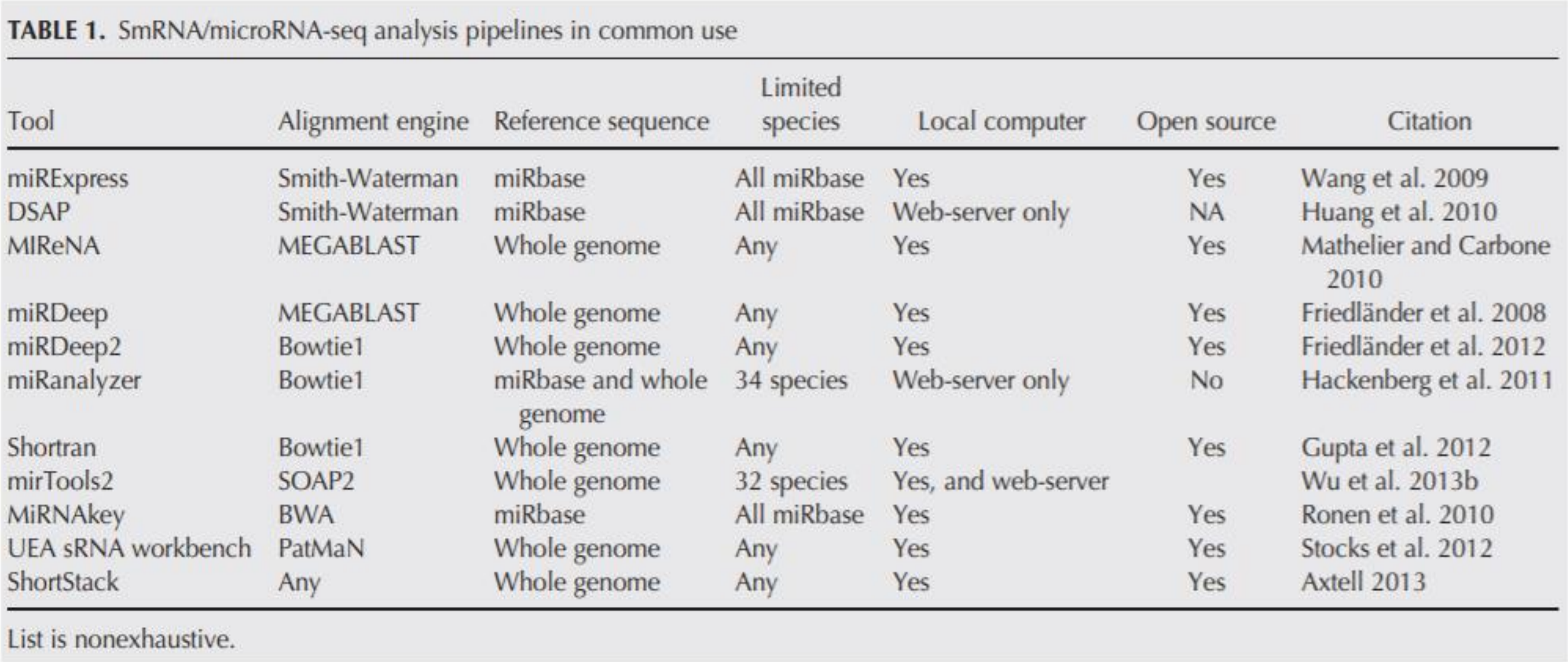


TABLE 1. SmRNA/microRNA-seq analysis pipelines in common use

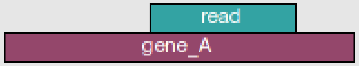
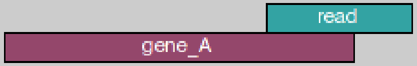
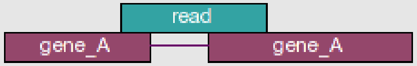
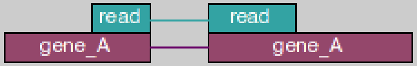
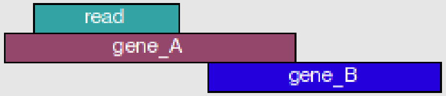
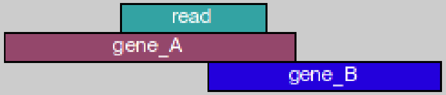
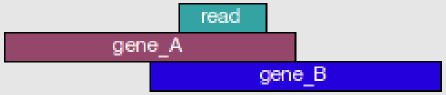
Tool	Alignment engine	Reference sequence	Limited species	Local computer	Open source	Citation
miRExpress	Smith-Waterman	miRbase	All miRbase	Yes	Yes	Wang et al. 2009
DSAP	Smith-Waterman	miRbase	All miRbase	Web-server only	NA	Huang et al. 2010
MIReNA	MEGABLAST	Whole genome	Any	Yes	Yes	Mathelier and Carbone 2010
miRDeep	MEGABLAST	Whole genome	Any	Yes	Yes	Friedländer et al. 2008
miRDeep2	Bowtie1	Whole genome	Any	Yes	Yes	Friedländer et al. 2012
miRanalyzer	Bowtie1	miRbase and whole genome	34 species	Web-server only	No	Hackenberg et al. 2011
Shortran	Bowtie1	Whole genome	Any	Yes	Yes	Gupta et al. 2012
mirTools2	SOAP2	Whole genome	32 species	Yes, and web-server		Wu et al. 2013b
MiRNAkey	BWA	miRbase	All miRbase	Yes	Yes	Ronen et al. 2010
UEA sRNA workbench	PatMaN	Whole genome	Any	Yes	Yes	Stocks et al. 2012
ShortStack	Any	Whole genome	Any	Yes	Yes	Axtell 2013

List is nonexhaustive.

READ COUNT

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

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

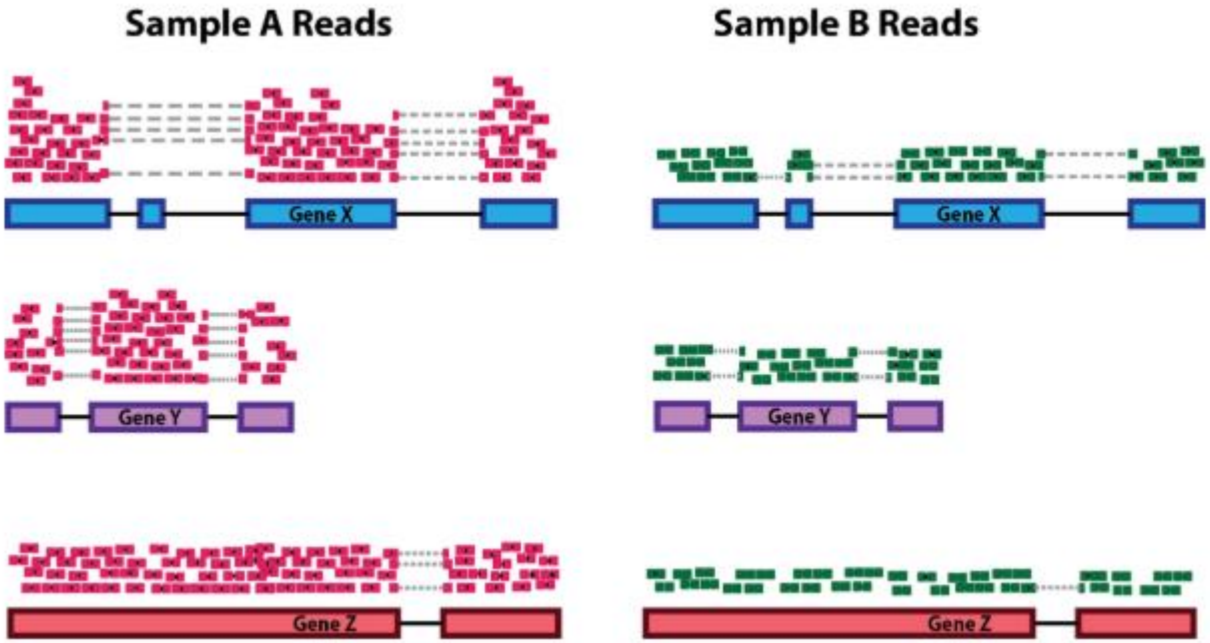
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

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)

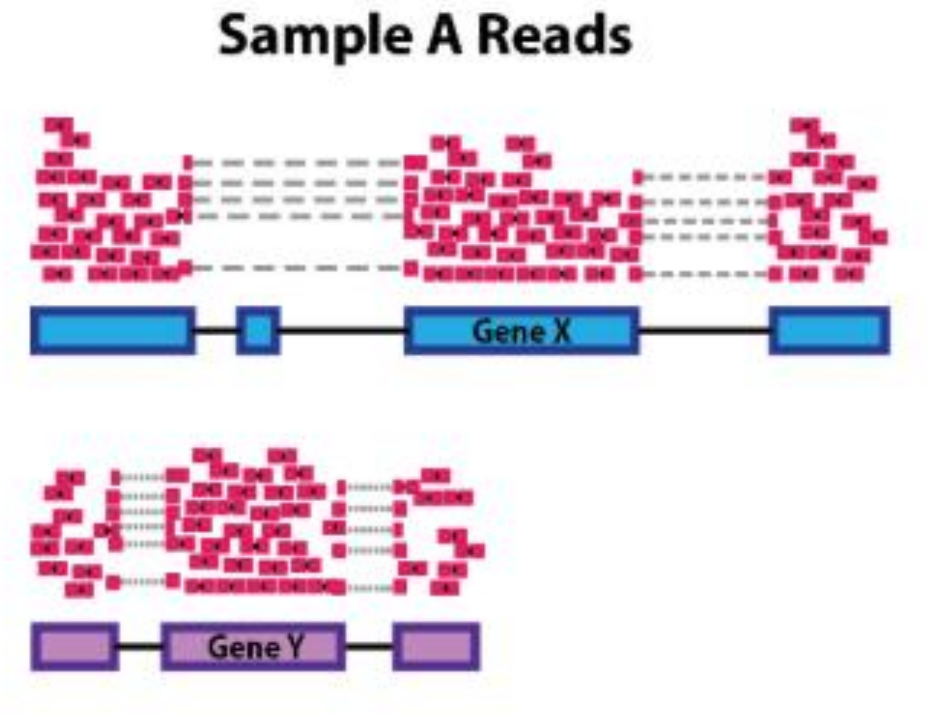
Main factors during normalization

Sequencing depth



Main factors during normalization

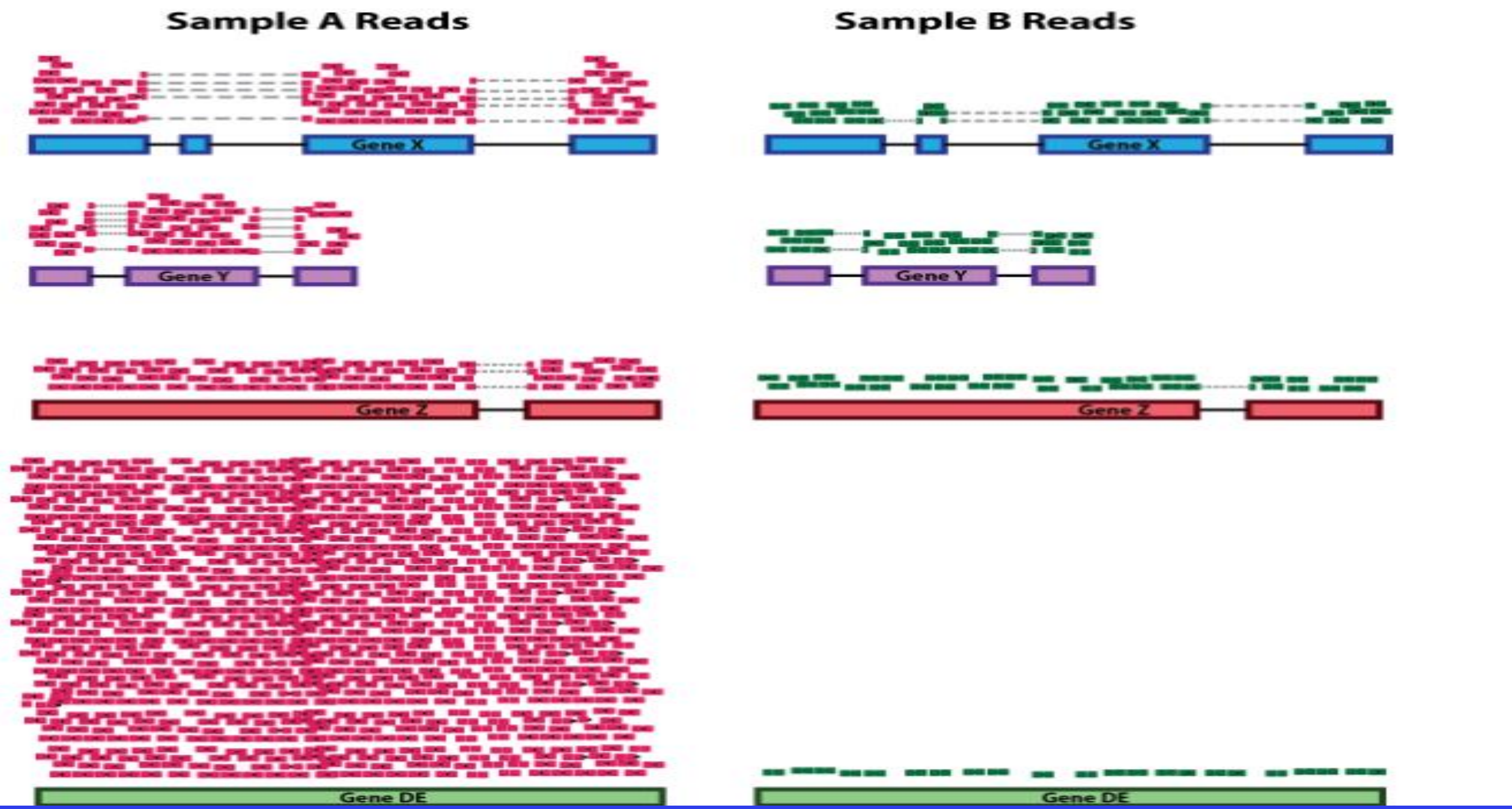
Gene length



Main factors during normalization

RNA composition

Anders and Huber , 2010 Genome Biol.



NORMALIZATION

Common normalization methods

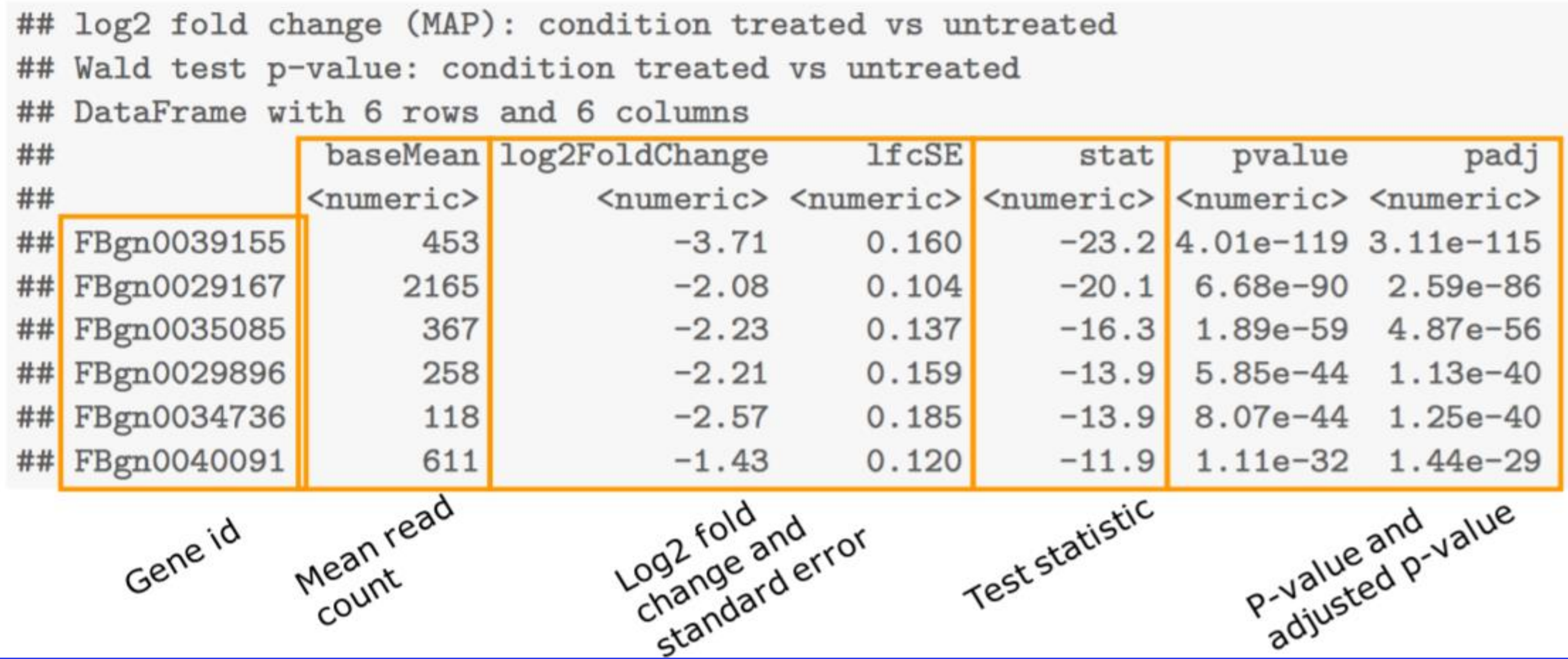
Normalization method	Description	Accounted factors	Recommendations for use
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for 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

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

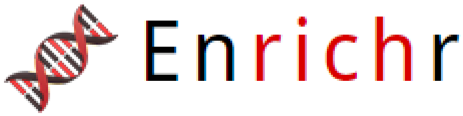
Gene Ontology project provides an ontology of **defined terms** representing gene product properties. The ontology covers three domains:

•**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.

Some GO and pathway analyses websites



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

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

https://david.ncifcrf.gov/

https://cytoscape.org/

Useful links

Deseq2 vignette:
http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

RNA-seq the teenage years
https://www.nature.com/articles/s41576-019-0150-2.pdf

Pseudoalignment: Kallisto
https://www.nature.com/articles/nbt.3519

You can learn more about NGS and its application in animal breeding and conservation with the course **25334 Genomic methods in breeding and management of aquatic living resources** (fall 2022)