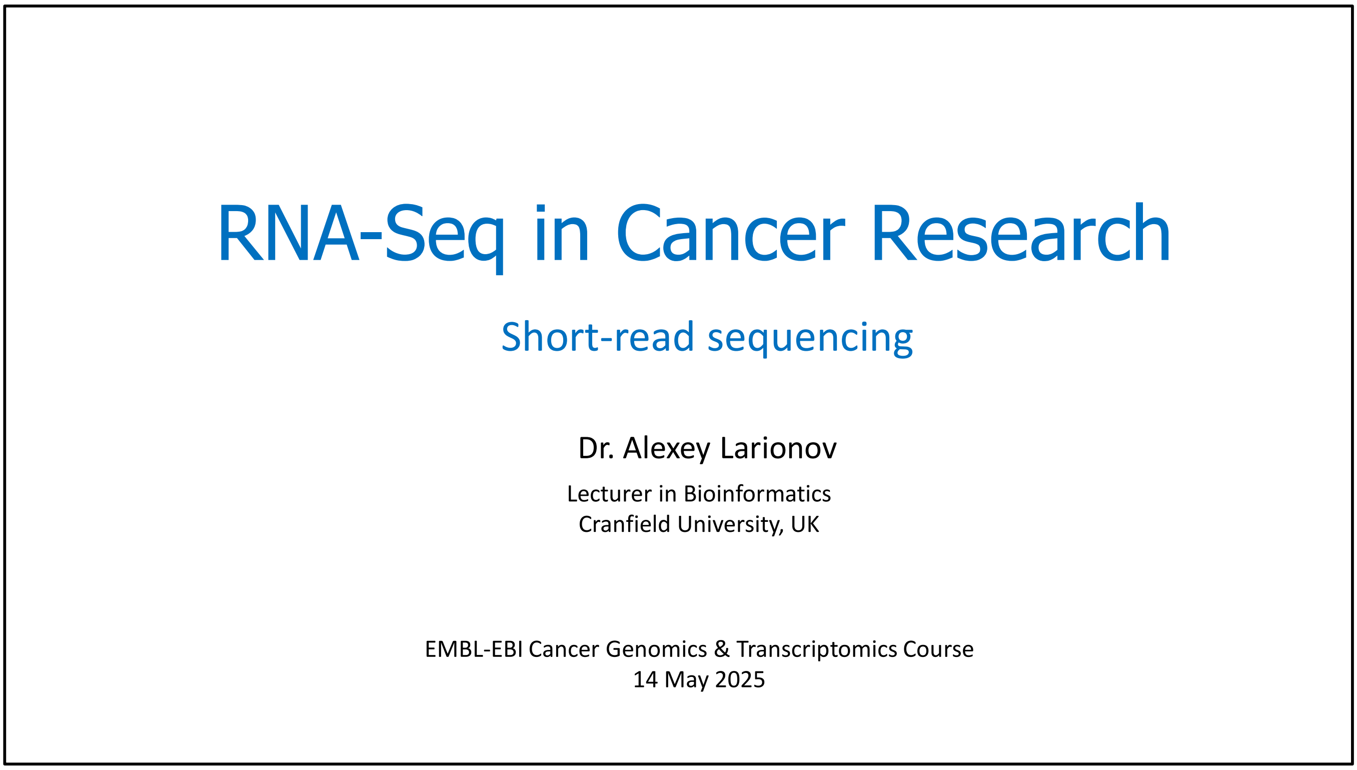


RNA-Seq in Cancer Research

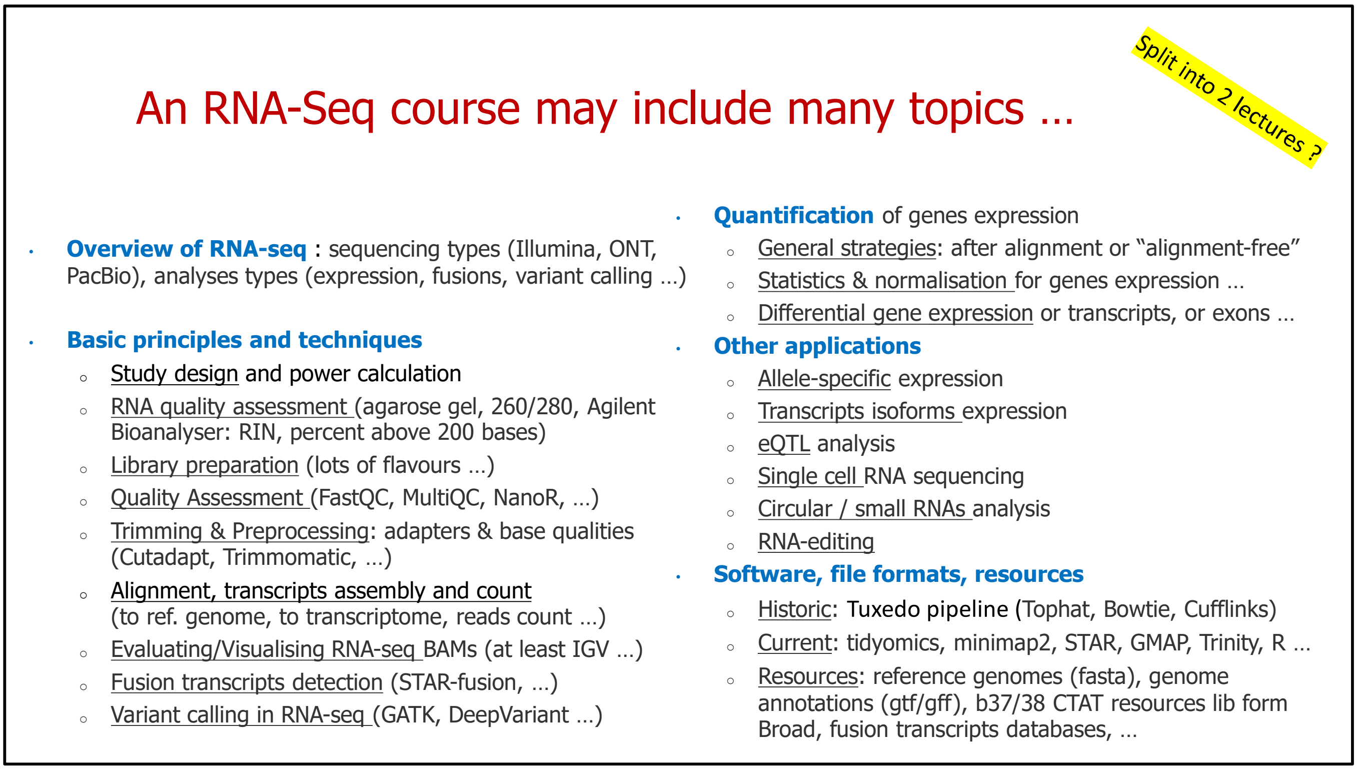
Short-read sequencing

Dr. Alexey Larionov

Lecturer in Bioinformatics
Cranfield University, UK

EMBL-EBI Cancer Genomics & Transcriptomics Course
14 May 2025

Hello, My name is Alexey Larionov, and this is my lecture about bulk RNA-sequencing at the EBI Cancer Genomics course.

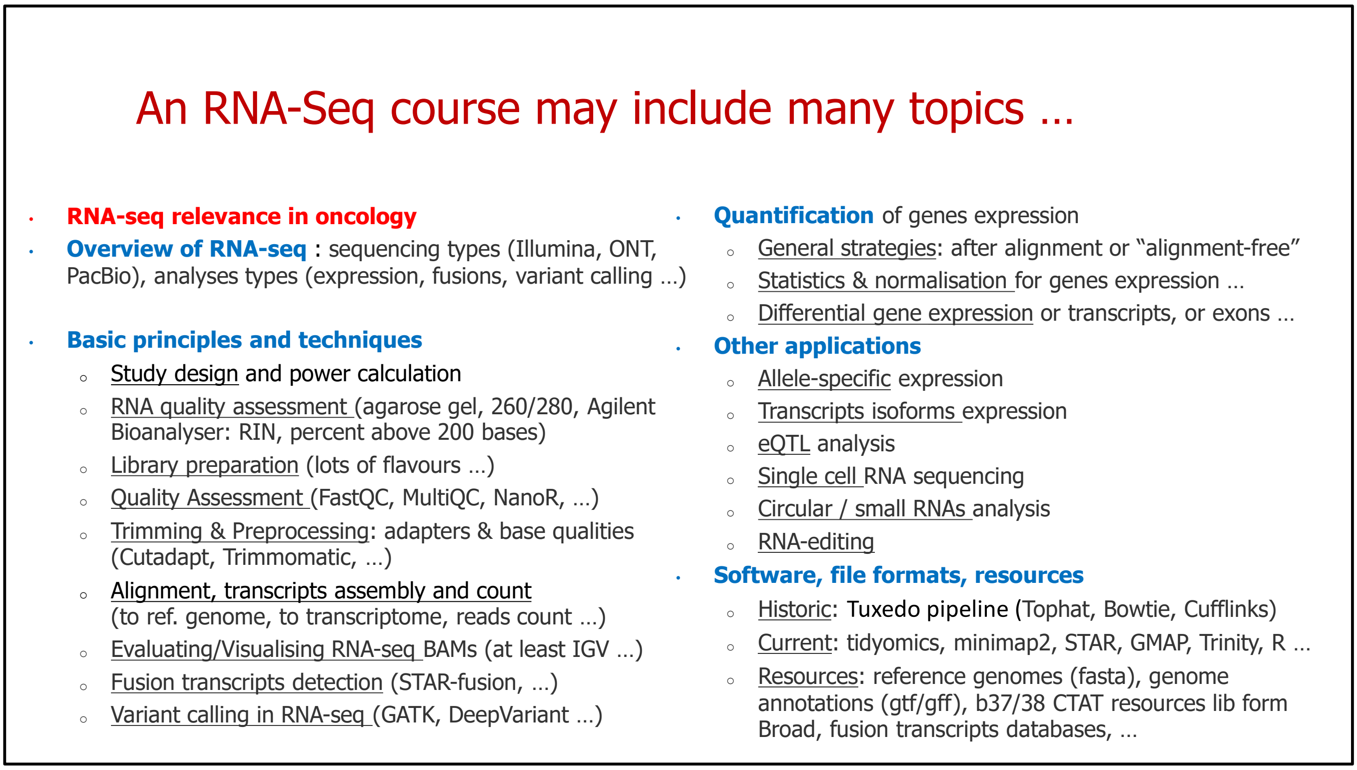


An RNA-Seq course may include many topics …

Split into 2 lectures ?

- **Overview of RNA-seq** : sequencing types (Illumina, ONT, PacBio), analyses types (expression, fusions, variant calling …)
- **Basic principles and techniques**
  - Study design and power calculation
  - RNA quality assessment (agarose gel, 260/280, Agilent Bioanalyser: RIN, percent above 200 bases)
  - Library preparation (lots of flavours …)
  - Quality Assessment (FastQC, MultiQC, NanoR, …)
  - Trimming & Preprocessing: adapters & base qualities (Cutadapt, Trimmomatic, …)
  - Alignment, transcripts assembly and count (to ref. genome, to transcriptome, reads count …)
  - Evaluating/Visualising RNA-seq BAMs (at least IGV …)
  - Fusion transcripts detection (STAR-fusion, …)
  - Variant calling in RNA-seq (GATK, DeepVariant …)
- **Quantification** of genes expression
  - General strategies: after alignment or "alignment-free"
  - Statistics & normalisation for genes expression …
  - Differential gene expression or transcripts, or exons …
- **Other applications**
  - Allele-specific expression
  - Transcripts isoforms expression
  - eQTL analysis
  - Single cell RNA sequencing
  - Circular / small RNAs analysis
  - RNA-editing
- **Software, file formats, resources**
  - Historic: Tuxedo pipeline (Tophat, Bowtie, Cufflinks)
  - Current: tidyomics, minimap2, STAR, GMAP, Trinity, R …
  - Resources: reference genomes (fasta), genome annotations (gtf/gff), b37/38 CTAT resources lib form Broad, fusion transcripts databases, …

To be honest, we could spend the whole week on this course talking about RNA-sequencing only: there are so many interesting directions, tools, and resources in RNA-sequencing, that it was very hard to prioritize things for a one-day introduction.



An RNA-Seq course may include many topics …

- **RNA-seq relevance in oncology**
- **Overview of RNA-seq** : sequencing types (Illumina, ONT, PacBio), analyses types (expression, fusions, variant calling …)
- **Basic principles and techniques**
  - Study design and power calculation
  - RNA quality assessment (agarose gel, 260/280, Agilent Bioanalyser: RIN, percent above 200 bases)
  - Library preparation (lots of flavours …)
  - Quality Assessment (FastQC, MultiQC, NanoR, …)
  - Trimming & Preprocessing: adapters & base qualities (Cutadapt, Trimmomatic, …)
  - Alignment, transcripts assembly and count (to ref. genome, to transcriptome, reads count …)
  - Evaluating/Visualising RNA-seq BAMs (at least IGV …)
  - Fusion transcripts detection (STAR-fusion, …)
  - Variant calling in RNA-seq (GATK, DeepVariant …)
- **Quantification** of genes expression
  - General strategies: after alignment or "alignment-free"
  - Statistics & normalisation for genes expression …
  - Differential gene expression or transcripts, or exons …
- **Other applications**
  - Allele-specific expression
  - Transcripts isoforms expression
  - eQTL analysis
  - Single cell RNA sequencing
  - Circular / small RNAs analysis
  - RNA-editing
- **Software, file formats, resources**
  - Historic: Tuxedo pipeline (Tophat, Bowtie, Cufflinks)
  - Current: tidyomics, minimap2, STAR, GMAP, Trinity, R …
  - Resources: reference genomes (fasta), genome annotations (gtf/gff), b37/38 CTAT resources lib form Broad, fusion transcripts databases, …

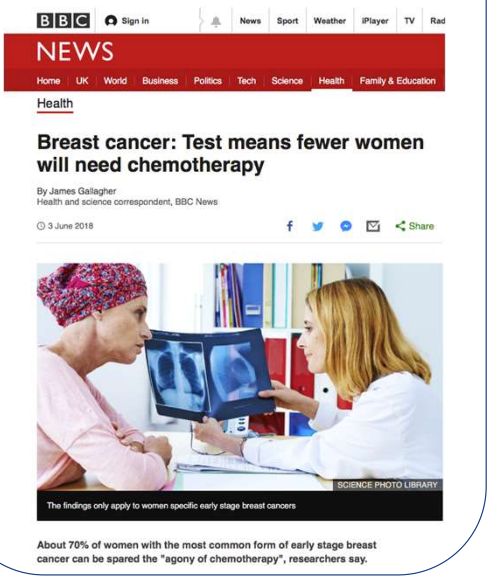
However, for this course we should focus on what is the most relevant to oncology.

Transcriptomic in oncology clinics

TAILORx study

Oncotype-Dx in Breast Cancer

Gene expression test allows to stratify tumours to avoid unnecessary chemotherapy



Health

Breast cancer: Test means fewer women will need chemotherapy

By James Gallagher
Health and science correspondent, BBC News

3 June 2018

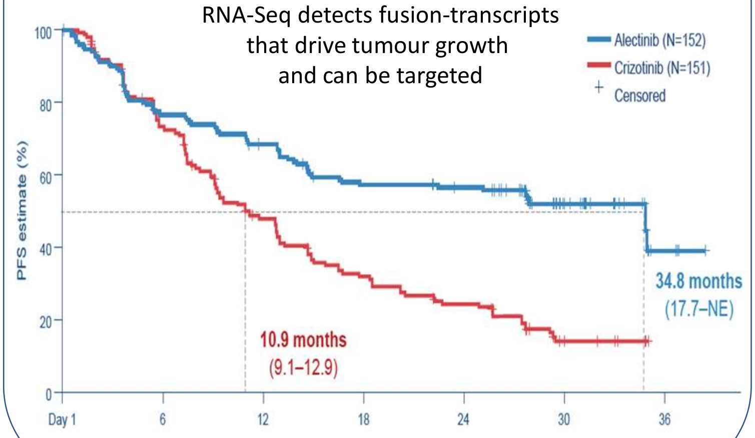
The findings only apply to women specific early stage breast cancers

About 70% of women with the most common form of early stage breast cancer can be spared the "agony of chemotherapy", researchers say.

ALEX study

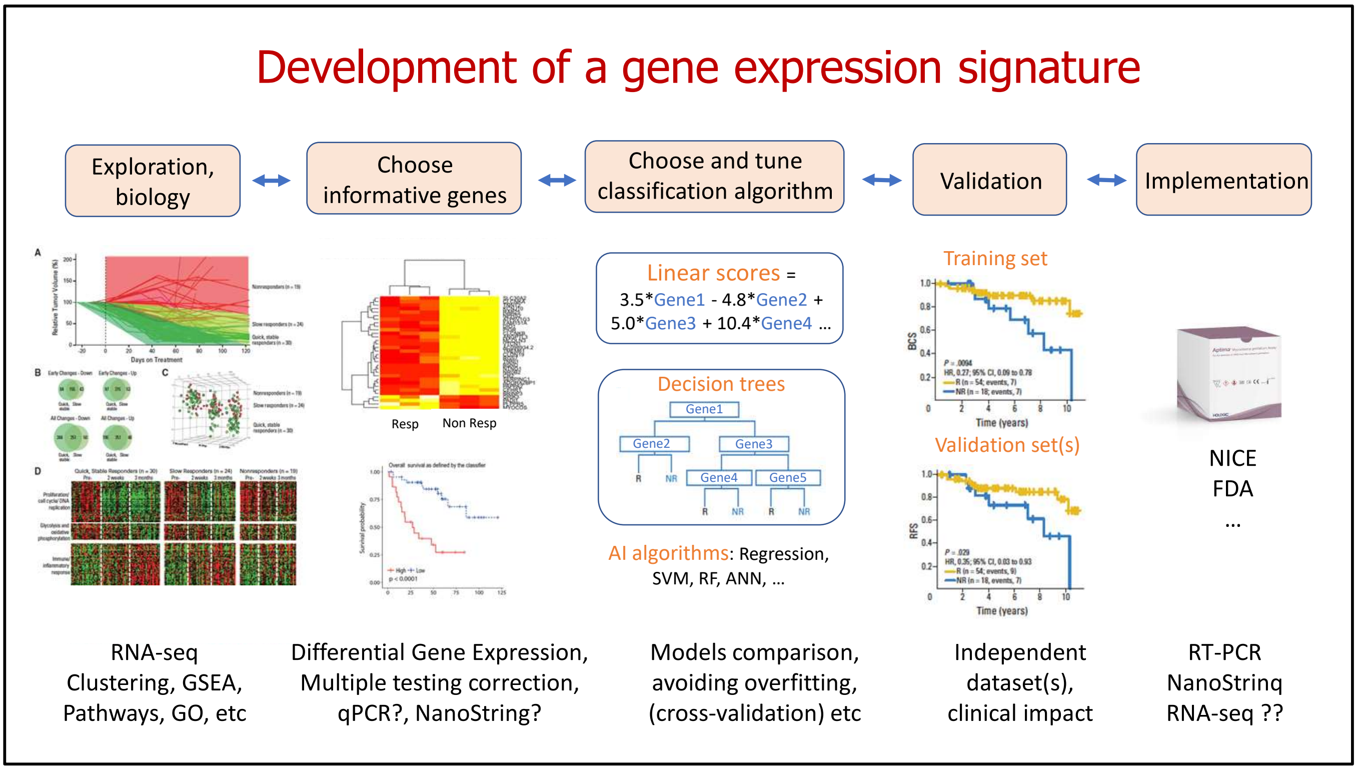
ALK-fusion in Lung Cancer

RNA-Seq detects fusion-transcripts that drive tumour growth and can be targeted



From this perspective, two fields are clearly coming forward:
1) the gene expression signatures and
2) the fusions detection
because they have direct clinical implications in oncology.

For instance, Oncotype-DX, which you can see on the left side of the slide, is a gene expression signature that may help in selecting breast cancer therapy .
Another example here is a fusion gene in lung cancer, which can be targeted by specific drugs. Such fusions, at the moment, are often detected by RNAseq.



Development of a gene expression signature

Exploration, biology ↔ Choose informative genes ↔ Choose and tune classification algorithm ↔ Validation ↔ Implementation

Linear scores = 3.5*Gene1 - 4.8*Gene2 + 5.0*Gene3 + 10.4*Gene4 …

Decision trees

AI algorithms: Regression, SVM, RF, ANN, …

Training set

Validation set(s)

NICE
FDA
…

RNA-seq
Clustering, GSEA,
Pathways, GO, etc

Differential Gene Expression,
Multiple testing correction,
qPCR?, NanoString?

Models comparison,
avoiding overfitting,
(cross-validation) etc

Independent
dataset(s),
clinical impact

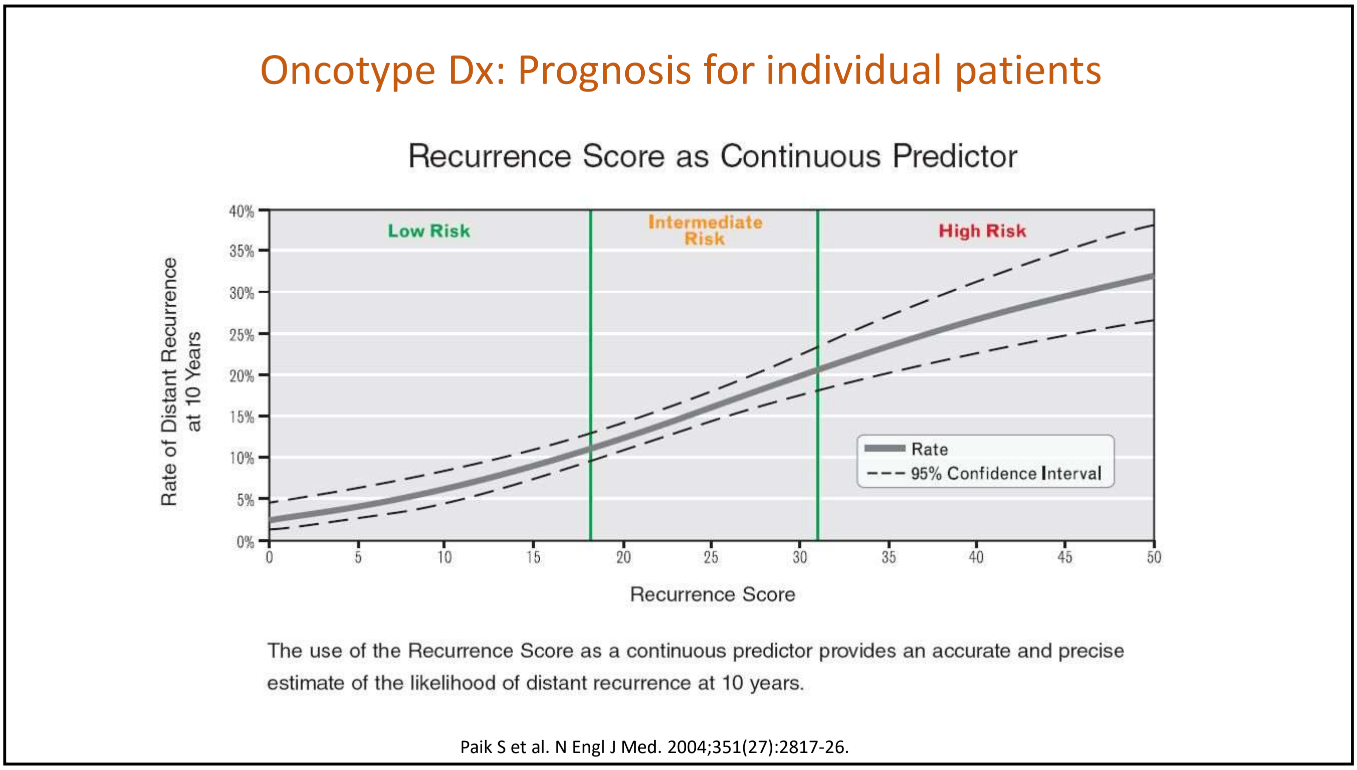
RT-PCR
NanoStrinq
RNA-seq ??

Development of a gene expression signature is a long process, and we will only touch on the beginning of it.

It starts with some exploration for understanding the biology and finding differentially expressed genes, for instance: between responders and non-responders to some treatment.

Then the most informative genes are used in some classification algorithm, producing a classifier, which should be validated on independent datasets and clinical trials before it can be used in clinic.

At the end, it might be more practical to measure the selected genes by RT-PCR or by NanoString, in contrast to the initial exploratory steps that usually include gene expression measurement by RNA-seq.



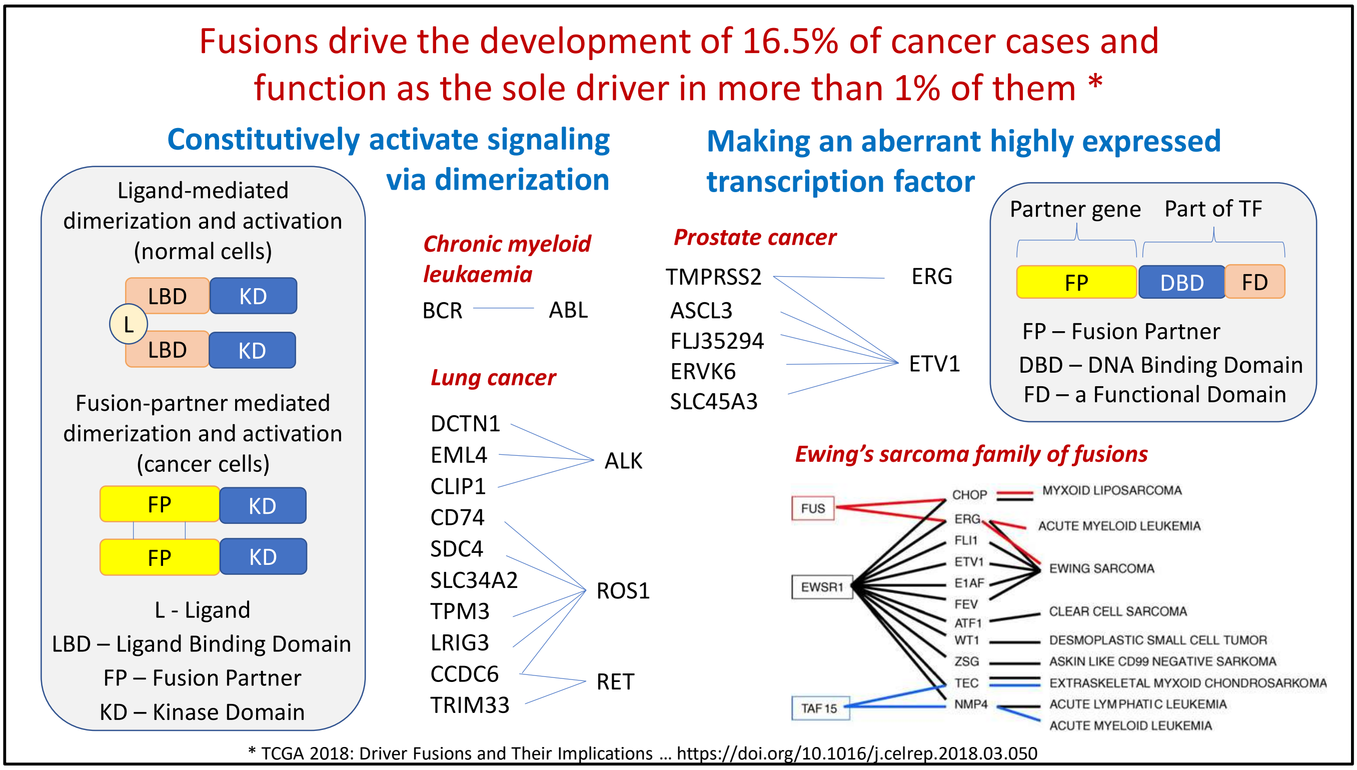
Oncotype Dx: Prognosis for individual patients

Recurrence Score as Continuous Predictor

The use of the Recurrence Score as a continuous predictor provides an accurate and precise estimate of the likelihood of distant recurrence at 10 years.

Paik S et al. N Engl J Med. 2004;351(27):2817-26.

And, if you are lucky and persistent enough, you may end up with something like this: a gene expression score that predicts clinical outcomes in individual patients.



Fusions drive the development of 16.5% of cancer cases and function as the sole driver in more than 1% of them *

Constitutively activate signaling via dimerization

Ligand-mediated dimerization and activation (normal cells)

Fusion-partner mediated dimerization and activation (cancer cells)

L - Ligand
LBD – Ligand Binding Domain
FP – Fusion Partner
KD – Kinase Domain

Chronic myeloid leukaemia

BCR — ABL

Lung cancer

DCTN1, EML4, CLIP1 — ALK
CD74, SDC4, SLC34A2, TPM3, LRIG3 — ROS1
CCDC6, TRIM33 — RET

Making an aberrant highly expressed transcription factor

Prostate cancer

TMPRSS2 — ERG
TMPRSS2, ASCL3, FLJ35294, ERVK6, SLC45A3 — ETV1

FP – Fusion Partner
DBD – DNA Binding Domain
FD – a Functional Domain

Ewing's sarcoma family of fusions

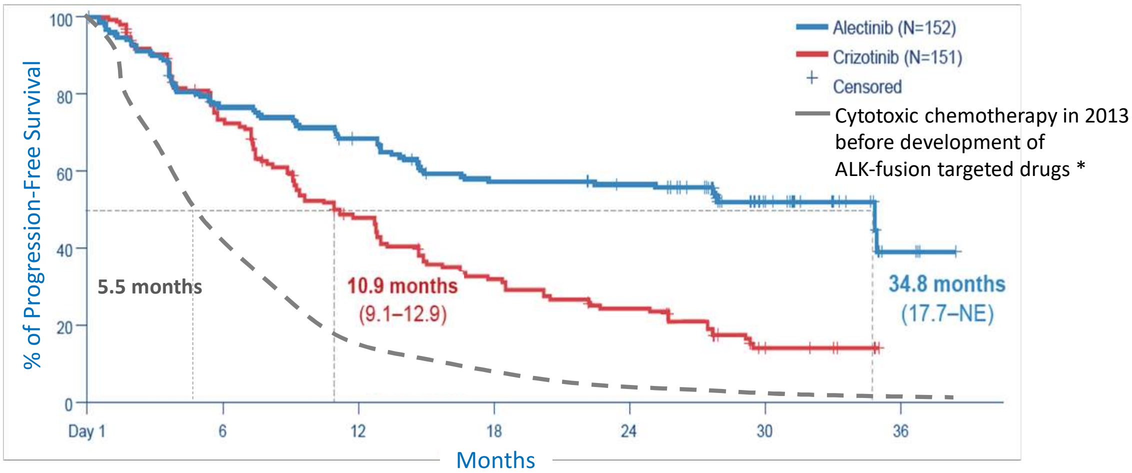
* TCGA 2018: Driver Fusions and Their Implications … https://doi.org/10.1016/j.celrep.2018.03.050

Fusion genes may contribute to the growth of about 15% of cancers.

There are many known cancer-related gene fusions that either
1) activate some oncogenic signaling through ligand-independent dimerization, or
2) create a new overexpressed transcription factor, leading to cell proliferation.

An important point here is that these fusion proteins are completely absent in normal cells. So, pharmacologically targeting such fusion protein we may stop the tumor growth, without affecting any normal cells.

Targeted treatments in advanced non-small-cell lung cancer driven by ALK-fusion



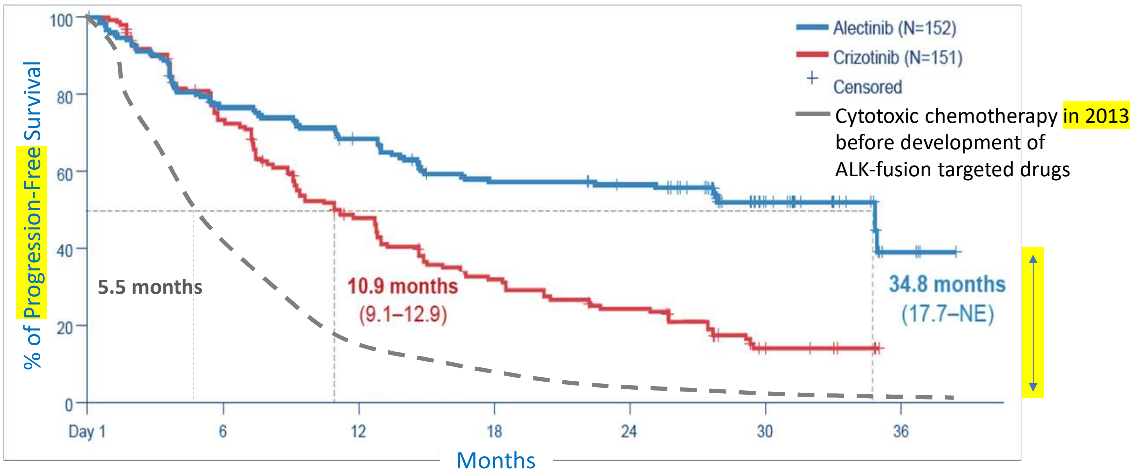
Camidge et al. JCO 2018, June 1, suppl. Abstract 9043

* Laporte et al. BMJ Open 2013;3:e001802

And this actually works.

Here you can see Kaplan Meyer curves that show proportions of patients that survived after certain time.

Targeted treatments in advanced non-small-cell lung cancer driven by ALK-fusion



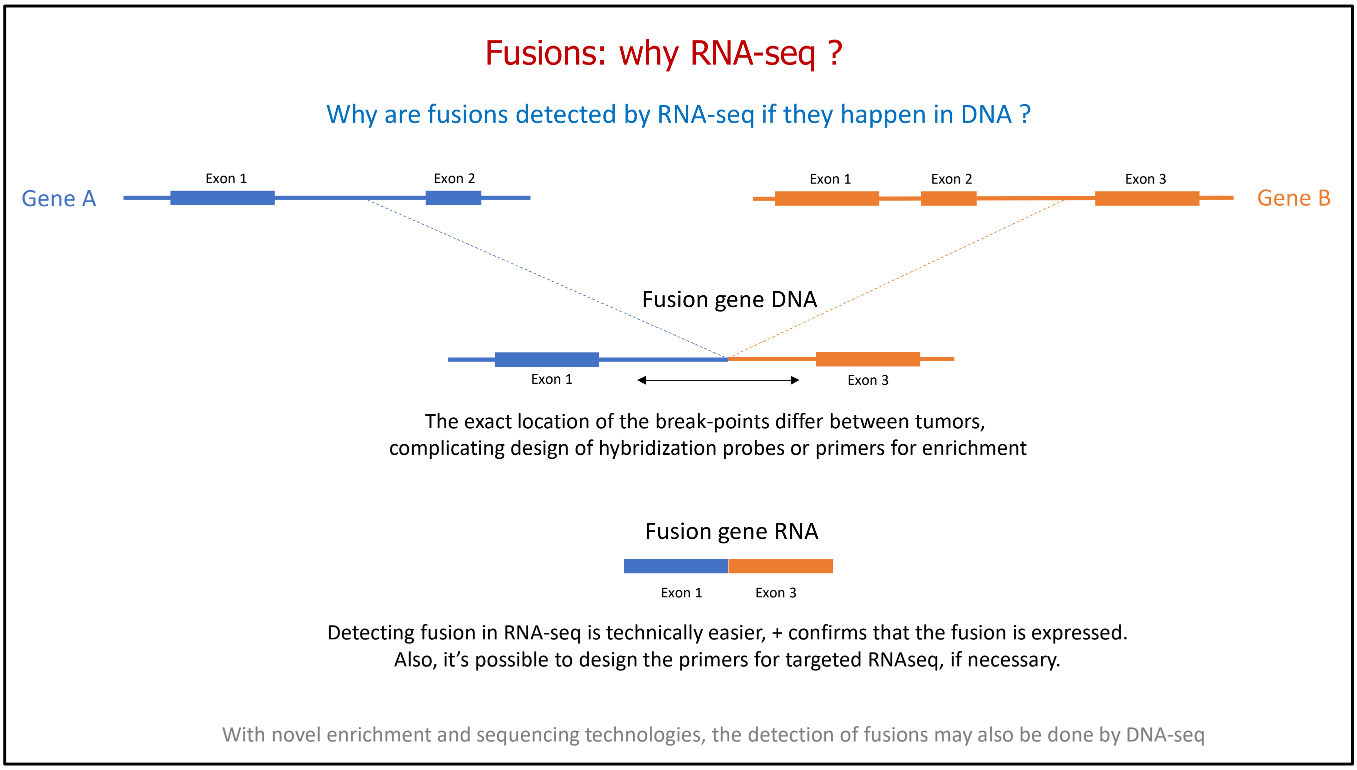
Camidge et al. JCO 2018, June 1, suppl. Abstract 9043

Laporte et al. BMJ Open 2013;3:e001802

These plots show the remarkable effect of targeting a fusion gene in lung cancer, illustrating that fusion detection dramatically changed clinical outcomes for some cancers over recent years.

Thus, in 2013, before development of the targeted treatments, in half of the patients the tumor progressed within less than half a year.

In 2018, if the fusion was detected and the appropriate targeted treatment was used, nearly half of the patients lived for almost 3 years without progression. Importantly: this is an advanced, already metastatic (!) cancer.



Fusions: why RNA-seq ?

Why are fusions detected by RNA-seq if they happen in DNA ?

The exact location of the break-points differ between tumors, complicating design of hybridization probes or primers for enrichment

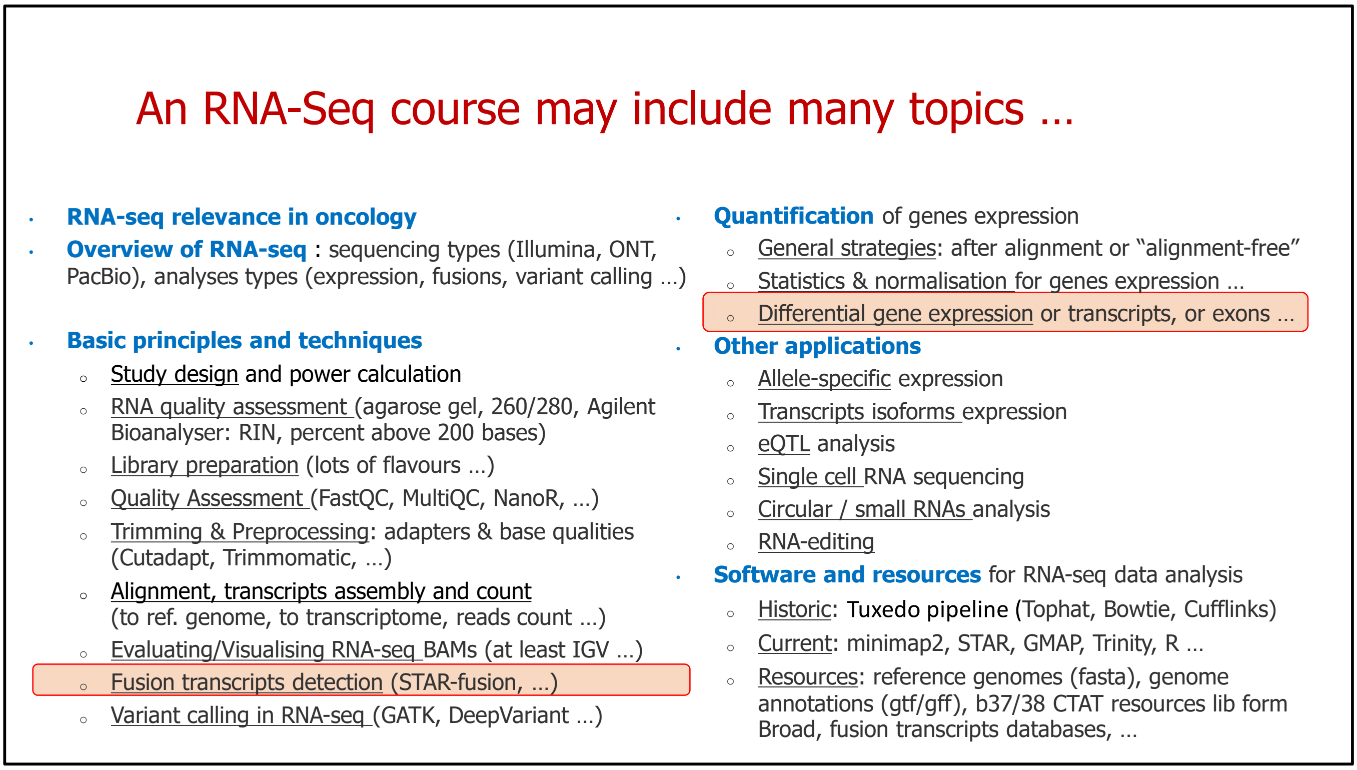
Detecting fusion in RNA-seq is technically easier, + confirms that the fusion is expressed. Also, it's possible to design the primers for targeted RNAseq, if necessary.

With novel enrichment and sequencing technologies, the detection of fusions may also be done by DNA-seq

Here you may ask: “OK, targeting fusions is great, but why are we talking about this in RNA-seq lecture, if these fusions happen in DNA?“

The answer is purely practical: because currently RNA-seq is much more convenient to detect such oncogenic fusions.

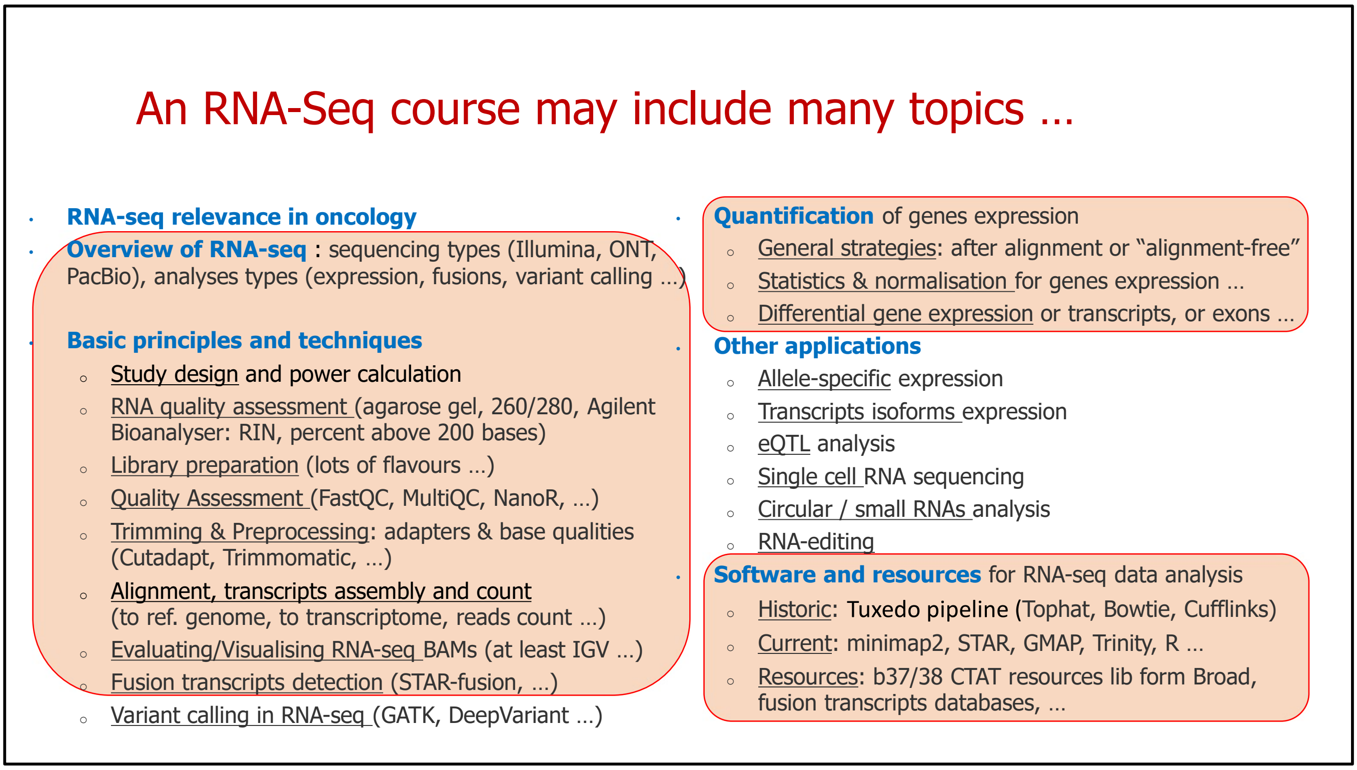
With new enrichment or sequencing technologies, the detection of fusions may change from RNA- to DNA-seq. But for-now oncogenic fusions are usually detected by RNA-seq.



An RNA-Seq course may include many topics …

- **RNA-seq relevance in oncology**
- **Overview of RNA-seq** : sequencing types (Illumina, ONT, PacBio), analyses types (expression, fusions, variant calling …)
- **Basic principles and techniques**
  - Study design and power calculation
  - RNA quality assessment (agarose gel, 260/280, Agilent Bioanalyser: RIN, percent above 200 bases)
  - Library preparation (lots of flavours …)
  - Quality Assessment (FastQC, MultiQC, NanoR, …)
  - Trimming & Preprocessing: adapters & base qualities (Cutadapt, Trimmomatic, …)
  - Alignment, transcripts assembly and count (to ref. genome, to transcriptome, reads count …)
  - Evaluating/Visualising RNA-seq BAMs (at least IGV …)
  - Fusion transcripts detection (STAR-fusion, …)
  - Variant calling in RNA-seq (GATK, DeepVariant …)
- **Quantification** of genes expression
  - General strategies: after alignment or "alignment-free"
  - Statistics & normalisation for genes expression …
  - Differential gene expression or transcripts, or exons …
- **Other applications**
  - Allele-specific expression
  - Transcripts isoforms expression
  - eQTL analysis
  - Single cell RNA sequencing
  - Circular / small RNAs analysis
  - RNA-editing
- **Software and resources** for RNA-seq data analysis
  - Historic: Tuxedo pipeline (Tophat, Bowtie, Cufflinks)
  - Current: minimap2, STAR, GMAP, Trinity, R …
  - Resources: reference genomes (fasta), genome annotations (gtf/gff), b37/38 CTAT resources lib form Broad, fusion transcripts databases, …

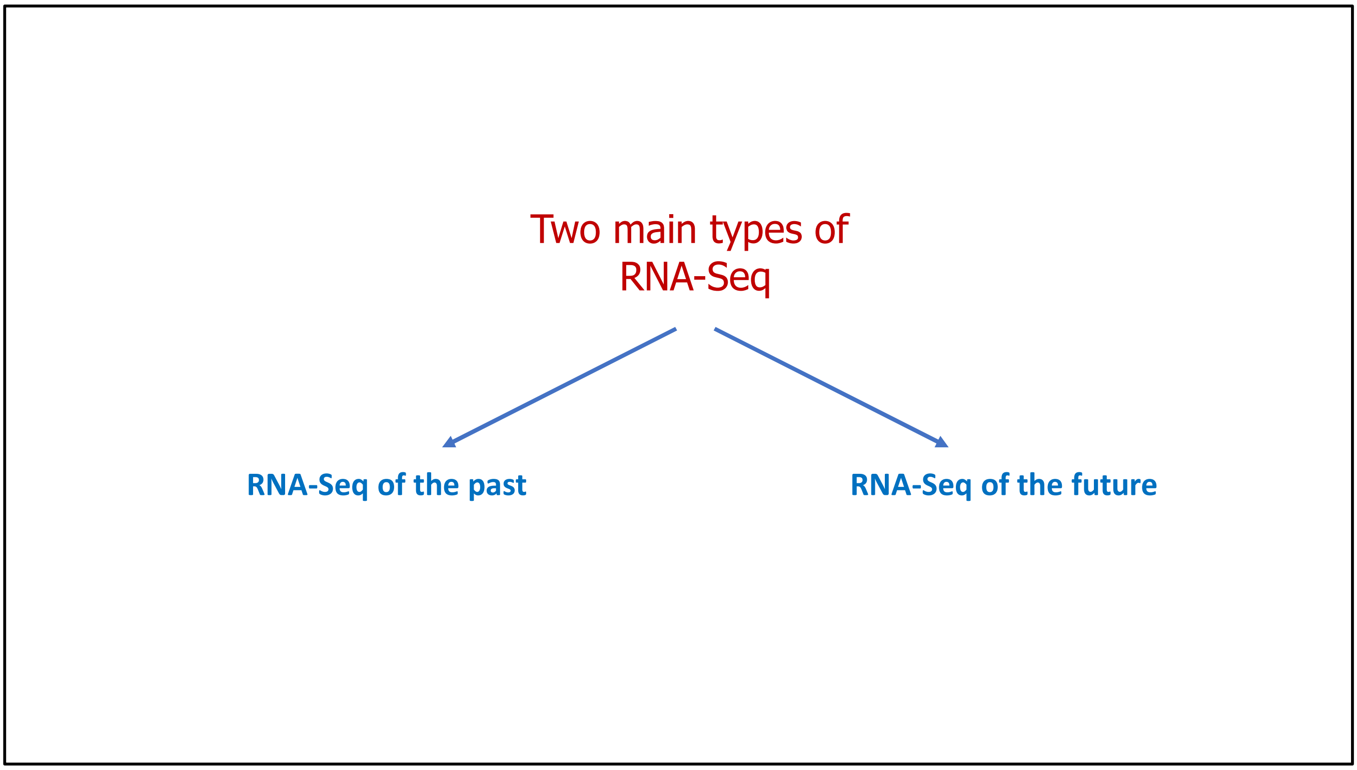
So, I hope that I have explained why I will focus on the differential gene expression and the fusions detection.



An RNA-Seq course may include many topics …

- **RNA-seq relevance in oncology**
- **Overview of RNA-seq** : sequencing types (Illumina, ONT, PacBio), analyses types (expression, fusions, variant calling …)
- **Basic principles and techniques**
  - Study design and power calculation
  - RNA quality assessment (agarose gel, 260/280, Agilent Bioanalyser: RIN, percent above 200 bases)
  - Library preparation (lots of flavours …)
  - Quality Assessment (FastQC, MultiQC, NanoR, …)
  - Trimming & Preprocessing: adapters & base qualities (Cutadapt, Trimmomatic, …)
  - Alignment, transcripts assembly and count (to ref. genome, to transcriptome, reads count …)
  - Evaluating/Visualising RNA-seq BAMs (at least IGV …)
  - Fusion transcripts detection (STAR-fusion, …)
  - Variant calling in RNA-seq (GATK, DeepVariant …)
- **Quantification** of genes expression
  - General strategies: after alignment or "alignment-free"
  - Statistics & normalisation for genes expression …
  - Differential gene expression or transcripts, or exons …
- **Other applications**
  - Allele-specific expression
  - Transcripts isoforms expression
  - eQTL analysis
  - Single cell RNA sequencing
  - Circular / small RNAs analysis
  - RNA-editing
- **Software and resources** for RNA-seq data analysis
  - Historic: Tuxedo pipeline (Tophat, Bowtie, Cufflinks)
  - Current: minimap2, STAR, GMAP, Trinity, R …
  - Resources: b37/38 CTAT resources lib form Broad, fusion transcripts databases, …

Well, don't worry, we will also discuss other fields, covering the techniques and methods related to the gene expression measurement and fusions detection.

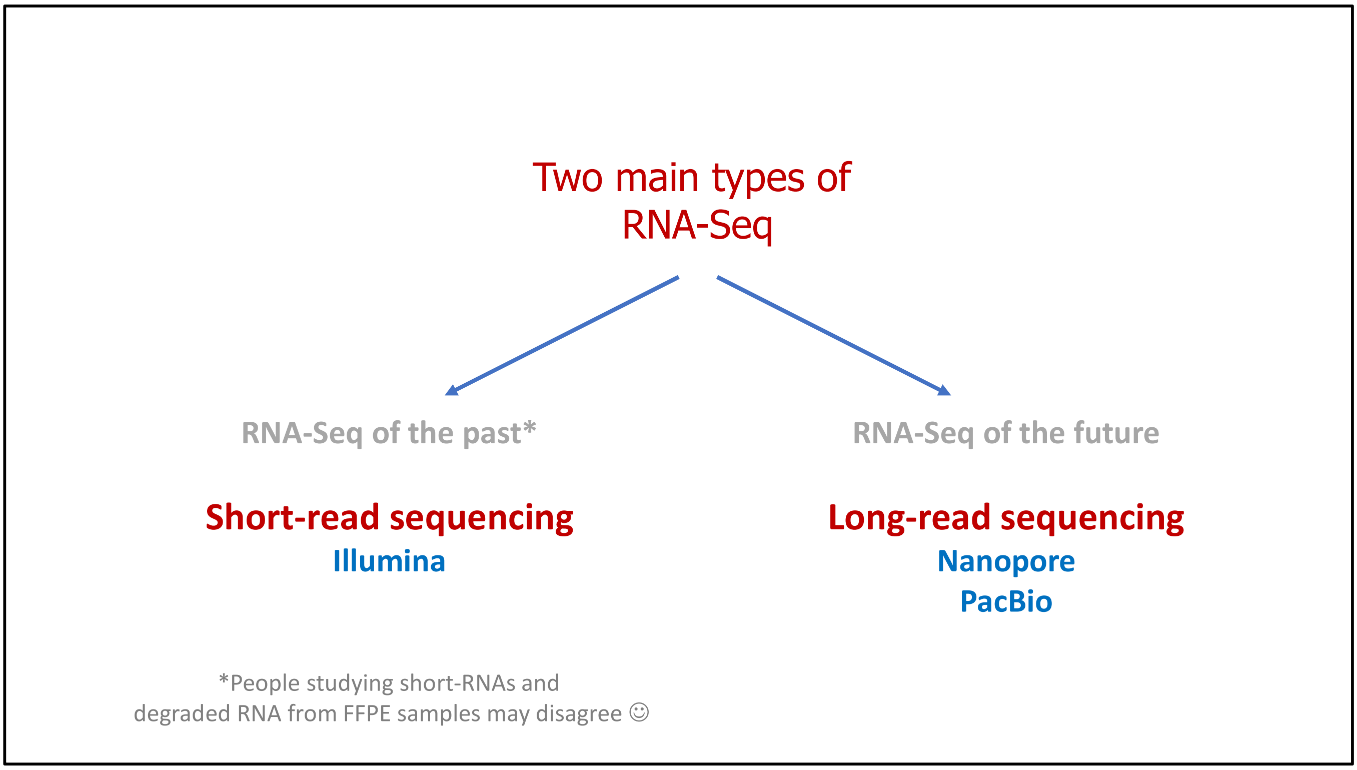


Two main types of RNA-Seq

RNA-Seq of the past

RNA-Seq of the future

Let’s start with a broad overview of RNA-sequencing technologies.



Two main types of RNA-Seq

RNA-Seq of the past*

Short-read sequencing

Illumina

RNA-Seq of the future

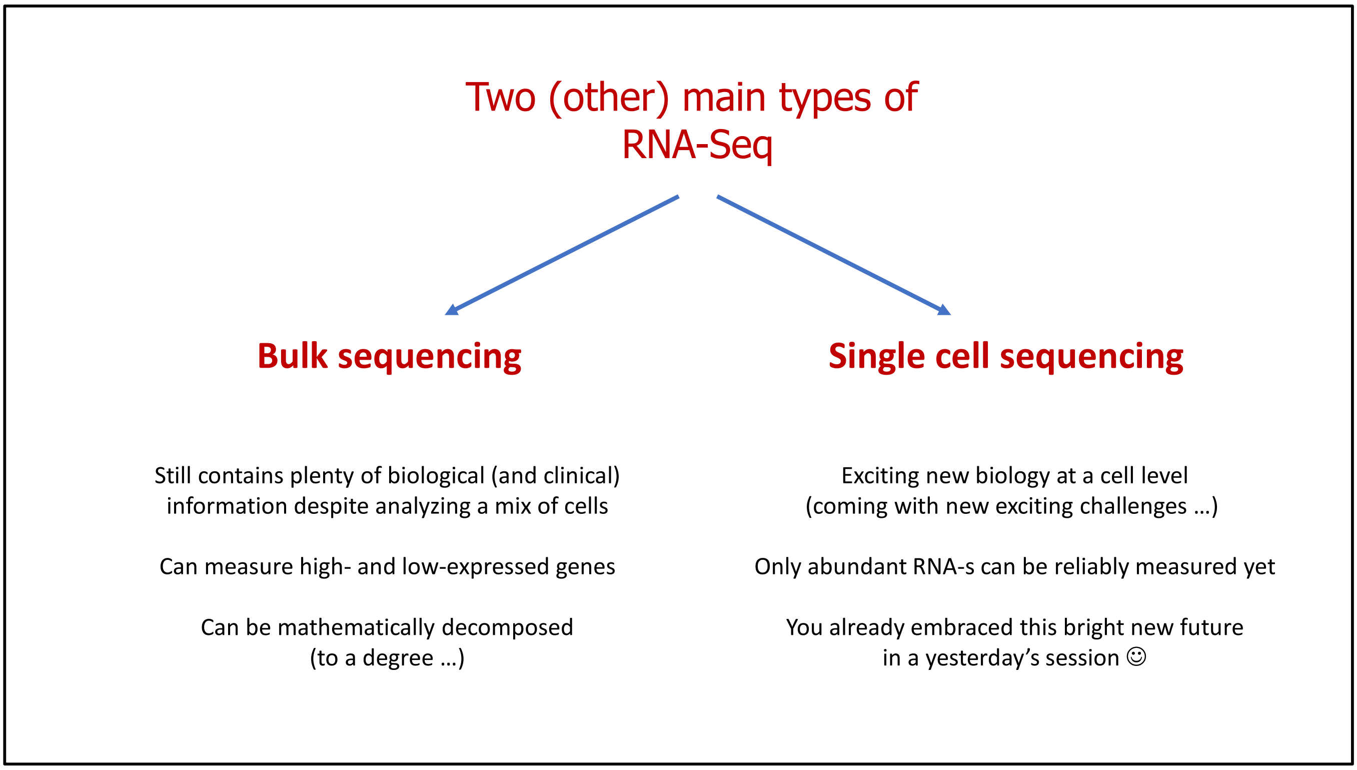
Long-read sequencing

Nanopore

PacBio

*People studying short-RNAs and degraded RNA from FFPE samples may disagree ☺

There are two main types of RNA-sequencing: ***short-read*** and ***long-read*** sequencing.



Two (other) main types of RNA-Seq

Bulk sequencing

Still contains plenty of biological (and clinical) information despite analyzing a mix of cells

Can measure high- and low-expressed genes

Can be mathematically decomposed (to a degree …)

Single cell sequencing

Exciting new biology at a cell level (coming with new exciting challenges …)

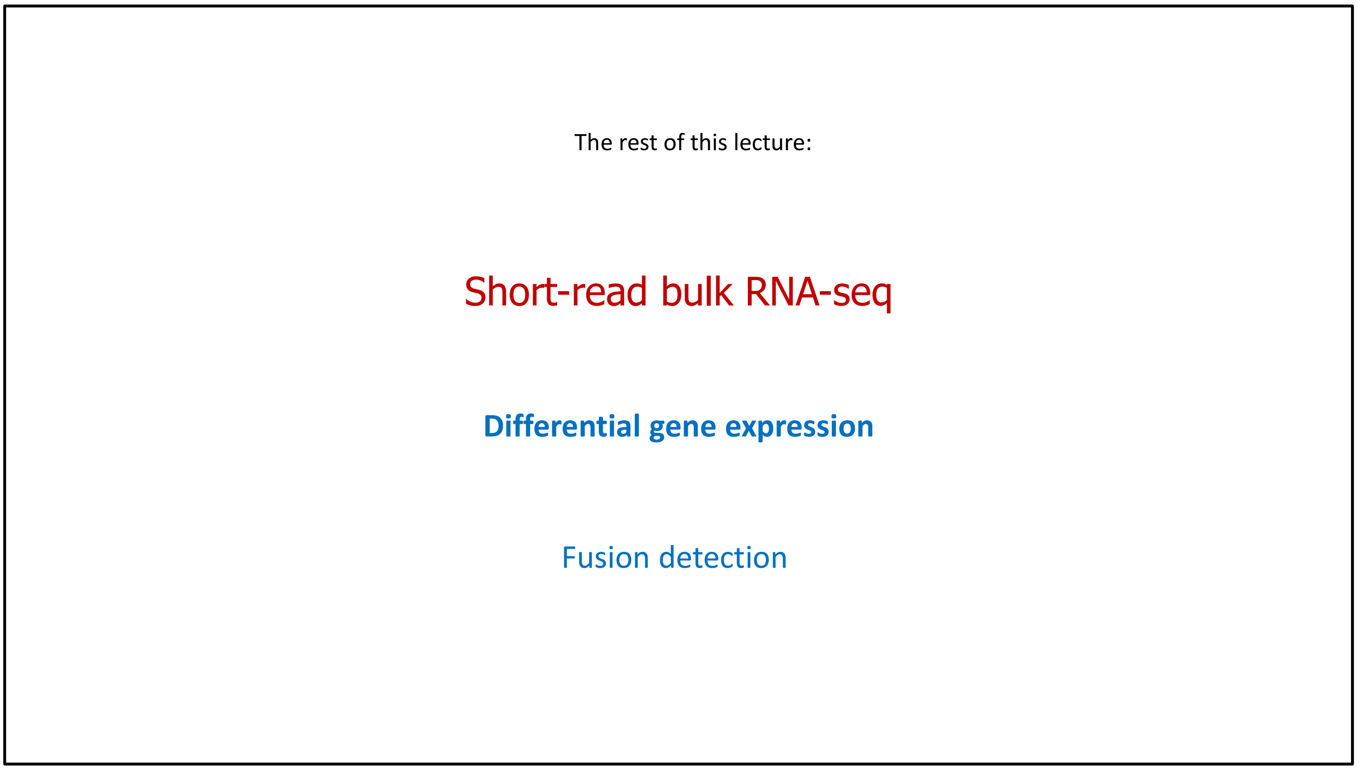
Only abundant RNA-s can be reliably measured yet

You already embraced this bright new future in a yesterday's session ☺

Another dimension in RNA-seq analysis is that it could be done in ***bulk*** samples, consisting of many cells (or even many cell types), and it could be done at a ***single-cell*** level. For now, only abundant RNA-s can be reliably analyzed at a single cell level. However, it seems that even the abundant RNA-s are sufficient to distinguish cell types and to study interesting biology.

Measuring RNA in single cells may sounds like scientific fiction. However, people do it, and you have a separate session about the single-cell analysis during this course.

Importantly, despite all the promises of the single-cell analysis, it doesn't mean that we should disregard the bulk sequencing. First of all, bulk sequencing is easy to apply to clinical samples, like solid tumor biopsies, and it still contains plenty of biological and clinical information despite (or because) it it represents a mix of different cells. Also, the bulk sequencing allows to analyze even the low-expressed genes.



The rest of this lecture:

Short-read bulk RNA-seq

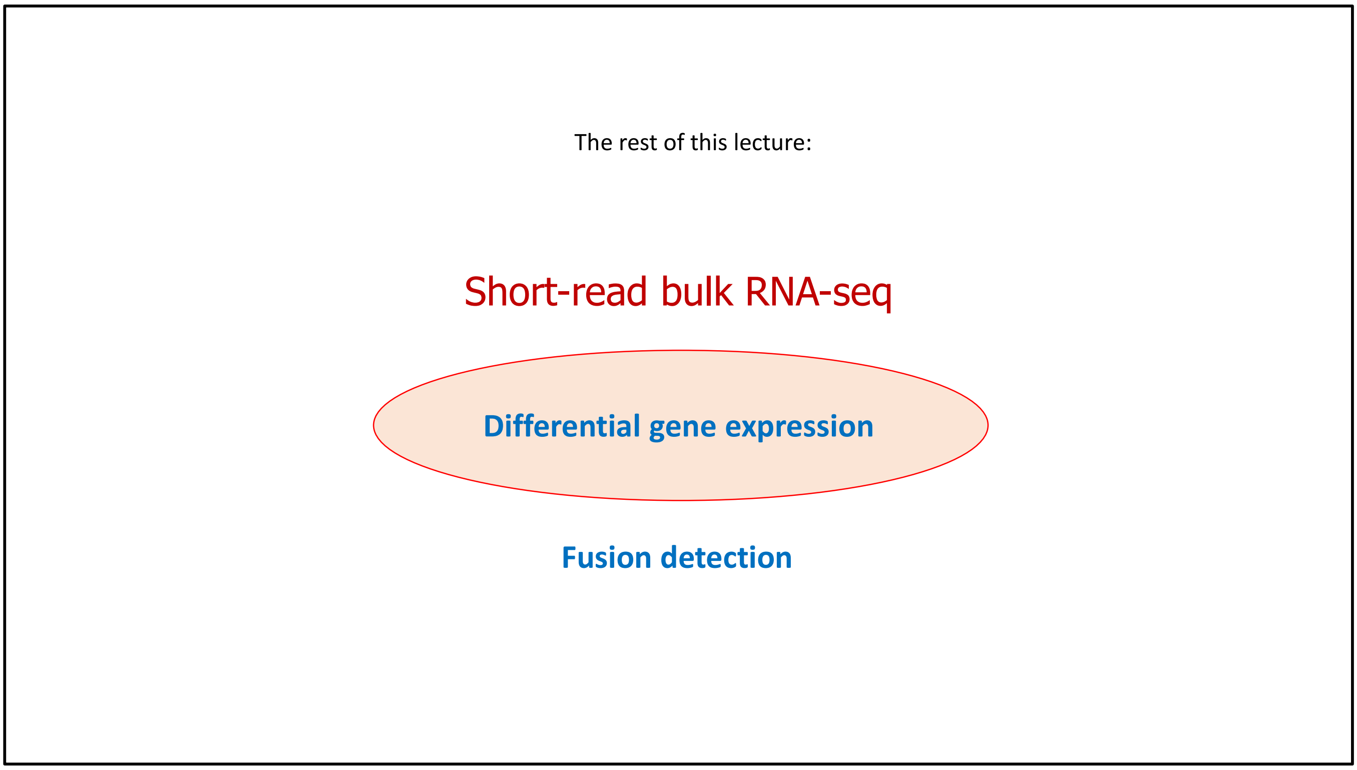
Differential gene expression

Fusion detection

So, the ***single-cells*** and ***long-reads*** are great, and they will be discussed in the other parts of the course. However, during this session we will focus on ***bulk short-reads*** RNA-seq only.

Importantly, many concepts currently used in long-reads or single-cell analysis were first developed for bulk short-reads data.

Also, whatever methods we will use in the future, there are already many good papers and resources made using short-reads bulk RNA seq. So, it’s important to know how to analyze such data.



The rest of this lecture:

Short-read bulk RNA-seq

Differential gene expression

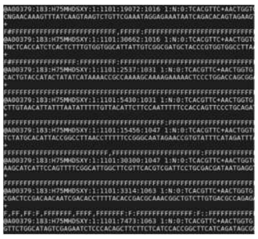
Fusion detection

And, as I said earlier, I will focus on the differential gene expression.

I will also touch upon the fusion detection; although to a lesser extent: because long-reads RNAseq is much better for fusion detection.

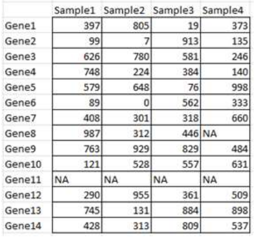
Typical gene expression analysis steps

Low level analysis → Statistical analysis → Functional analysis



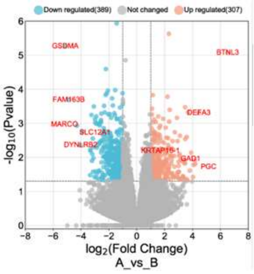
```
@A00379:183:H75MHDSXY:1:1101:19072:1016 1:N:0:TCACGTTC+AACTGGT
CNGAACAAAGTTTATCAAGTAAGTCTGTTCGAAATAGGAGAAATAATCAGACACAGTAGAAG
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00379:183:H75MHDSXY:1:1101:30662:1016 1:N:0:TCACGTTC+AACTGGT
TNCTCACCATCTCACTCTTTGTGGTGGCATTATTGTCGGCGATGCTACCCGTGGTGGCCTTA
+
F#FFFFFFFFFFFFFFFFF:FFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00379:183:H75MHDSXY:1:1101:2537:1031 1:N:0:TCACGTTC+AACTGGTG
CACTGTACCATACTATATCATAAAACCGCCAAAAGCAAAAGAAAAACTCCCTGGACCAGCGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00379:183:H75MHDSXY:1:1101:5430:1031 1:N:0:TCACGTTC+AACTGGTG
CTTGTAACATTATTTAATATTTTTGTTACATTCTTCCAATTTTTCCACCAGTTCCCTGCAGA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFF
@A00379:183:H75MHDSXY:1:1101:15456:1047 1:N:0:TCACGTTC+AACTGGT
TCTATGCACATTACCGGCCTTAACCTTTTTCCGGGCAATAGAACCGTGTATTTCATAGATTT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFF
@A00379:183:H75MHDSXY:1:1101:30300:1047 1:N:0:TCACGTTC+AACTGGT
AAGCATCATTCCAGTTTTCGGCATTGGCTTCGTTCACGTCGATTCCTGCGACGATAATGAGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00379:183:H75MHDSXY:1:1101:3314:1063 1:N:0:TCACGTTC+AACTGGTG
CGACTCCGACAACAATCGACACCTTTTACACCGACGCAAACGGCTGTCTTGTGACGCCAGAG
+
F,FF,FF:F,FFFFFFF,FFFF,FFFFFFF:F:FFFFFFFFFFFFFF:F::FFFFFFFFFFF
@A00379:183:H75MHDSXY:1:1101:7473:1063 1:N:0:TCACGTTC+AACTGGTG
GTTCTGGCATAGTCGAGAATCTCCCACAGCTTCTTCTCATCCACCGGCTTCATCAGATAGCG
```

Raw data
(FASTQ files)

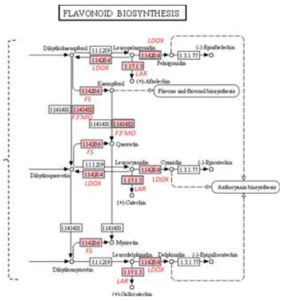


	Sample1	Sample2	Sample3	Sample4
Gene1	397	805	19	373
Gene2	99	7	913	135
Gene3	626	780	581	246
Gene4	748	224	384	140
Gene5	579	648	76	998
Gene6	89	0	562	333
Gene7	408	301	318	660
Gene8	987	312	446	NA
Gene9	763	929	829	484
Gene10	121	528	557	631
Gene11	NA	NA	NA	NA
Gene12	290	955	361	509
Gene13	745	131	884	898
Gene14	428	313	809	537

Counts
per sample
(transcripts or genes)



Differentially expressed genes,
clustering, predicting response or
prognosis

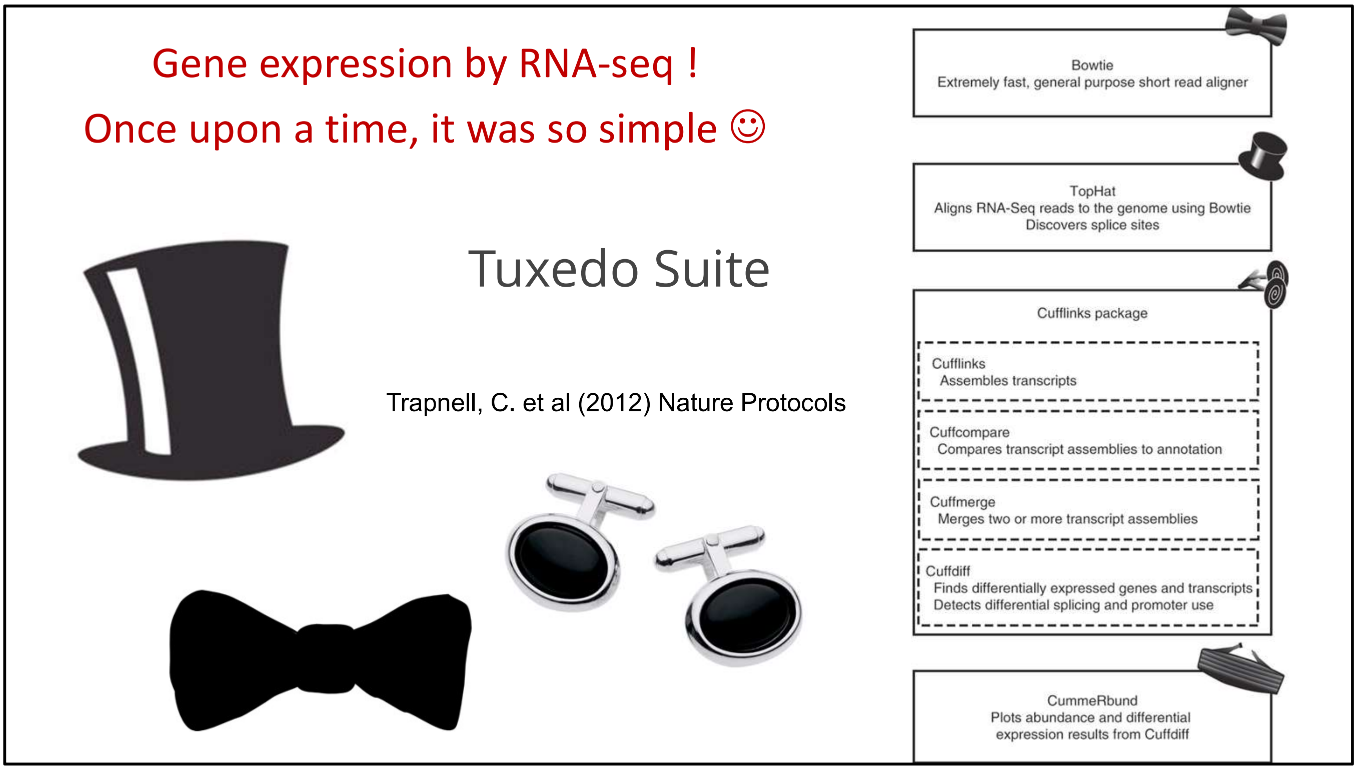


Gene sets enrichment,
Fitting to pathways, GOs

At a very high level, gene expression data analysis may be split into three steps:

1) Processing from raw data to transcripts / genes counts. This involves bioinformatics tools specialised at such tasks as alignment and counting.
2) Identifying differentially expressed genes, samples and genes clustering. This step usually involves highly specialised statistical tests and even machine-learning algorithms.
3) Functional interpretation of the results, like finding activated pathways, or enriched gene-ontology terms. This step depends on large biological databases and statistical tests for enrichment.

This session will mainly focus on the first two steps. Responding to the participants feedback from the previous years, I have added some elements of functional analysis. However, this is a very broad field, which is not entirely specific for transcriptomic (e.g. you may apply it to mutational data or to proteomics). So, I will only briefly touch upon the functional analysis.



Gene expression by RNA-seq !

Once upon a time, it was so simple ☺

Tuxedo Suite

Trapnell, C. et al (2012) Nature Protocols

Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

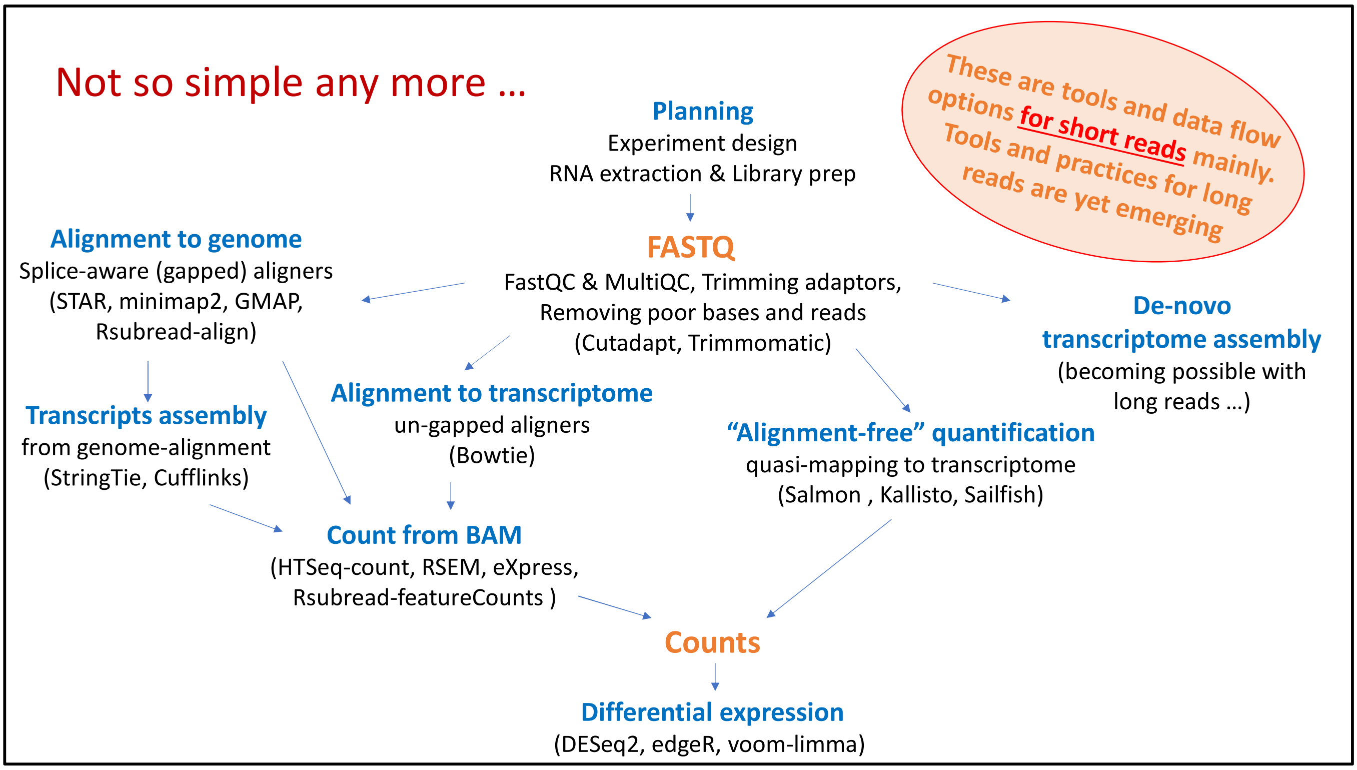
Cuffmerge
Merges two or more transcript assemblies

Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

CummeRbund
Plots abundance and differential expression results from Cuffdiff

Gene expression! Just about a dozen years ago it was so simple (I am joking :)

I mean that for quite a long time this brilliant pipeline dominated the RNA-seq world: providing an integrated solution from alignment to differential expression !



Not so simple any more …

These are tools and data flow options **for short reads** mainly. Tools and practices for long reads are yet emerging

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

“Alignment-free” quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

Over the years it has become much more complicated: the field has become crowded with so many tools and algorithms!

And if you think that I am over-complicating here, then look at the next slide…

Diversity of methods and tools for RNA-Seq data analysis

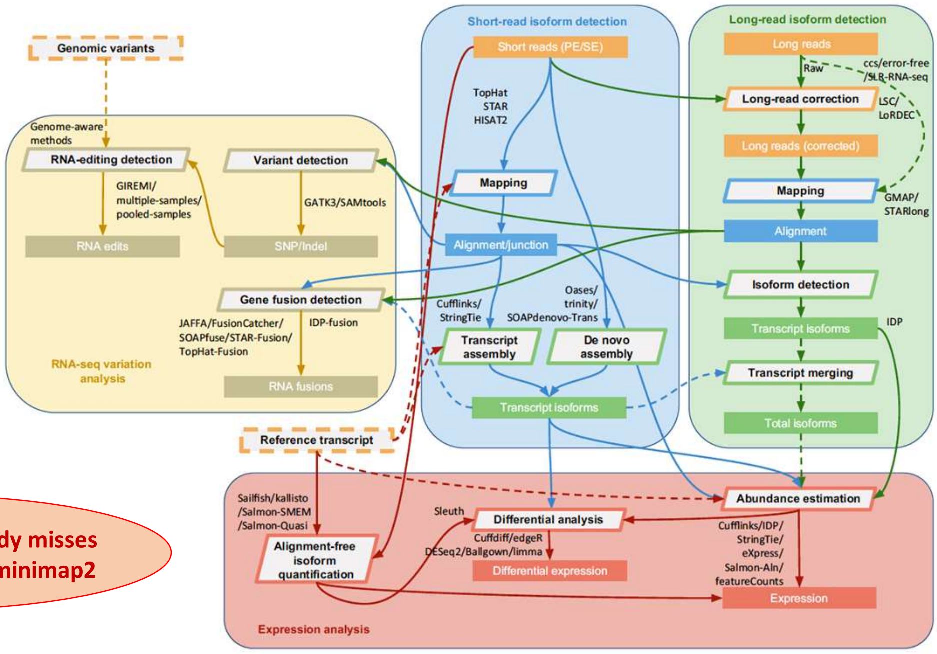
The authors examine

39 analysis tools …
~120 combinations
~490 analyses …

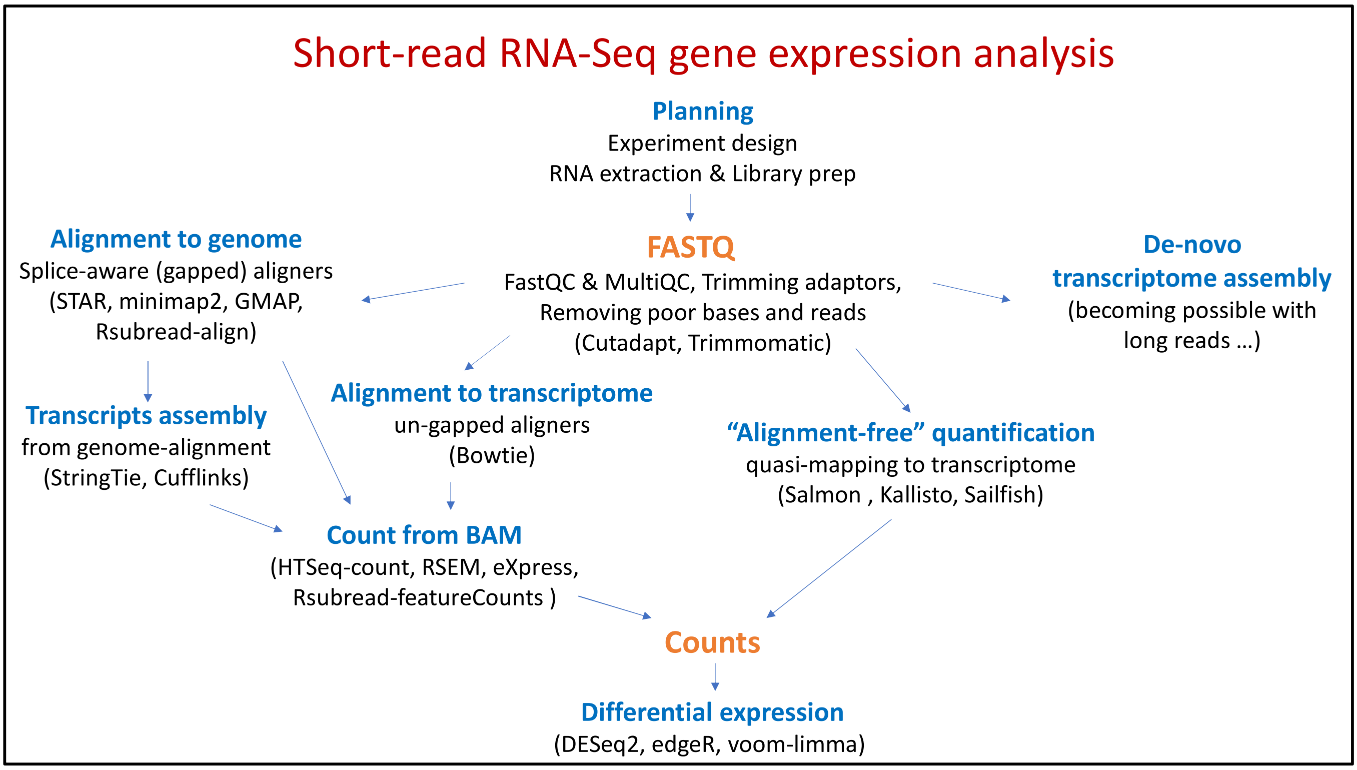
with a variety of
data types

Sahraeian *et al* 2017
Nature Communications 8: 59
DOI: 10.1038/s41467-017-00050-4

This paper (of 2017) already misses
some key new tools, like minimap2



This is from a very nice paper that was written several years ago. It already misses many tools introduced since that. And it had already identified tens of reputable tools and compared more than a hundred different trajectories for RNA-seq data analysis.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

“Alignment-free” quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

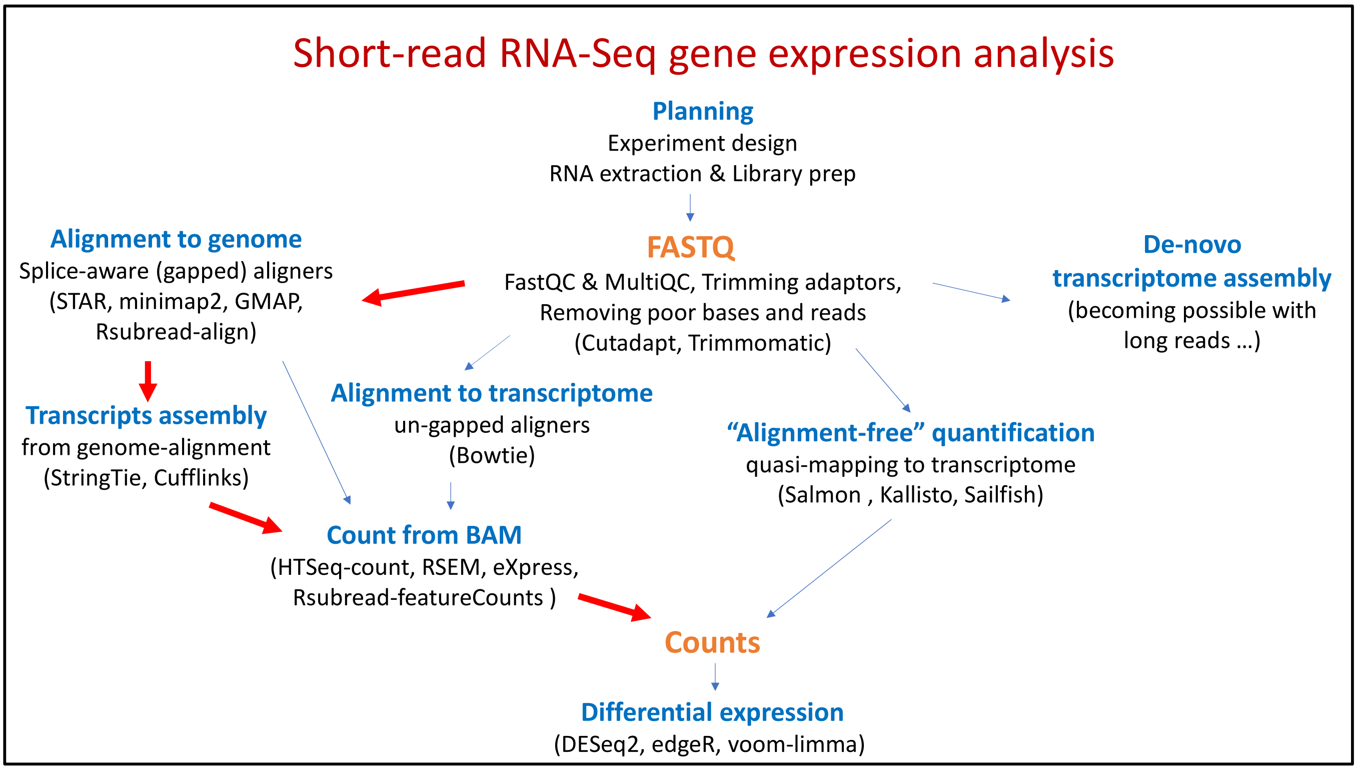
Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

For easier navigation through the main choices, I prefer to simplify the picture.

This slide illustrates the main trajectories how to get from the short-reads FASTQ file to the transcripts/genes Counts.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

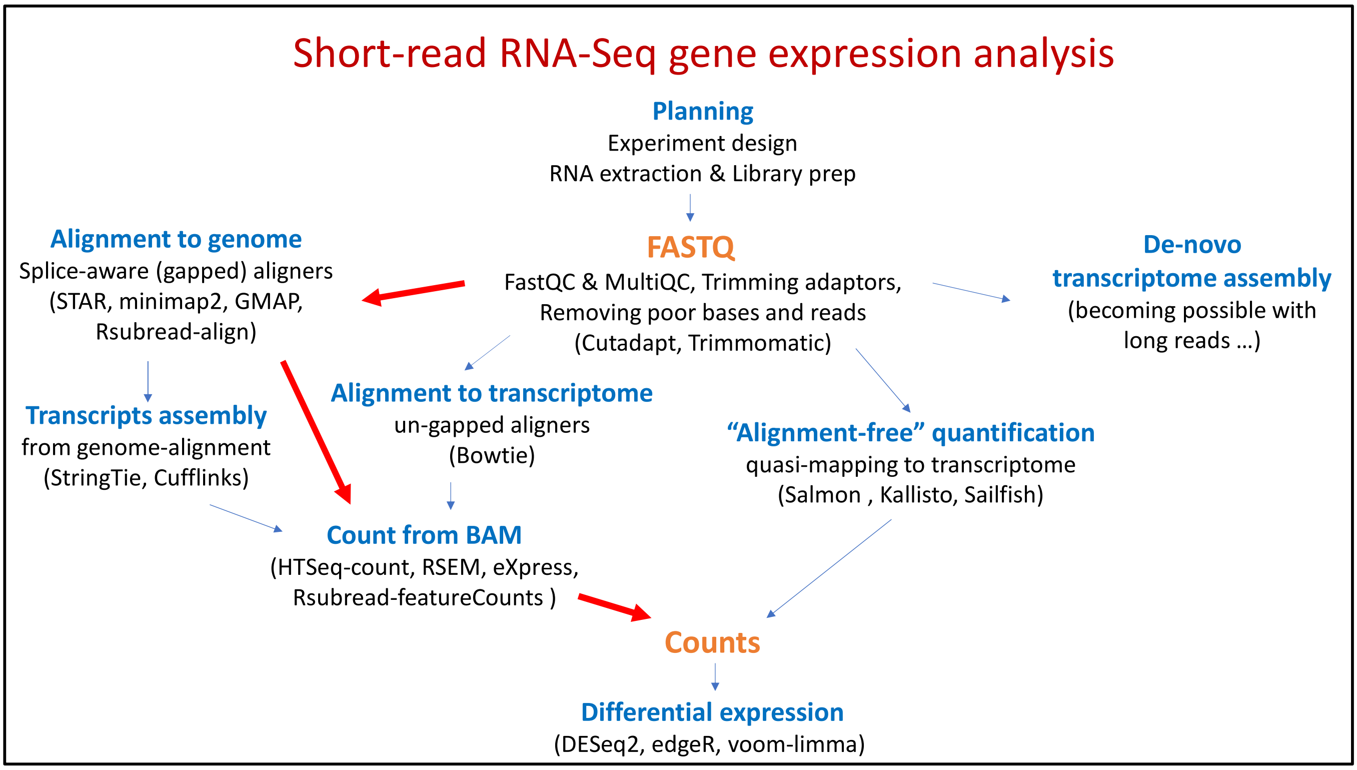
“Alignment-free” quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

In context of human cancer research, the longest trajectory includes:
(1) Alignment to Reference Genome,
(2) Transcriptome Assembly, which might potentially discover new transcripts,
(3) followed by the Transcript Count from BAM.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

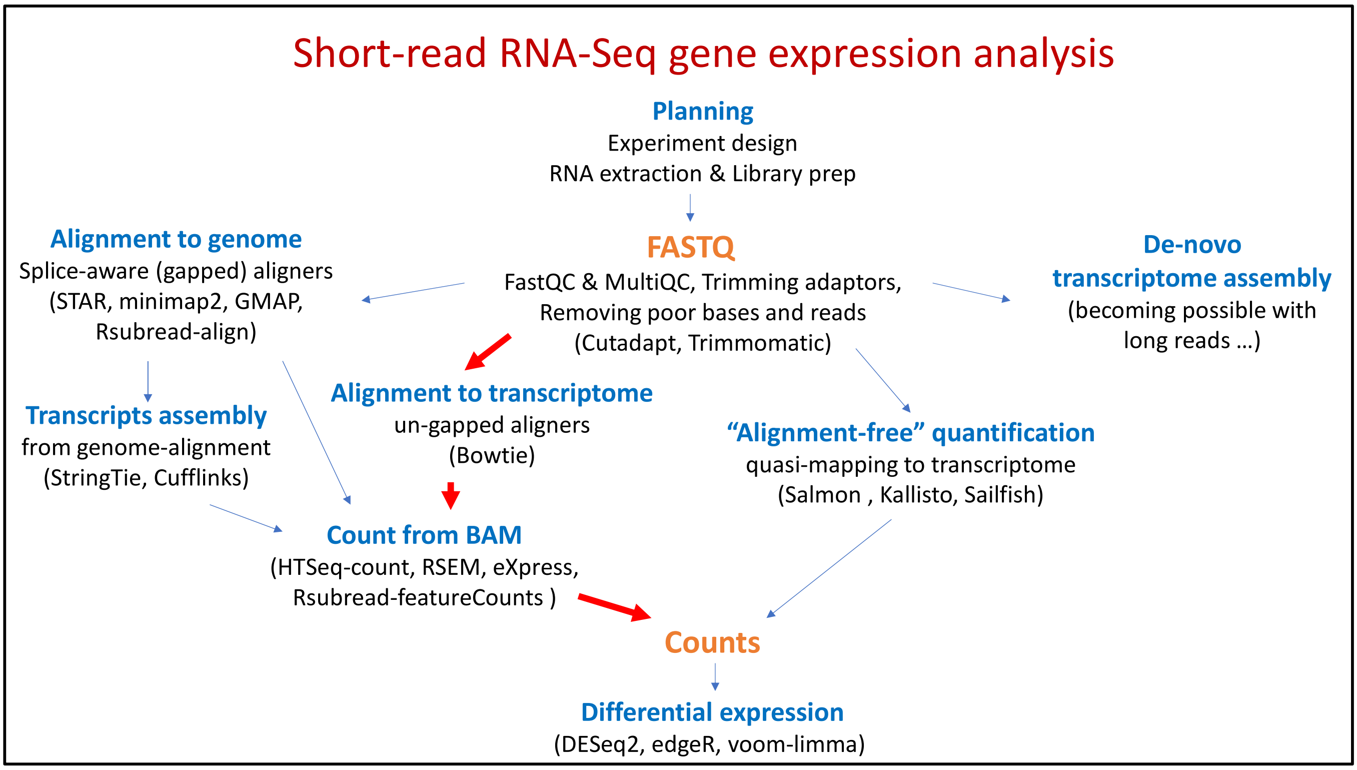
“Alignment-free” quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

A shorter path skips the transcript assembly, assuming that we already have a good annotated transcriptome for the studied species (i.e. human transcriptome).
So, this path goes directly from the Genome Alignment to the Counts, counting against the known features already annotated in the genome.
The features could be known genes, transcripts, exons or something else.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

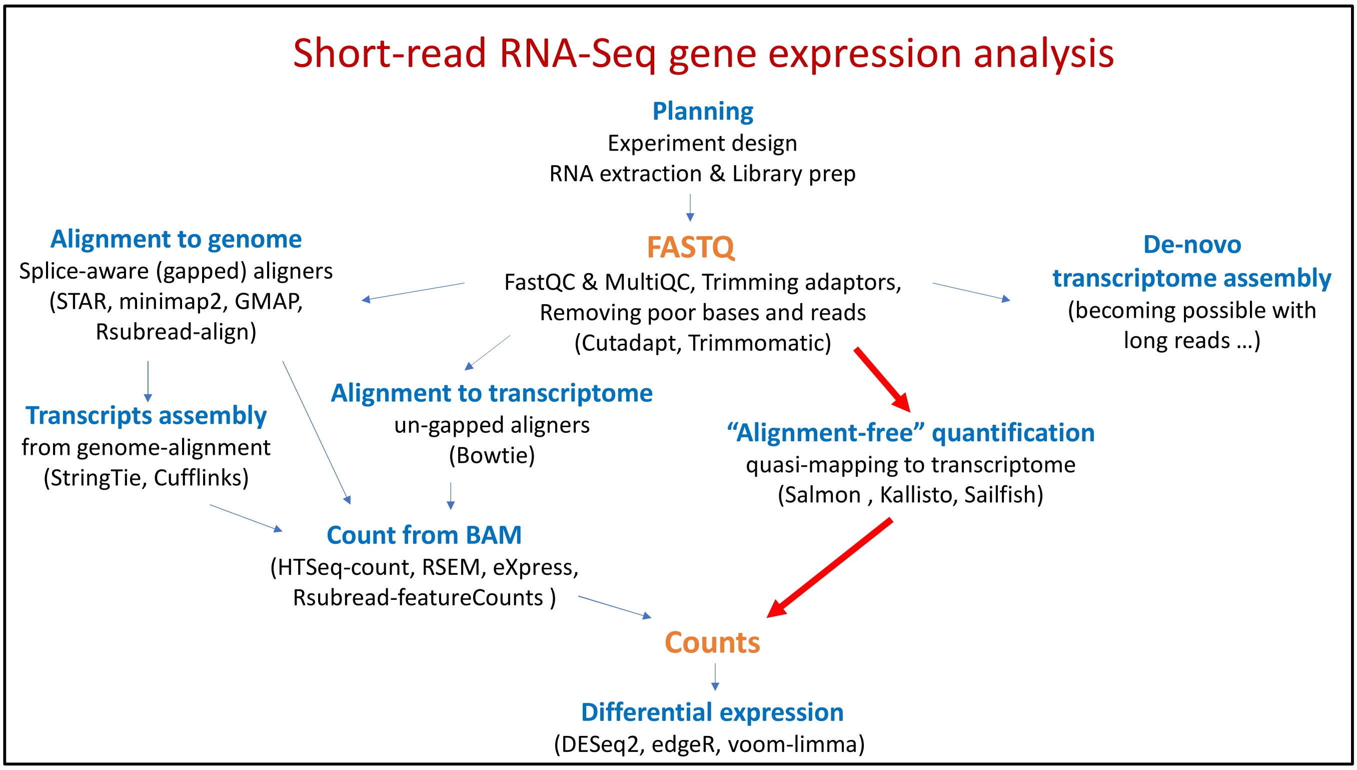
“Alignment-free” quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

Another path goes through alignment to transcriptome.
It does not require a reference genome, and it doesn’t require a gapped, splice-aware aligner.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

“Alignment-free” quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

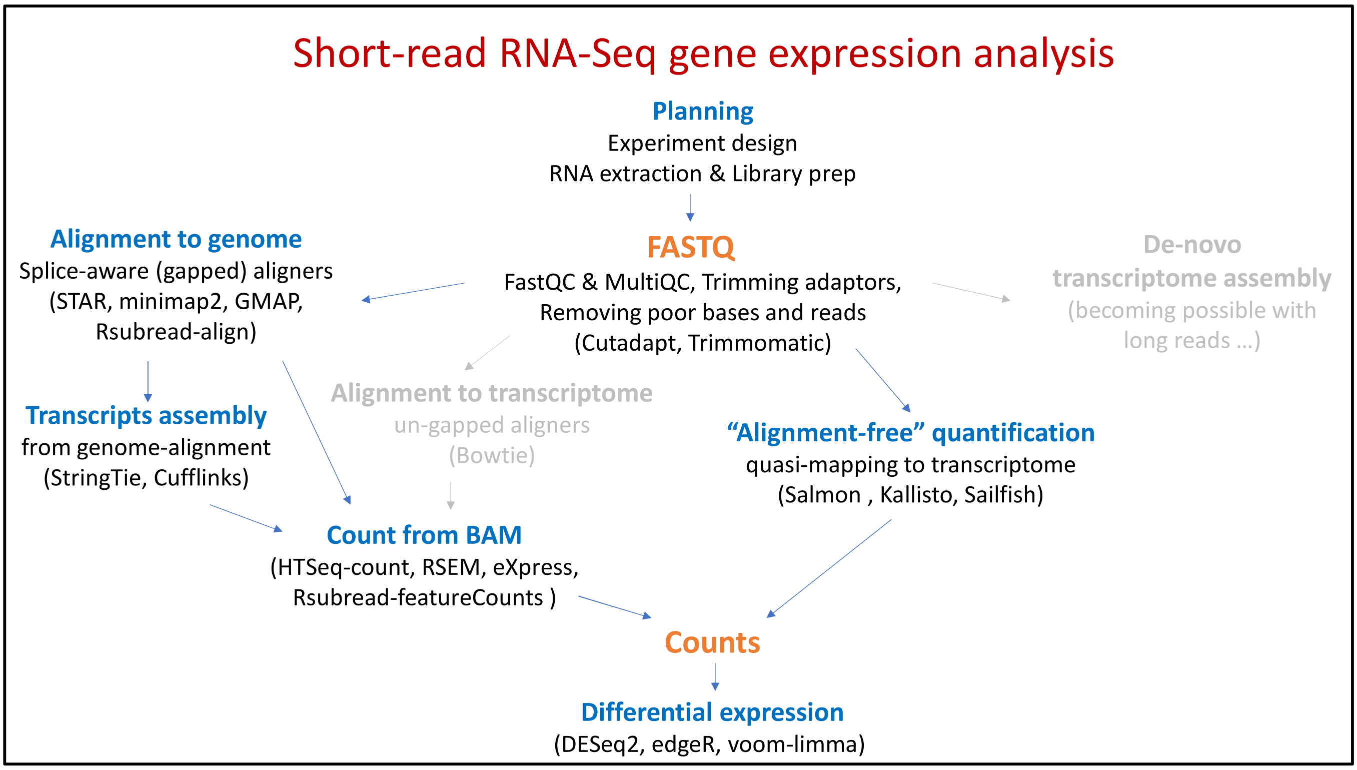
Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

The shortest path, employing “alignment-free” tools, jumps directly from the FASTQ to the Counts.

The alignment-free tools do not make a proper alignment into BAM files. Instead, they use some simple and quick “quasi-mapping” to get the counts against known transcripts.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

“Alignment-free” quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

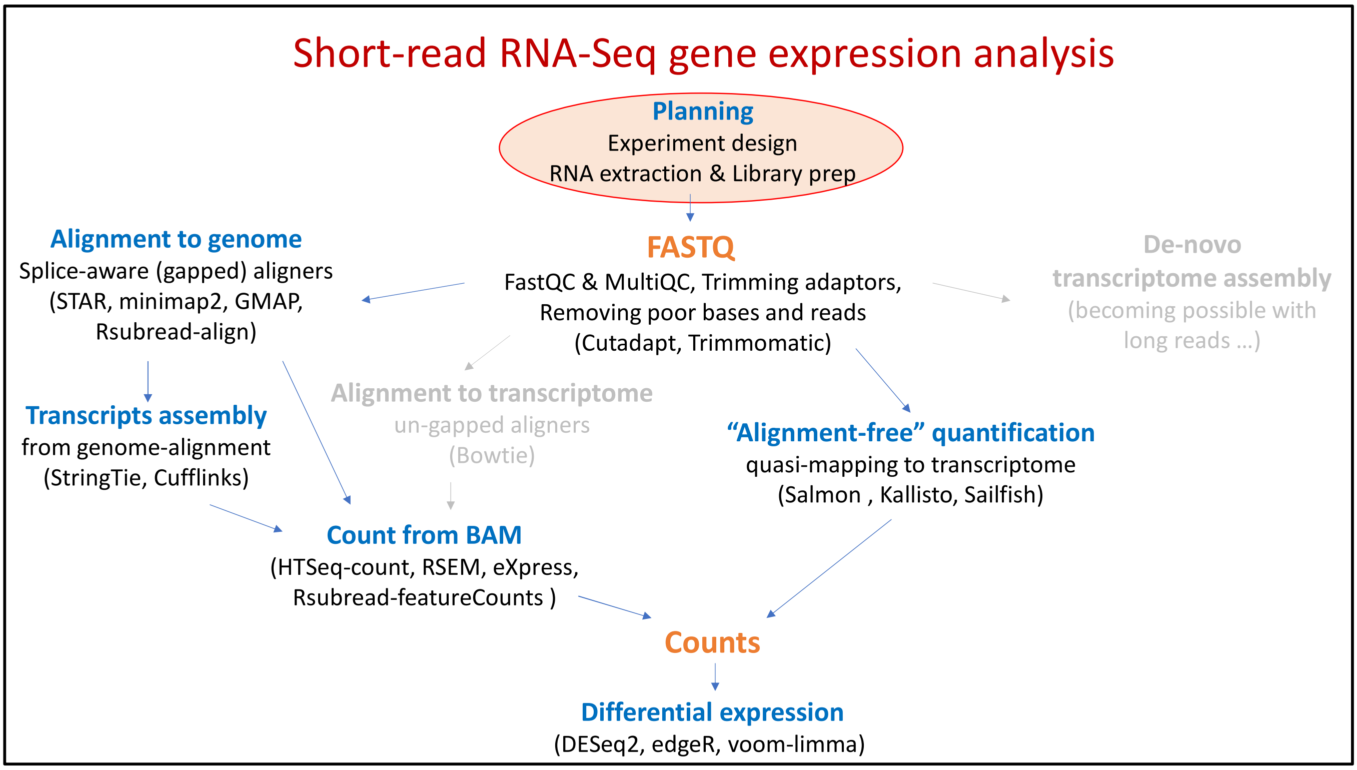
Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

In this talk, I will briefly go through the each of the mentioned steps, ecxcept the alignment to transcriptome. It was successfully done in many studies before the development of the alignment-free methods. But now the alignment-free methods can count against the known transcriptome much easier and faster.

Also, I will not consider *De-novo* transcriptome assembly, because, realistically, it should be done only with long reads nowadays.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

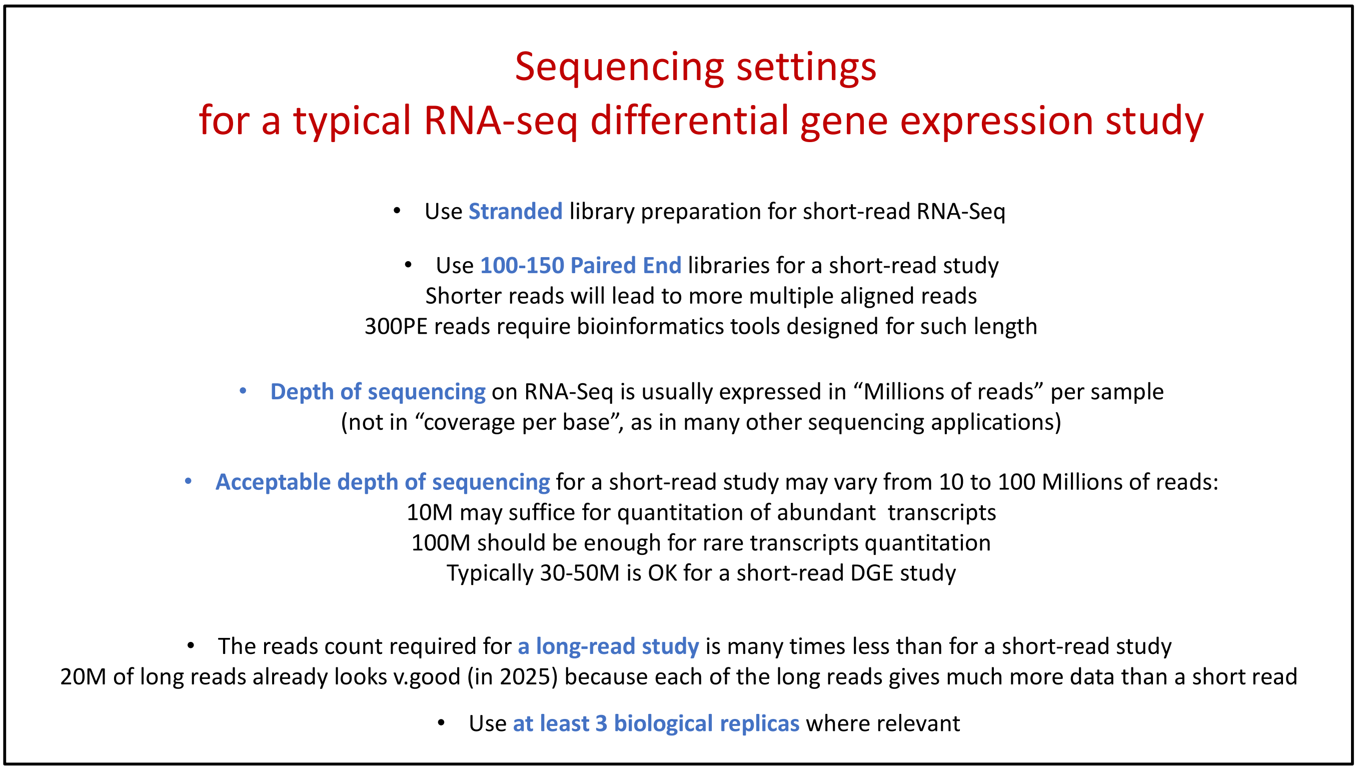
“Alignment-free” quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

However, before discussing the clever splice-aware aligners or quasi-mapping, we shall briefly look at the experiment design, RNA quality assessment and Illumina library preparation for RNA-seq.



Sequencing settings for a typical RNA-seq differential gene expression study

- Use **Stranded** library preparation for short-read RNA-Seq

- Use **100-150 Paired End** libraries for a short-read study
Shorter reads will lead to more multiple aligned reads
300PE reads require bioinformatics tools designed for such length

- **Depth of sequencing** on RNA-Seq is usually expressed in “Millions of reads” per sample
(not in “coverage per base”, as in many other sequencing applications)

- **Acceptable depth of sequencing** for a short-read study may vary from 10 to 100 Millions of reads:
10M may suffice for quantitation of abundant transcripts
100M should be enough for rare transcripts quantitation
Typically 30-50M is OK for a short-read DGE study

- The reads count required for **a long-read study** is many times less than for a short-read study
20M of long reads already looks v.good (in 2025) because each of the long reads gives much more data than a short read

- Use **at least 3 biological replicas** where relevant

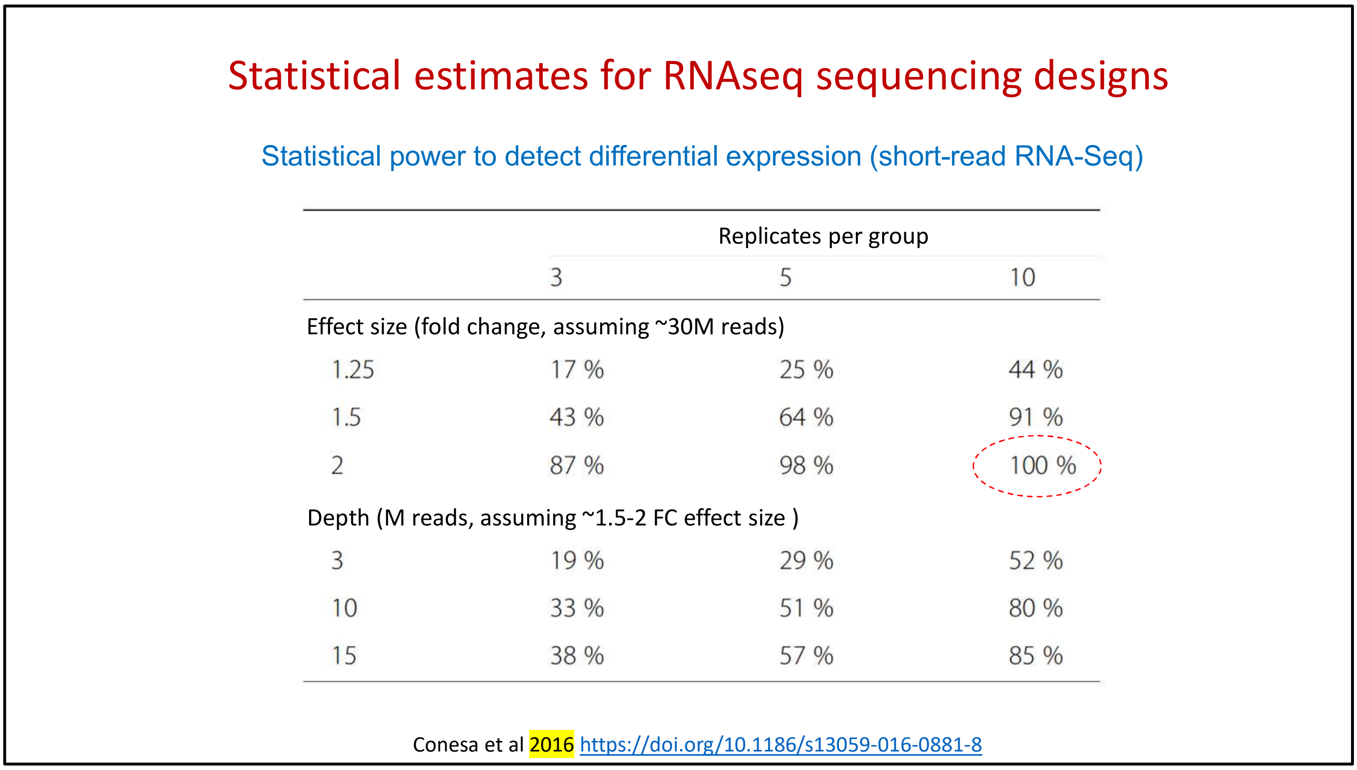
To my knowledge, only stranded library preparation is currently used for RNAseq (we will discuss it shortly).

100-150 paired end libraries are well compatible with most of the short-read RNA-seq bioinformatics tools.

Please note, the depth of sequencing in RNAseq is expressed in “millions of reads” per sample, not in “coverage per base” as in DNA-sequencing.

Just some years ago, a typical depth of sequencing for a short-read RNA-seq gene expression experiment was about 30-50 M reads. Now, after the cost of Illumina sequencing has significantly dropped, 100M+ reads is becoming quite common. The higher depth allows better quantification of low-abundance genes.

Of course, always plan for a sufficient number of biological replicates for statistical analysis.



Statistical estimates for RNAseq sequencing designs

Statistical power to detect differential expression (short-read RNA-Seq)

	Replicates per group		
	3	5	10
Effect size (fold change, assuming ~30M reads)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Depth (M reads, assuming ~1.5-2 FC effect size)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

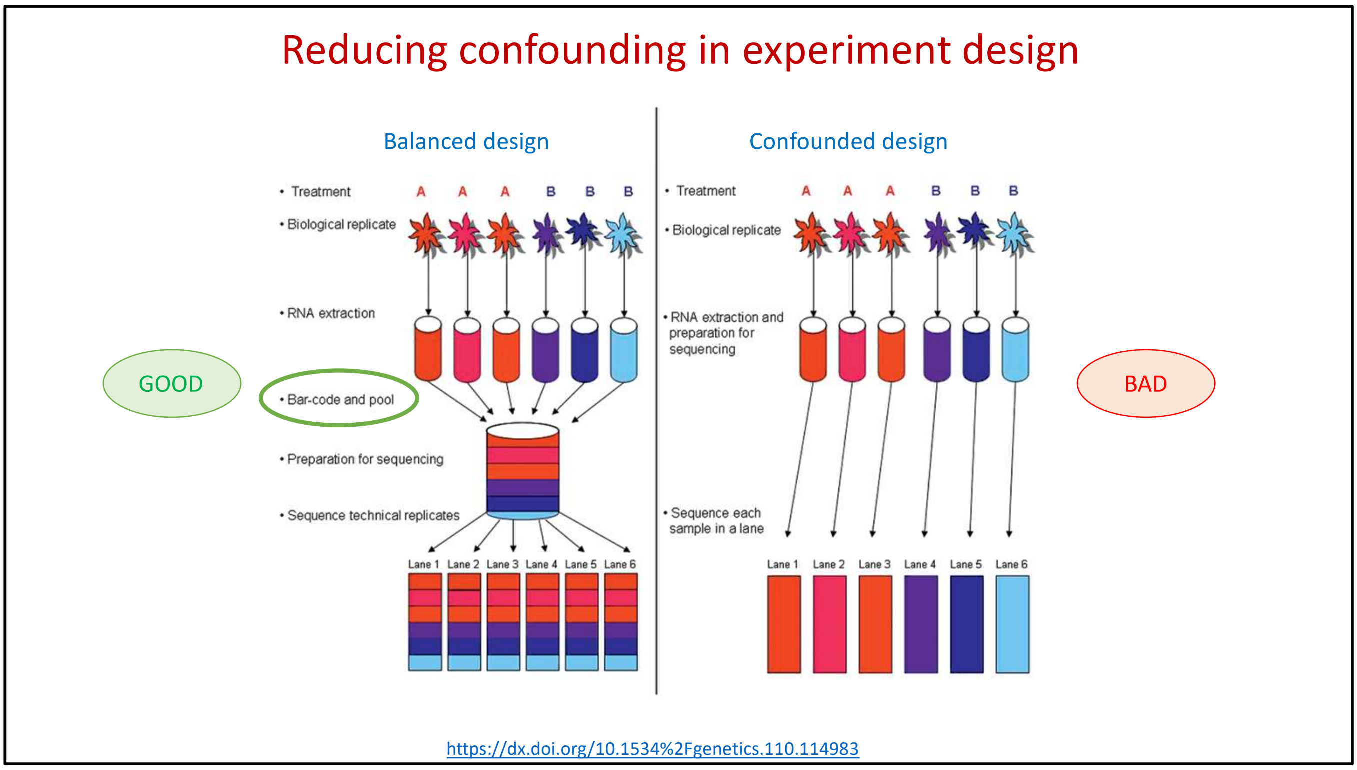
Conesa et al 2016 https://doi.org/10.1186/s13059-016-0881-8

When sequencing was not as affordable as now, there were studies with some statistical modelling to estimate the depth of sequencing or number of replicas to detect a specific fold change in differential expression. It might be noted that, for instance, 30M reads and 10 biological replicas are sufficient for detecting all genes with at least 2-fold change. Of course, any such models are based on some assumptions that are not always correct, and the short-reads sequencing is now MUCH less expensive than 10 years ago.

At the moment (2025), a starting point to consider for a short-reads RNAseq differential ***gene*** expression experiment may look like this:

(1) At least 30-50 millions reads per sample (100M is often affordable)
(2) As many BIOLOGICAL REPLICAS as affordable and accepted in your specific field

At least 100M reads would be expected for new ***isoforms*** identification and quantification, or for *De-Novo* transcriptome assembly. However, such tasks should require long reads nowdays.



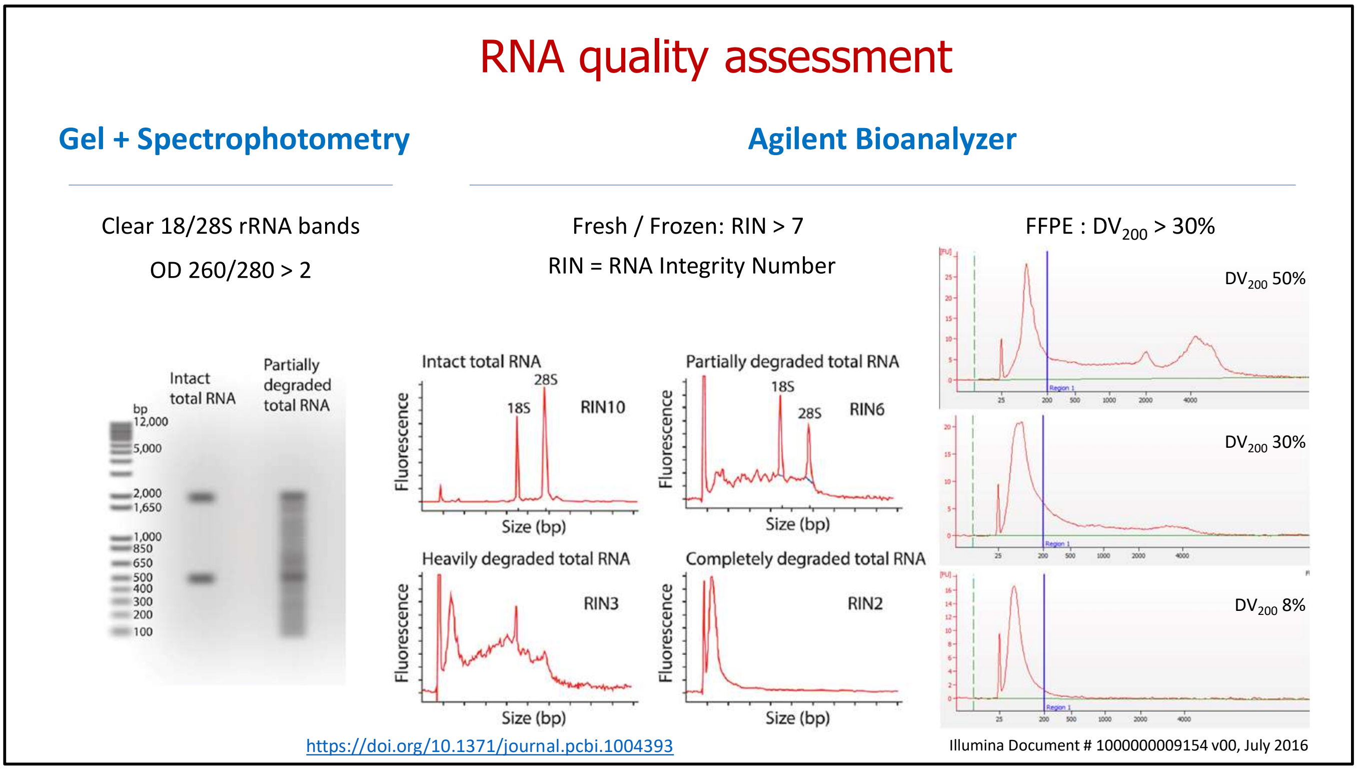
Reducing confounding in experiment design

https://dx.doi.org/10.1534%2Fgenetics.110.114983

It is very important to consider multiplexing, placing the compared samples on the same lanes.

If it is not possible to multiplex all the samples on a single lane, it is important to place an equal proportion of cases and controls on each lane (batch).

For instance:
If you have a study with 240 samples, which include 25% of cases and 75% of controls, you may use 10 lanes of sequencing placing 24 samples per lane, with 6 cases and 18 controls per lane.
Of course, the number of samples per lane (24 in this example) may vary depending on the required depth of sequencing.



RNA quality assessment

Gel + Spectrophotometry

Clear 18/28S rRNA bands

OD 260/280 > 2

Agilent Bioanalyzer

Fresh / Frozen: RIN > 7

RIN = RNA Integrity Number

FFPE : DV_{200} > 30%

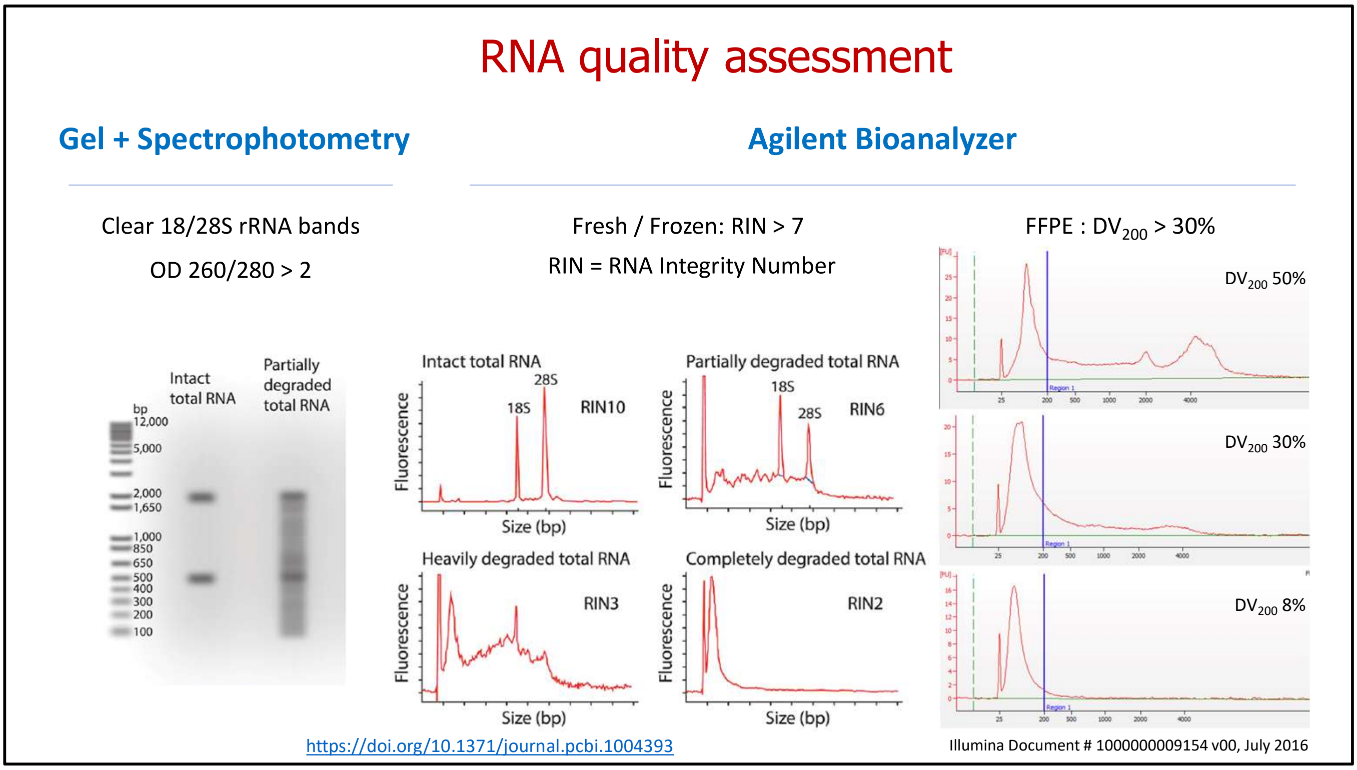
https://doi.org/10.1371/journal.pcbi.1004393

Illumina Document # 1000000009154 v00, July 2016

The RNA assessment should consider ***quantity***, ***purity*** and ***integrity***.

The ***quantity*** can be assessed by spectrophotometer (e.g. Nanodrop) or spectrofluorometer (e.g. Orbit). The spectrofluorometer is more accurate. In many cases (cell lines, blood cells, fresh-frozen surgical biopsies) at least 100ng (e.g. 10ul of 10ng/ul) of total RNA is easily available, which is sufficient for PCR-free RNAseq library preparation. However, low abundance samples (e.g. needle biopsies or FFPE samples) may require PCR-based library prep kits.

The ***purity*** of RNA, in terms of absence of the protein contamination, can be evaluated with a spectrophotometer by measuring 260 / 280 ratio. The traces of phenol, which sometime may be used in RNA extraction, could be detected by an additional absorption peak at about 230.



RNA quality assessment

Gel + Spectrophotometry

Clear 18/28S rRNA bands

OD 260/280 > 2

Agilent Bioanalyzer

Fresh / Frozen: RIN > 7

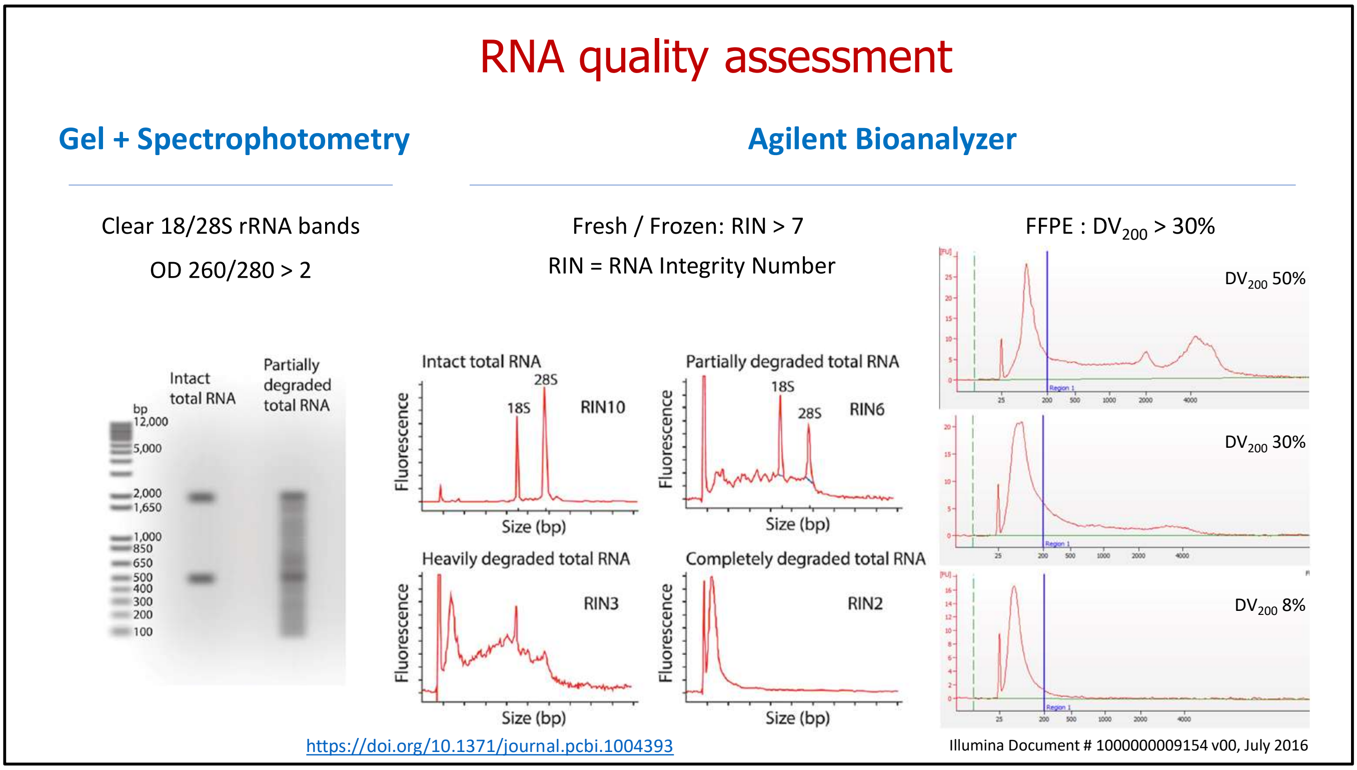
RIN = RNA Integrity Number

FFPE : DV$_{200}$ > 30%

https://doi.org/10.1371/journal.pcbi.1004393

Illumina Document # 1000000009154 v00, July 2016

Integrity of RNA is usually assessed by the integrity of *Ribosomal* RNA, which constitutes up to 90% of the total RNA. There should be two clear visible peaks of 18 and 28S ribosomal RNAs in the non-degraded total human RNA. Nowadays, capillary electrophoresis machines (available from Agilent or QIAGEN) is preferred over the big gels. Using an ML algorithm the capillary electrophoresis machine software calculates a score, RNA-integrity number (RIN). RIN 10 corresponds to the perfect RNA with fully preserved 18 and 28 peaks. The RNA integrity number number is going down when RNA is degraded. If you study cell lines or fresh-frozen tissue, you should aim at an RNA integrity number above 7.



RNA quality assessment

Gel + Spectrophotometry

Clear 18/28S rRNA bands

OD 260/280 > 2

Agilent Bioanalyzer

Fresh / Frozen: RIN > 7

RIN = RNA Integrity Number

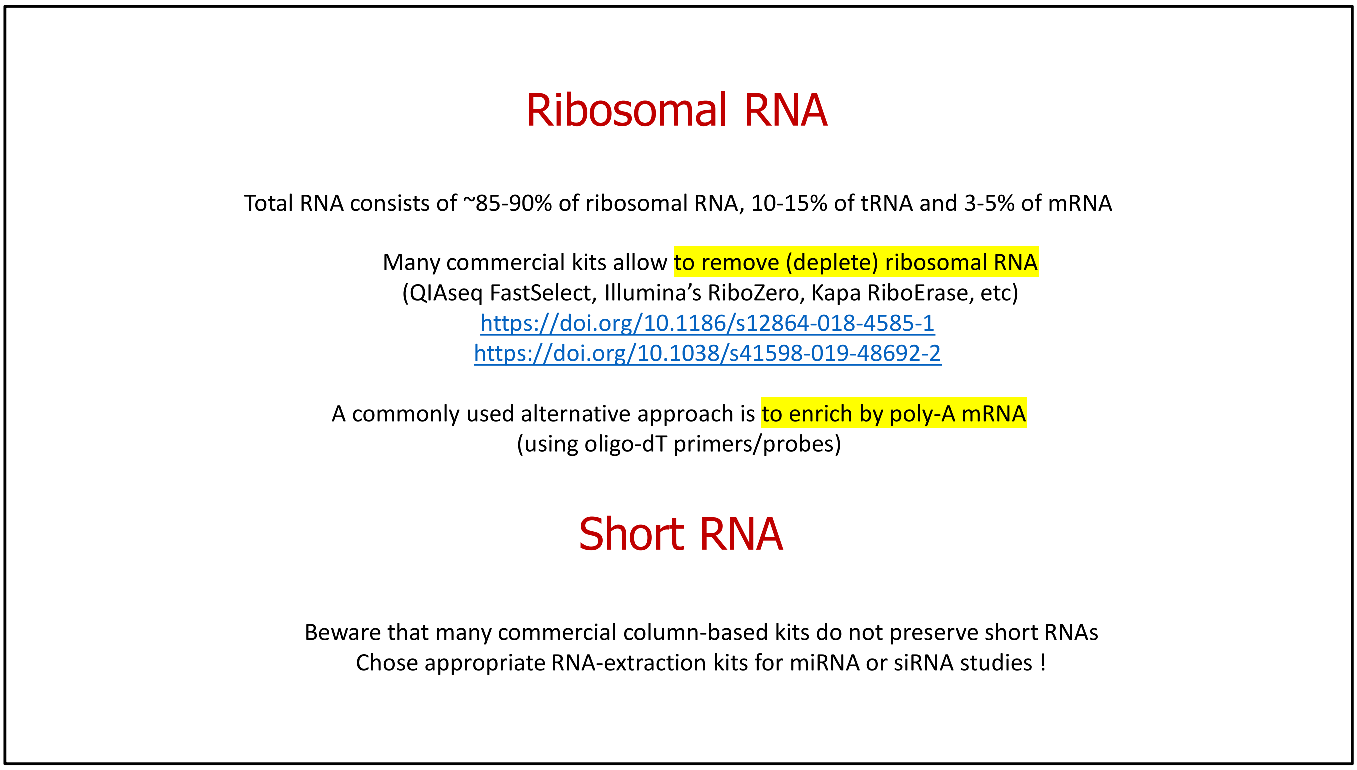
FFPE : DV_{200} > 30%

https://doi.org/10.1371/journal.pcbi.1004393

Illumina Document # 1000000009154 v00, July 2016

The short reads RNA-seq may also be done on degraded samples, such as formalin-fixed paraffin-embedded (FFPE) archival biopsies. DV200 index was suggested to formalize RNA integrity assessment in such samples : the samples can be used if at least 30% of the RNAs are longer than 200 base pairs.

Although, it is possible to use partially degraded RNA in RNA-seq, the quality of RNA should be similar between different samples.   In other words, the gene expressions in samples with high-quality RNA should not be directly compared to the gene expressions from samples with low RNA quality.



Ribosomal RNA

Total RNA consists of ~85-90% of ribosomal RNA, 10-15% of tRNA and 3-5% of mRNA

Many commercial kits allow to remove (deplete) ribosomal RNA
(QIAseq FastSelect, Illumina's RiboZero, Kapa RiboErase, etc)
https://doi.org/10.1186/s12864-018-4585-1
https://doi.org/10.1038/s41598-019-48692-2

A commonly used alternative approach is to enrich by poly-A mRNA
(using oligo-dT primers/probes)

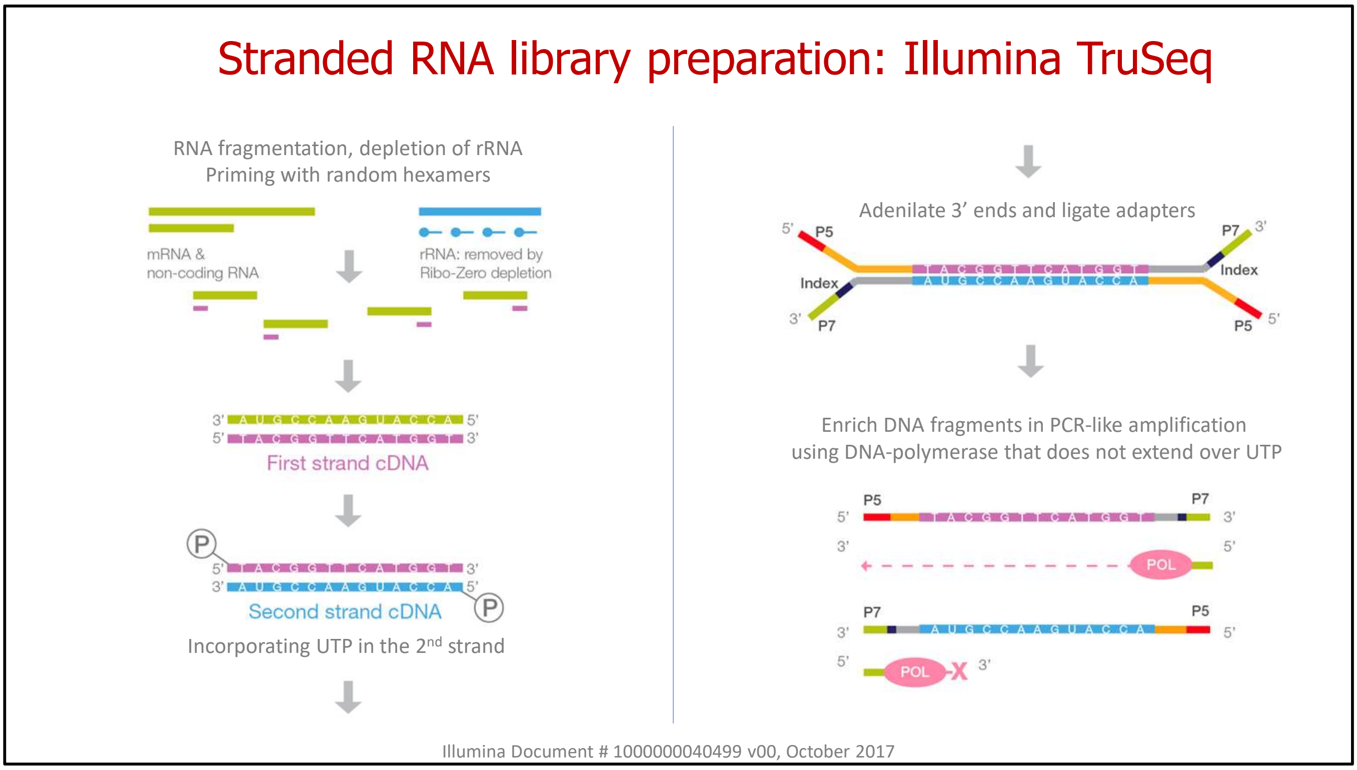
Short RNA

Beware that many commercial column-based kits do not preserve short RNAs
Chose appropriate RNA-extraction kits for miRNA or siRNA studies !

After extraction of RNA, we usually are not interested in most of it: i.e. we usually don't study ribosomal RNA, which constitutes up to 90% of total RNA.

There are many kits that remove ribosomal RNA, or enrich mRNA before library preparation. Using poly-A based enrichment kits, be aware that some human genes do not have poly-A tails on their mRNA (e.g. histones https://www.nature.com/articles/nrg2438)

Another practical point to note is that many RNA extraction column-based kits do not preserve short RNAs. It's even good if you are only interested in protein-coding genes. But be careful if you study, for instance, micro-RNAs.



Stranded RNA library preparation: Illumina TruSeq

RNA fragmentation, depletion of rRNA
Priming with random hexamers

mRNA &
non-coding RNA
rRNA: removed by
Ribo-Zero depletion
3' A U G C C A A G U A C C A 5'
5' T A C G G T T C A T G G T 3'
First strand cDNA
P
5' T A C G G T T C A T G G T 3'
3' A U G C C A A G U A C C A 5'
Second strand cDNA
P

Incorporating UTP in the 2nd strand

Adenilate 3' ends and ligate adapters

5' P5
P7 3'
Index
T A C G G T T C A T G G T
A U G C C A A G U A C C A
Index
3' P7
P5 5'

Enrich DNA fragments in PCR-like amplification
using DNA-polymerase that does not extend over UTP

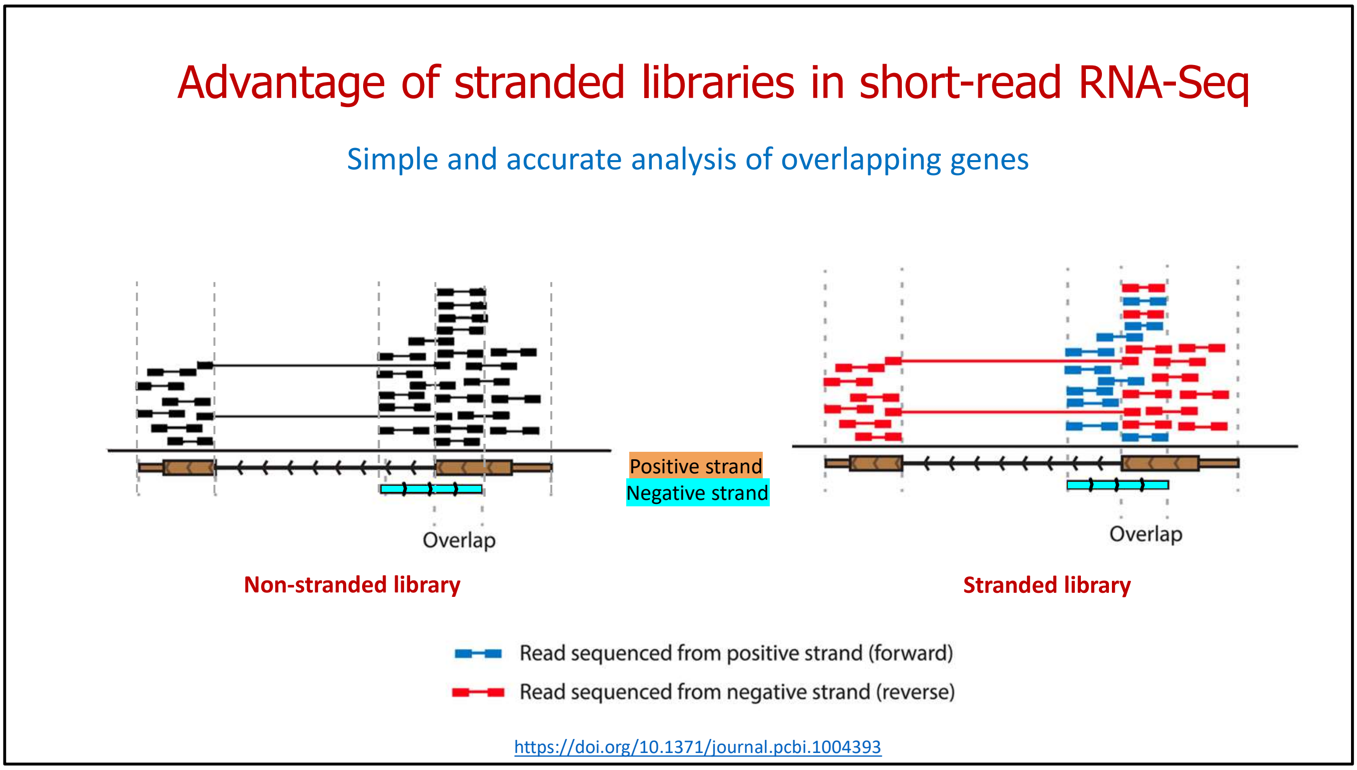
P5
P7
5'
T A C G G T T C A T G G T
3'
3'
5'
POL
P7
P5
3'
A U G C C A A G U A C C A
5'
5'
POL
X
3'

Illumina Document # 1000000040499 v00, October 2017

Finally, a couple of practical notes about library preparation: (i) using stranded kits and (ii) using UMIs.

Currently, all main Illumina RNAseq library preparation kits are stranded. However, beware that some old RNAseq datasets might be generated with non-stranded kits.

Stranded kits discard non-coding complementary strands, which could be generated during the library preparation. This slide illustrates an Illumina kit for stranded library preparation. It incorporates UTP during the second-strand synthesis of cDNA. Then it uses DNA-polymerase that does not extend through UTP – so only one strand is amplified and taken for sequencing.



Advantage of stranded libraries in short-read RNA-Seq

Simple and accurate analysis of overlapping genes

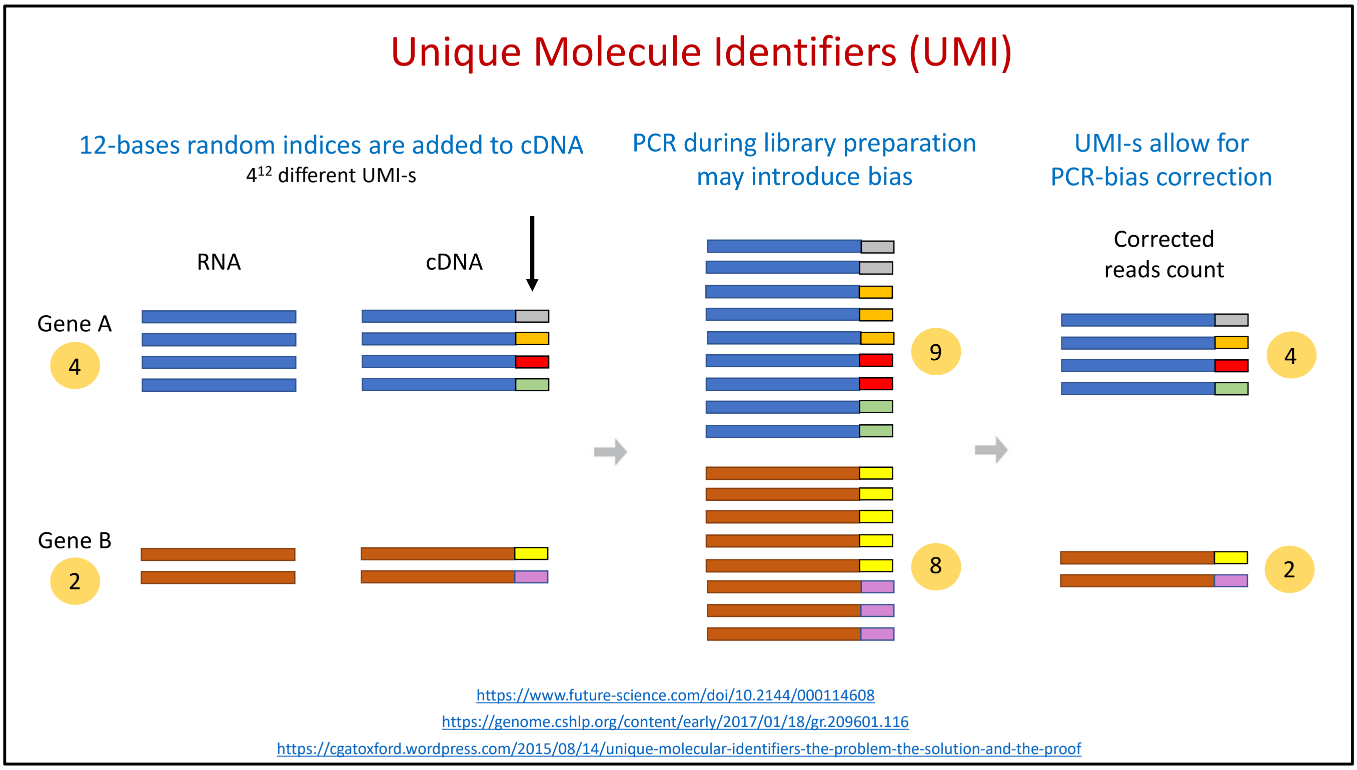
🟦 Read sequenced from positive strand (forward)

🟥 Read sequenced from negative strand (reverse)

https://doi.org/10.1371/journal.pcbi.1004393

Why is the strand information important?

Consider two overlapping genes located on the opposite strands. In the non-stranded library, you would not be able to distinguish reads belonging to such genes.



Unique Molecule Identifiers (UMI)

12-bases random indices are added to cDNA

4^{12} different UMI-s

PCR during library preparation may introduce bias

UMI-s allow for PCR-bias correction

https://www.future-science.com/doi/10.2144/000114608

https://genome.cshlp.org/content/early/2017/01/18/gr.209601.116

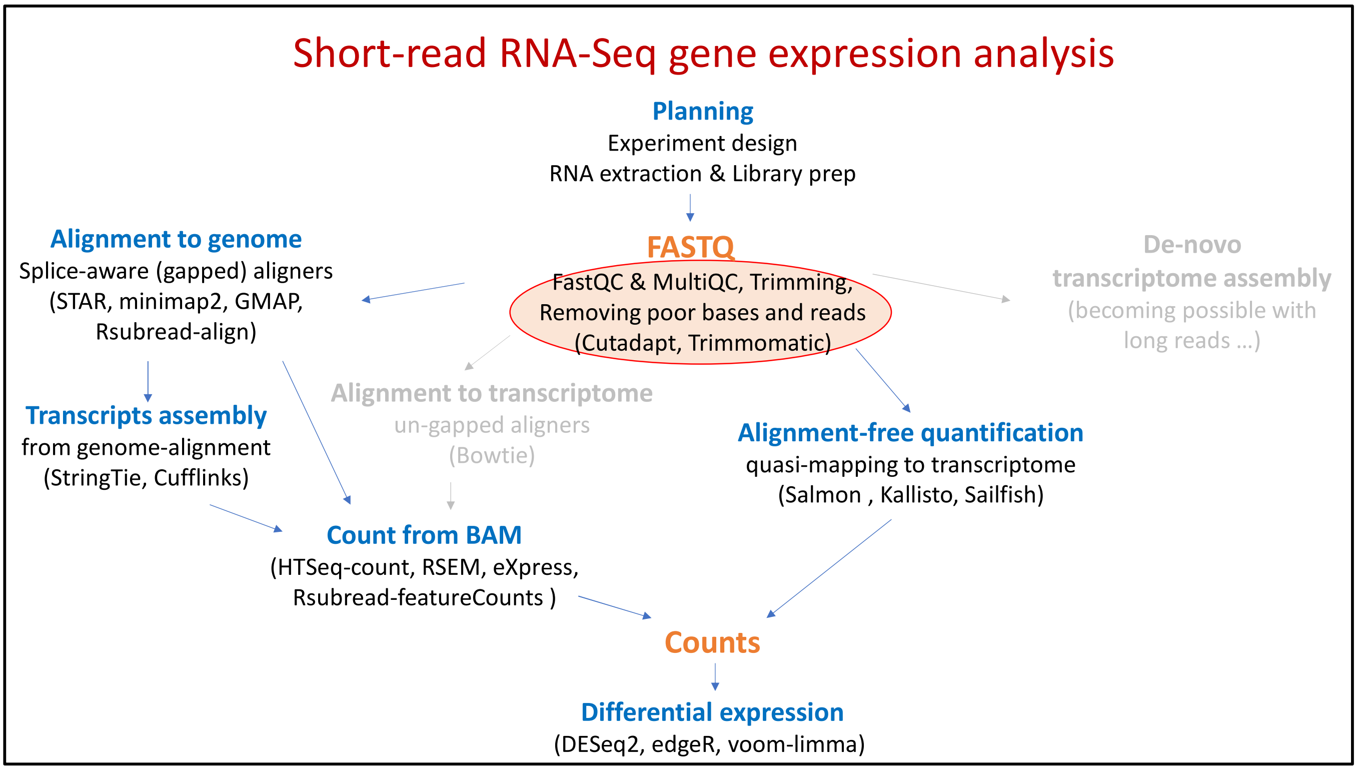
https://cgatoxford.wordpress.com/2015/08/14/unique-molecular-identifiers-the-problem-the-solution-and-the-proof

Always consider adding Unique Molecular Identifiers, if you use PCR during the library preparation.

Unique molecular identifiers (UMI) are just short random barcodes added to each sequenced fragment before the library preparation.

If a library preparation kits includes PCR, the amplification efficiency may vary between different fragments. Presence of the UMI-s allows to correct for the bias potentially introduced by PCR.

Typically, the number of PCR cycles is kept to minimum during the library prep, so the PCR bias is not very strong. However, using UMI is the only way to account for this bias in PCR-based ***RNA***-seq library prep. During ***DNA*** sequencing, the PCR duplicates are usually removed during the deduplication step. This step is not used in ***RNA***-seq data analysis because it may introduce more errors than it removes: https://www.nature.com/articles/srep25533).



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

Alignment-free quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

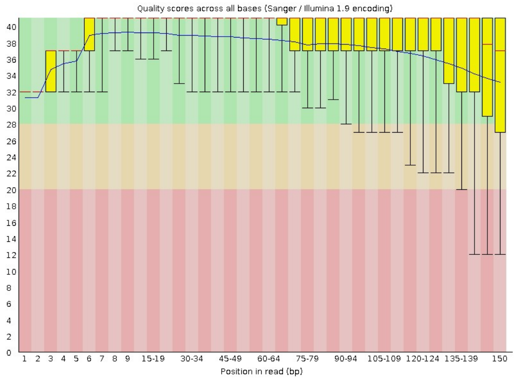
I am not going to discuss here how to use FastQC or how to trim the adaptors. I am sure you have already practiced (or will practice) FastQC during this course.

From FastQC to MultiQC

https://multiqc.info

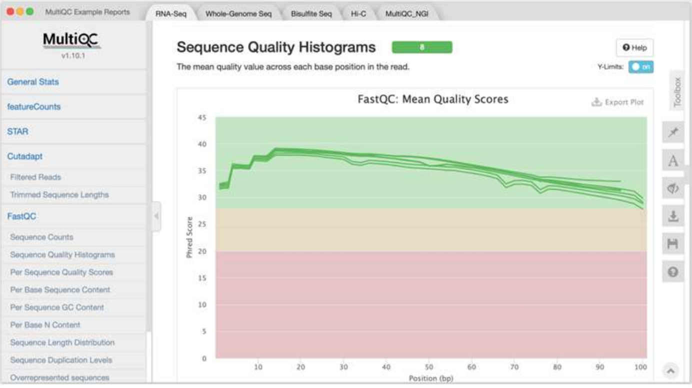
FastQC

outputs results per one sample



MultiQC

aggregates FastQC results from many samples



Sequence Quality Histograms 8

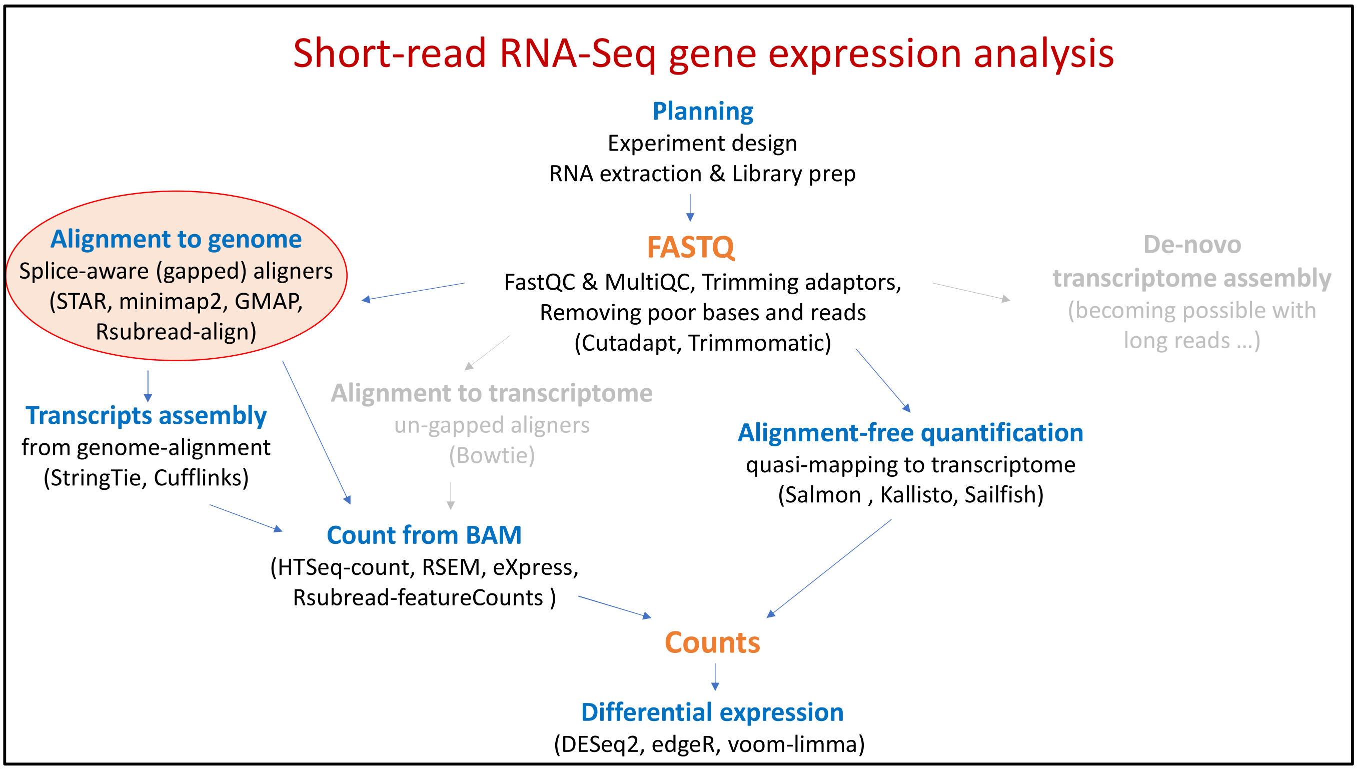
The mean quality value across each base position in the read.

FastQC: Mean Quality Scores

MultiQC can extract metrics from logs/outputs of many bioinformatics tools: FastQC, picard, flagstat, cutadapt, STAR …

Just don’t forget that after FastQC you may use MultiQC to compare QC metrics from multiple samples. Importantly, MultiQC may be used to summarize results of many other genomics QC tools too.

Of course, long reads may require different tools for QC and pre-processing (we will discuss it in a different session).



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

Alignment-free quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

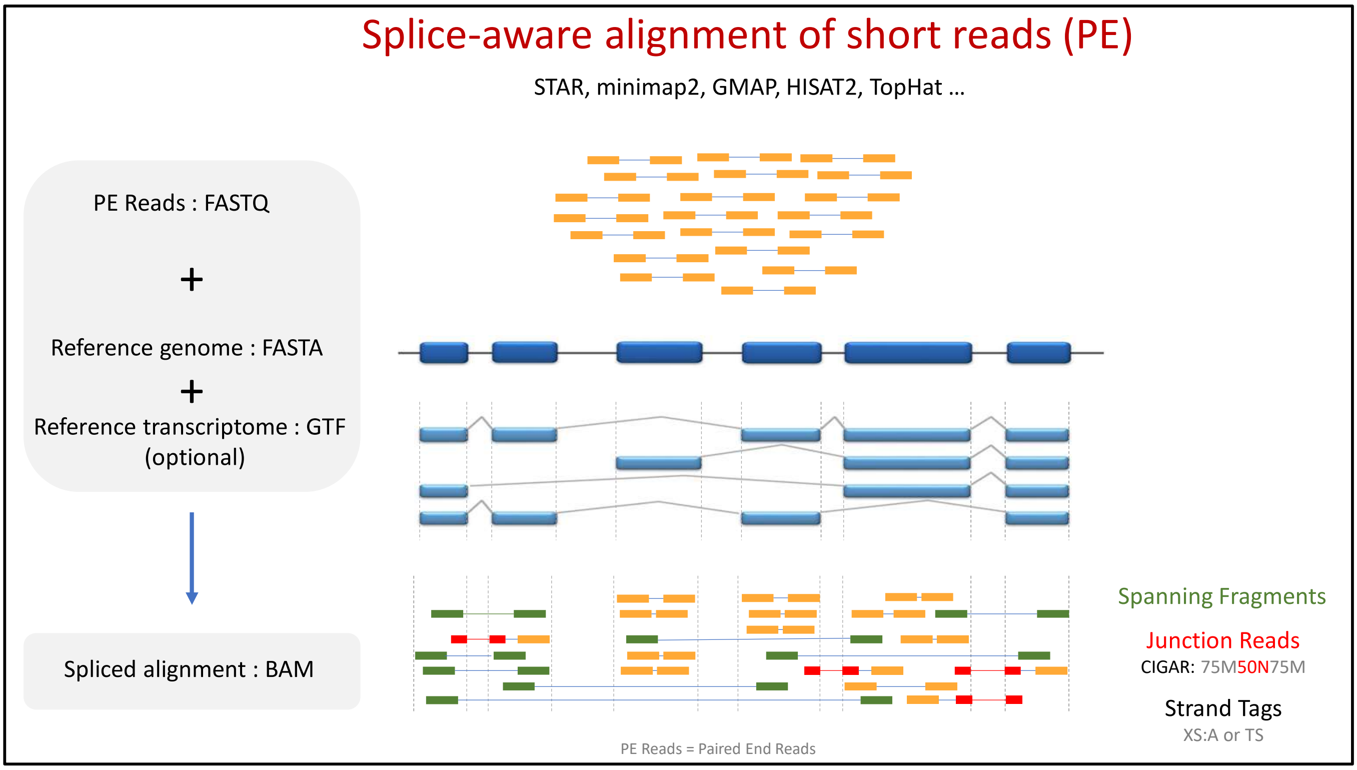
Counts

Differential expression
(DESeq2, edgeR, voom-limma)

At last, we are going to discuss the alignment!

Importantly, here we consider the alignment of RNA to a reference ***genome***. So, the aligner should be able to recognize gaps in RNA-seq reads, which are introduced by splicing.

There are multiple well-established splice-aware aligners that can be used at this step. You will see a similar pattern at each step of RNA-seq analysis, when multiple tools are available for each task (and I mean multiple good reputable tools). I can not give you a “secret list” of what tools are better. All I can do is to name some different tools and to explain my personal choices.



Splice-aware alignment of short reads (PE)

STAR, minimap2, GMAP, HISAT2, TopHat …

PE Reads : FASTQ

+

Reference genome : FASTA

+

Reference transcriptome : GTF (optional)

Spliced alignment : BAM

Spanning Fragments

Junction Reads

CIGAR: 75M50N75M

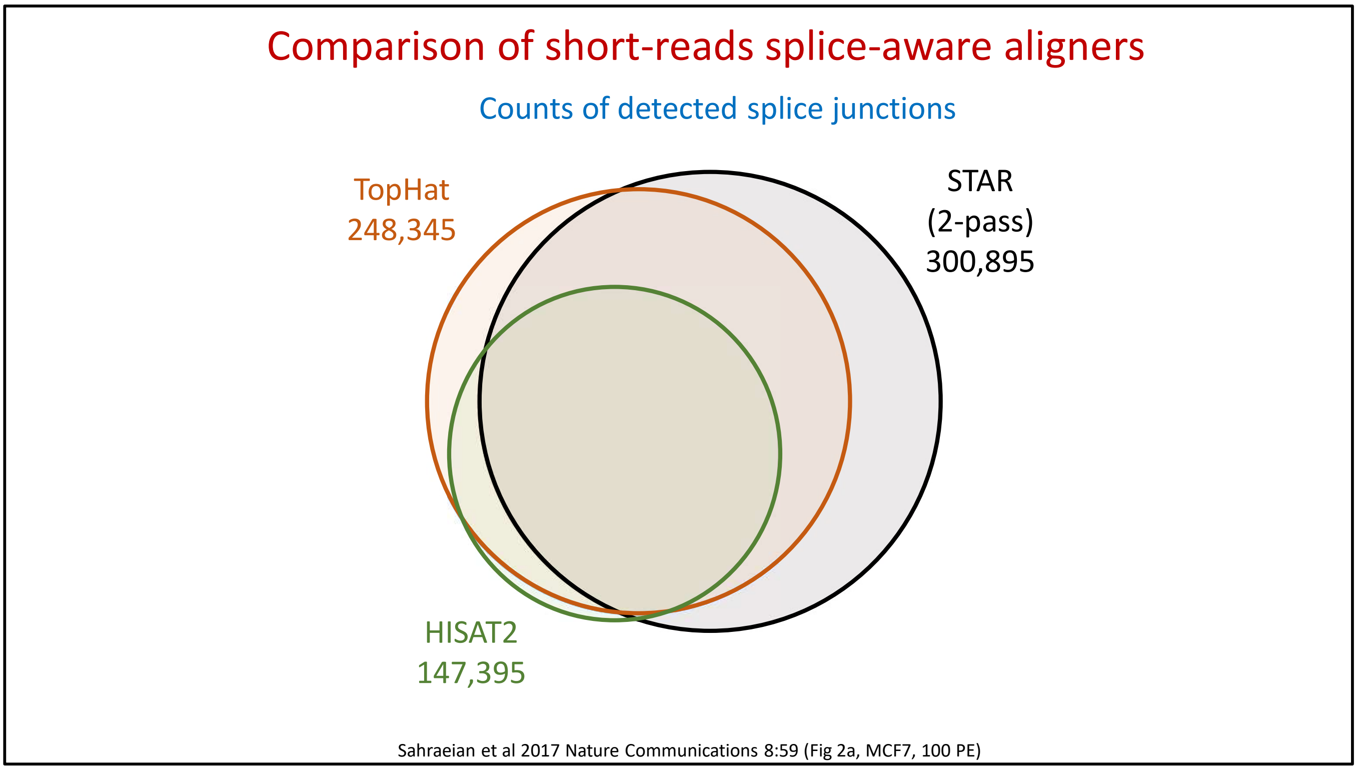
Strand Tags

XS:A or TS

PE Reads = Paired End Reads

Like the usual non-gapped aligners, the splice-aware aligners take the reads in FASTQ format and map them to the known sequence of the reference genome. Also, the splice-aware aligners may use an additional GTF file with coordinates of known transcripts to help in finding already known splice junctions in the analyzed sequencing data.

The RNA-seq alignment is written into a BAM file, which has some RNA-specific features (in addition to the standard features available in DNA-seq BAM files):
1) the “N” letters are added to the CIGAR string, in the places of introns, spliced out from RNA
2) additional tags may be added for the strand information



Comparison of short-reads splice-aware aligners

Counts of detected splice junctions

TopHat
248,345

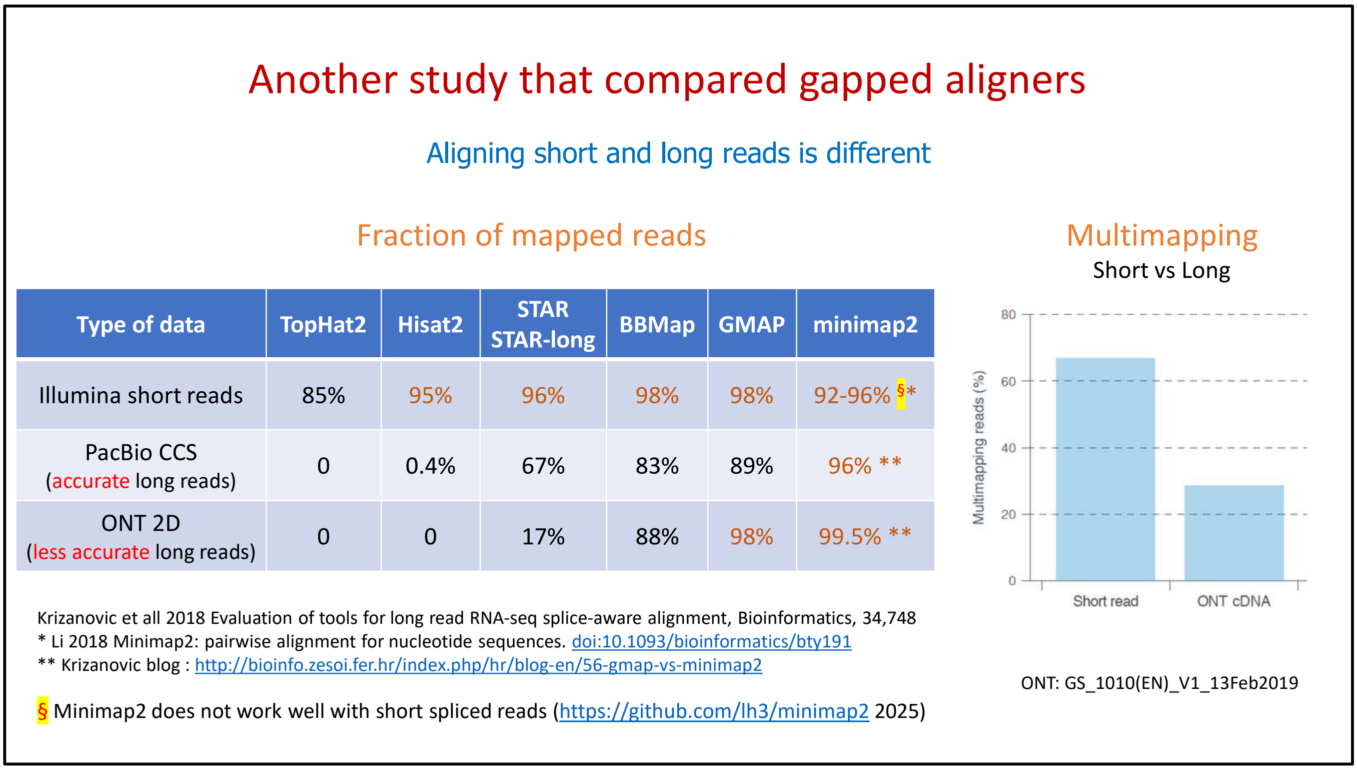
STAR
(2-pass)
300,895

HISAT2
147,395

Sahraeian et al 2017 Nature Communications 8:59 (Fig 2a, MCF7, 100 PE)

Having many tools at each step, of course, there are many studies that compare the tools.

This is a comparison between 3 gapped aligners, showing ***the number of splice junctions*** they could detected in some test dataset. It seems that TopHat and STAR detected more splice junctions than HISAT2 in this study. From this plot, we can not say which aligner is better: if you prefer a more conservative approach HISAT2 is better. If you prefer higher sensitivity, STAR may look better.



Another study that compared gapped aligners

Aligning short and long reads is different

Fraction of mapped reads

Type of data	TopHat2	Hisat2	STAR STAR-long	BBMap	GMAP	minimap2
Illumina short reads	85%	95%	96%	98%	98%	92-96% §*
PacBio CCS (accurate long reads)	0	0.4%	67%	83%	89%	96% **
ONT 2D (less accurate long reads)	0	0	17%	88%	98%	99.5% **

Krizanovic et all 2018 Evaluation of tools for long read RNA-seq splice-aware alignment, Bioinformatics, 34,748
* Li 2018 Minimap2: pairwise alignment for nucleotide sequences. doi:10.1093/bioinformatics/bty191
** Krizanovic blog : http://bioinfo.zesoi.fer.hr/index.php/hr/blog-en/56-gmap-vs-minimap2

§ Minimap2 does not work well with short spliced reads (https://github.com/lh3/minimap2 2025)

Multimapping

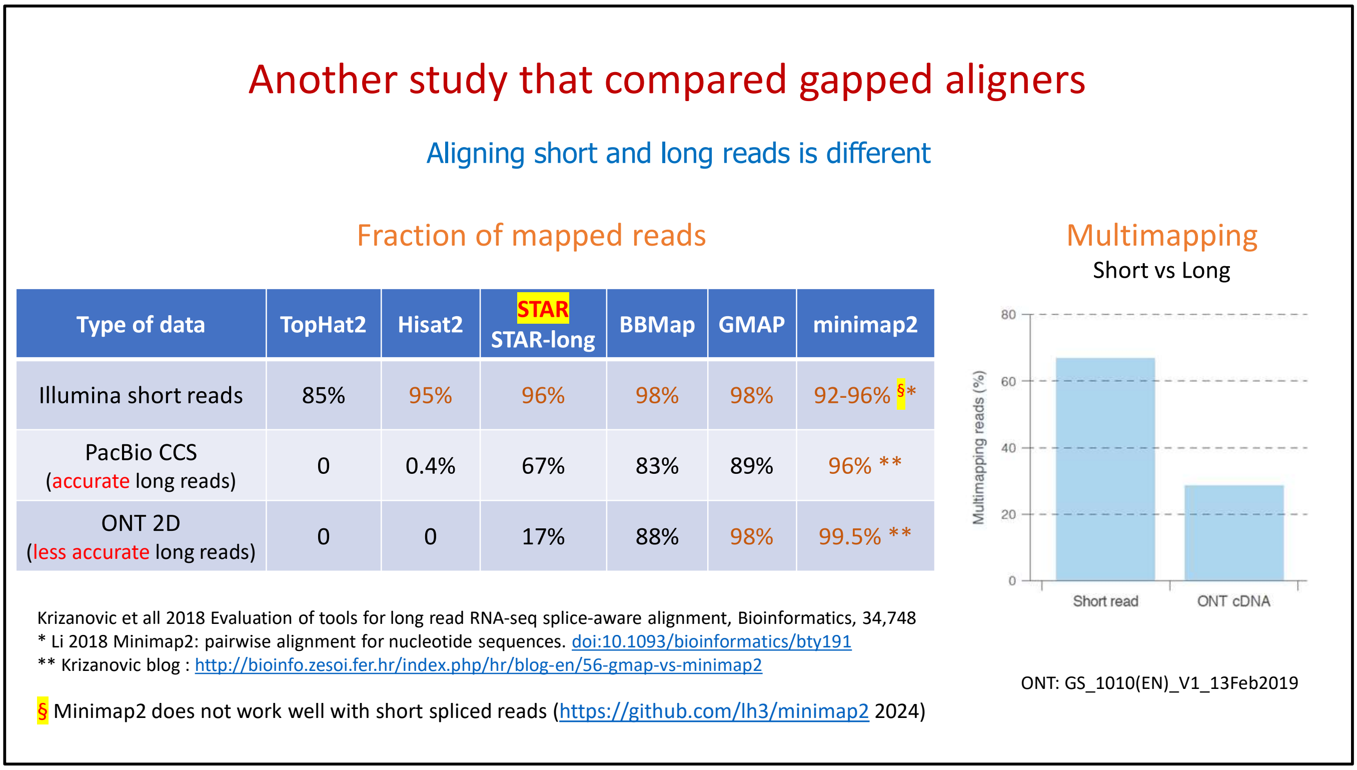
Short vs Long

ONT: GS_1010(EN)_V1_13Feb2019

This is another study that compared gapped aligners by ***the fraction of mapped reads***.

In this comparison, most of aligners showed a reasonably similar performance on the short reads except for TopHat2, which was clearly inferior to other aligners.

Also, you can see on this slide that multimapping is a real problem in the short-reads alignment, and how the long-reads sequencing solves this problem.



Another study that compared gapped aligners

Aligning short and long reads is different

Fraction of mapped reads

Type of data	TopHat2	Hisat2	STAR STAR-long	BBMap	GMAP	minimap2
Illumina short reads	85%	95%	96%	98%	98%	92-96% §*
PacBio CCS (accurate long reads)	0	0.4%	67%	83%	89%	96% **
ONT 2D (less accurate long reads)	0	0	17%	88%	98%	99.5% **

Krizanovic et all 2018 Evaluation of tools for long read RNA-seq splice-aware alignment, Bioinformatics, 34,748
* Li 2018 Minimap2: pairwise alignment for nucleotide sequences. doi:10.1093/bioinformatics/bty191
** Krizanovic blog : http://bioinfo.zesoi.fer.hr/index.php/hr/blog-en/56-gmap-vs-minimap2

§ Minimap2 does not work well with short spliced reads (https://github.com/lh3/minimap2 2024)

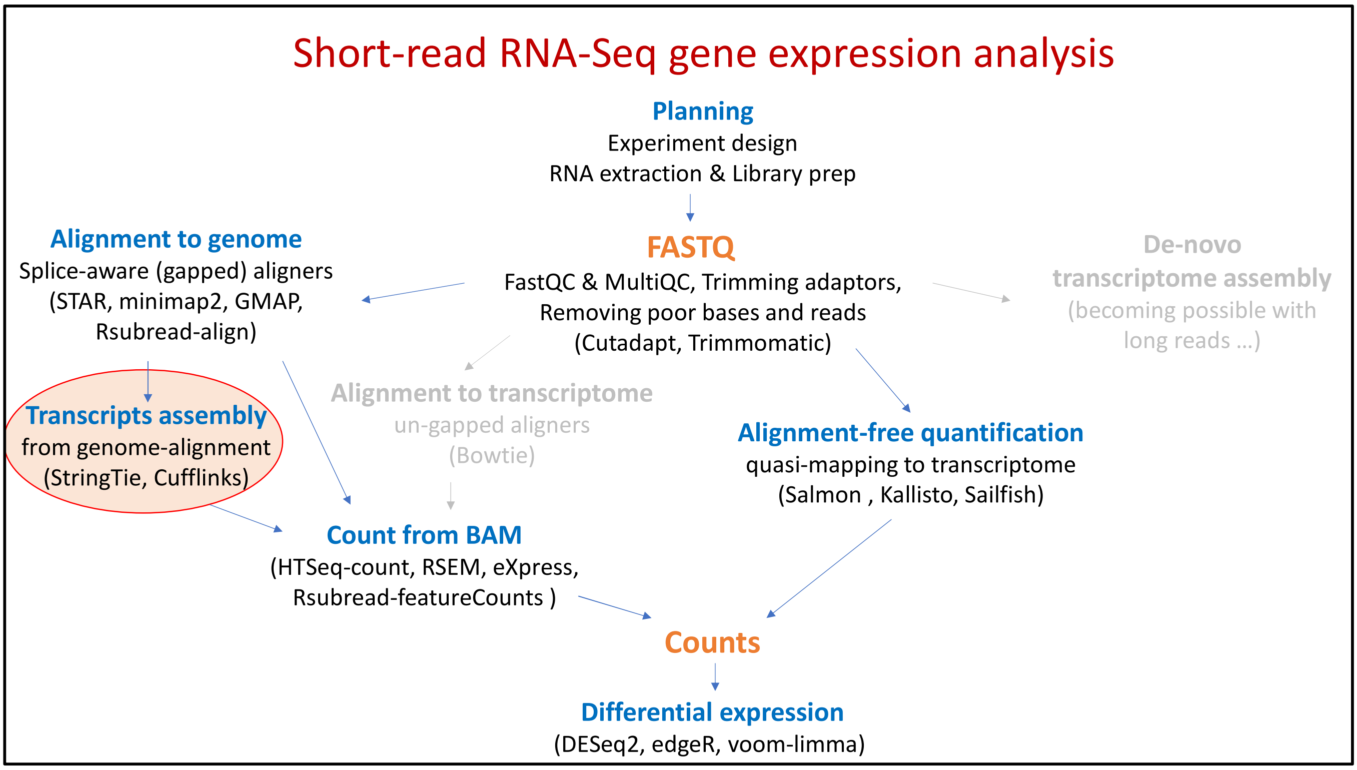
Multimapping

Short vs Long

ONT: GS_1010(EN)_V1_13Feb2019

Considering that STAR is coming from a very reputable team, with active updates, with a good support, with a large community of users,
at the moment, STAR is my personal preference for short reads RNAseq alignment.

Despite some studies reported use of minimap2 for Illumina RNAseq data, the minimap2 GitHub page does not advise using it for short-reads gapped reads (in 2025). However, minimap2 is the current aligner of choice for the long-reads RNAseq data (and it is very popular for non-gapped alignment).



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with long reads ...)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

Alignment-free quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

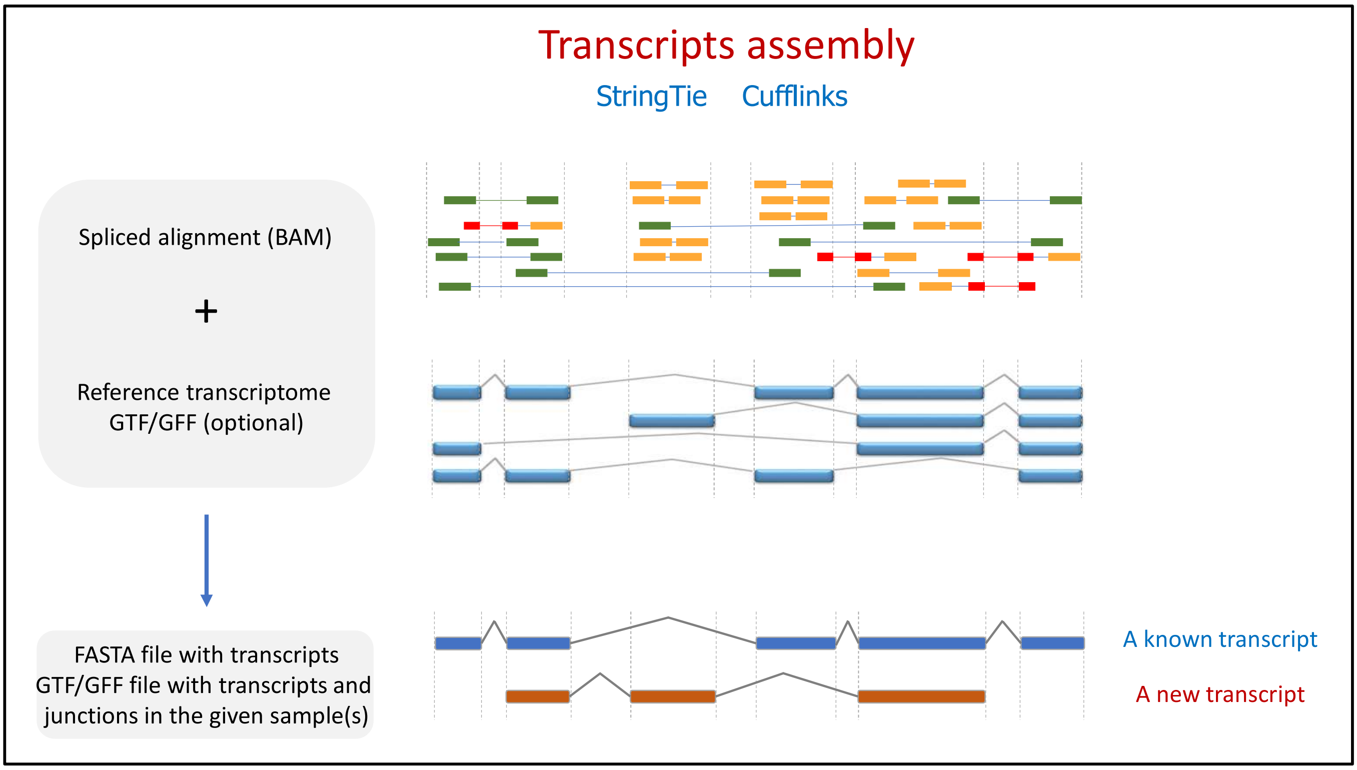
Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

Now let’s consider the Transcriptome Assembly.

It aspires to infer transcripts from the short-read RNA-seq data aligned to the reference genome ...



Transcripts assembly

StringTie Cufflinks

Spliced alignment (BAM)

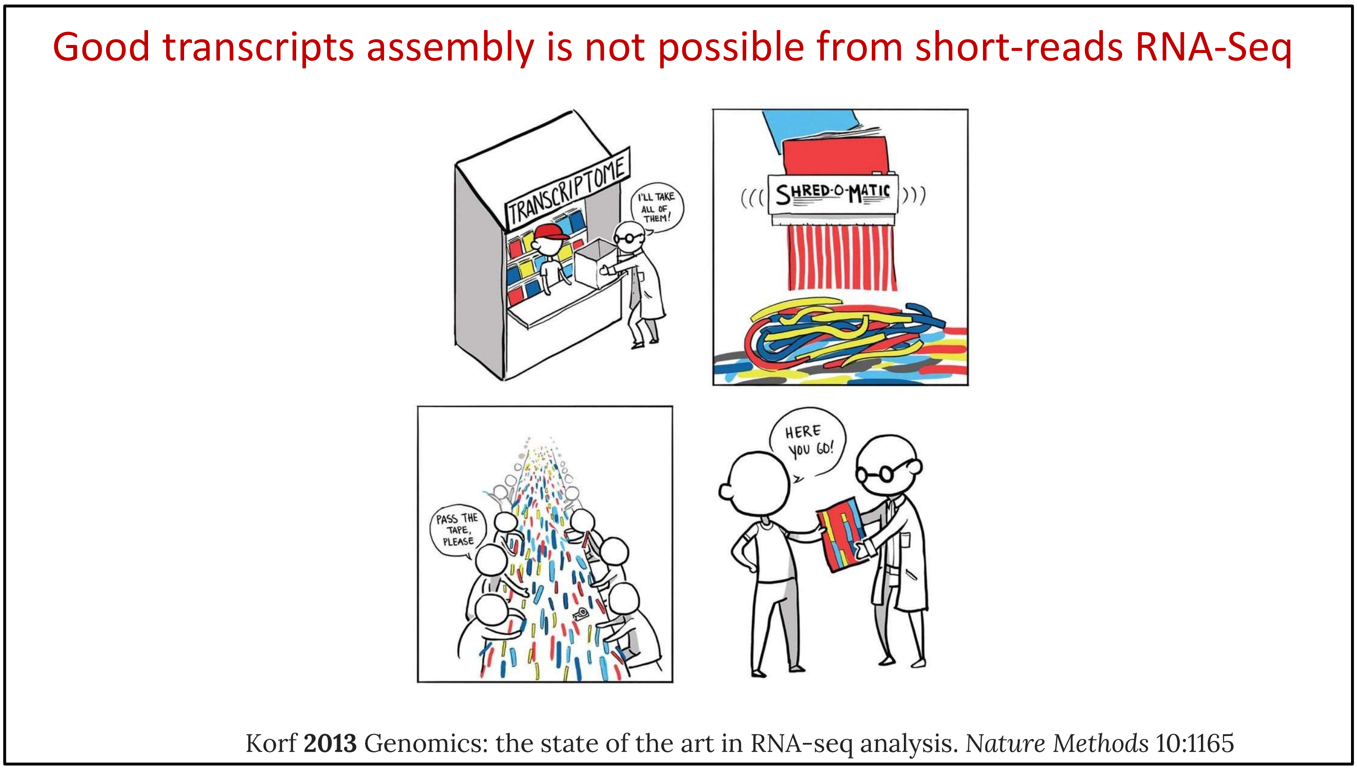
+

Reference transcriptome GTF/GFF (optional)

FASTA file with transcripts
GTF/GFF file with transcripts and junctions in the given sample(s)

The input files include the aligned BAM with an optional addition of GTF/GFF with coordinates of already known transcripts.

The outputs could be FASTA and/or GTF/GFF files with the transcripts detected in the sample(s), possibly including the new transcripts, not known previously in the reference transcriptome.



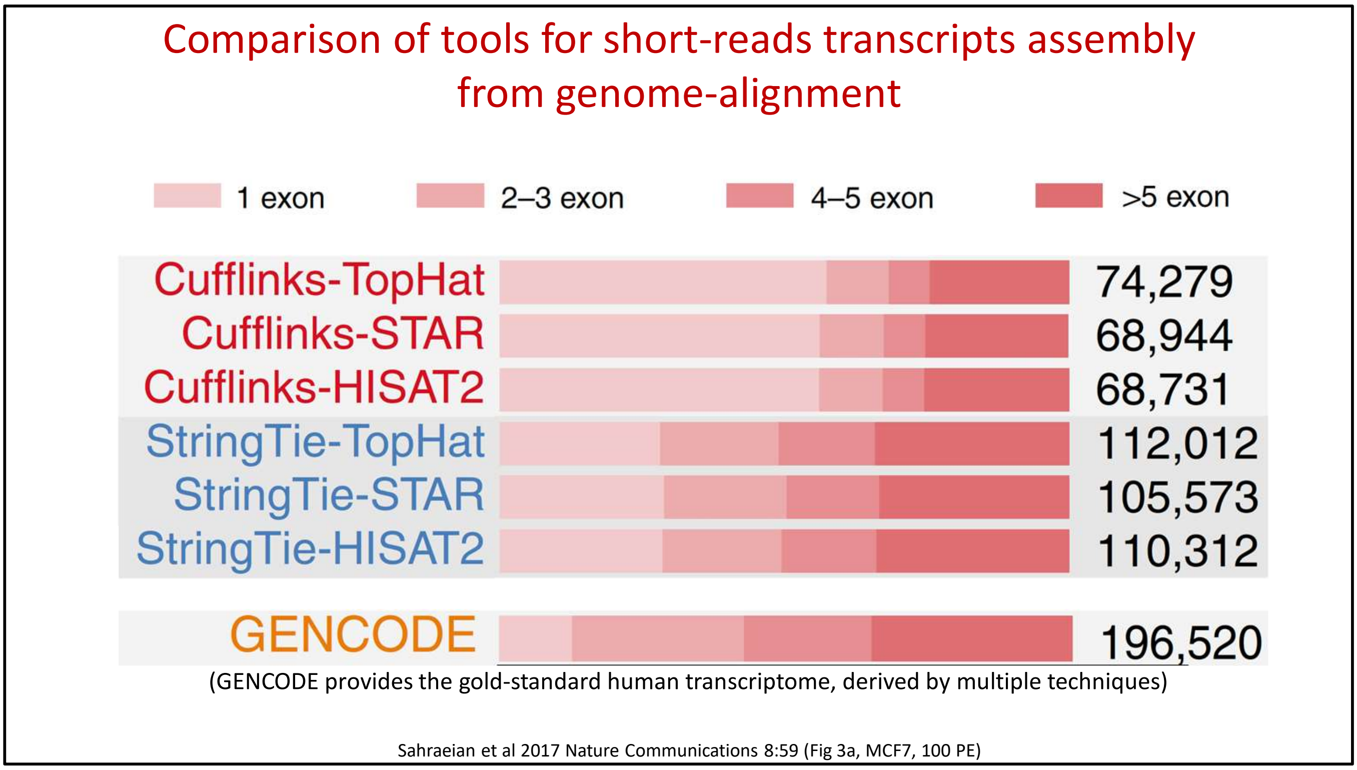
Good transcripts assembly is not possible from short-reads RNA-Seq

Korf **2013** Genomics: the state of the art in RNA-seq analysis. *Nature Methods* 10:1165

Again, there are multiple tools for the transcriptome assembly, and many comparisons between the different tools…

This slide shows a good summary of all these comparisons.

In short: it is not possible to make a good transcriptome assembly from the short-reads RNAseq data, whatever tool you use.



Comparison of tools for short-reads transcripts assembly from genome-alignment

(GENCODE provides the gold-standard human transcriptome, derived by multiple techniques)

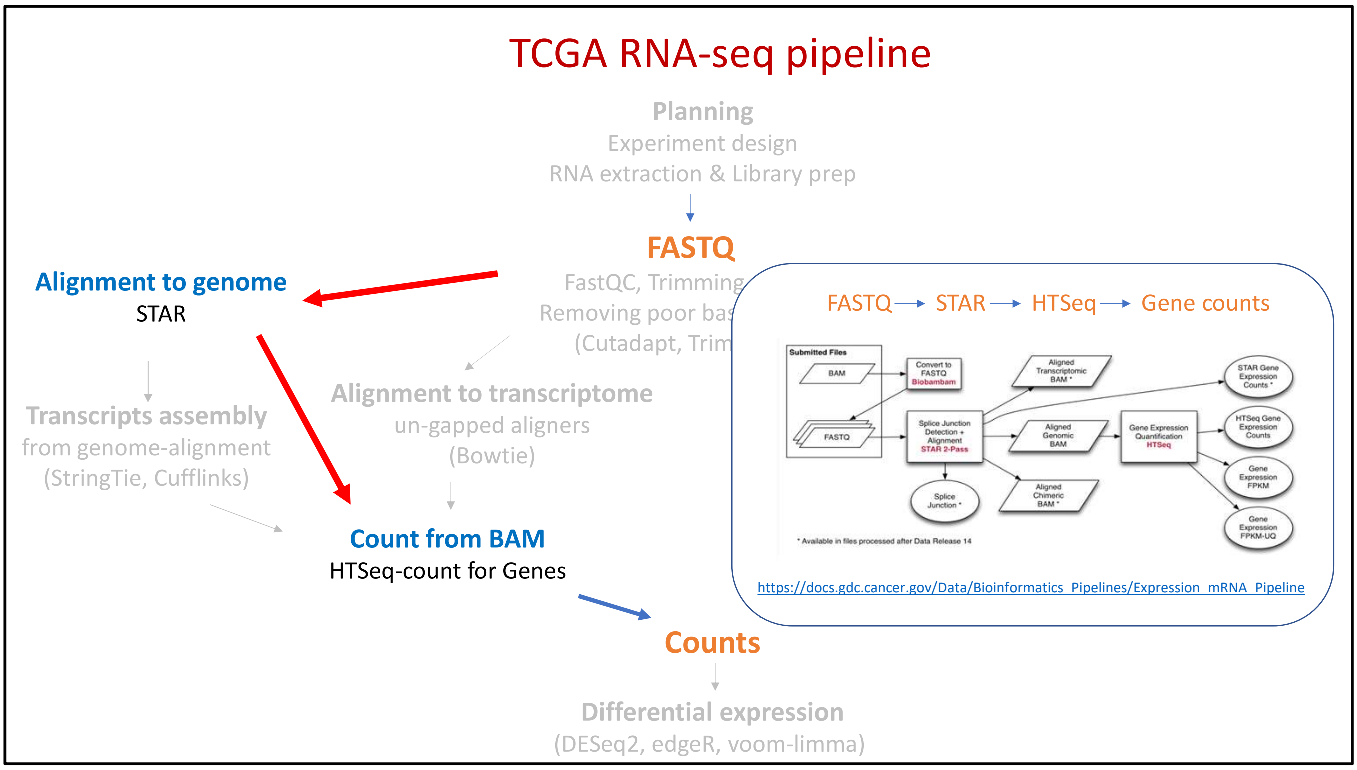
Sahraeian et al 2017 Nature Communications 8:59 (Fig 3a, MCF7, 100 PE)

If you prefer a more specific comparison, then you can see that even the best transcript assembly tools (e.g. Cufflings and StringTie) are quite discordant when using short reads.

And, in this study, none of these tools produced a result similar to the expected transcriptome from GenCode.

GenCode is a consortium that synthesises multiple types of data (long and short RNAseq, optical mapping etc) to provide the best currently possible transcriptome assemblies for human and mise.

Of course, transcriptomes were derived from the short reads before the long reads became available. However, nowadays transcriptome assembly and isoforms analysis should be done with the long reads.



TCGA RNA-seq pipeline

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC, Trimming
Removing poor bas
(Cutadapt, Trim

Alignment to genome
STAR

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

Count from BAM
HTSeq-count for Genes

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

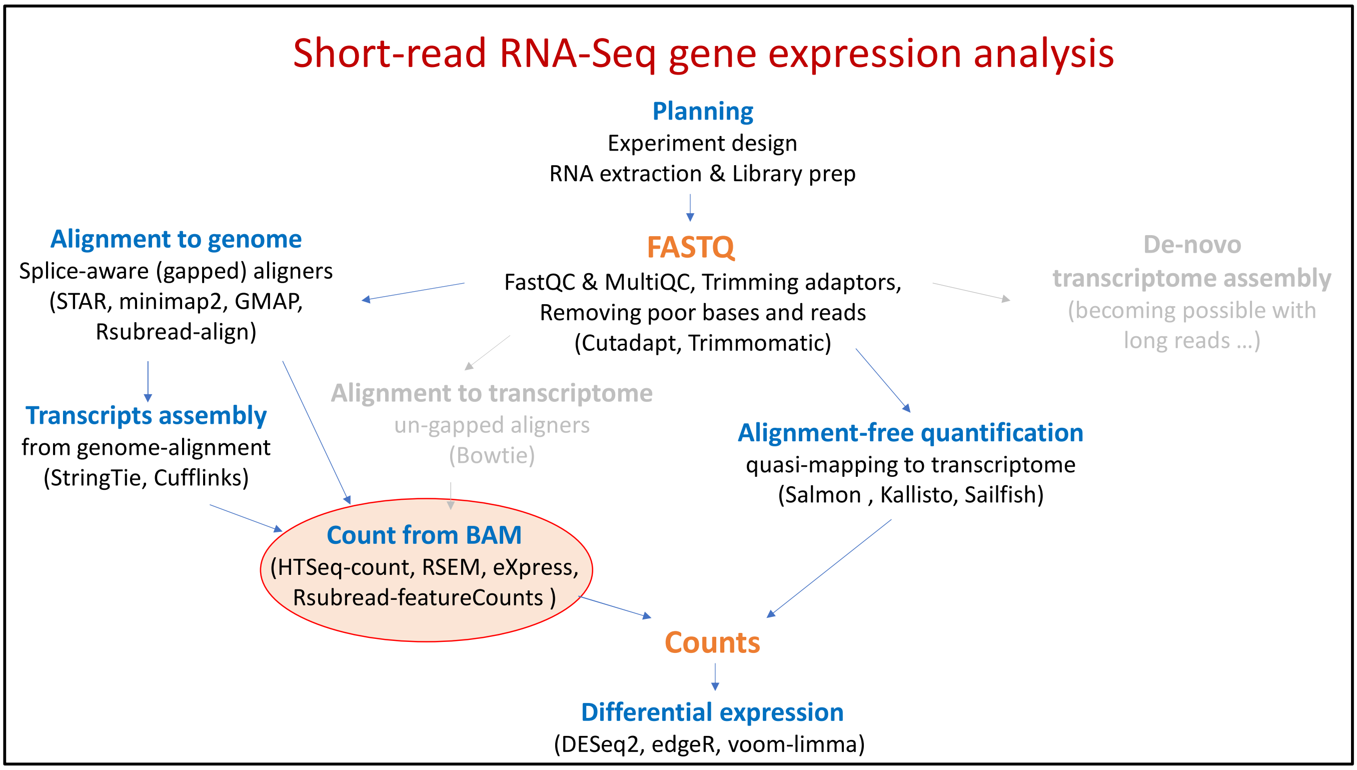
FASTQ → STAR → HTSeq → Gene counts

* Available in files processed after Data Release 14

https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline

So, it is not a surprise, that, for instance, TCGA (a very large international consortium that sequenced thousands of tumors) decided to count against known GenCode genes, rather than to experiment with the in-house transcript assembly.

Of course, this strategy is not applicable for most RNAseq studies in agri-food or ecology, where many species still don’t have such good transcriptomes available.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

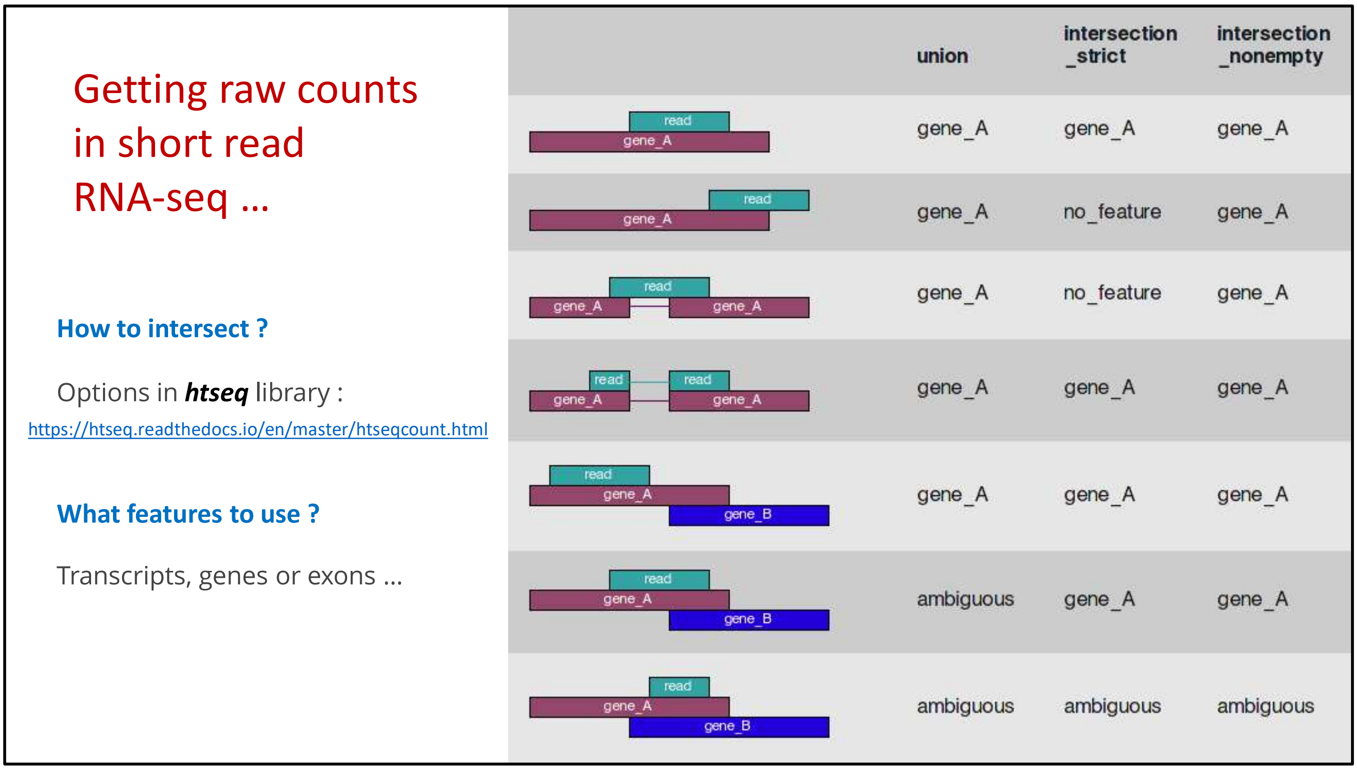
Alignment-free quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

You might think that counting reads overlapping the known genes should be a pretty straight forward task …



Getting raw counts in short read RNA-seq …

How to intersect ?

Options in ***htseq*** library :
https://htseq.readthedocs.io/en/master/htseqcount.html

What features to use ?

Transcripts, genes or exons …

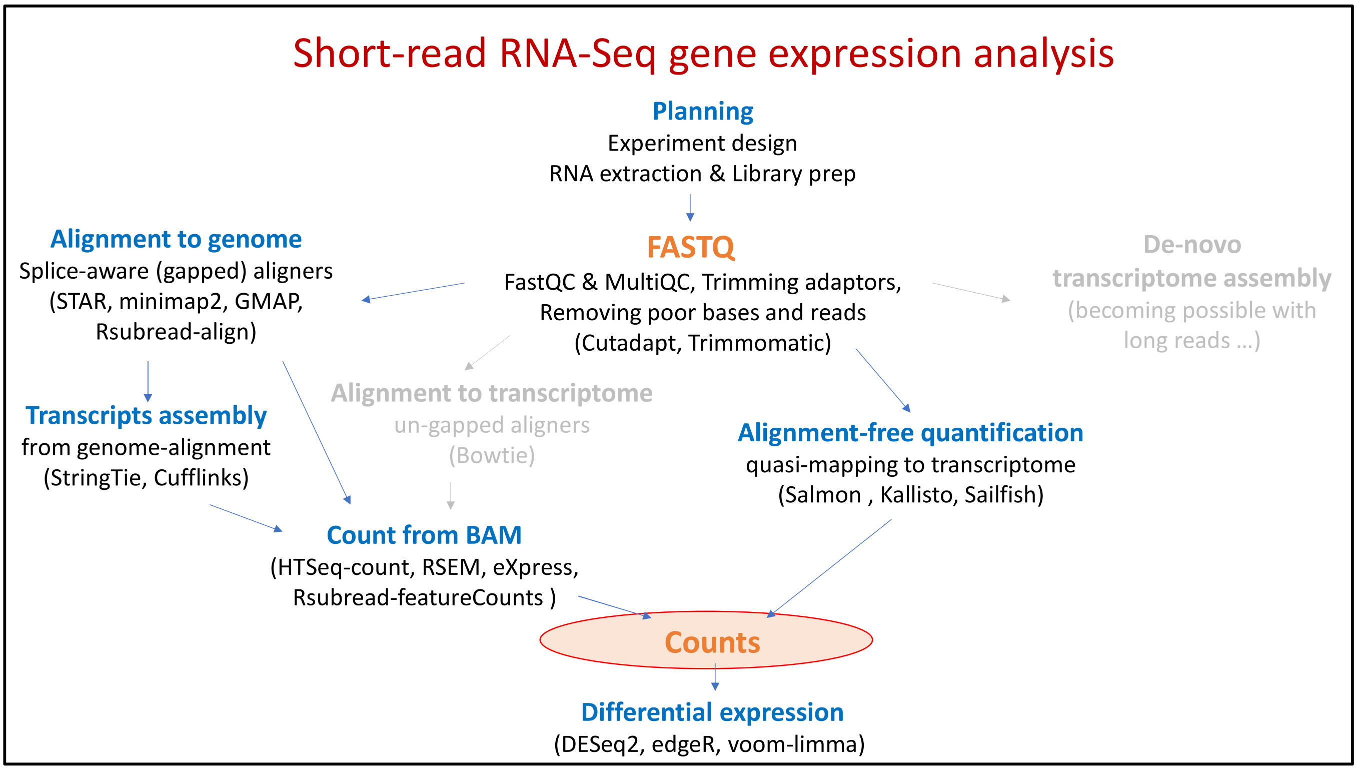
	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

But we still have multiple tools and should make many choices!

For instance:
- How to count the read if it overlaps the gene just partially?
- How to count reads that overlap two genes ?

Another question is what features to choose: genes, transcripts or exons?

Etc



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

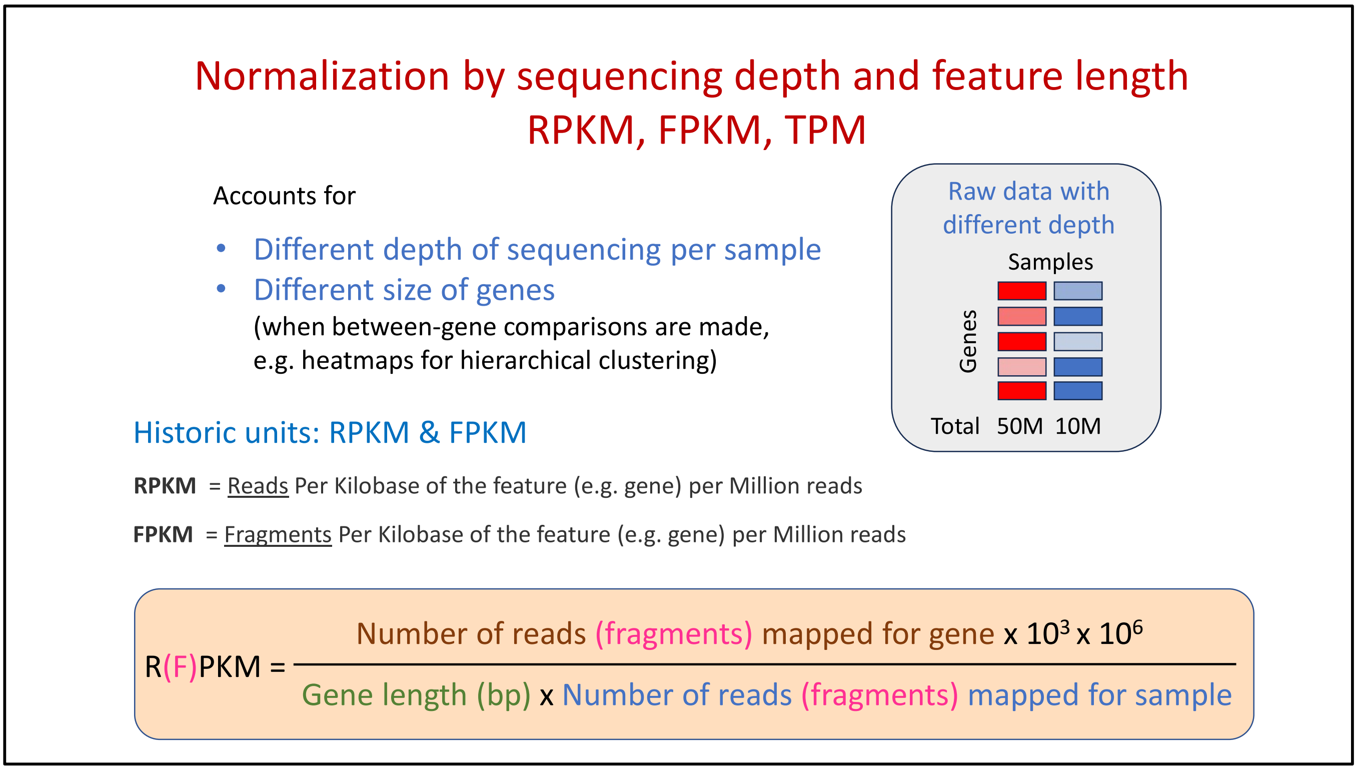
Alignment-free quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

Finally, when we have the counts, they could be represented in different ways.



Normalization by sequencing depth and feature length RPKM, FPKM, TPM

Accounts for

- Different depth of sequencing per sample
- Different size of genes
  (when between-gene comparisons are made, e.g. heatmaps for hierarchical clustering)

Raw data with different depth
Samples
Genes
Total 50M 10M

Historic units: RPKM & FPKM

RPKM = Reads Per Kilobase of the feature (e.g. gene) per Million reads

FPKM = Fragments Per Kilobase of the feature (e.g. gene) per Million reads

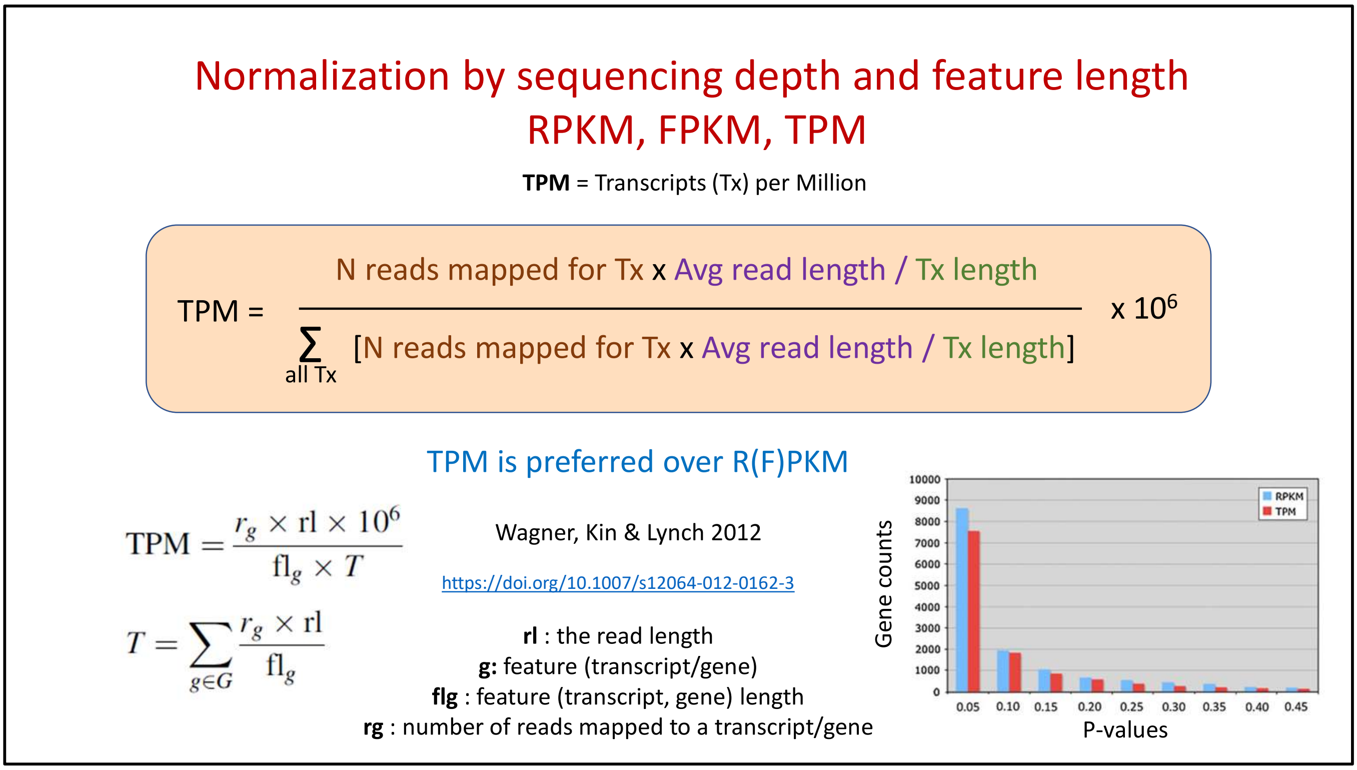
$$\text{R(F)PKM} = \frac{\text{Number of reads (fragments) mapped for gene} \times 10^3 \times 10^6}{\text{Gene length (bp)} \times \text{Number of reads (fragments) mapped for sample}}$$

The raw counts cannot be used for direct comparison between samples and genes.

To compare counts between samples and genes, the raw counts should be normalized by the library and gene sizes.

Two initial historic methods of normalizing and reporting the counts were called RPKM or FPKM, which stands for Reads (or Fragments) per Kilobase per Million. Meaning per kilobase of gene length and per million reads of the library depth.

Although prone to some biases, these units were used in many very reputable studies, which reported results well validated later by orthogonal methods.



Normalization by sequencing depth and feature length RPKM, FPKM, TPM

TPM = Transcripts (Tx) per Million

$$\text{TPM} = \frac{\text{N reads mapped for Tx} \times \text{Avg read length} / \text{Tx length}}{\sum_{\text{all Tx}} [\text{N reads mapped for Tx} \times \text{Avg read length} / \text{Tx length}]} \times 10^{6}$$

TPM is preferred over R(F)PKM

$$\text{TPM} = \frac{r_g \times \text{rl} \times 10^6}{\text{fl}_g \times T}$$

$$T = \sum_{g \in G} \frac{r_g \times \text{rl}}{\text{fl}_g}$$

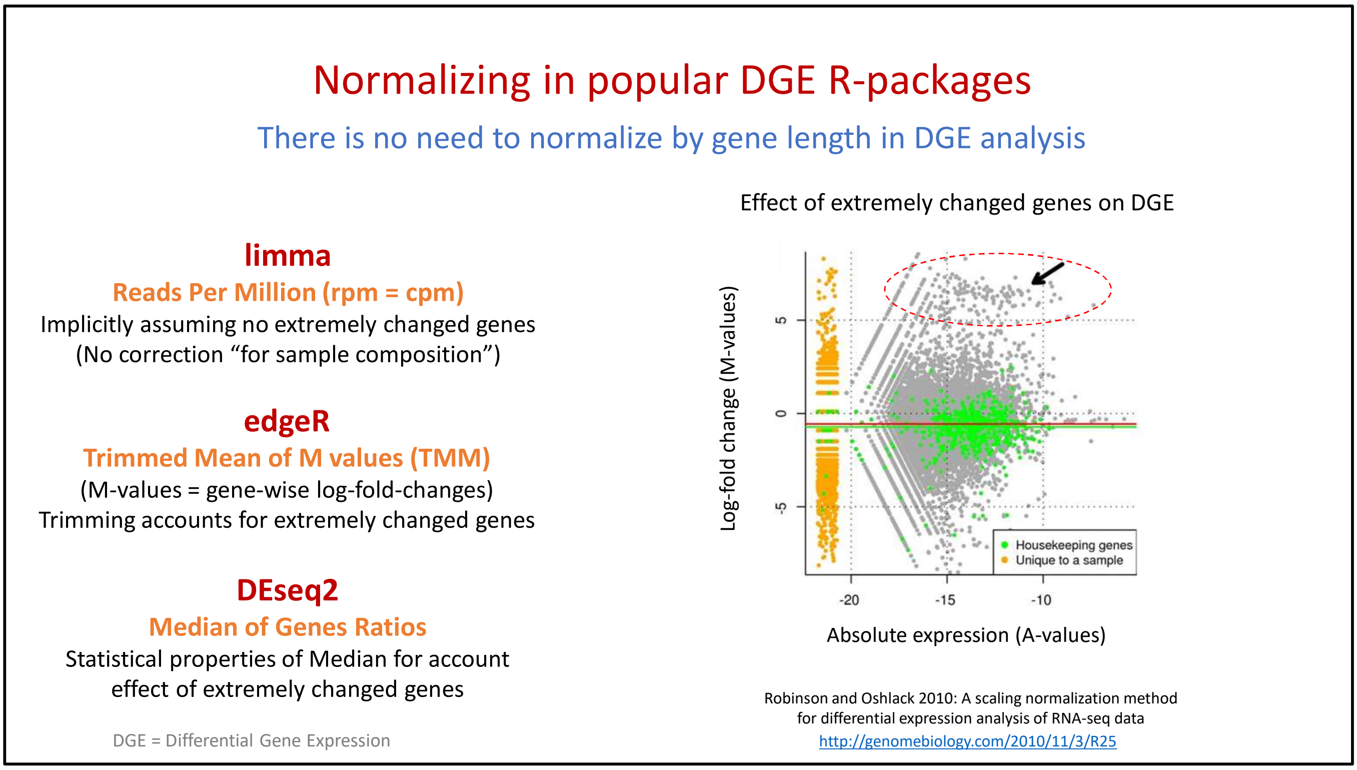
Wagner, Kin & Lynch 2012

https://doi.org/10.1007/s12064-012-0162-3

rl : the read length
g: feature (transcript/gene)
flg : feature (transcript, gene) length
rg : number of reads mapped to a transcript/gene

One of the criticisms is that R(F)PKM may slightly inflate the significance when used for differential gene expression analysis. In essence, this bias is caused by a possible differential transcript use, leading to variation in the transcripts' length between samples.

So, the currently preferred unit to report ***normalized*** counts is TPM (Transcripts per Million). However, the above bias attributed to R(F)PKM is not very strong (definitely, it is not as strong as the wording used by Wagner *et al* in their paper :)



Normalizing in popular DGE R-packages

There is no need to normalize by gene length in DGE analysis

limma

Reads Per Million (rpm = cpm)

Implicitly assuming no extremely changed genes
(No correction “for sample composition”)

edgeR

Trimmed Mean of M values (TMM)

(M-values = gene-wise log-fold-changes)
Trimming accounts for extremely changed genes

DEseq2

Median of Genes Ratios

Statistical properties of Median for account
effect of extremely changed genes

DGE = Differential Gene Expression

Effect of extremely changed genes on DGE

Robinson and Oshlack 2010: A scaling normalization method
for differential expression analysis of RNA-seq data
http://genomebiology.com/2010/11/3/R25

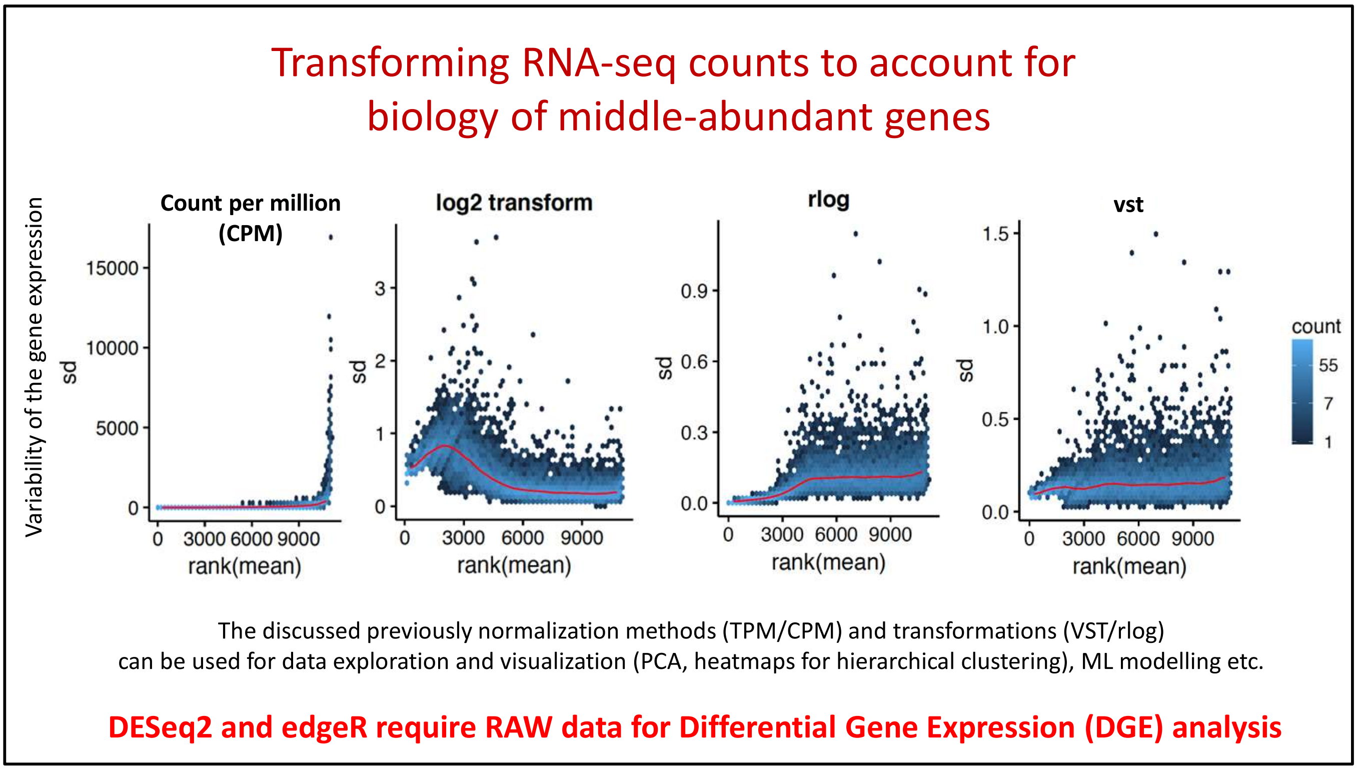
Importantly, the most popular R libraries for Differential Gene Expression analysis (limma, DESeq2 and edgeR) use their own customized normalizing procedures. So, they require the RAW data for the analysis.

Because differential expression analysis does not require comparison between genes, the DGE normalization methods don’t normalize by the gene length.

Limma adopts the most simplistic approach and just normalises per library depth, using Counts Per Million (CPM).

DESeq2 and edgeR developers noted that if a sample has a small number of very highly expressed genes (e.g. liver abundantly producing albumins) it may appear like a relative decrease in expression of all other genes. This may lead to a bias slightly inflating the number of differentially expressed genes.

To avoid this bias, DESeq2 and edgeR normalise not by the total count, but by a sort of median (median negates effect of gross outliers) or by some trimmed mean (mean after removal of extremely expressed genes). This “median ratio” and “trimmed mean” are also called “normalization factors” in these R libraries vocabulary.



Transforming RNA-seq counts to account for biology of middle-abundant genes

The discussed previously normalization methods (TPM/CPM) and transformations (VST/rlog)
can be used for data exploration and visualization (PCA, heatmaps for hierarchical clustering), ML modelling etc.

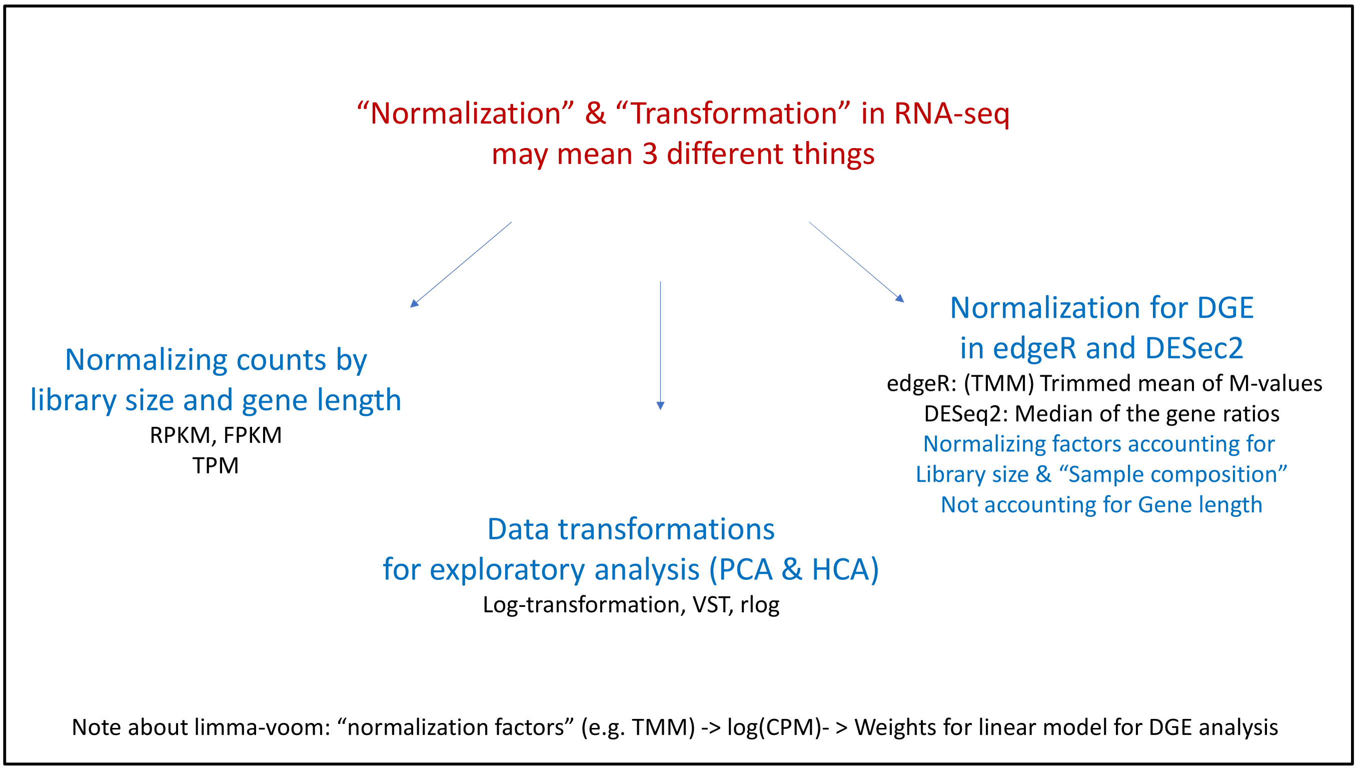
DESeq2 and edgeR require RAW data for Differential Gene Expression (DGE) analysis

In addition to the counts' normalization, some additional transformations are often used for exploratory analysis, such as PCA or HCA.

PCA (Principal Component Analysis) or HCA (Hierarchical Clustering Analysis) are used to visualize groups of similar samples, if any. To detect how similar the samples are to each other, we calculate the distances between samples based on the differences in gene expressions.

The left plot on this slide, shows the counts already normalised by depth of sequencing (CPM, as used in limma). You can see that the largest variation is observed in the most highly abundant genes. So, if we use such non-transformed data for PCA or HCA, the samples similarity would be mainly based on the expressions of these highly abundant genes.

To account for expressions of the less-abundant genes the normalised counts could be log-transformed (the second left plot on the slide). However, the simple log-transformation may put an undue weight to the least-abundant genes. To avoid this, rlog and VST (Variance-Stabilising Transformation) are used, as shown in the two right plots on the slide. DESeq2 provides in-build functions for applying VST or rlog to the TPM normalised abundancies: you will try these functions during the practical session.



“Normalization” & “Transformation” in RNA-seq
may mean 3 different things

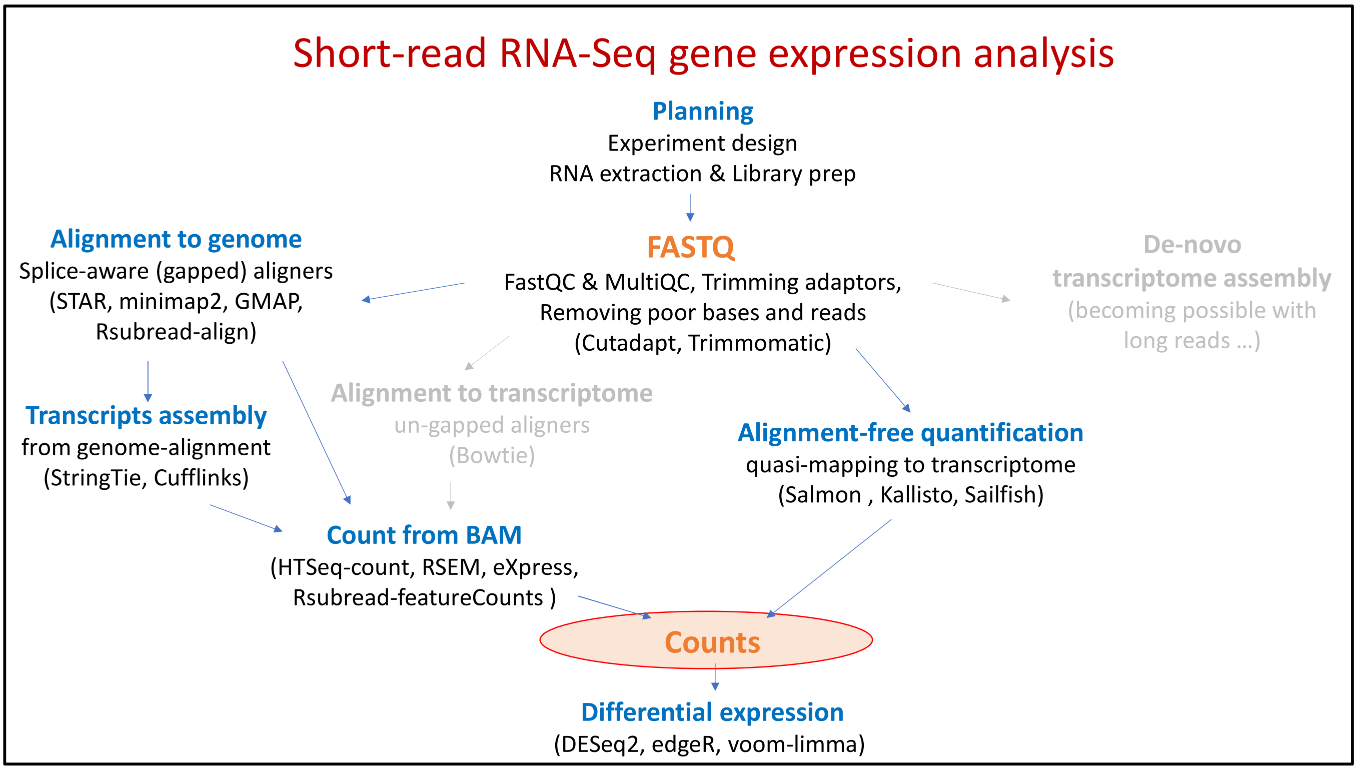
Normalizing counts by
library size and gene length
RPKM, FPKM
TPM

Data transformations
for exploratory analysis (PCA & HCA)
Log-transformation, VST, rlog

Normalization for DGE
in edgeR and DESec2
edgeR: (TMM) Trimmed mean of M-values
DESeq2: Median of the gene ratios
Normalizing factors accounting for
Library size & “Sample composition”
Not accounting for Gene length

Note about limma-voom: “normalization factors” (e.g. TMM) -> log(CPM)- > Weights for linear model for DGE analysis

This is a simplified summary for the normalization and transformation methods used in RNAseq data analysis.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

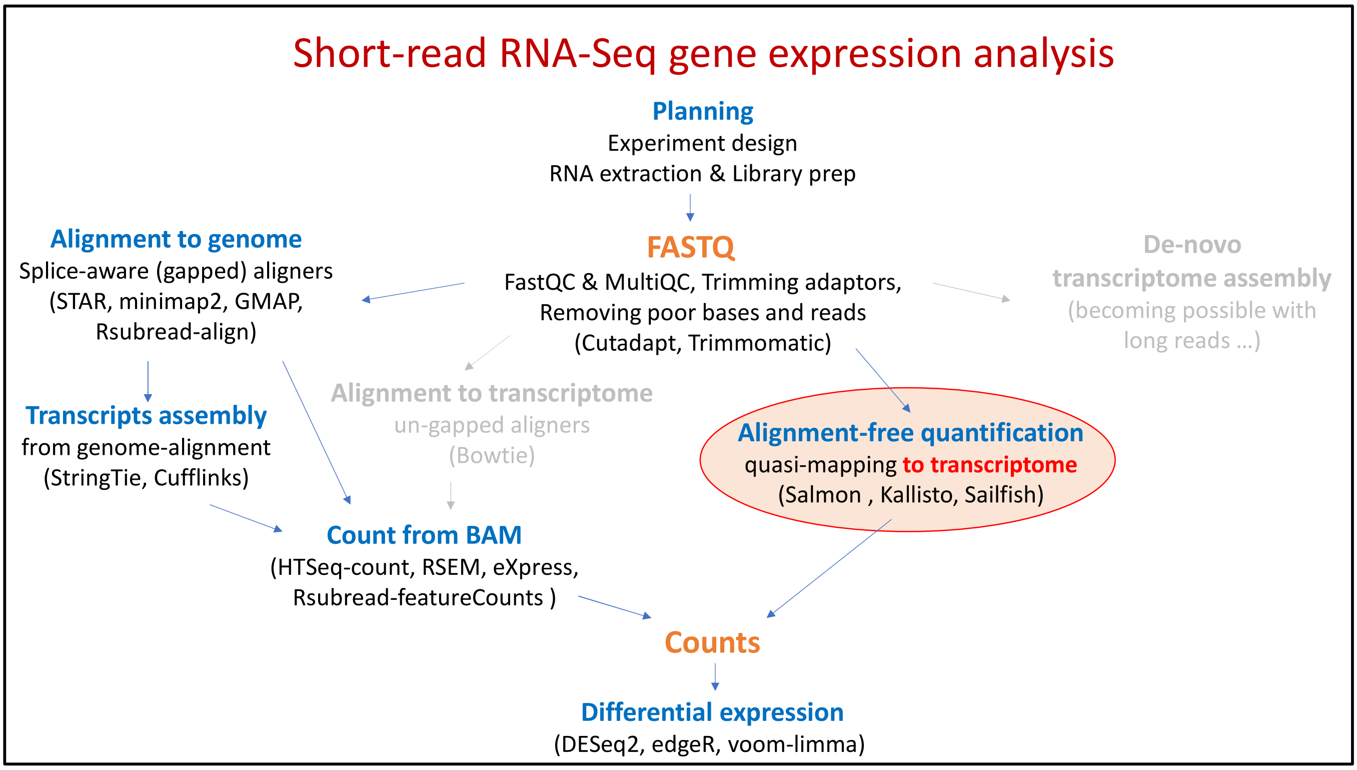
Alignment-free quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

Now, lets come back to our high-level overview of the analysis.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

Alignment-free quantification
quasi-mapping **to transcriptome**
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

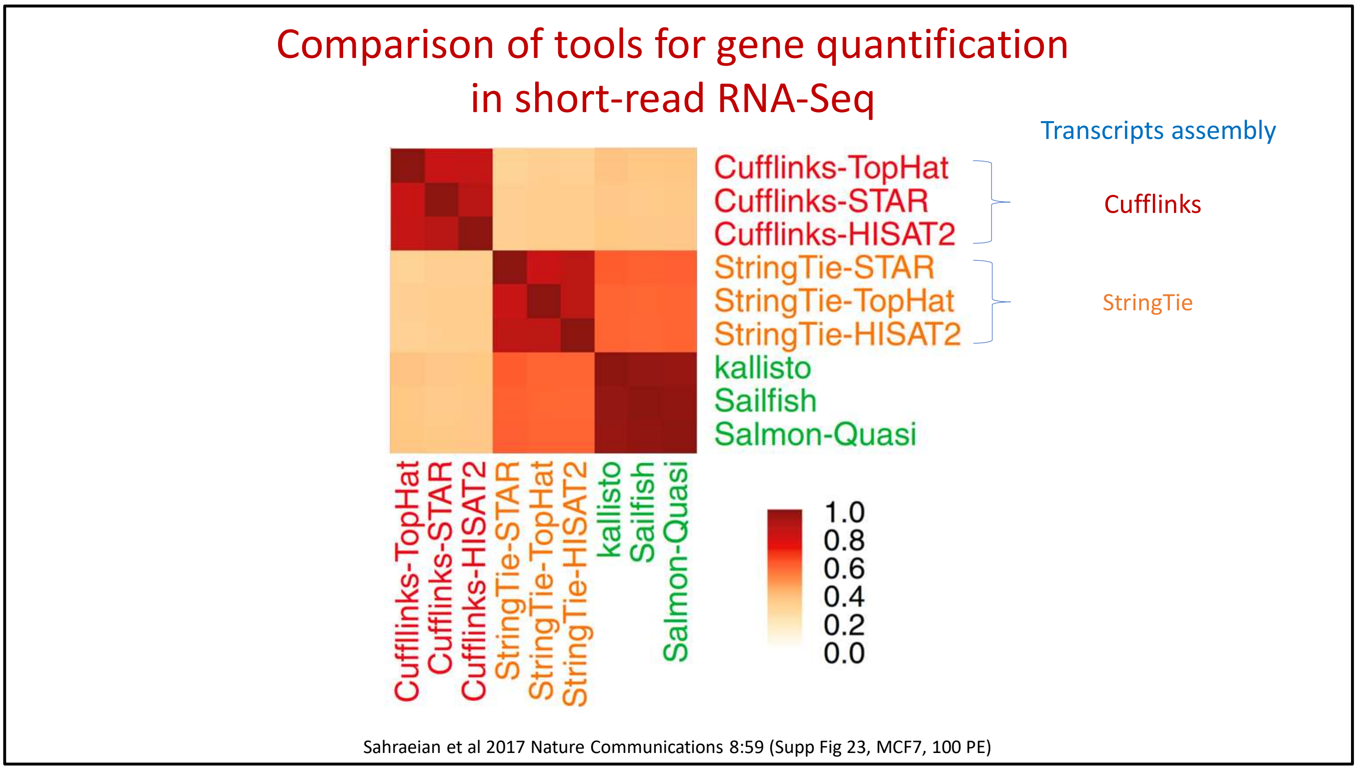
Counts

Differential expression
(DESeq2, edgeR, voom-limma)

After we considered the ways of getting the Counts from BAM-s, let's see how we can get the counts directly from FASTQ files.

If you are only interested in expression of already known genes, which is often the case in human cancer research, then there is no need in formal alignment to BAMs. It could be done much easier and faster using the "Alignment-free" quasi-mapping tools. Somehow, it happened that these tools are named following a marine (or sea-food?) theme: Salmon, Kallisto and Sailfish.

These tools do "quick and dirty" mapping of reads to known transcripts, without generating BAM files, using some common-sense heuristics if a read maps into two (or more) different genes. They do it extremely fast and output the counts per transcript (raw and TPM normalized).

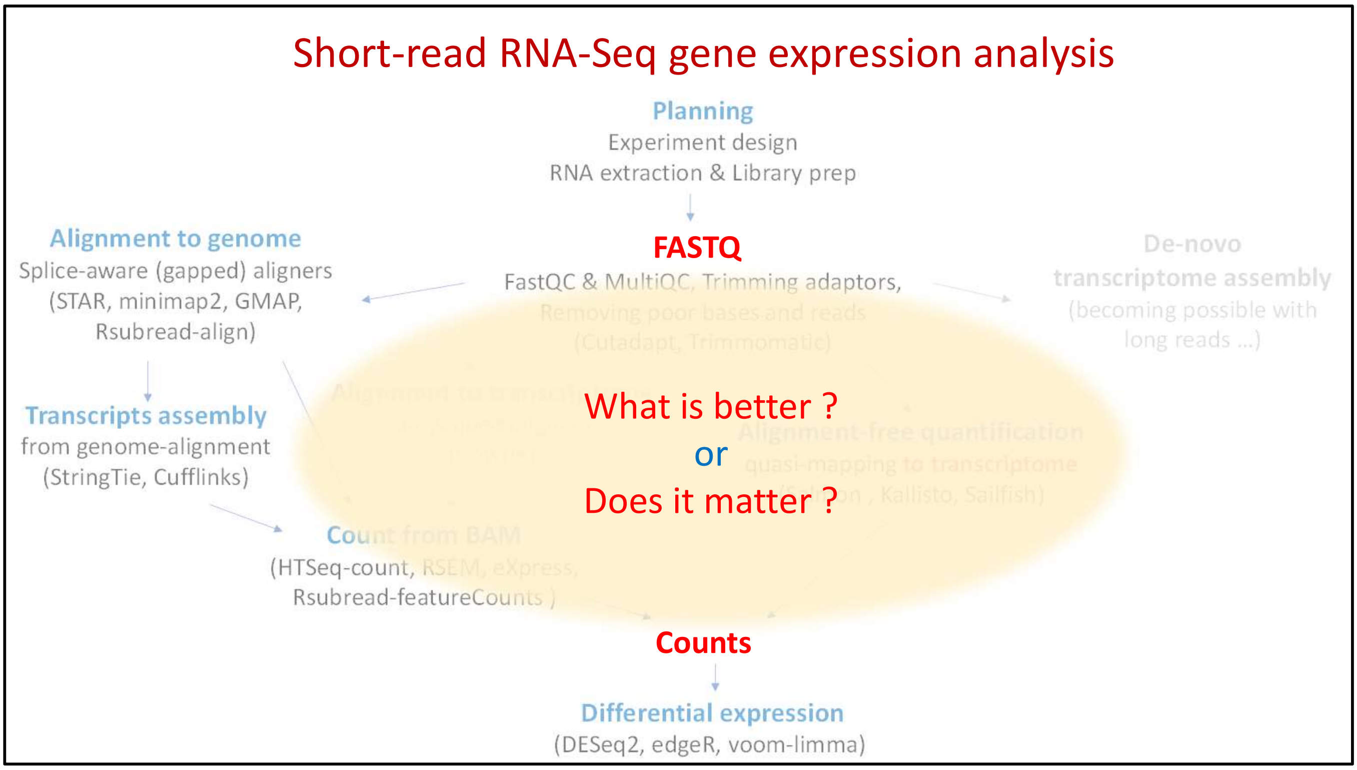


Comparison of tools for gene quantification in short-read RNA-Seq

Sahraeian et al 2017 Nature Communications 8:59 (Supp Fig 23, MCF7, 100 PE)

Quite encouragingly, the results of the different alignment-tree tools are very concordant.

For comparison, this heatmap shows that introducing a ***Transcript assembly*** step leads to the much more discordant results, depending on the used tool.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads …)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome

Alignment-free quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

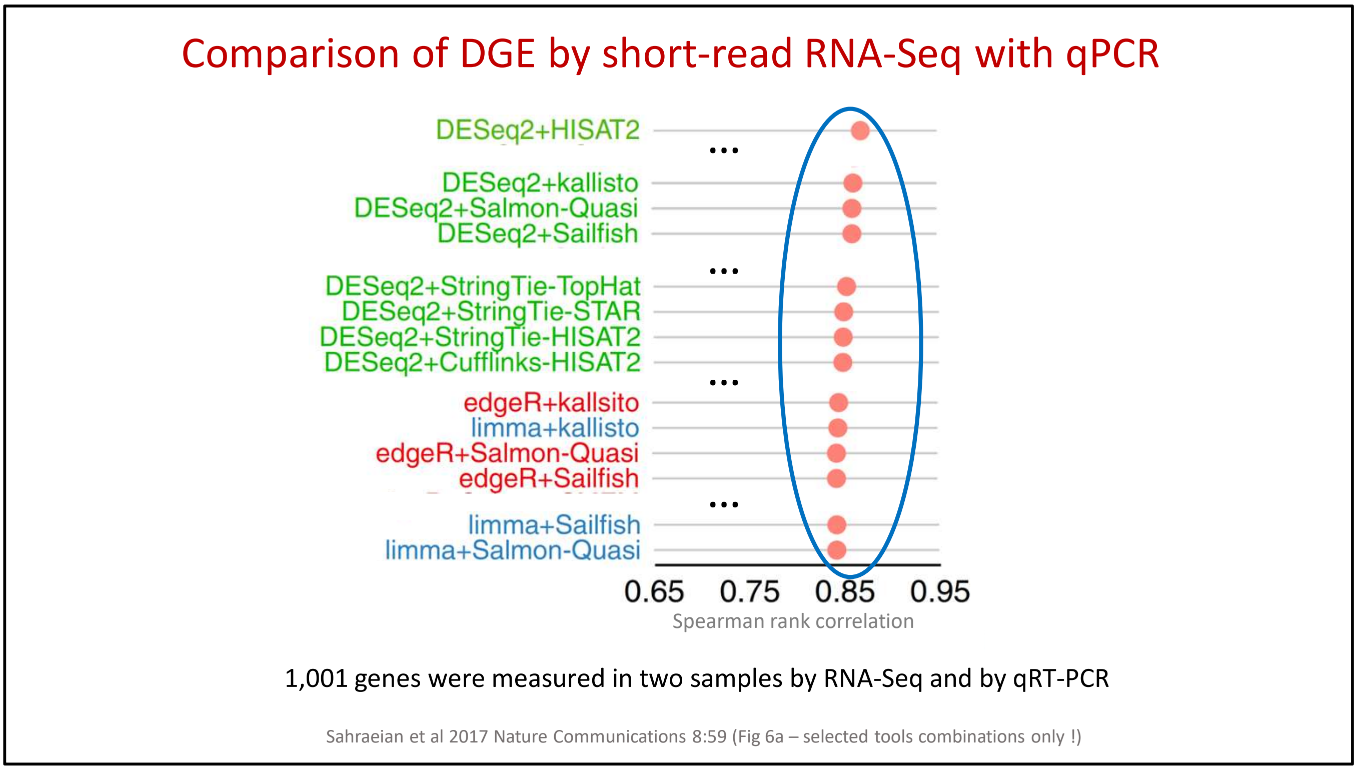
What is better ?
or
Does it matter ?

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

Now, after we considered multiple trajectories and tools for getting transcript counts from the short-read RNA-seq FASTQ files, the natural questions are:

1) What is better?
or
2) Does it make any difference ?



Comparison of DGE by short-read RNA-Seq with qPCR

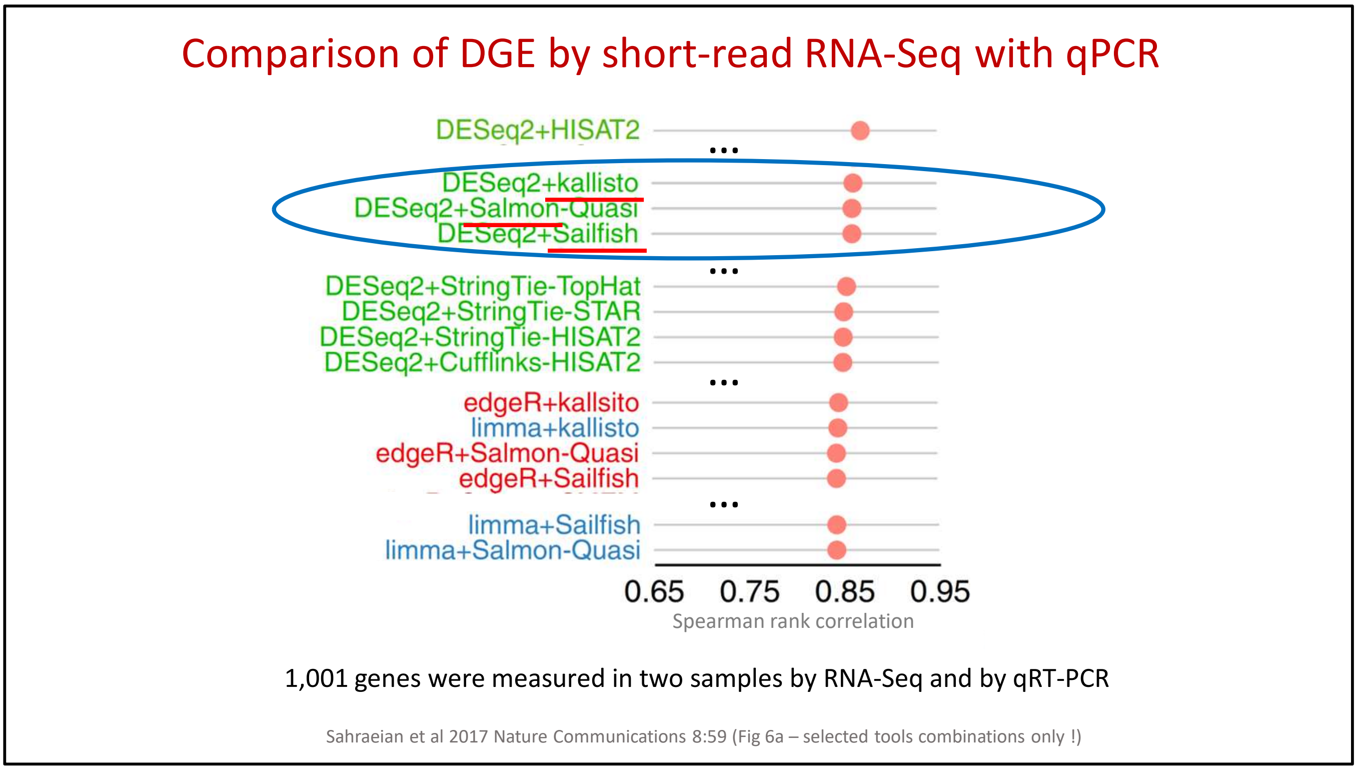
1,001 genes were measured in two samples by RNA-Seq and by qRT-PCR

Sahraeian et al 2017 Nature Communications 8:59 (Fig 6a – selected tools combinations only !)

To answer this question, we would need to compare RNAseq with another experimental method of measuring gene expressions. Of course, this already has been done. This plot shows the experiment where a thousand genes were measured in two samples by RNA-seq and by RT-PCR. Then the results of Differential Gene Expression (DGE) analysis based on RT-PCR were compared with the results of different RNA-seq pipelines.

On this figure you can see that even the uncertainty introduced by transcript assembly has little effect on the order of differentially expressed genes, if the same pipeline was applied to the compared samples.

So, if it happened that historically your bioinformatics core prefers using STAR and HTSeq-count for DGE analysis: don’t panic, they just follow the example of TCGA ☺

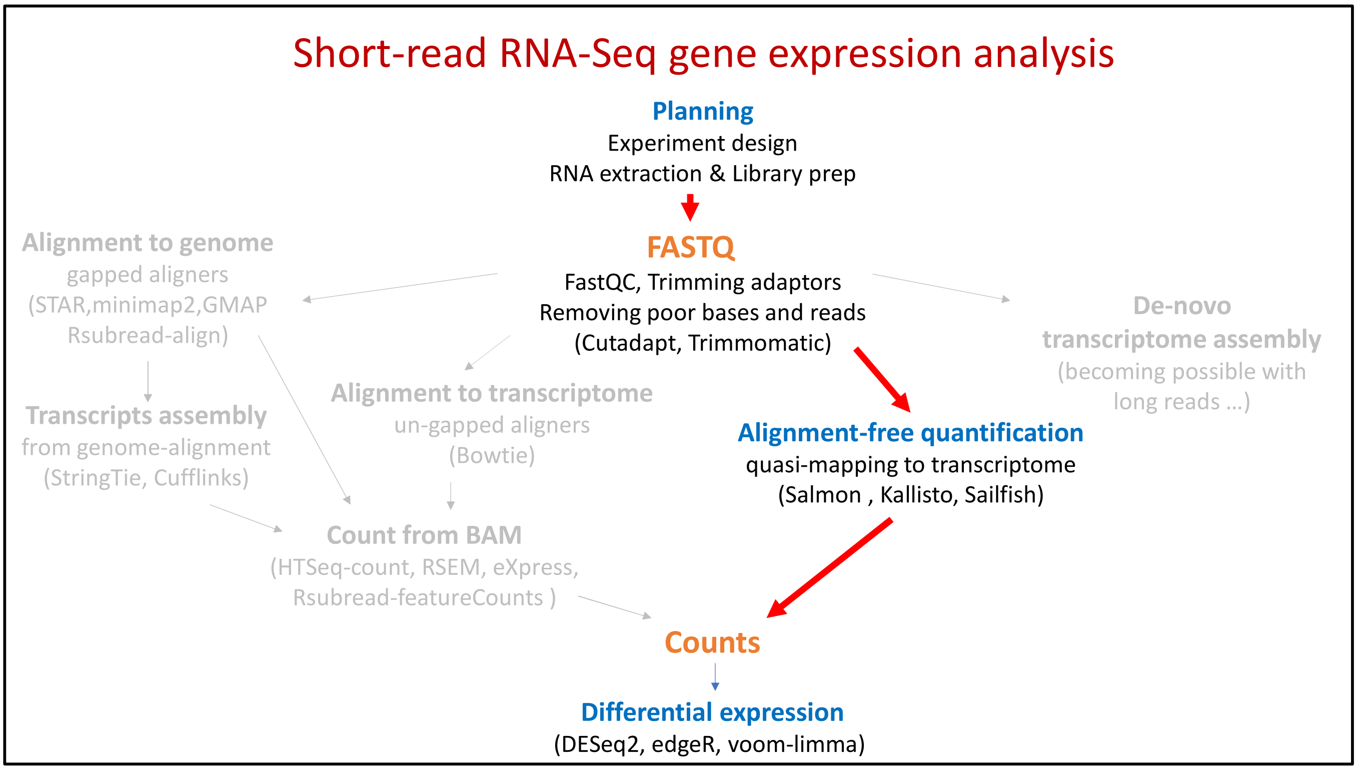


Comparison of DGE by short-read RNA-Seq with qPCR

1,001 genes were measured in two samples by RNA-Seq and by qRT-PCR

Sahraeian et al 2017 Nature Communications 8:59 (Fig 6a – selected tools combinations only !)

On the other hand, you can see that the Alignment-free quasi-mapping showed a very good correlation to qPCR.



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC, Trimming adaptors
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
gapped aligners
(STAR,minimap2,GMAP
Rsubread-align)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Alignment-free quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

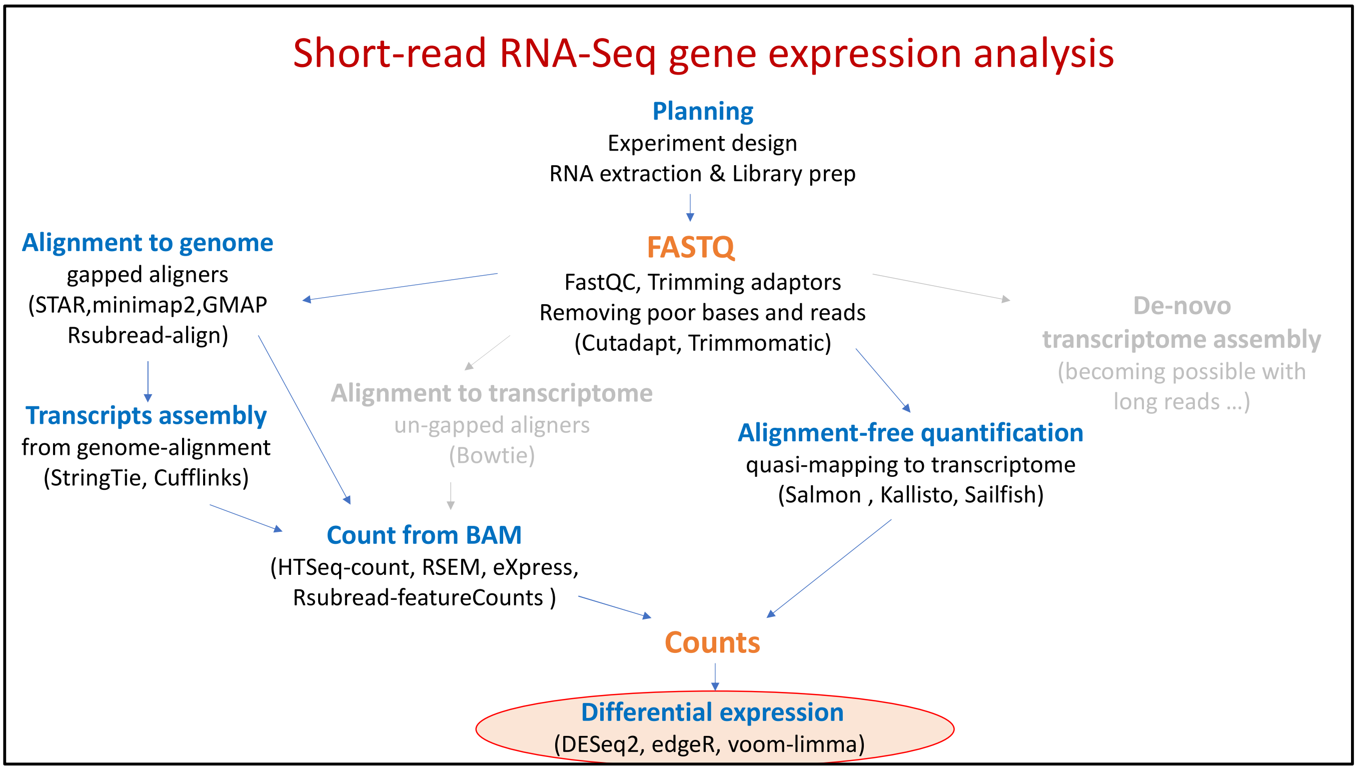
De-novo transcriptome assembly
(becoming possible with
long reads …)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

So, for a simple Differential Gene Expression (DGE) analysis for *already known genes* there is no need in gapped alignment, complicated transcripts assembly and counting from BAMs. Virtually the same results could be achieved quicker by the quasi-mapping tools.

This is why I personally prefer the quasi-mapping tools: why to use a complicated multi-step pipeline, including the long and computationally demanding alignment step, if you can do the same in one step and quickly?



Short-read RNA-Seq gene expression analysis

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC, Trimming adaptors
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
gapped aligners
(STAR,minimap2,GMAP
Rsubread-align)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners
(Bowtie)

Alignment-free quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

De-novo transcriptome assembly
(becoming possible with
long reads …)

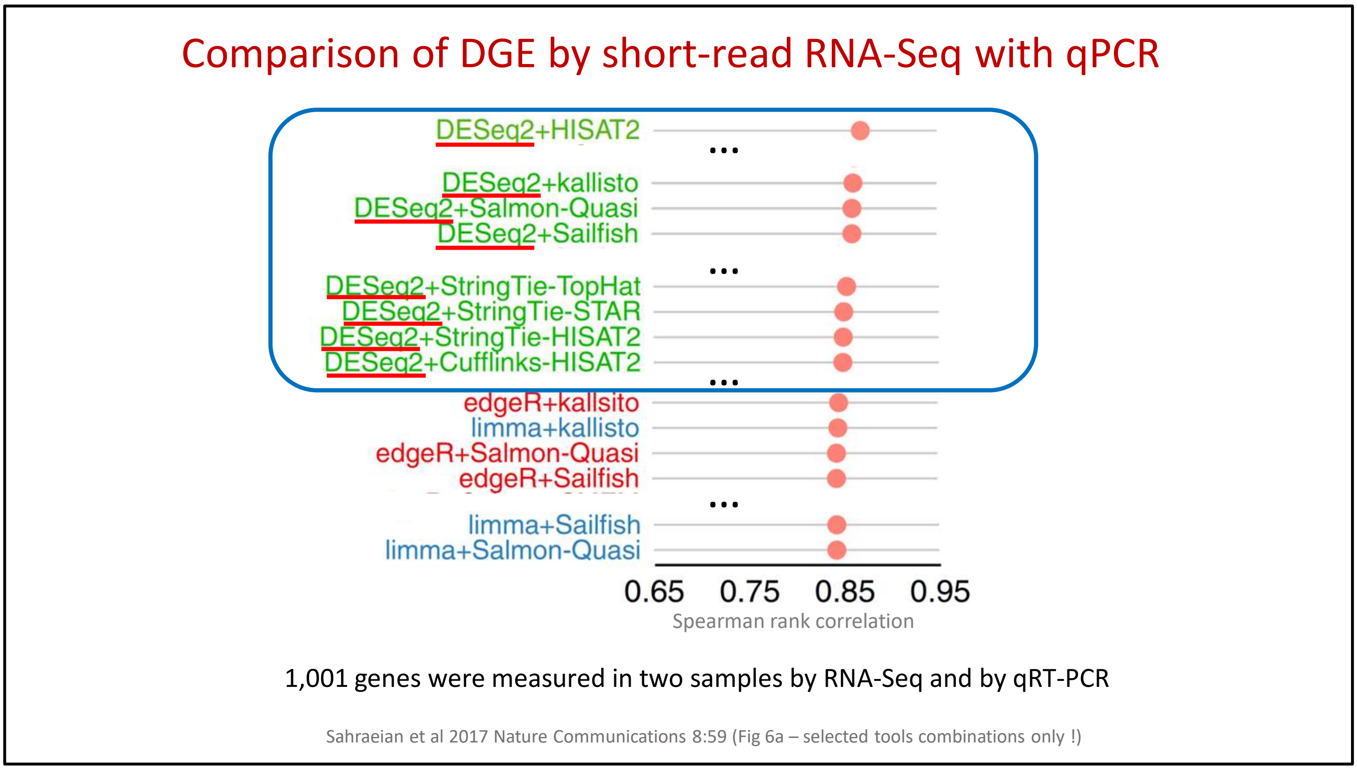
Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

OK, after we have finally obtained the counts, let’s now use them for the differential gene expression analysis!

Again, there are several tools to choose from !

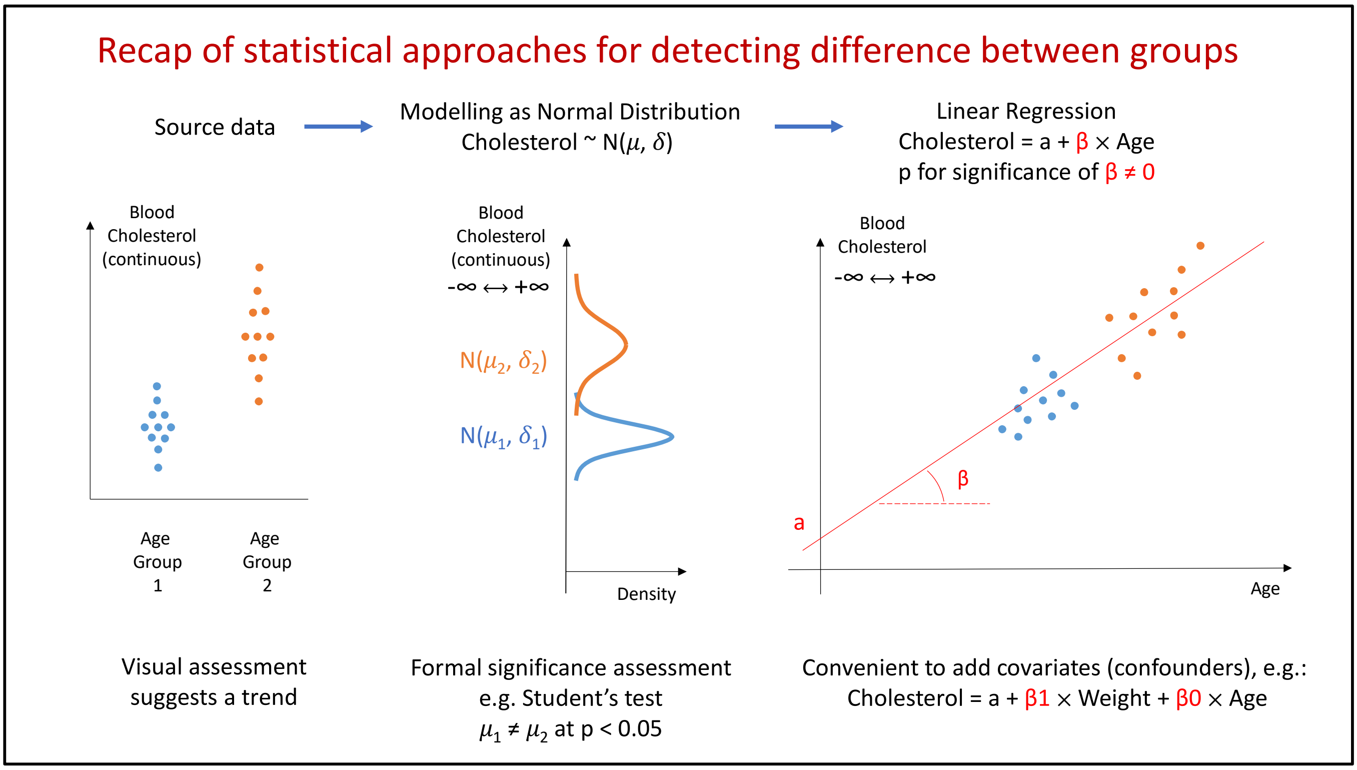


Comparison of DGE by short-read RNA-Seq with qPCR

1,001 genes were measured in two samples by RNA-Seq and by qRT-PCR

Sahraeian et al 2017 Nature Communications 8:59 (Fig 6a – selected tools combinations only !)

Luckily, the same study that I have already cited many times, suggested that using DESeq2 may be a reasonable choice.



Recap of statistical approaches for detecting difference between groups

Source data → Modelling as Normal Distribution Cholesterol ~ $N(\mu, \delta)$ → Linear Regression Cholesterol = a + β × Age p for significance of $\beta \neq 0$

Visual assessment suggests a trend

Formal significance assessment e.g. Student’s test $\mu_1 \neq \mu_2$ at $p < 0.05$

Convenient to add covariates (confounders), e.g.: Cholesterol = a + β1 × Weight + β0 × Age

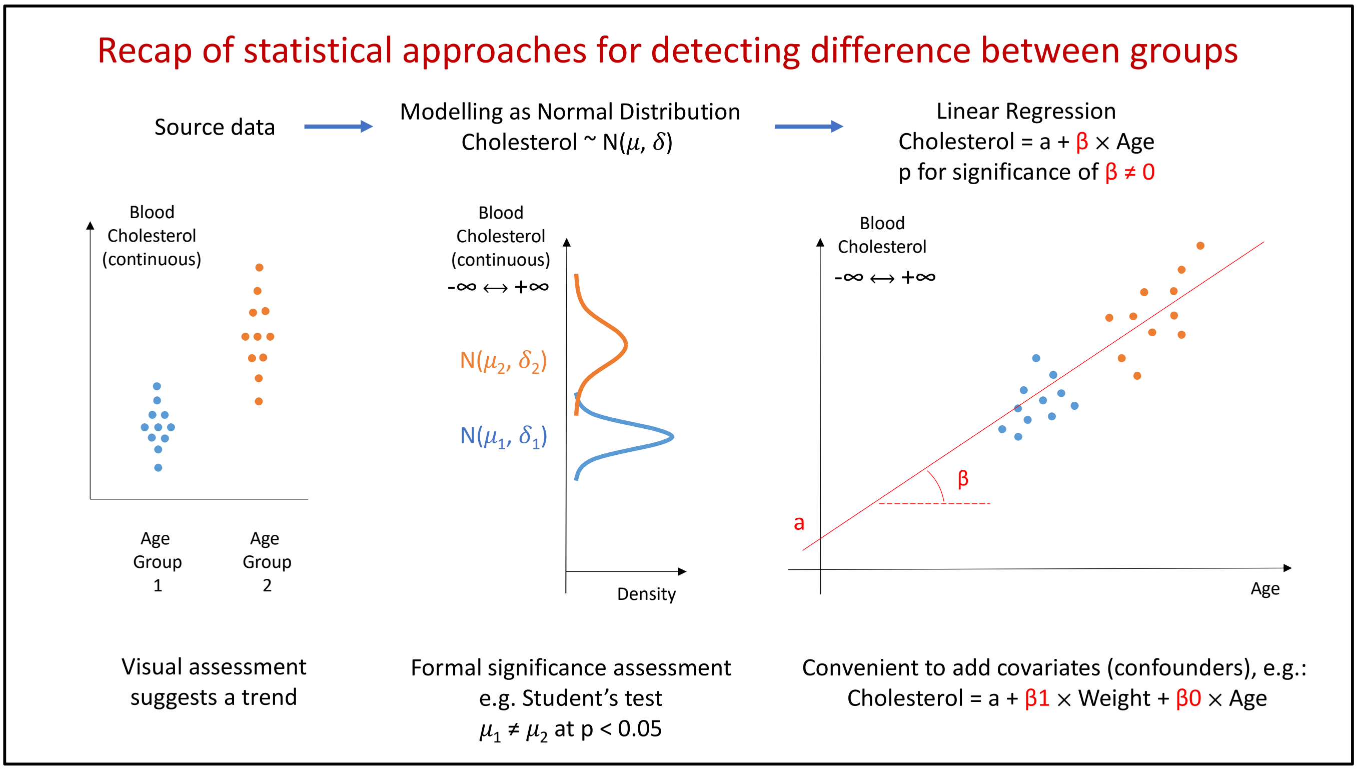
Don’t worry, I will not go deep into statistical details. I will just try to explain the intuition behind what is done by DESeq2.

Let's start with an analogy: let’s consider how we would compare, for instance, blood cholesterol level in younger and older people.

Usually, we start with some plot to visualize the data. Here you can see the dot-plot on the left.

Then we often assume that the observed values were sampled from normal distribution(s), in other words, we describe or “model” the data using normal distribution (as shown in the middle). At this point we already can apply some standard statistical tests, like the Student’s t-test, which, in essence, compares mean values in the two groups.

The same mean values can also be compared using a regression framework, as shown at the right. There is nothing special in this example yet: its just a standard linear regression. If, on average, the blood cholesterol in young people is lower than in the old, then the regression line will have some slope. Mathematically, this means that the coefficient beta in the regression equation (shown above the plot) will be significantly different from zero.



Recap of statistical approaches for detecting difference between groups

Source data → Modelling as Normal Distribution Cholesterol ~ N(μ, δ) → Linear Regression Cholesterol = a + β × Age p for significance of β ≠ 0

Visual assessment suggests a trend

Formal significance assessment e.g. Student's test $\mu_1 \neq \mu_2$ at $p < 0.05$

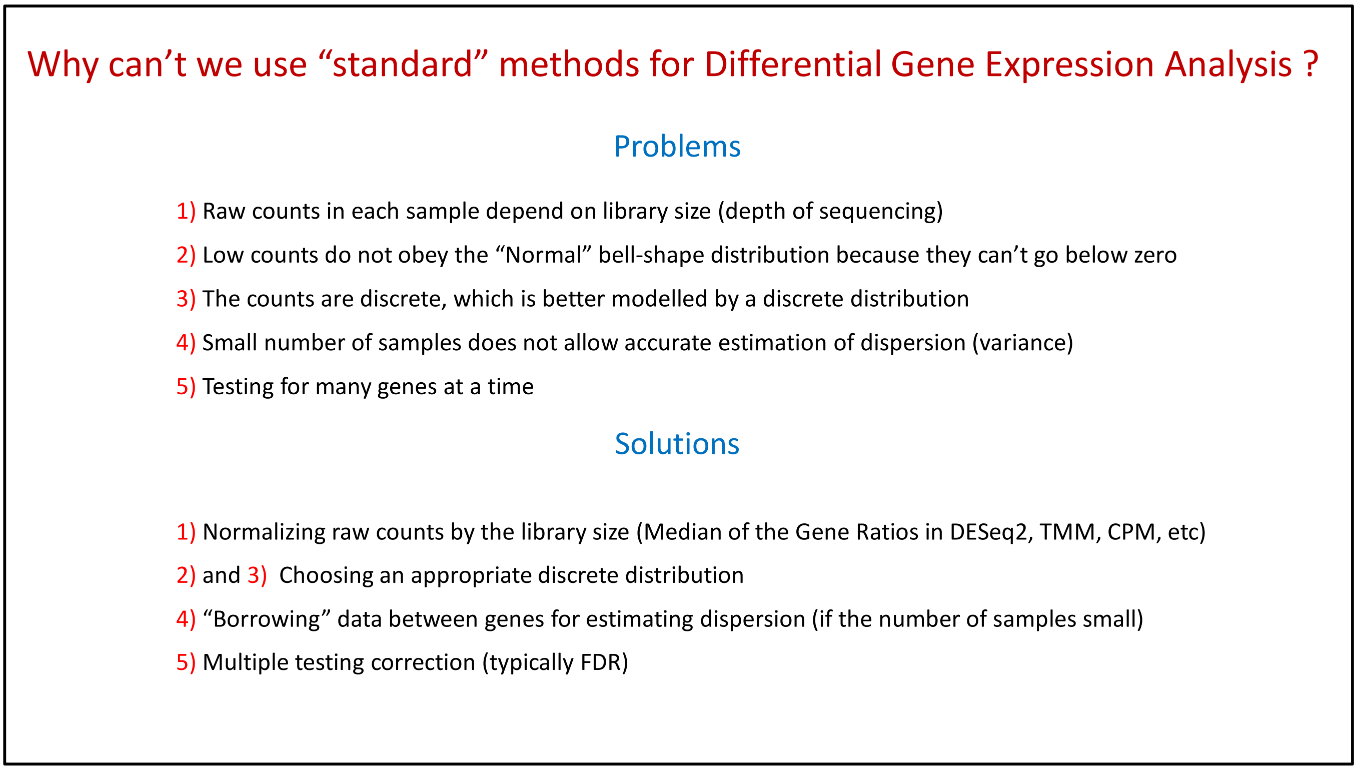
Convenient to add covariates (confounders), e.g.: Cholesterol = a + β1 × Weight + β0 × Age

At this point you may ask: “Why to complicate things by regression modelling, if we could use a simple t-test?”

The answer is shown in the model BELOW the plot. In contrast to the t-test, the regression modelling allows accounting for confounders. For instance, if, on average, the compared age groups had slightly different weight, the regression modelling would allow to control for the weight difference by adding it into the equation.

Thus, in the model below the plot, beta-zero will show the association of cholesterol with age, after excluding the possible confounding effect of the weight.

Again, there is nothing special here so far: its just a recap of ***standard*** statistical methods, which are commonly used to detect differences between two groups.



Why can’t we use “standard” methods for Differential Gene Expression Analysis ?

Problems

1) Raw counts in each sample depend on library size (depth of sequencing)
2) Low counts do not obey the “Normal” bell-shape distribution because they can’t go below zero
3) The counts are discrete, which is better modelled by a discrete distribution
4) Small number of samples does not allow accurate estimation of dispersion (variance)
5) Testing for many genes at a time

Solutions

1) Normalizing raw counts by the library size (Median of the Gene Ratios in DESeq2, TMM, CPM, etc)

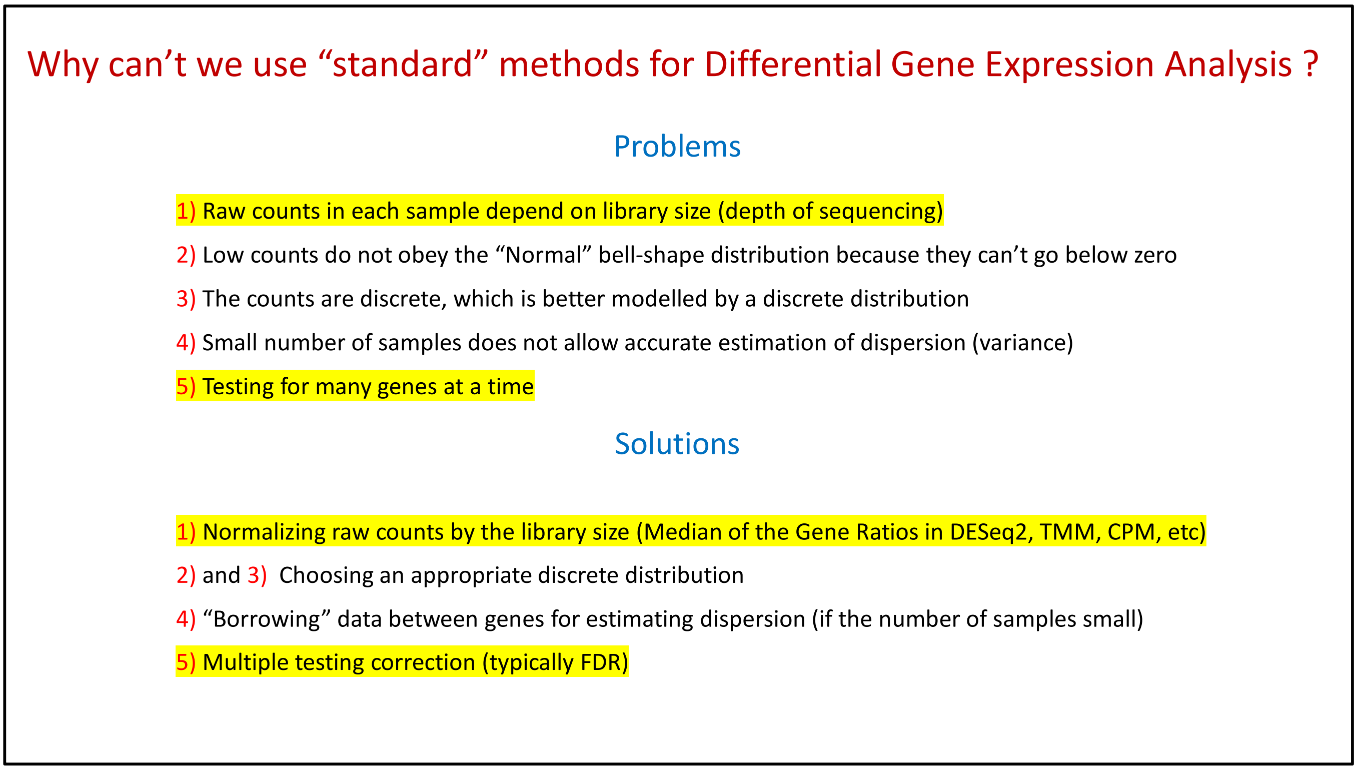
2) and 3) Choosing an appropriate discrete distribution

4) “Borrowing” data between genes for estimating dispersion (if the number of samples small)

5) Multiple testing correction (typically FDR)

So, why can’t we just apply these standard methods to the differential gene expression in RNA-seq counts?

Well, there are several problems here.



Why can’t we use “standard” methods for Differential Gene Expression Analysis ?

Problems

1) Raw counts in each sample depend on library size (depth of sequencing)

2) Low counts do not obey the “Normal” bell-shape distribution because they can’t go below zero

3) The counts are discrete, which is better modelled by a discrete distribution

4) Small number of samples does not allow accurate estimation of dispersion (variance)

5) Testing for many genes at a time

Solutions

1) Normalizing raw counts by the library size (Median of the Gene Ratios in DESeq2, TMM, CPM, etc)

2) and 3) Choosing an appropriate discrete distribution

4) “Borrowing” data between genes for estimating dispersion (if the number of samples small)

5) Multiple testing correction (typically FDR)

Some problems are relatively simple:

For instance, it’s obvious that to compare genes between samples we need to normalize for the total number of reads (= library size) in the samples.

It’s also pretty clear that testing for many genes at a time we should apply some multiple testing correction.

That’s about problems 1 and 5 on this slide.

Why can’t we use “standard” methods for Differential Gene Expression Analysis ?

Problems

1) Raw counts in each sample depend on library size (depth of sequencing)

2) Low counts do not obey the “Normal” bell-shape distribution because they can’t go below zero

3) The counts are discrete, which is better modelled by a discrete distribution

4) Small number of samples does not allow accurate estimation of dispersion (variance)

5) Testing for many genes at a time

Solutions

1) Normalizing raw counts by the library size (Median of the Gene Ratios in DESeq2, TMM, CPM, etc)

2) and 3) Choosing an appropriate discrete distribution

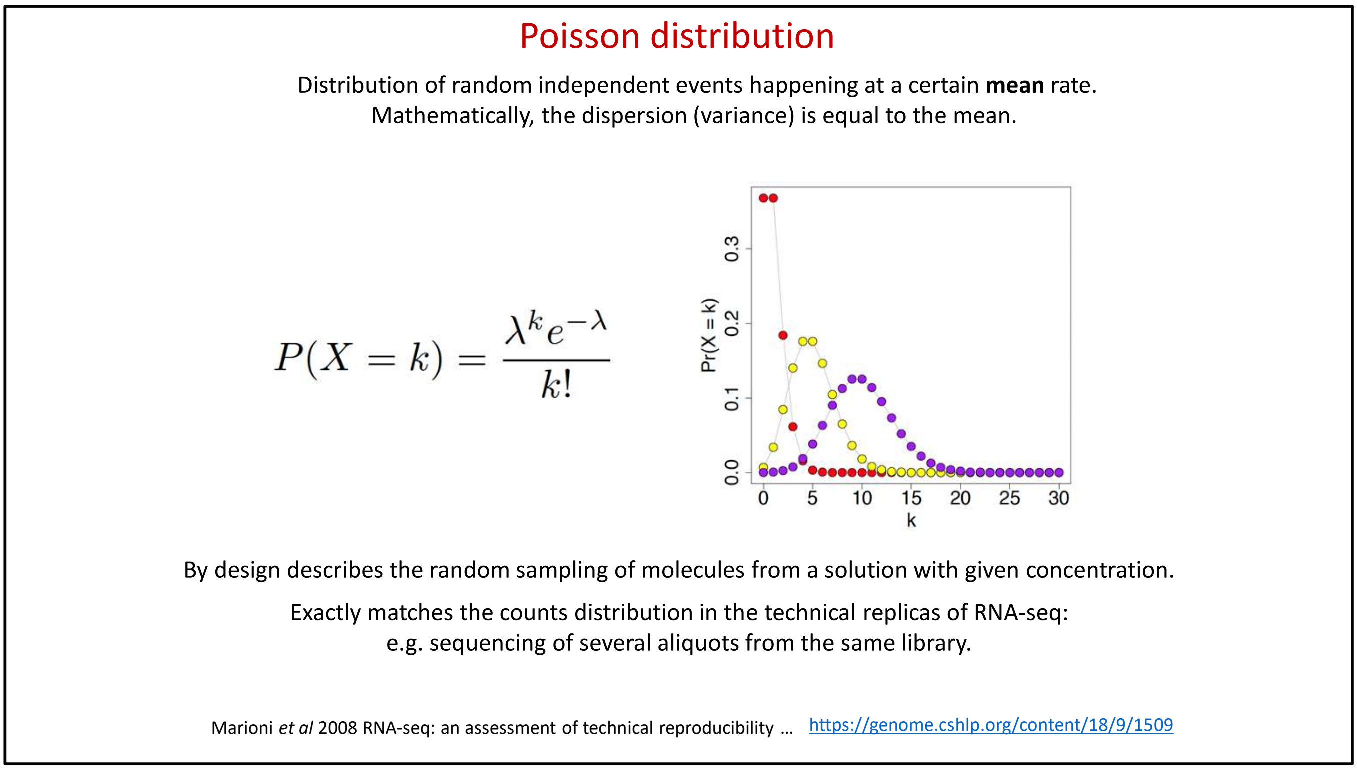
4) “Borrowing” data between genes for estimating dispersion (if the number of samples small)

5) Multiple testing correction (typically FDR)

Problems 2 and 3 are about the distribution that would properly describe (=“model”) the counts. The normal distribution is not good for modeling the counts because it includes negative values, and it is continuous. In contrast, the counts are always positive, and they are discrete.

You may say that blood cholesterol also can’t be negative. Yes, that’s correct, but it never goes anywhere close to zero either. So, within the range of real blood cholesterol values, the approximation by the Normal is OK. This is not the case for RNA-seq counts, which often may be quite close to zero.

So, a different distribution is needed. And, of course, this should be a discrete distribution too.



Poisson distribution

Distribution of random independent events happening at a certain **mean** rate.
Mathematically, the dispersion (variance) is equal to the mean.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Pr(X = k)
0.0
0.1
0.2
0.3
0
5
10
15
20
25
30
k

By design describes the random sampling of molecules from a solution with given concentration.

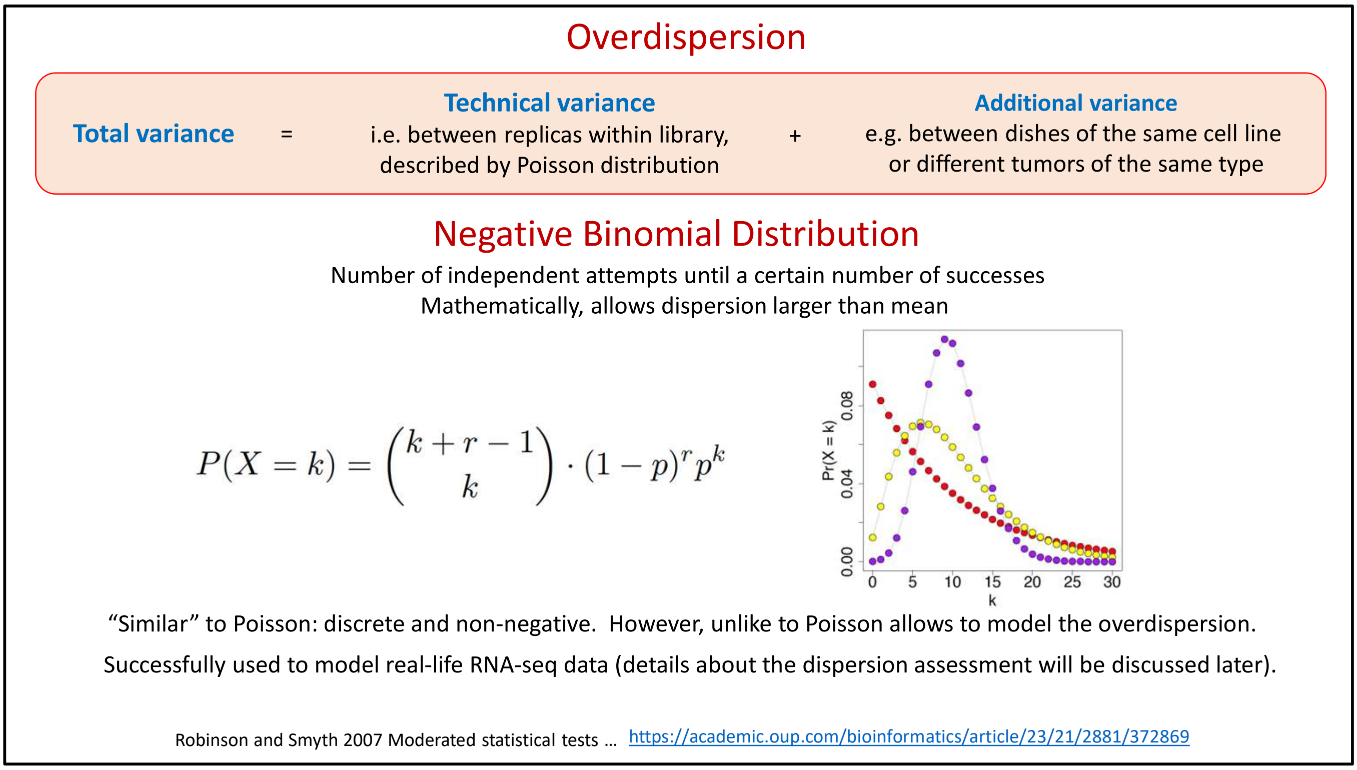
Exactly matches the counts distribution in the technical replicas of RNA-seq:
e.g. sequencing of several aliquots from the same library.

Marioni *et al* 2008 RNA-seq: an assessment of technical reproducibility … https://genome.cshlp.org/content/18/9/1509

And such distribution exists: it’s the Poisson distribution.

It by design perfectly describes the random sampling of molecules from a solution with given concentration.

It perfectly matches the distribution of empirically observed counts in TECHNICAL replicates in RNA-seq: when we sequence multiple aliquots from the same library.



Overdispersion

Total variance = **Technical variance** i.e. between replicas within library, described by Poisson distribution + **Additional variance** e.g. between dishes of the same cell line or different tumors of the same type

Negative Binomial Distribution

Number of independent attempts until a certain number of successes
Mathematically, allows dispersion larger than mean

$$P(X=k) = \binom{k+r-1}{k} \cdot (1-p)^r p^k$$

“Similar” to Poisson: discrete and non-negative. However, unlike to Poisson allows to model the overdispersion.

Successfully used to model real-life RNA-seq data (details about the dispersion assessment will be discussed later).

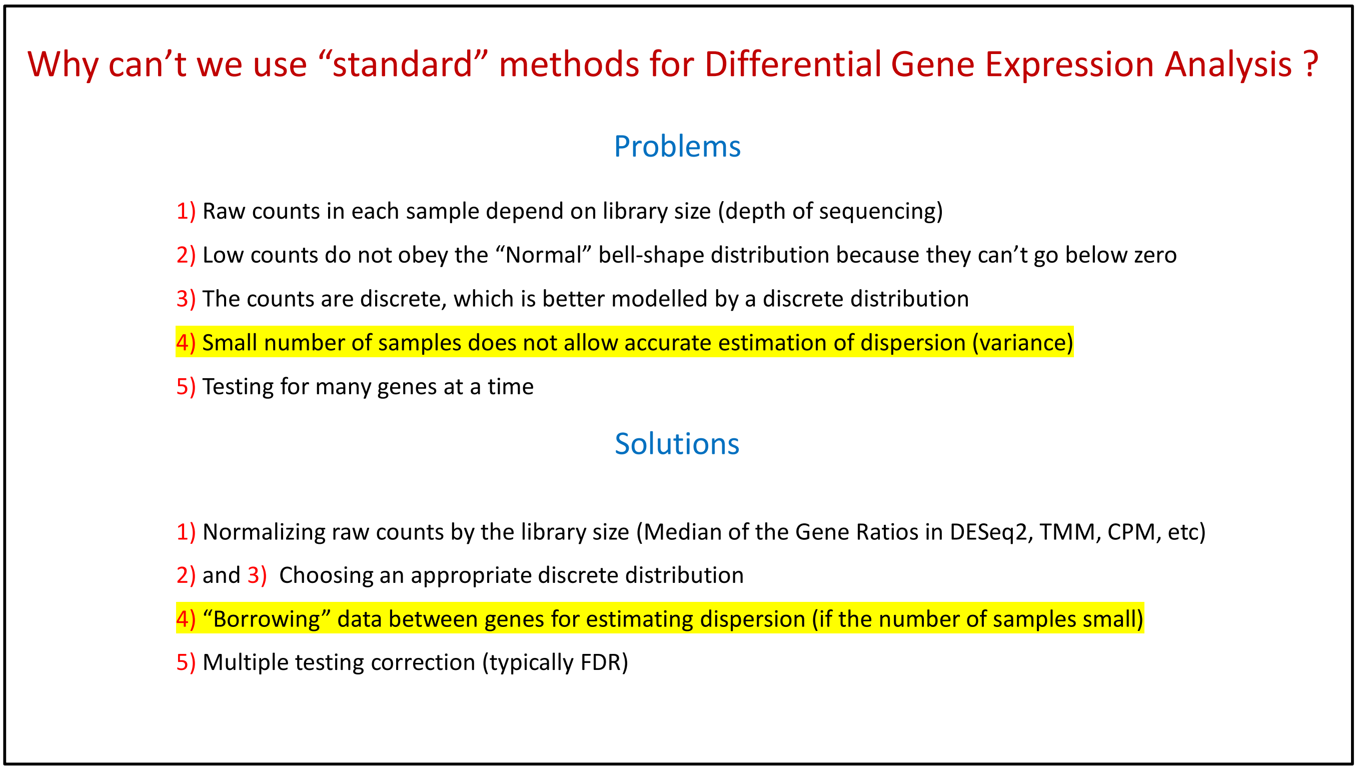
Robinson and Smyth 2007 Moderated statistical tests … https://academic.oup.com/bioinformatics/article/23/21/2881/372869

Well, in real life we usually are more interested in differences between BIOLOGICAL replicates (rather than in comparing aliquots from the same sample :)

For instance, we may look for differences between several tumors vs several normal samples, etc. Obviously, the dispersion of counts between different tumors should be higher than in several technical replicas from the same tumor.

It happened, that Poisson distribution is not good for fitting this additional dispersion, because mathematically the variance (=dispersion) in Poisson distribution is linked to it’s mean.

This is why some clever statisticians decided to use Negative Binomial distribution instead of the Poisson: its very similar to Poisson in terms of being discrete and non-negative; in addition, mathematically, its variance is not linked to the mean. So, it can describe the *overdispersion* observed in RNA-seq counts from real life biological samples.



Why can’t we use “standard” methods for Differential Gene Expression Analysis ?

Problems

1) Raw counts in each sample depend on library size (depth of sequencing)

2) Low counts do not obey the “Normal” bell-shape distribution because they can’t go below zero

3) The counts are discrete, which is better modelled by a discrete distribution

4) Small number of samples does not allow accurate estimation of dispersion (variance)

5) Testing for many genes at a time

Solutions

1) Normalizing raw counts by the library size (Median of the Gene Ratios in DESeq2, TMM, CPM, etc)

2) and 3) Choosing an appropriate discrete distribution

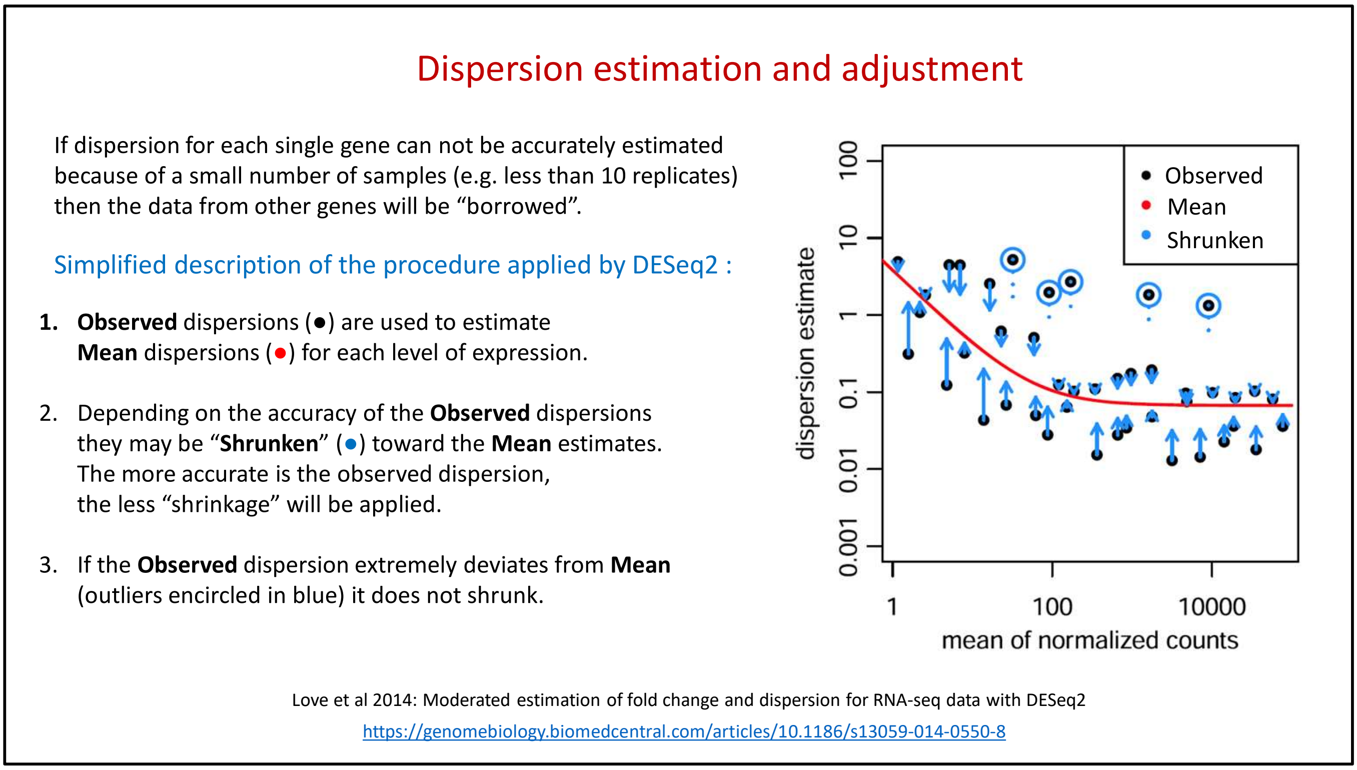
4) “Borrowing” data between genes for estimating dispersion (if the number of samples small)

5) Multiple testing correction (typically FDR)

Finally, the problem 4 is, essentially, about the small number of replicas in studies.

Initially, RNA-sequencing was very expensive. It was not uncommon, when an RNA-seq experiment studying some treatment in cell lines, would include just 3 dishes with *vs* 3 dishes without a treatment. This means that for each gene we would have just 6 counts: 3 in each group to compare.

You may remember from basic statistics that with a small number of observations the error of the mean is high, which may prevent getting any statistical significance when comparing means from small numbers of observations (especially, after the multiple testing correction…)



Dispersion estimation and adjustment

If dispersion for each single gene can not be accurately estimated because of a small number of samples (e.g. less than 10 replicates) then the data from other genes will be "borrowed".

Simplified description of the procedure applied by DESeq2 :

1. **Observed** dispersions (●) are used to estimate **Mean** dispersions (●) for each level of expression.
2. Depending on the accuracy of the **Observed** dispersions they may be "**Shrunken**" (●) toward the **Mean** estimates. The more accurate is the observed dispersion, the less "shrinkage" will be applied.
3. If the **Observed** dispersion extremely deviates from **Mean** (outliers encircled in blue) it does not shrunk.

Love et al 2014: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

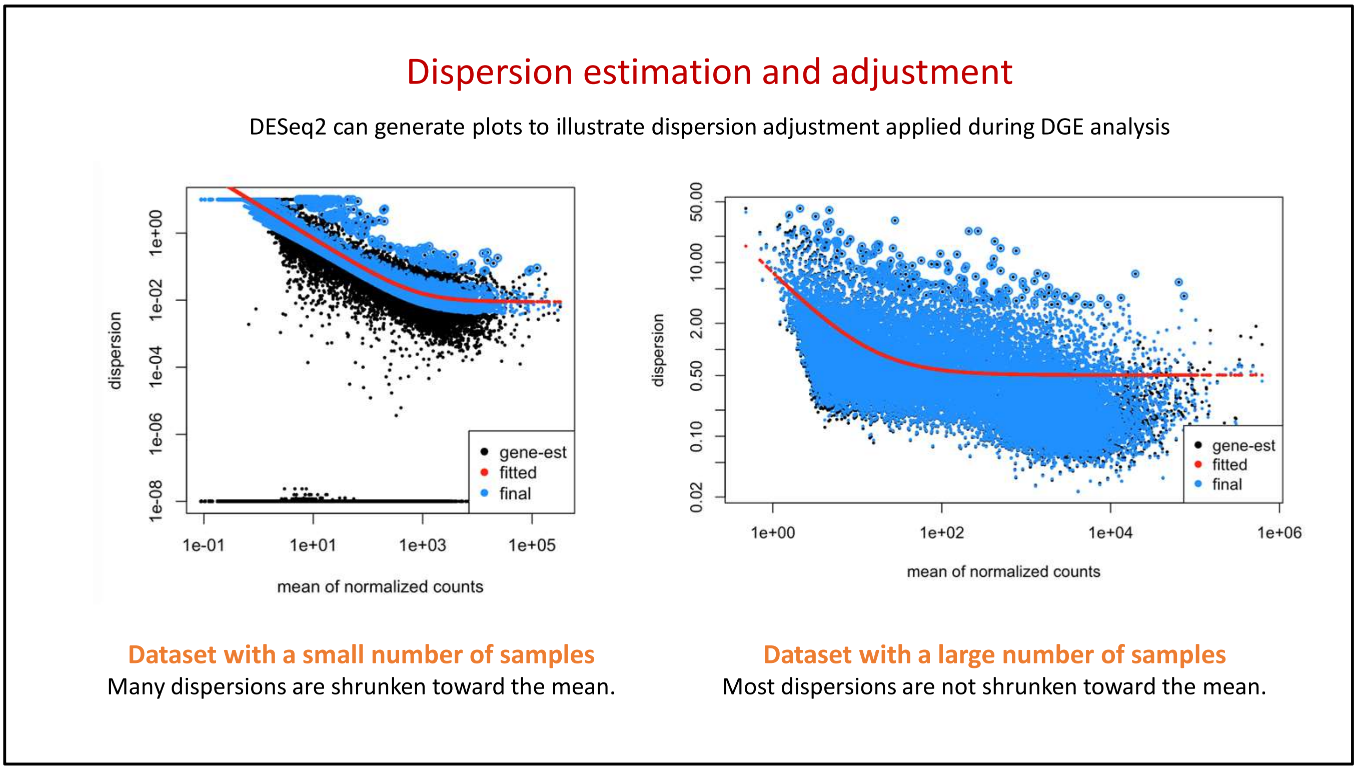
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8

To solve this problem, it was suggested that the dispersion could be similar in genes with similar levels of expression. So, unless the gene is a clear outlier, its OBSERVED dispersion can be shifted towards the MEAN dispersion in all the genes with the similar level of expression.

This picture illustrates the dispersion adjustment implemented in DESeq2:

The black dots show actually OBSERVED dispersions, the red line shows the MEAN dispersion for different levels of gene expression, and the blue arrows show the dispersion estimates used in the analysis, which are are SHRUNKEN toward the mean.

For some genes, which have exceptionally high dispersion, the dispersion is not adjusted. This, effectively may prevent any significance detection in such genes, if the number of replicas is small. Such gross outliers are encircled in blue on this picture.



Dispersion estimation and adjustment

DESeq2 can generate plots to illustrate dispersion adjustment applied during DGE analysis

Dataset with a small number of samples
Many dispersions are shrunken toward the mean.

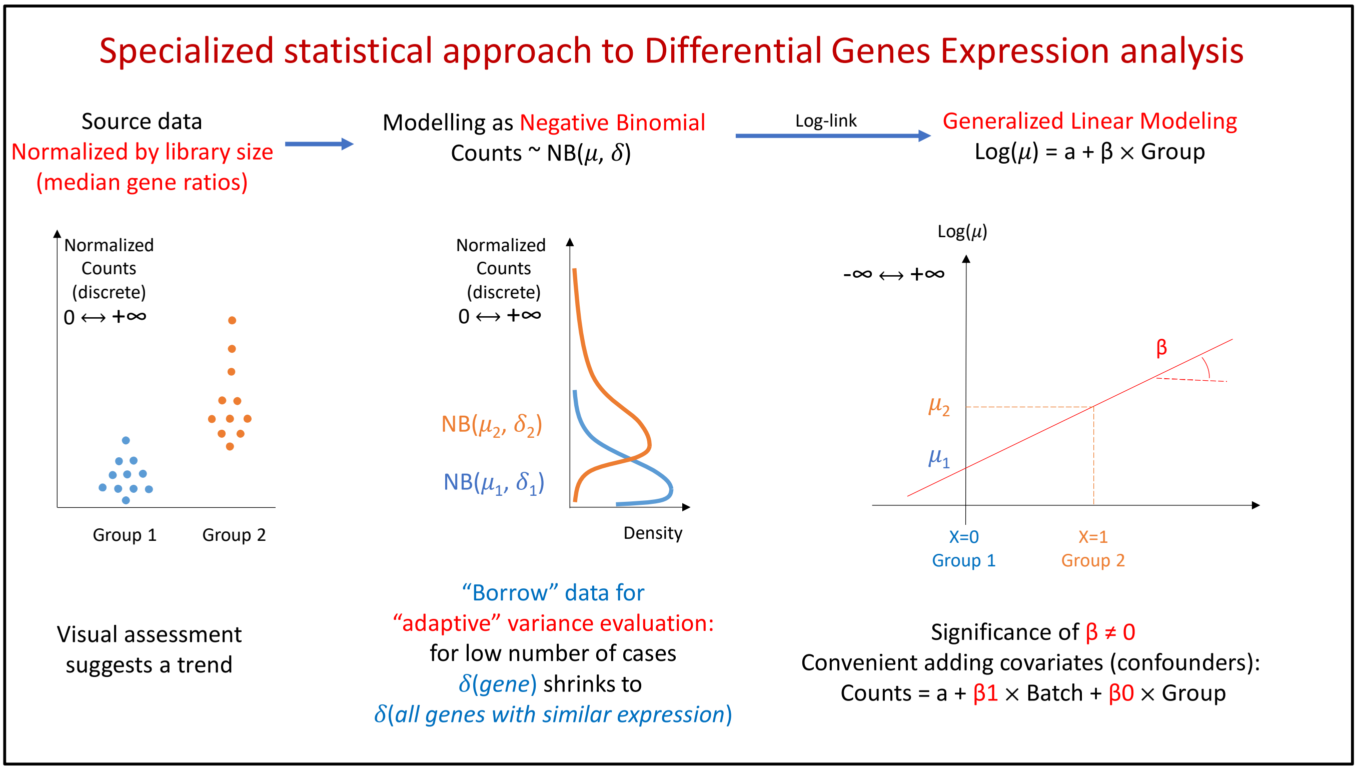
Dataset with a large number of samples
Most dispersions are not shrunken toward the mean.

The degree of dispersion shrinkage in DESeq2 heuristically depends on the number of replicas.

The left panel shows a plot for an experiment with low number of replicas. You can see that most of the observed dispersions (black dots: “gene-est”) have been changed to the estimates (shown in blue: “final”) by shrinking toward the mean (shown by the red line: “fitted”). You can also see that some genes with exceptionally high dispersions were kept unchanged (encircled in blue), to avoid inflating the significance for such genes.

It is claimed that the dispersion adjustment implemented in DESeq2 allows to compare gene expressions between the groups containing as little as 3 replicas each.

Of course, there is another way of solving this problem: just getting more replicas. The right panel shows the dispersion adjustment plot for an analysis using hundreds of samples from TCGA. If DEseq2 detects a large number of samples, it decides that no dispersion adjustment is needed. You can see that in the right plot the final dispersion estimates used in the analysis (blue dots) are virtually overlapping the actually observed dispersions (the black dots).



Specialized statistical approach to Differential Genes Expression analysis

Source data
Normalized by library size
(median gene ratios)

Modelling as Negative Binomial
Counts ~ NB(μ, δ)

Log-link

Generalized Linear Modeling
Log(μ) = a + β × Group

Visual assessment
suggests a trend

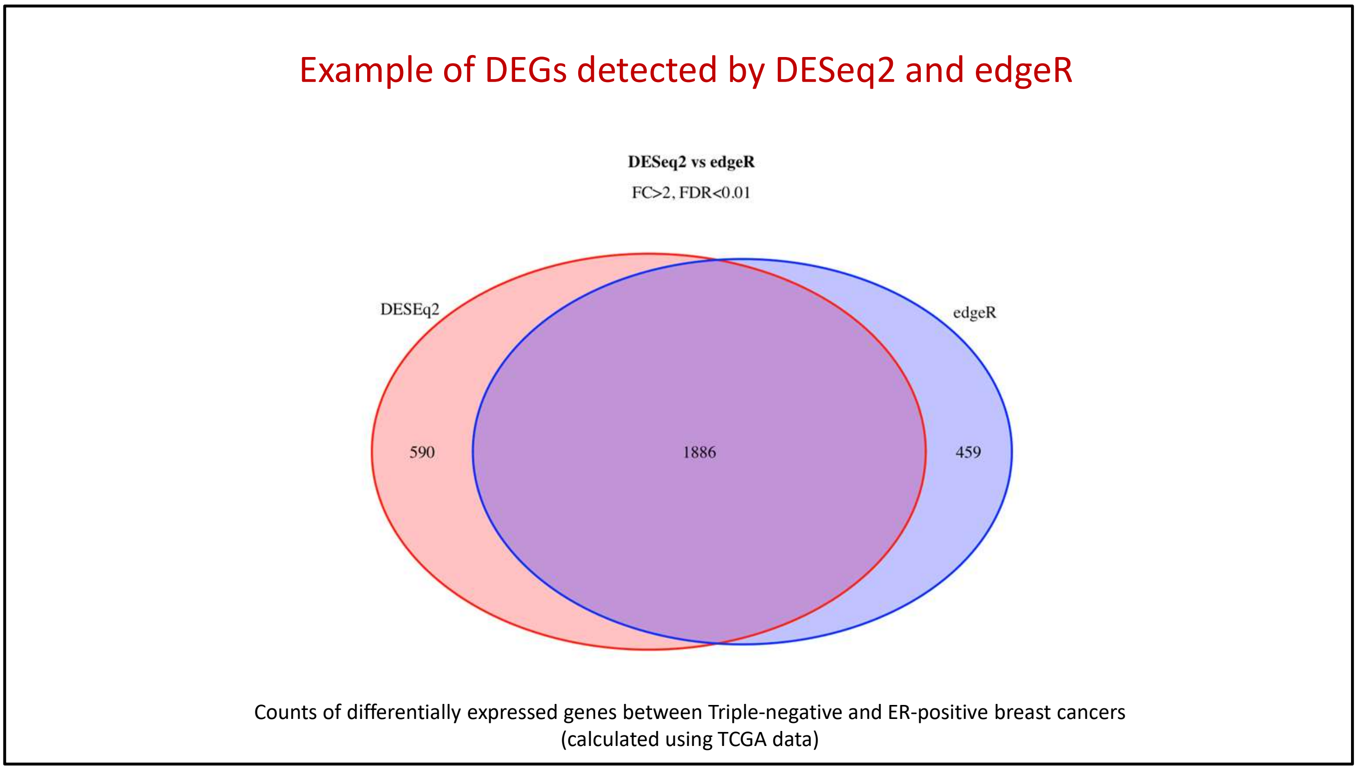
“Borrow” data for
“adaptive” variance evaluation:
for low number of cases
δ(*gene*) shrinks to
δ(*all genes with similar expression*)

Significance of β ≠ 0
Convenient adding covariates (confounders):
Counts = a + β1 × Batch + β0 × Group

This is a summary of the statistical approaches implemented in DESeq2:

As we discussed,
- DESeq2 models the counts with Negative Binomial distribution, and
- It adjusts the variance in genes by shrinking it to the mean

Then, using a technique called “Generalized Linear Modelling” DESeq2 implements the significance testing using the regression framework to allow the correction for confounders. For instance, a batch correction can be done just by a simple inclusion of the batch variable into the model.



Example of DEGs detected by DESeq2 and edgeR

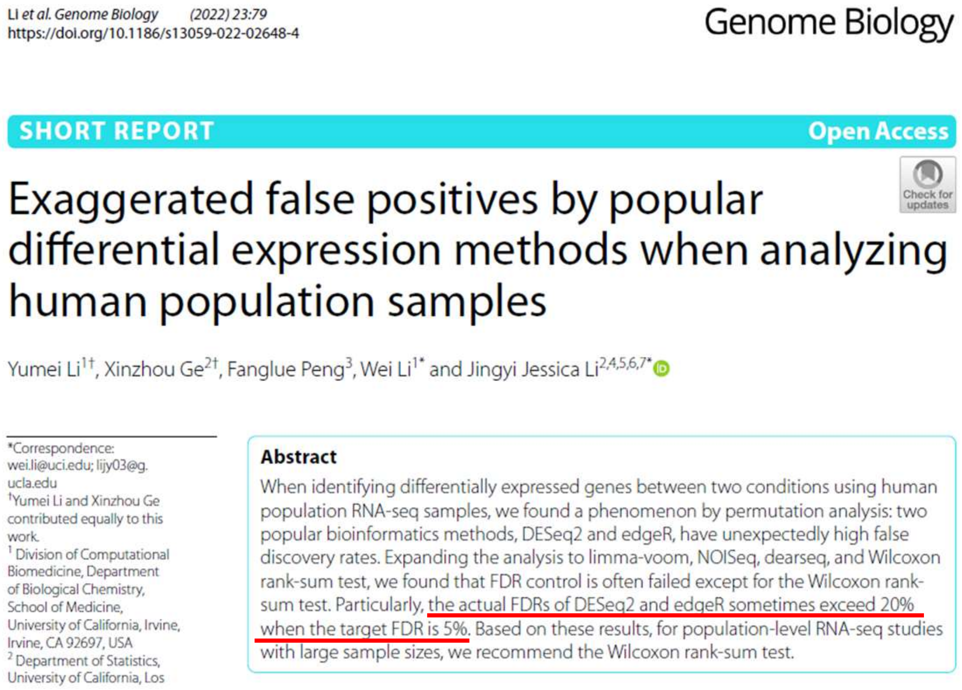
Counts of differentially expressed genes between Triple-negative and ER-positive breast cancers (calculated using TCGA data)

Finally, it should be noted that edgeR uses methods similar to DESeq2, with just some minor technical differences in the normalization and in the dispersion adjustment.

This plot illustrates the concordance between DESeq2 and edgeR. As you can see, the majority of genes are detected by both methods in this example.

At the same time, you can see that even with quite stringent settings (FC>2, FDR<0.01) about a quota of genes in each method does not overlap with the other.

A pinch of salt …



Li et al. Genome Biology (2022) 23:79
https://doi.org/10.1186/s13059-022-02648-4

Genome Biology

SHORT REPORT

Open Access

Exaggerated false positives by popular differential expression methods when analyzing human population samples

Yumei Li[1†], Xinzhou Ge[2†], Fanglue Peng[3], Wei Li[1*] and Jingyi Jessica Li[2,4,5,6,7*]

*Correspondence: wei.li@uci.edu; lijy03@g.ucla.edu
†Yumei Li and Xinzhou Ge contributed equally to this work.
[1] Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA
[2] Department of Statistics, University of California, Los

Abstract

When identifying differentially expressed genes between two conditions using human population RNA-seq samples, we found a phenomenon by permutation analysis: two popular bioinformatics methods, DESeq2 and edgeR, have unexpectedly high false discovery rates. Expanding the analysis to limma-voom, NOISeq, dearseq, and Wilcoxon rank-sum test, we found that FDR control is often failed except for the Wilcoxon rank-sum test. Particularly, the actual FDRs of DESeq2 and edgeR sometimes exceed 20% when the target FDR is 5%. Based on these results, for population-level RNA-seq studies with large sample sizes, we recommend the Wilcoxon rank-sum test.

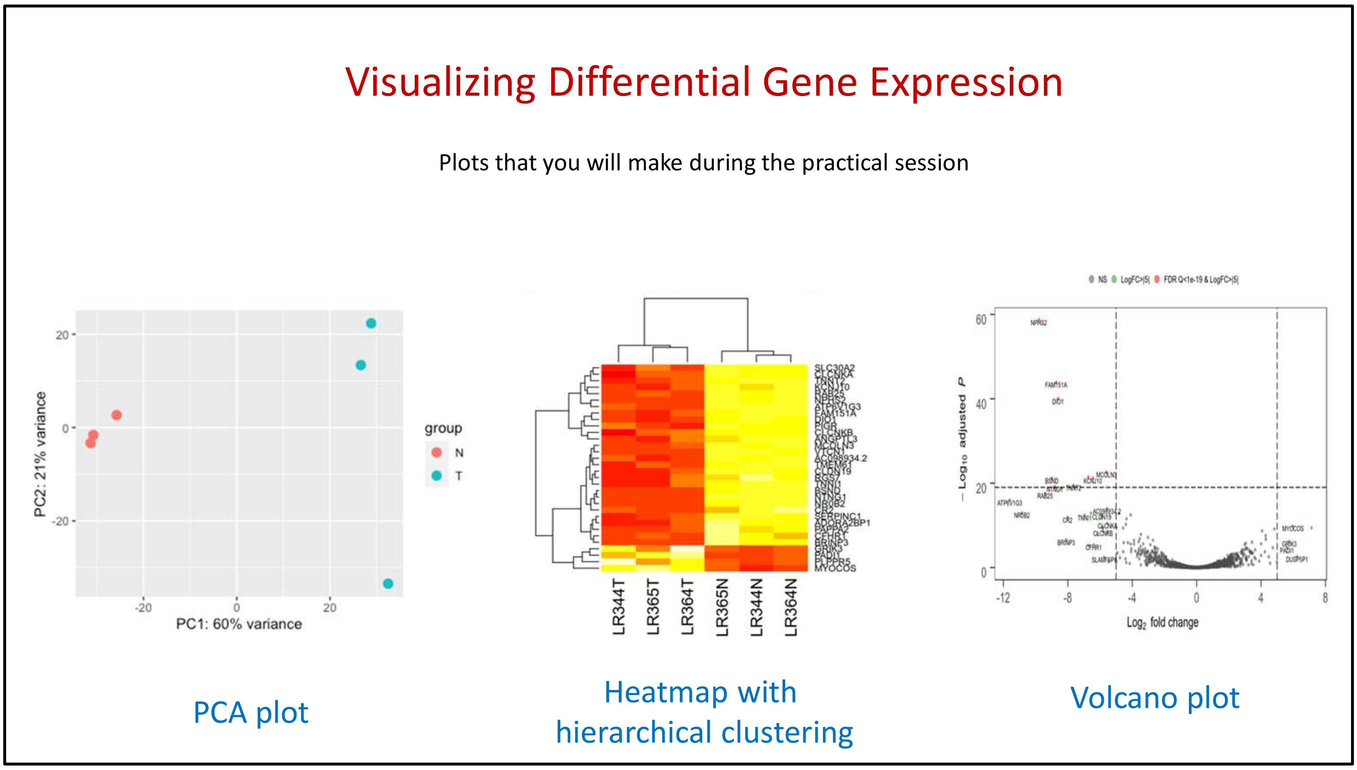
This brings us to the last point about statistics in RNAseq DGE analysis.

The described previously approaches represent a powerful and mature framework that has been used for many years, and obtained many valid results confirmed by independent follow-up studies.

At the same time, some key elements in this framework were developed at the time when RNA-seq was very expensive. There was a pressure to get meaningful results from as little data as possible. So, some assumptions underlying statistical methods (e.g. the dispersion shrinkage) were a compromise dictated by feasibility.

Now, when RNA-seq is more affordable, using larger numbers of biological replicas and higher depth of sequencing is becoming feasible. This allows using more conventional non-parametric statistical methods.

Of course, as with any experimental method, the key DGE findings should be validated by orthogonal techniques (e.g. by NanoString or qRT-PCR).



Visualizing Differential Gene Expression

Plots that you will make during the practical session

PCA plot

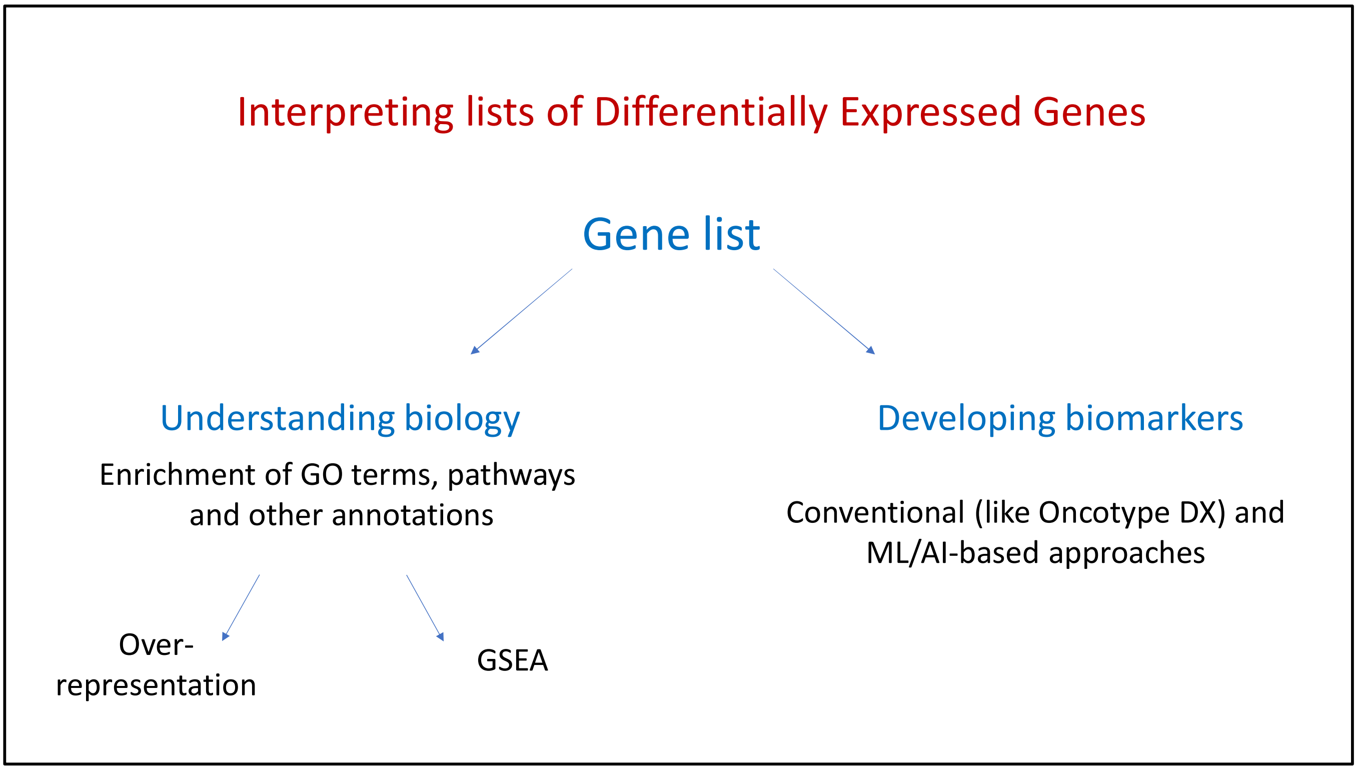
Heatmap with hierarchical clustering

Volcano plot

One picture says a thousand words ☺

These are the most common visualization techniques used in the gene expression analysis:
- PCA
- Heatmaps
- Volcano plots

We will make some of such plots during our practical session.



Interpreting lists of Differentially Expressed Genes

Gene list

Understanding biology

Enrichment of GO terms, pathways and other annotations

Over-representation

GSEA

Developing biomarkers

Conventional (like Oncotype DX) and ML/AI-based approaches

What can be done with the gene lists obtained by the Differential Gene Expression (DGE) analysis?

The two most common tasks are
1) Biological/Functional interpretation of the DGE analysis results
2) Using the Differentially Expressed Genes (DEGs) for biomarker development

Biomarker development doesn't have to be based on the DGE analysis, and it is a very wide field that is beyond this introductory session.

So, here we will only discuss the biological/functional interpretation of the gene lists. Two popular ways of dung this are ***Over-representation analysis*** and ***Gene Set Enrichment Analysis*** (GSEA).

Functional annotations of the genes

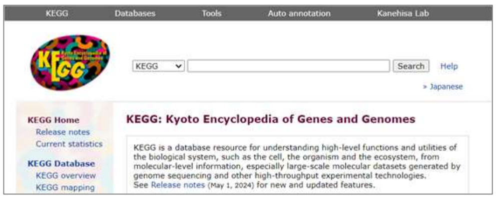
Gene Ontology

Current release 2024-06-17: 42,093 GO terms | 7,624,863 annotations
1,537,645 gene products | 5,409 species (see statistics)

THE GENE ONTOLOGY RESOURCE

https://geneontology.org

KEGG Pathways



KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
See Release notes (May 1, 2024) for new and updated features.

https://www.genome.jp/kegg

etc

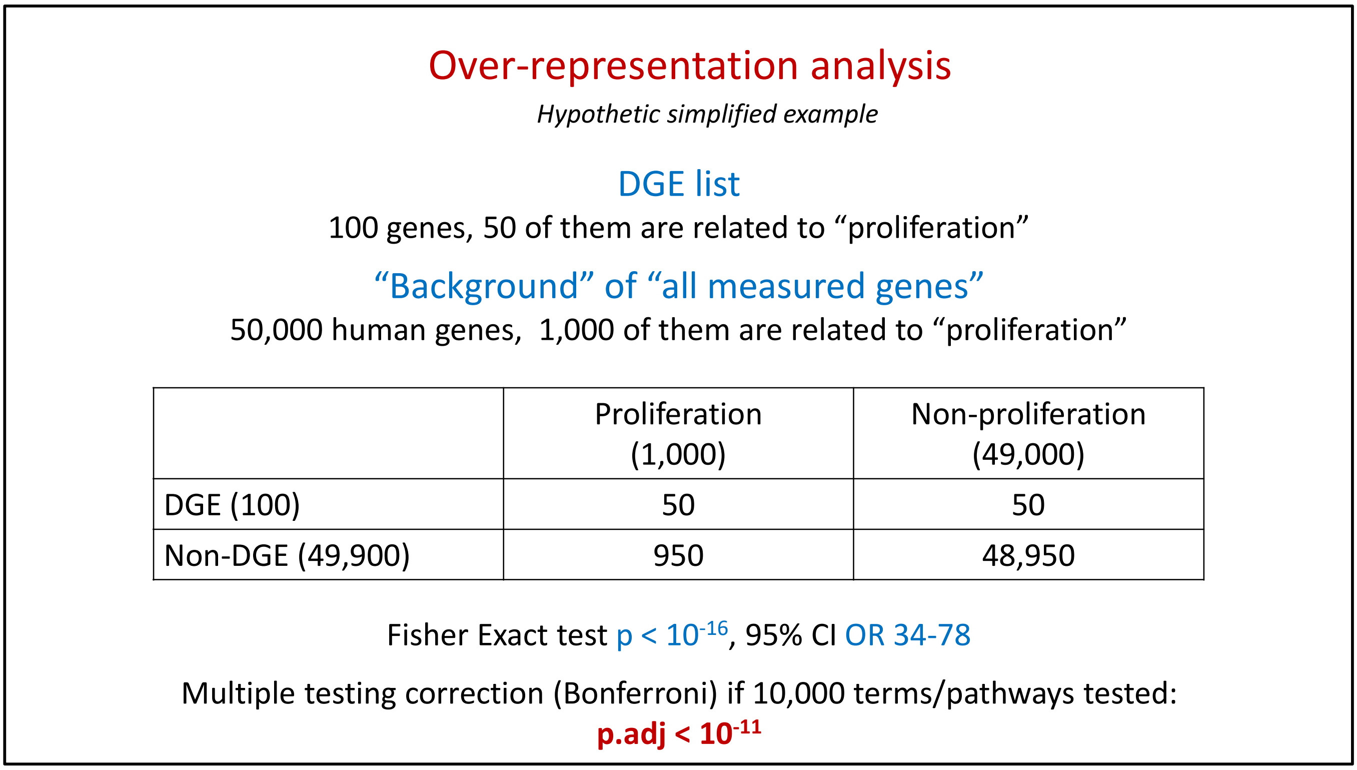
Reactome

https://reactome.org

MSigDB

https://www.gsea-msigdb.org/gsea/msigdb

Both Over-representation analysis and GSEA depend on pre-existing functional annotations of individual genes. There are many databases describing gene functions. Some very popular examples include Gene Ontology, KEGG, Reactome, MSigDB etc. You may explore these resources following the links provided on the slide.



Over-representation analysis

Hypothetic simplified example

DGE list

100 genes, 50 of them are related to “proliferation”

“Background” of “all measured genes”

50,000 human genes, 1,000 of them are related to “proliferation”

	Proliferation (1,000)	Non-proliferation (49,000)
DGE (100)	50	50
Non-DGE (49,900)	950	48,950

Fisher Exact test $p < 10^{-16}$, 95% CI OR 34-78

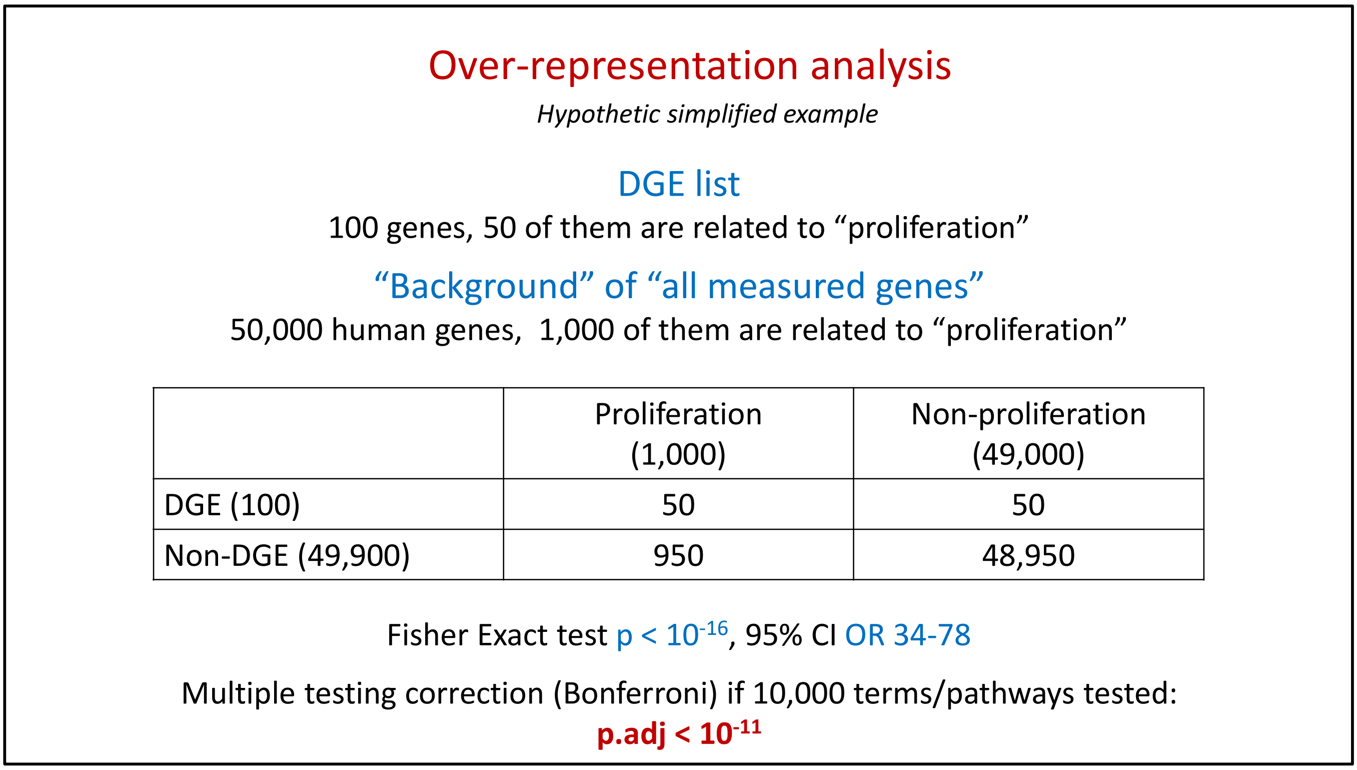
Multiple testing correction (Bonferroni) if 10,000 terms/pathways tested:
$p.adj < 10^{-11}$

This slide illustrates the over-representation analysis.

Assume that you obtained 100 differentially expressed genes, for instance between cancer and normal cells.
Then assume that 50 of these genes are related to proliferation.
Can you conclude that the proliferation genes are overrepresented in the DGE list?
(in other words: “is this DGE list enriched by proliferation genes?”)

It depends on whether the proportion of proliferation genes is higher within DEGs than within the background list of all the genes included into the analysis.

This can be assessed by a simple Fisher exact test (as shown on the slide), or by other appropriate statistical tests.



Over-representation analysis

Hypothetic simplified example

DGE list

100 genes, 50 of them are related to “proliferation”

“Background” of “all measured genes”

50,000 human genes, 1,000 of them are related to “proliferation”

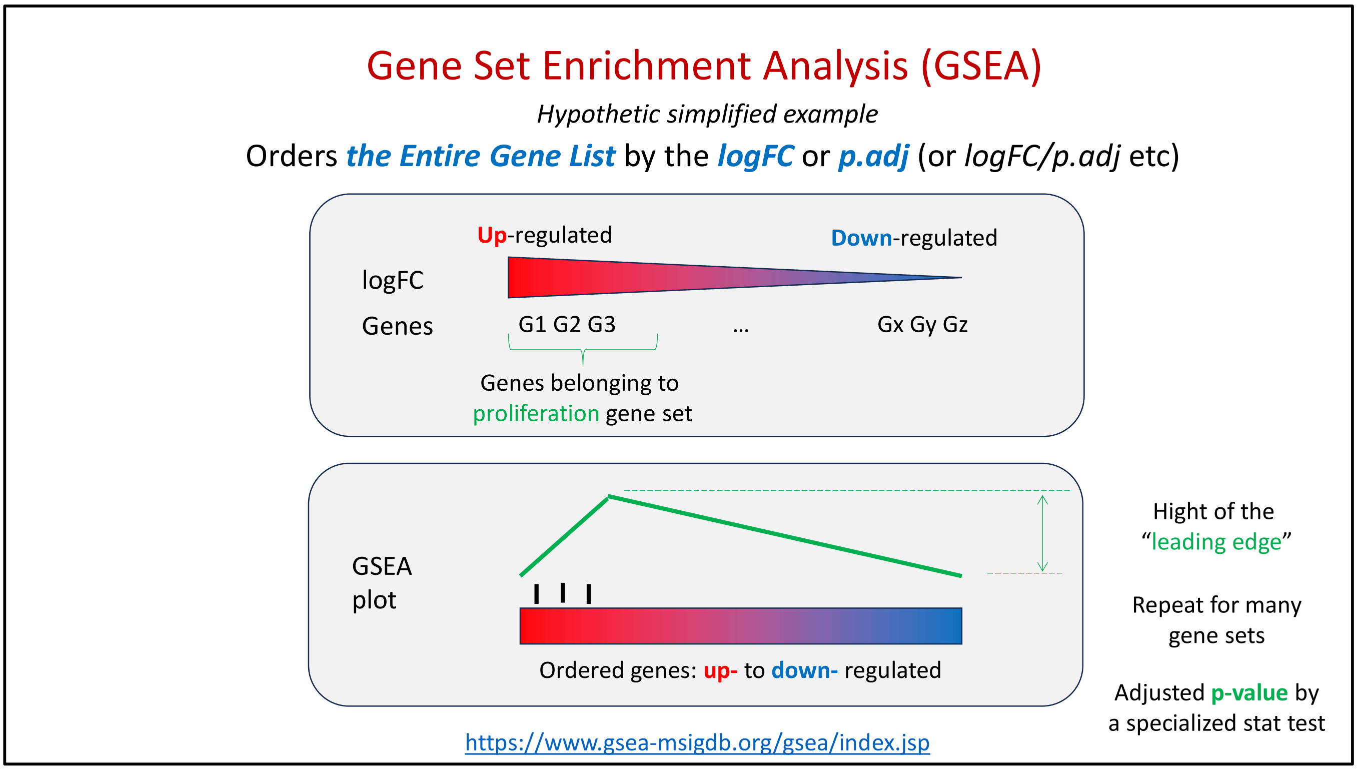
	Proliferation (1,000)	Non-proliferation (49,000)
DGE (100)	50	50
Non-DGE (49,900)	950	48,950

Fisher Exact test $p < 10^{-16}$, 95% CI OR 34-78

Multiple testing correction (Bonferroni) if 10,000 terms/pathways tested:
p.adj < 10^{-11}

Often, we test against many hypotheses, e.g. not only against the proliferation genes, but also against genes involved in apoptosis, glycolysis etc. Of course, in such case we must apply a multiple testing correction.

An important question is what to take as the “background” genes, against which we estimate the over-representation. The full genes list produced by DESeq2 could be a good starting point (of course, only including the genes present in the selected annotation database). Alternatively, you may use the whole list of the annotated genes from the database.



Gene Set Enrichment Analysis (GSEA)

Hypothetic simplified example

Orders ***the Entire Gene List*** by the ***logFC*** or ***p.adj*** (or *logFC/p.adj* etc)

Up-regulated
Down-regulated
logFC
Genes
G1 G2 G3
…
Gx Gy Gz
Genes belonging to proliferation gene set
GSEA plot
Ordered genes: **up-** to **down-** regulated

Hight of the "leading edge"

Repeat for many gene sets

Adjusted **p-value** by a specialized stat test

https://www.gsea-msigdb.org/gsea/index.jsp

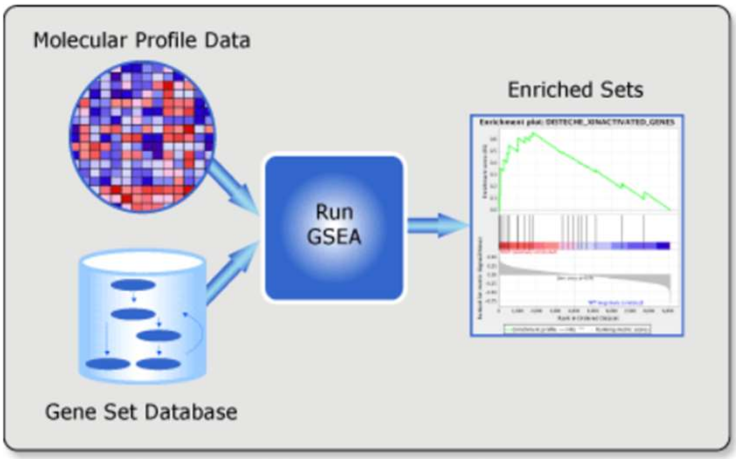
The Gene Set Enrichment Analysis (GSEA) uses the entire list of genes produced by DESeq2, not only the Differentially Expressed Genes.

First the entire list of genes is ordered by the fold change (or by the p-value, or by some index combining FC and p-value).

Then, the position of the proliferation genes is evaluated within the ordered gene list. If all proliferation genes are at the top of the ordered list, we may suggest that proliferation genes are up-regulated in our experiment. Special plotting and statistical tests allow to visualize the GSEA results and estimate the significance.

Of course, the multiple testing correction should be applied when testing against many gene sets.

Gene Set Enrichment Analysis (GSEA)

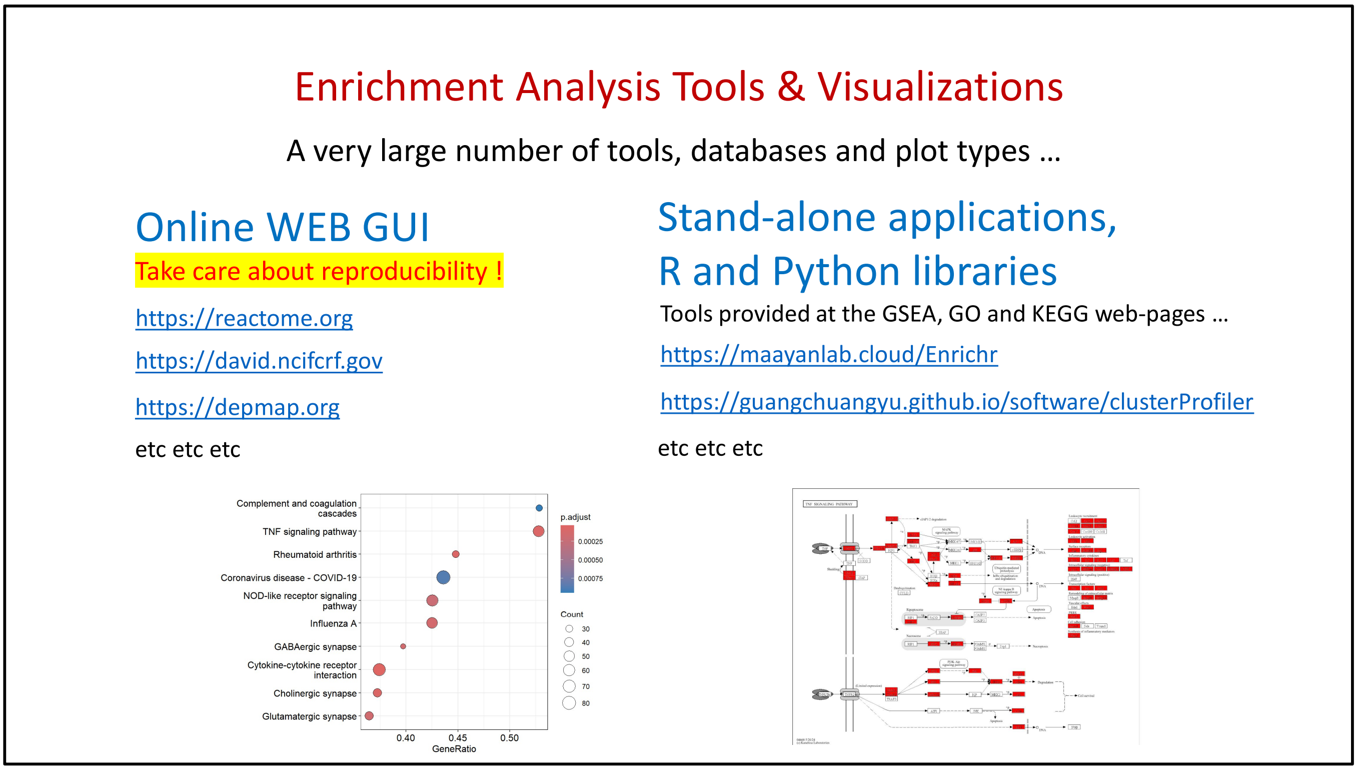


MSigDB

https://www.gsea-msigdb.org/gsea/index.jsp

You may read about the details of GSEA implementation on the method's web site.

Importantly, the GSEA web site also provides one of the most comprehensive databases, including many known Molecular Signatures and genes annotations.



Enrichment Analysis Tools & Visualizations

A very large number of tools, databases and plot types …

Online WEB GUI

Take care about reproducibility !

https://reactome.org

https://david.ncifcrf.gov

https://depmap.org

etc etc etc

Stand-alone applications, R and Python libraries

Tools provided at the GSEA, GO and KEGG web-pages …

https://maayanlab.cloud/Enrichr

https://guangchuangyu.github.io/software/clusterProfiler

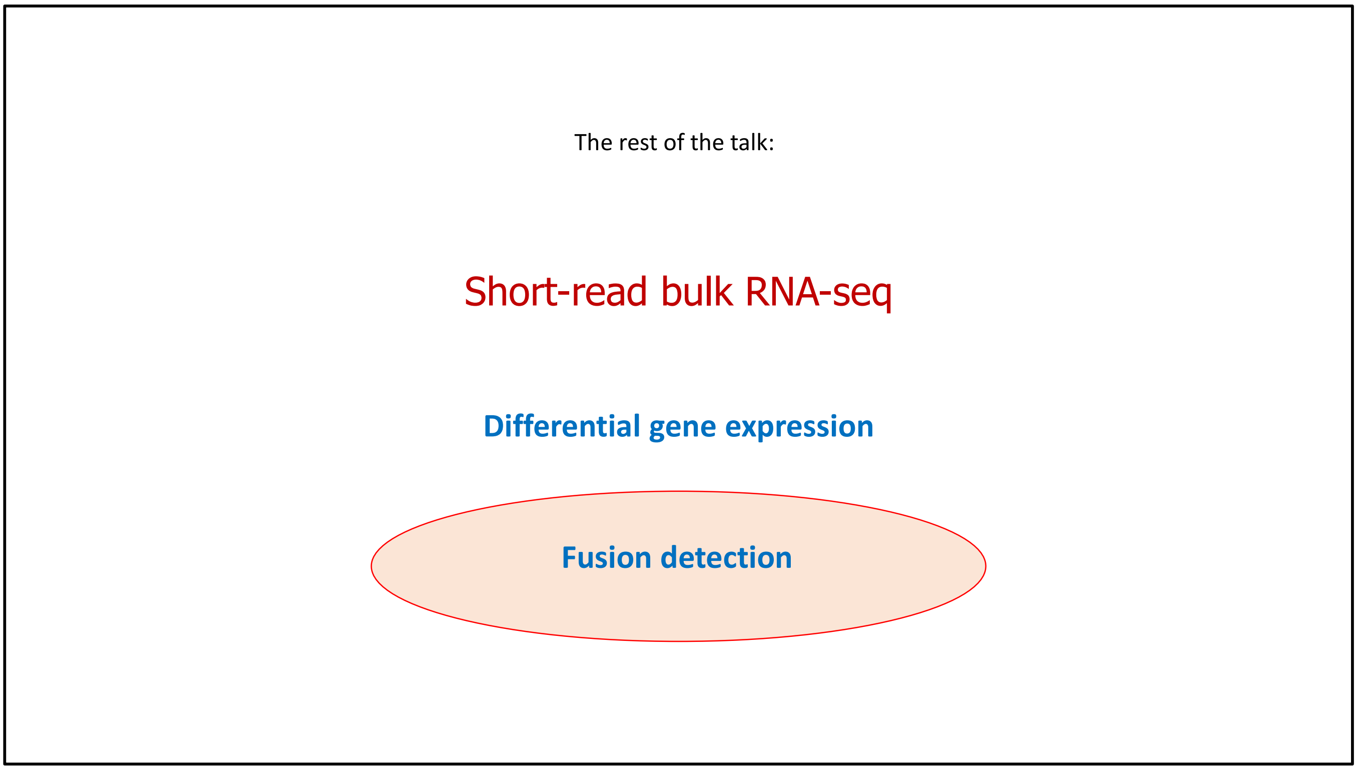
etc etc etc

There are many online tools and stand-alone software that can be used for the enrichment analysis.

Using interactive WEB tools, one should pay special attention to the versions of the used resources, selected thresholds etc. Otherwise, it may be difficult to reproduce the results.

The stand-alone tools or software packages usually provide logs (and scripts) required for the reproducible research.

Similarly, there are many ways of visualizing enrichment analysis. Here you can see a couple of plots that will be used during the practical session. The left dot-plot shows the degree of enrichment, p-value and the number of genes per set/pathway using position, size and color of the dots. The right plot illustrates a KEGG pathway, with the enriched genes highlighted in red.



The rest of the talk:

Short-read bulk RNA-seq

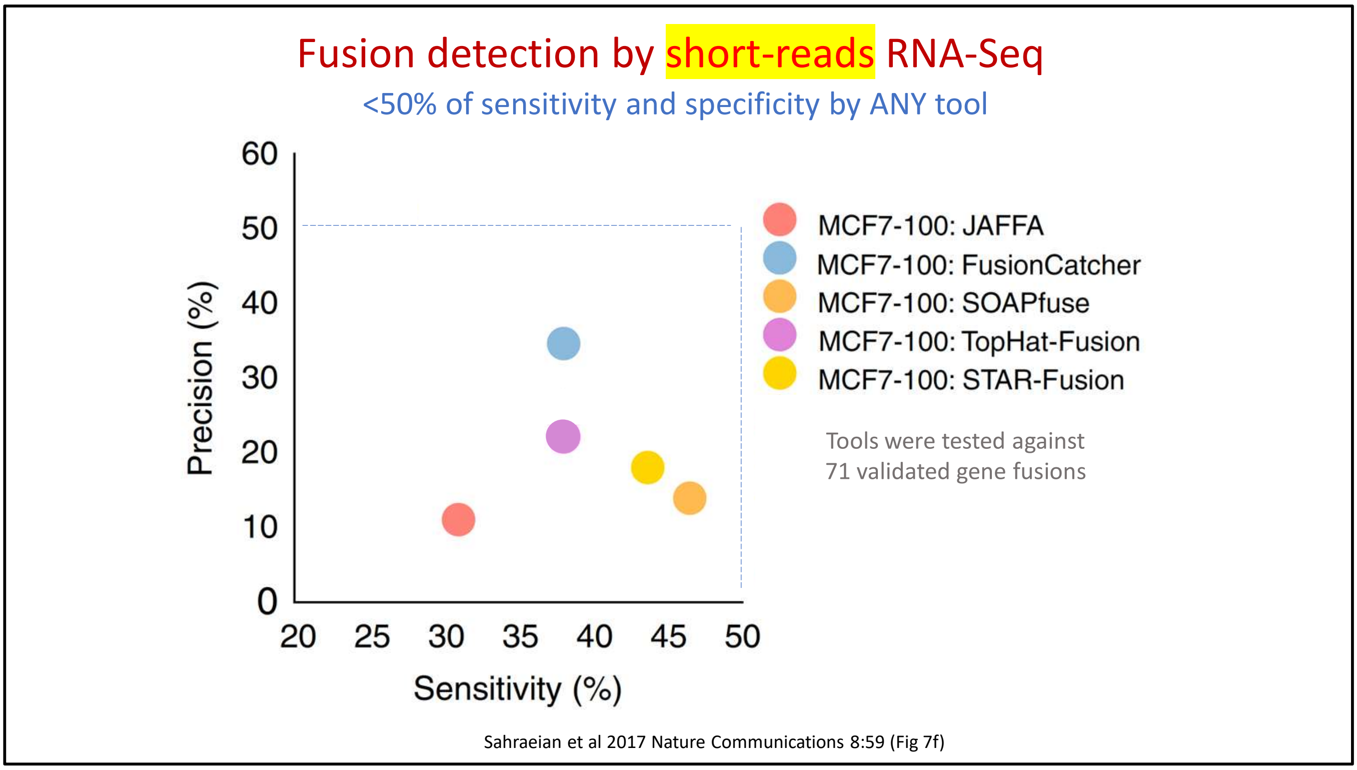
Differential gene expression

Fusion detection

This concludes the overview of the gene expression analysis in bulk short-read RNA-seq data.

Because of the high clinical relevance in oncology, I also will mention here fusion detection.

However, nowadays the fusion detection should be done using the long reads sequencing.



Fusion detection by short-reads RNA-Seq

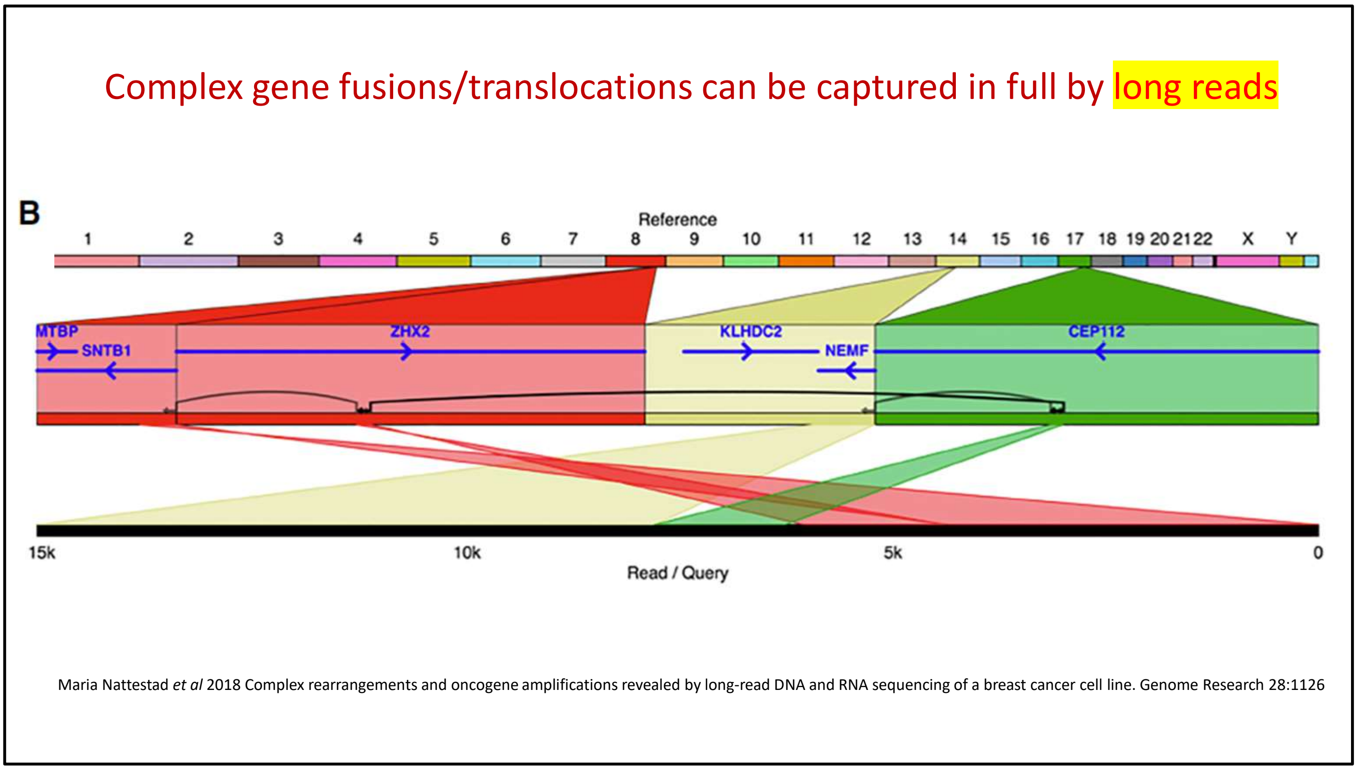
<50% of sensitivity and specificity by ANY tool

Tools were tested against 71 validated gene fusions

Sahraeian et al 2017 Nature Communications 8:59 (Fig 7f)

Although many clever people did their best in the fusion detection from the short reads, you can see that, out of the 5 top tools for fusion detection, none could achieve even 50% of sensitivity or precision in this study.

Of course, it could be slightly improved by deeper sequencing, and in our practical session we will run an example, when a clinically relevant fusion is detected from the short reads by STAR-fusion.

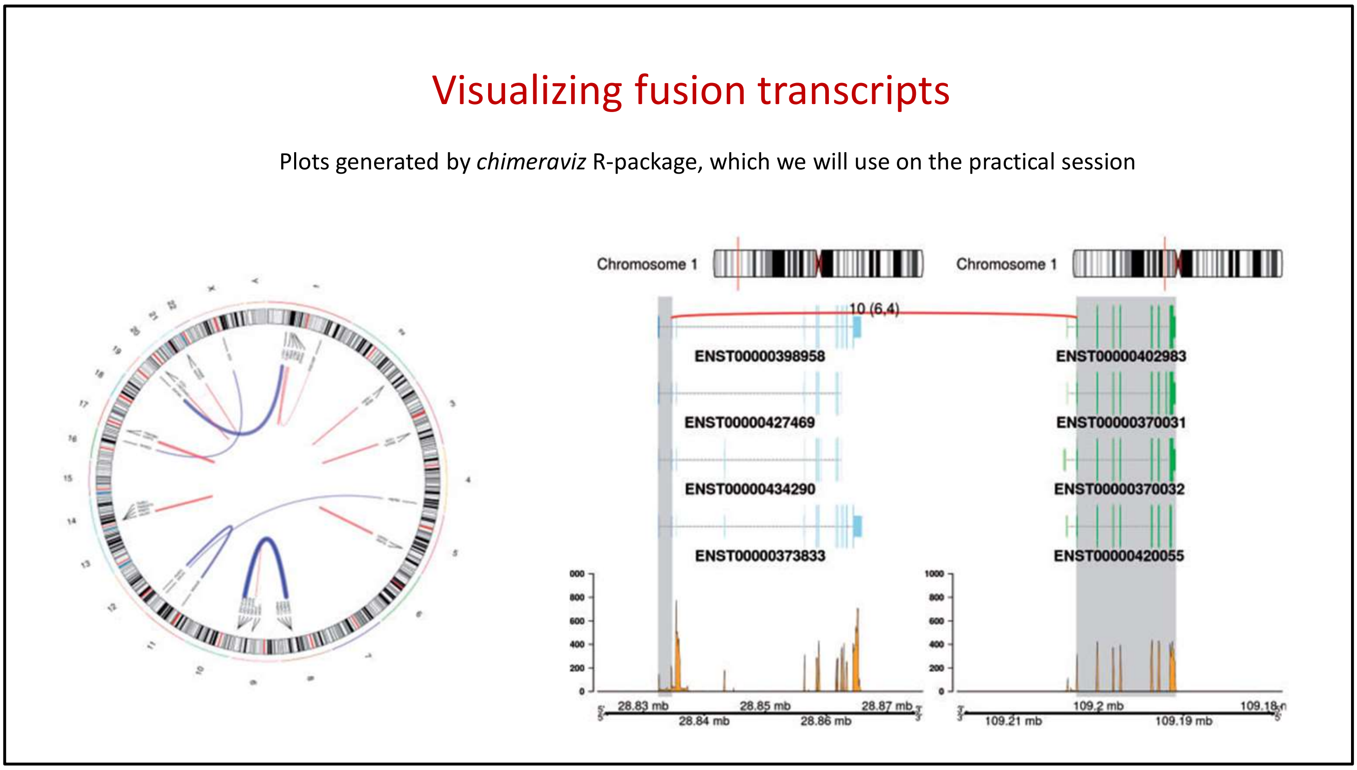


Complex gene fusions/translocations can be captured in full by long reads

Maria Nattestad *et al* 2018 Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Research 28:1126

However, the long reads open a whole new chapter in fusion (and other complex rearrangements) detection.

This field is growing as we speak, and I will not discuss it further in this lecture, focused on the short-reads data analysis.

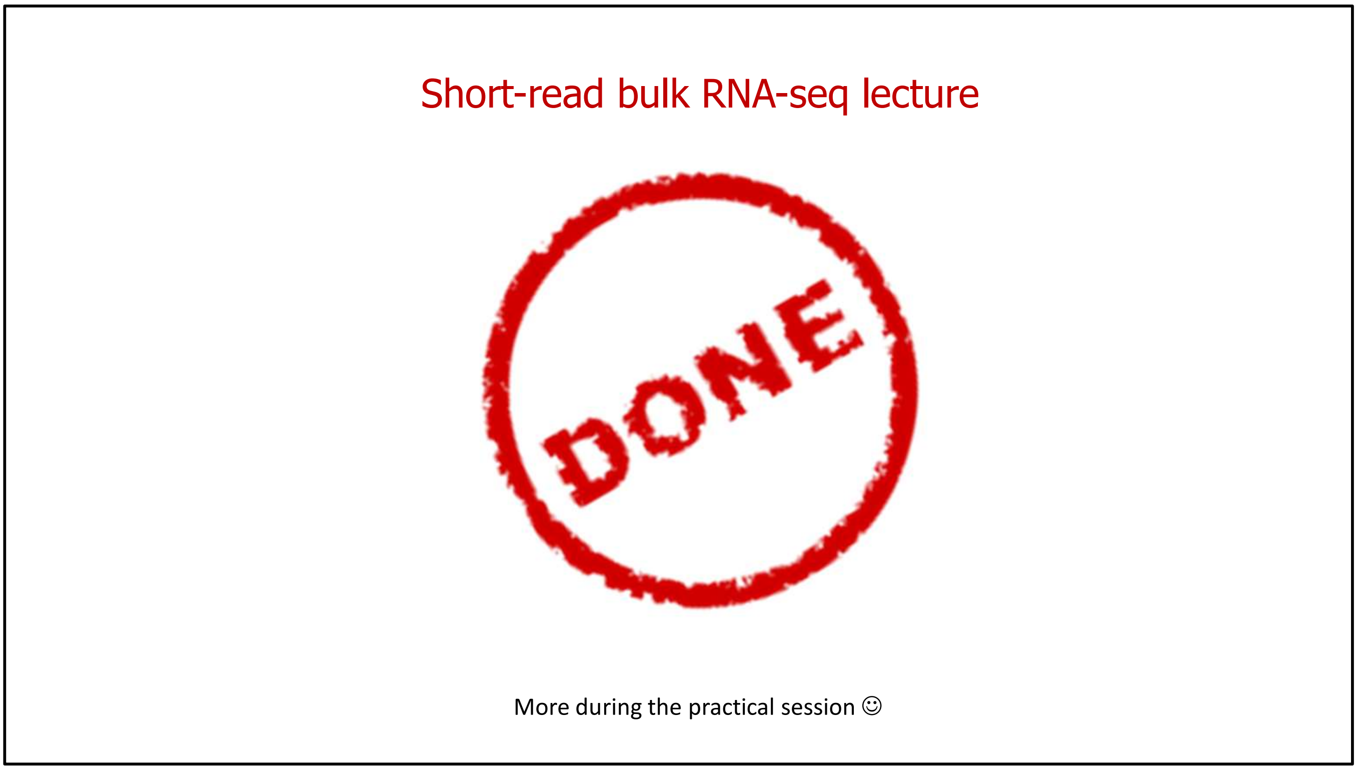


Visualizing fusion transcripts

Plots generated by *chimeraviz* R-package, which we will use on the practical session

Many types of plots are used to visualise gene fusions.

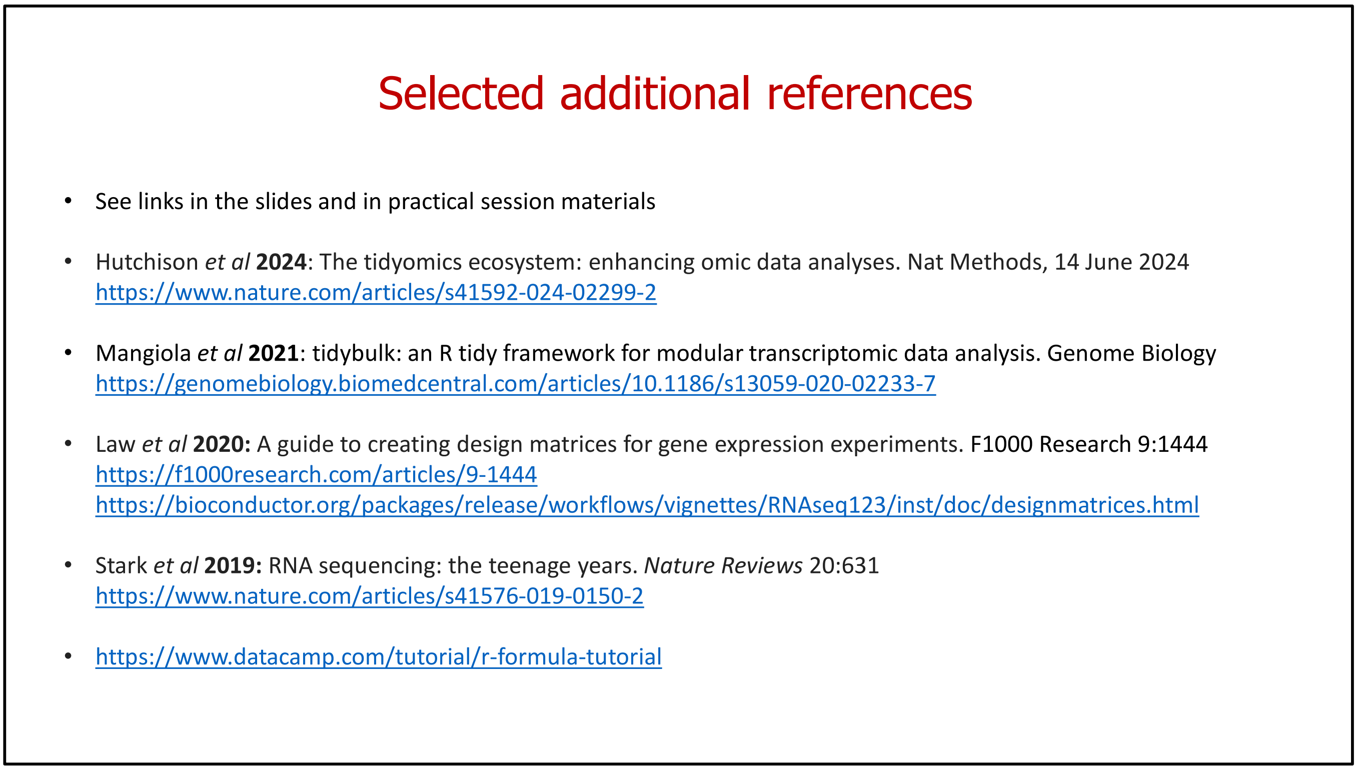
In our practical session we will use an R package called Chimraviz, which allows to show the detected fusions in a Circus plot (and in some other types plots).



Short-read bulk RNA-seq lecture

More during the practical session ☺

And this is it for this lecture.

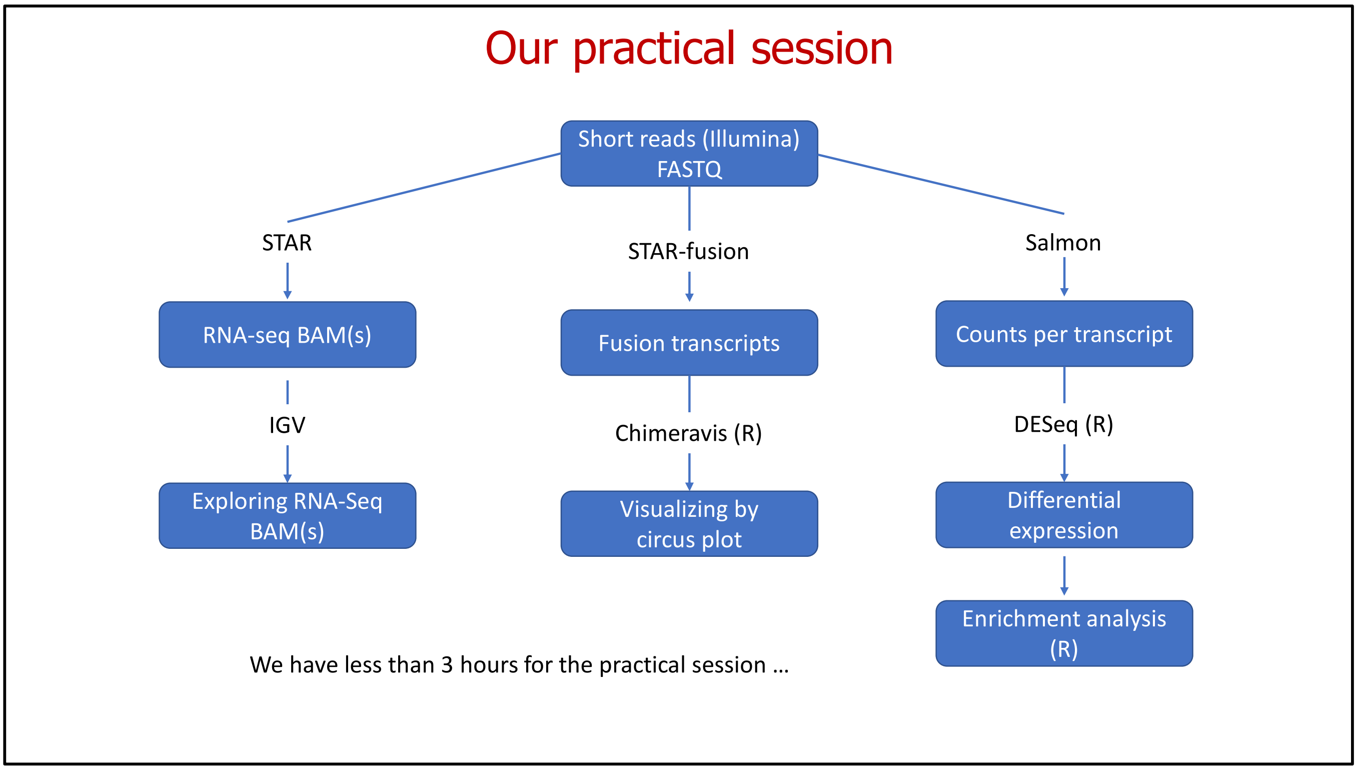


Selected additional references

- See links in the slides and in practical session materials
- Hutchison *et al* **2024**: The tidyomics ecosystem: enhancing omic data analyses. Nat Methods, 14 June 2024
  https://www.nature.com/articles/s41592-024-02299-2
- Mangiola *et al* **2021**: tidybulk: an R tidy framework for modular transcriptomic data analysis. Genome Biology
  https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02233-7
- Law *et al* **2020:** A guide to creating design matrices for gene expression experiments. F1000 Research 9:1444
  https://f1000research.com/articles/9-1444
  https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/designmatrices.html
- Stark *et al* **2019:** RNA sequencing: the teenage years. *Nature Reviews* 20:631
  https://www.nature.com/articles/s41576-019-0150-2
- https://www.datacamp.com/tutorial/r-formula-tutorial

Many references are available directly in the slides and in practical session martials.

These are just some selected additional references, showing a broader context, and touching upon selected points that we haven’t had time to discuss during the lecture.



Our practical session

Short reads (Illumina) FASTQ

STAR

RNA-seq BAM(s)

IGV

Exploring RNA-Seq BAM(s)

STAR-fusion

Fusion transcripts

Chimeravis (R)

Visualizing by circus plot

Salmon

Counts per transcript

DESeq (R)

Differential expression

Enrichment analysis (R)

We have less than 3 hours for the practical session …

Practical session plan.