Introduction to Transcriptomics (*RNA-seq*)

Natalia Rego

Bioinformatics Unit, Institut Pasteur de Montevideo
Evolutionary Genomics Lab, Faculty of Sciences, UdelaR
natalia@pasteur.edu.uy

Fundamentos y Herramientas Bioinformáticas para Análisis Genómicos, 2025

From the GENOME to the TRANSCRIPTOME

FROM THE GENOME TO THE TRANSCRIPTOME

GENOME

- The complete DNA sequence of an organism
- Relatively static; the seme in (almost) every cell
- Encodes *the potential* for gene expression

TRANSCRIPTOME

- The complete set of RNA molecules expressed in a cell or tissue at given time
- Dynamic; changes with cell type, condition, and environment

TRANSCRIPTOMICS is the study of the transcriptome – the collection of all RNA transcripts – to understand which genes are active, when, and to what extent

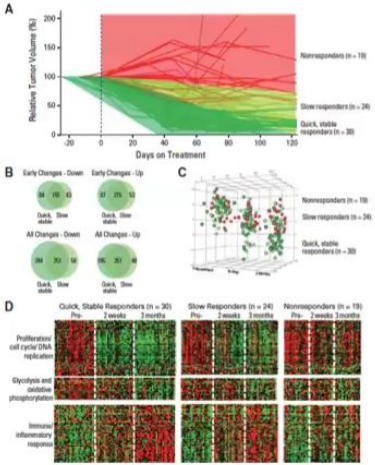
RNA-SEQ

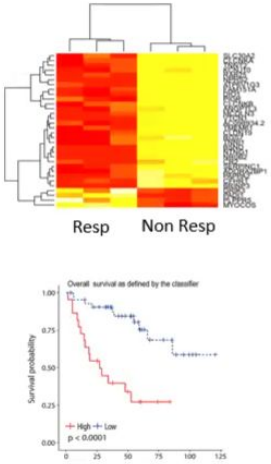
RNA sequencing uses high-throughput sequencing to quantify RNA molecules, detect novel transcripts, and compare expression patterns across samples

What question are we really asking when we perform an RNA-seq experiment?

Development of a gene expression signature for cancer diagnosis & treatment

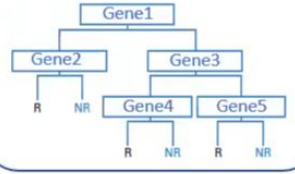
Exploration ↔ Choose informative genes ↔ Choose and tune classification algorithm ↔ Validation ↔ Implementation





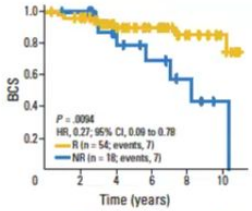
Linear scores =
3.5*Gene1 - 4.8*Gene2 +
5.0*Gene3 + 10.4*Gene4 ...

Decision trees

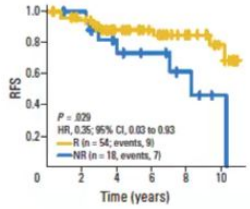


AI algorithms: Regression, SVM, RF, ANN, ...

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

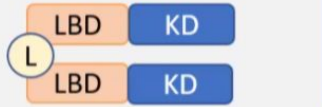
Independent
dataset(s),
clinical impact

RT-PCR
NanoStrinq
RNA-seq ??

Fusions drive the development of 16,5% cancer cases

Constitutively activating signaling

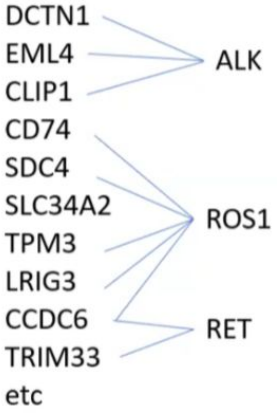
Ligand-mediated dimerization and activation



Fusion-partner mediated (dimerization and) activation

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

Lung cancer

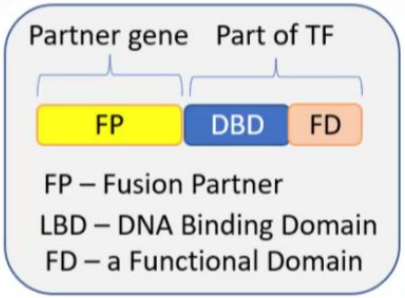


Chronic myeloid leukaemia

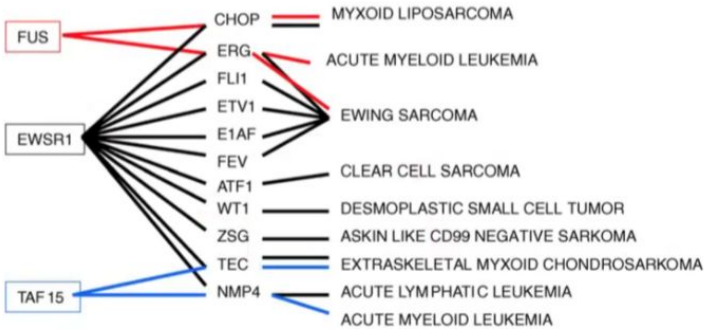
BCR — ABL

Making an aberrant highly expressed transcription factor

Prostate cancer



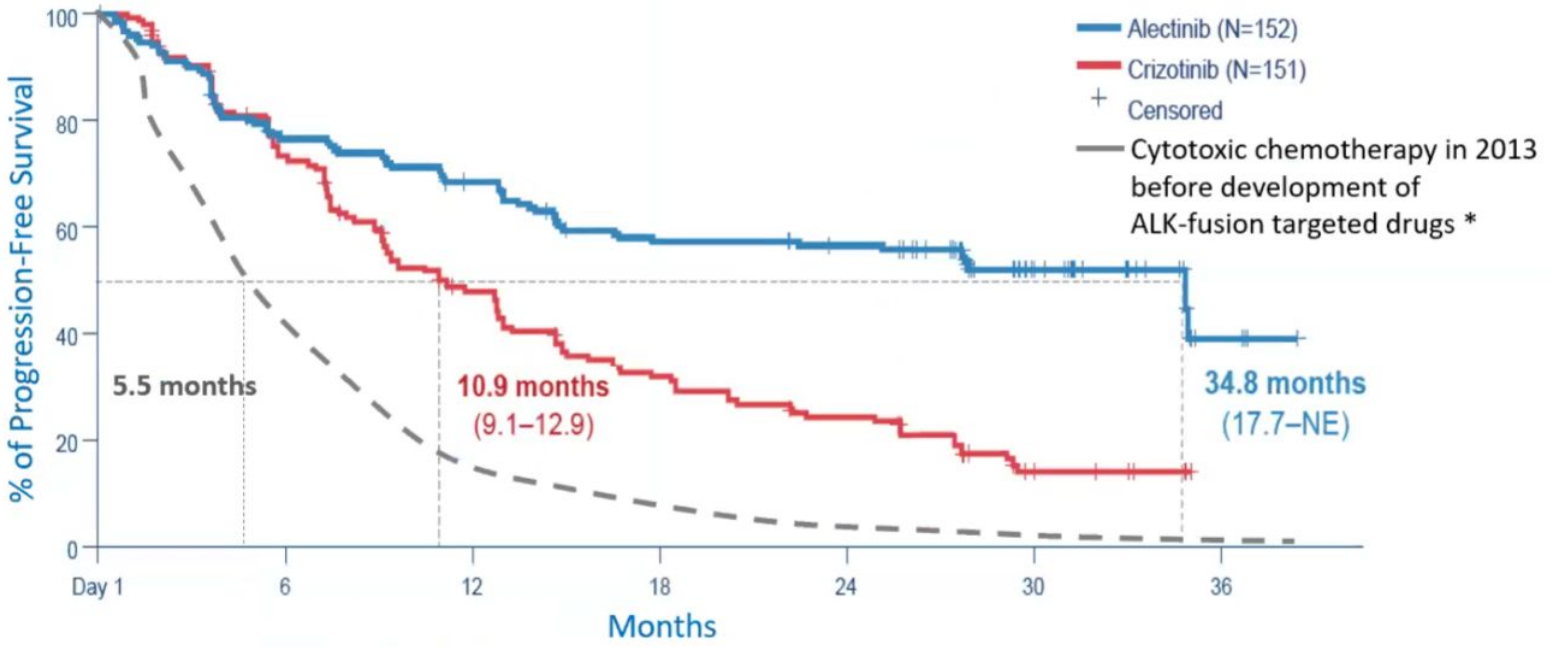
Ewing's sarcoma family of fusions



The TCGA et al 2018

Fusions drive the development of 16,5% cancer cases

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



Laporte et al, 2013; Camidge et al, 2018

What am I going to talk about today?

Flavours of Transcriptomics

SHORT-READ (Illumina) → **LONG-READ (ONT, PacBio)**

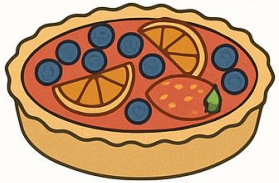
BULK RNA-SEQ

- Still contains plenty of biological (and clinical) information despite analyzing a mix of cells
- Can measure even low-expressed genes

SINGLE-CELL RNA-SEQ

- Exciting new biology at a cell level (comes with new exciting challenges)
- Only abundant RNAs may be reliably measured

SPATIAL



Can reveal the spatial organization of gene expression

A RNA-seq lecture may include many topics…

sequencing technologies
*Illumina
*ONT, PacBio

quantification of gene expression
*strategies after alignment of alignment-free
*statistics and normalisation
*differential gene expression, transcripts, exons

basic principles and techniques
*study design and power calculation
*RNA quality assessment (agarose gel, Agilent Bioanalyser: RIN, % above 200 bases)
*library preparation (total RNA, mRNA, short RNAs…)
*QC (FastQC, MultiQC, fastp…)
*trimming and preprocessing (Trimmomatic, Cutadapt, BBduk…)
*alignment, transcript assembly and count (to ref. genome, to transcriptome, read counts…)
*evaluating & visualizing RNAseq BAMs (IGV)
*fusion transcripts detection (STAR-fusion…)
*variant calling in RNAseq (GATK, DeepVariant…)

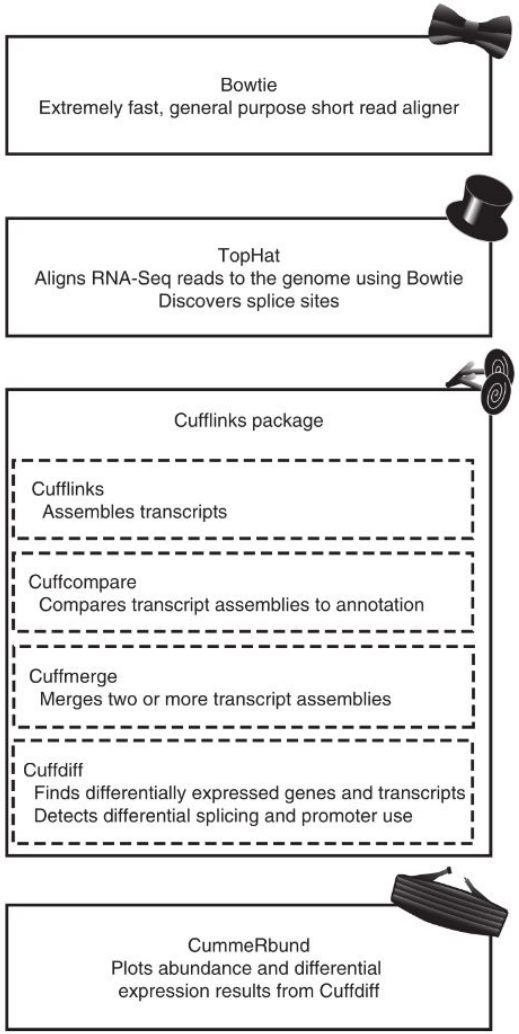
other applications
*allele-specific expression
*alternative splicing…
*eQTLs
*single-cell RNAseq
*small or circular ncRNAs
*RNA-editing

software, file formats, resources
*historic: tuxedo pipeline
*current: nextflow pipelines
*resources: reference genomes (fasta), genome annotations (gtf/gff), Genecode, MSigDB…

Once upon a time, RNA-seq was simple…

The Tuxedo protocol

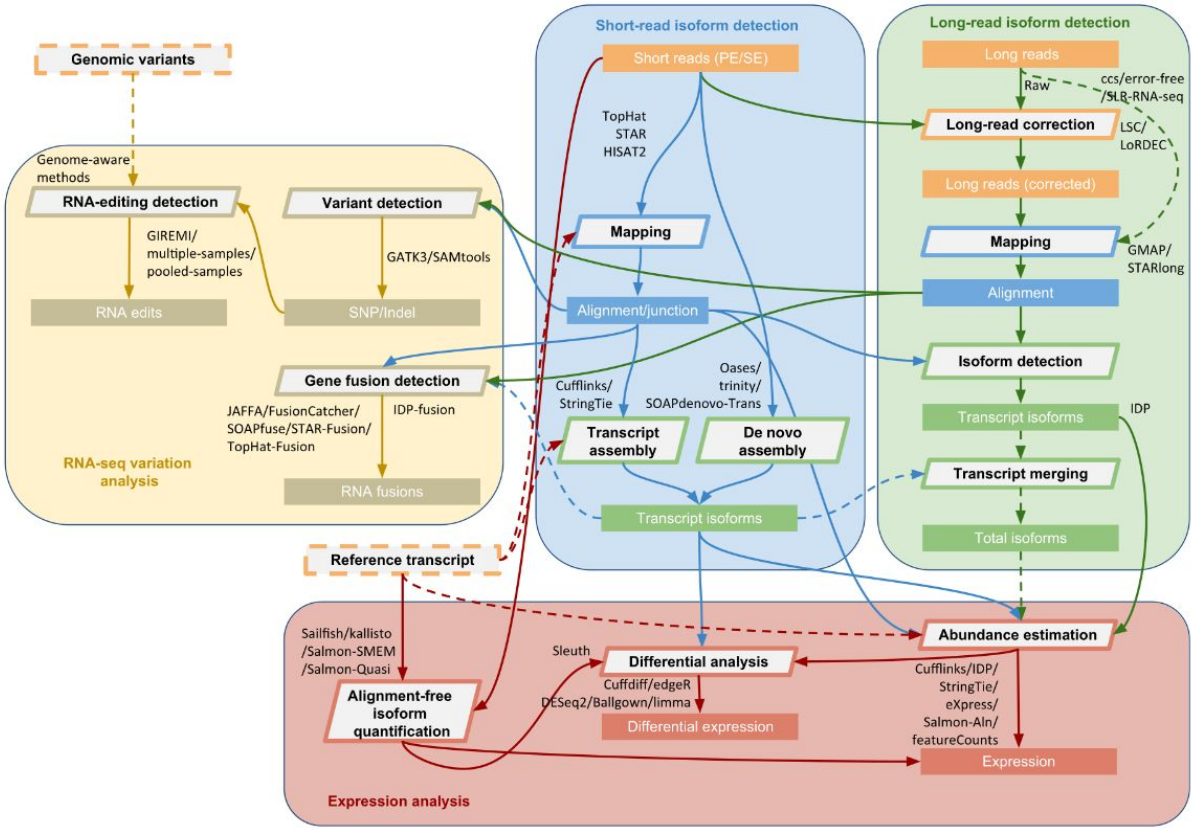
Trapnell et al, 2012



RNACocktail protocol

The authors examine:

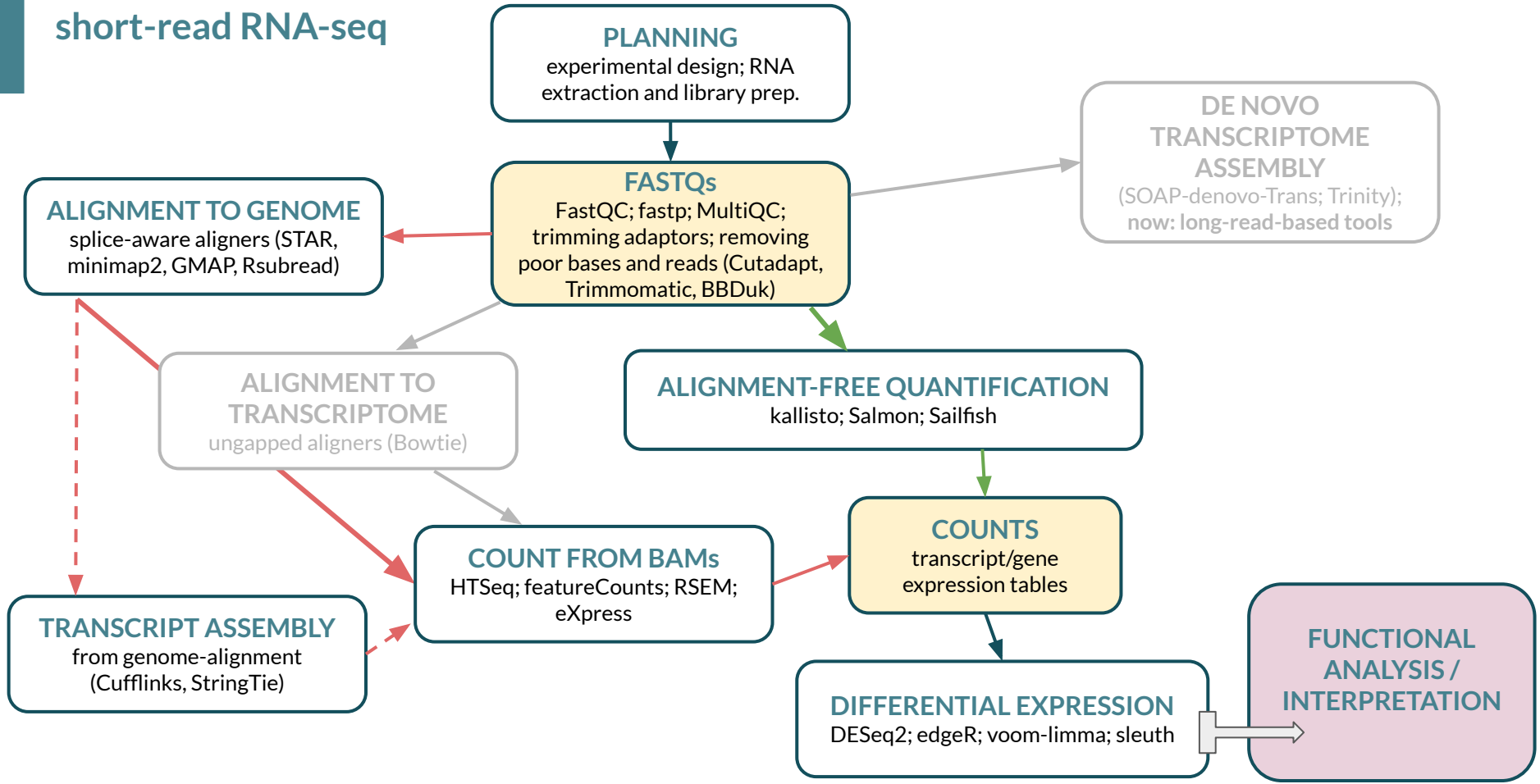
39 analysis tools
~ 120 combinations
~ 490 analysis on 15 different samples



Sahraeian et al, 2017

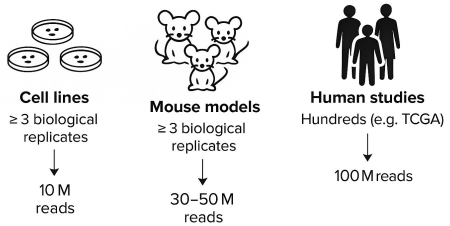
https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools

short-read RNA-seq



PLANNING: typical settings (Illumina)

- **stranded** library
- **100-150 paired-end** libraries
  - shorter reads increases multi-mapping
  - 300 PE too much
- depth of sequencing: **millions of reads per sample** (not coverage)
  - 10 to 100 millions reads
    - 10M may suffice for abundant transcripts
    - 100M may be enough for rare transcripts
    - typically 30-**50M**
    - for long reads, 10-20M reads looks good
- **at least 3 biological replicates per condition/group**
  - cell lines
  - animal models
  - humans (e.g. TCGA hundreds of patients per cancer type)



PLANNING: typical settings (Illumina)

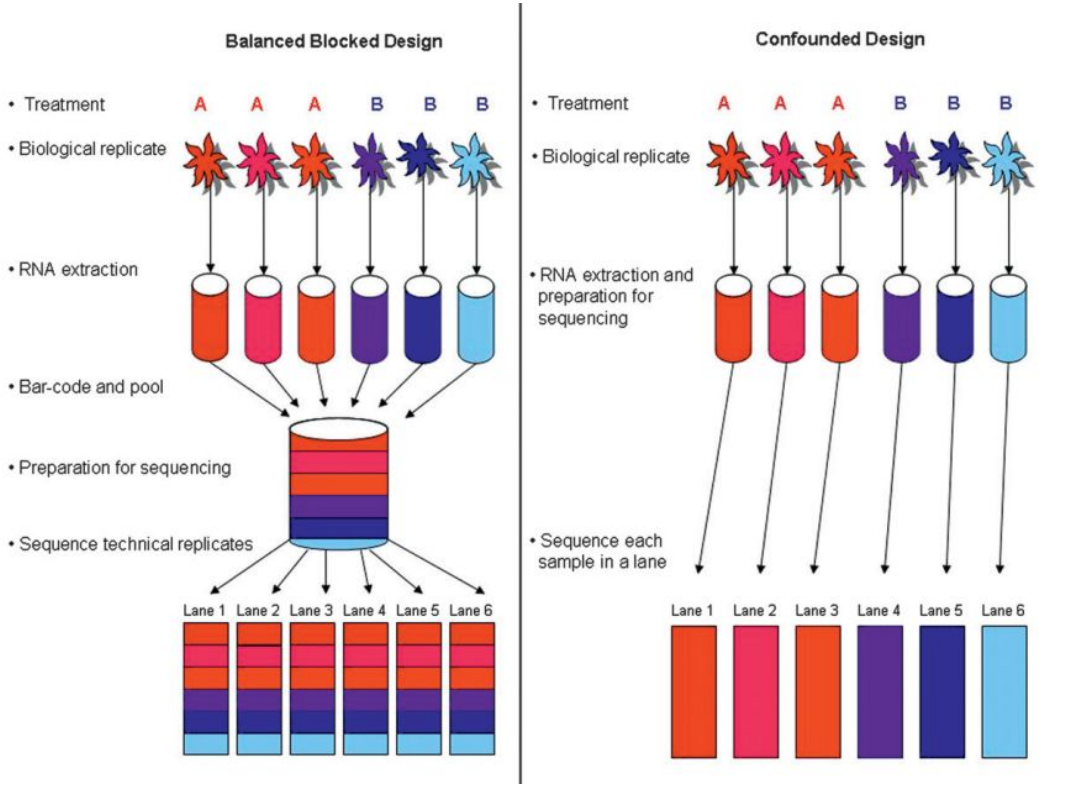
a priori estimates of experiment settings

Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Conesa et al, 2012

PLANNING: avoid confounding in experimental design

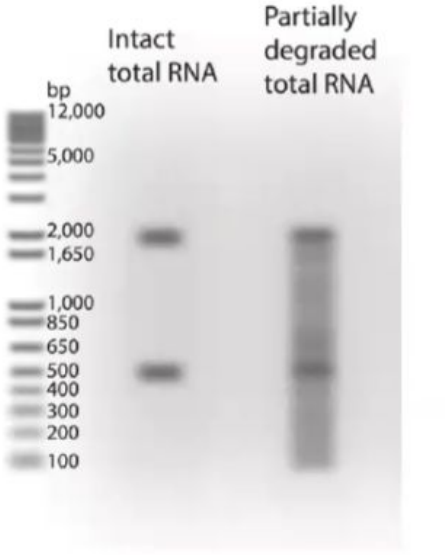


Auer & Doerge, 2010

PLANNING: RNA assessment

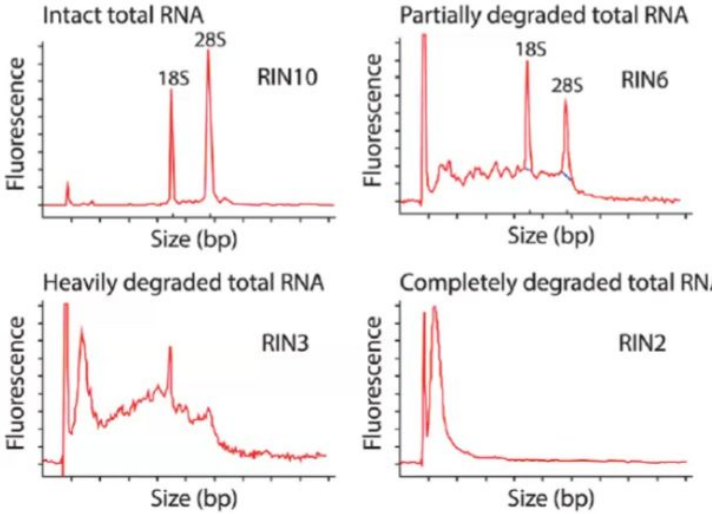
Clear 18/28S rRNA bands

OD 260/280 > 2

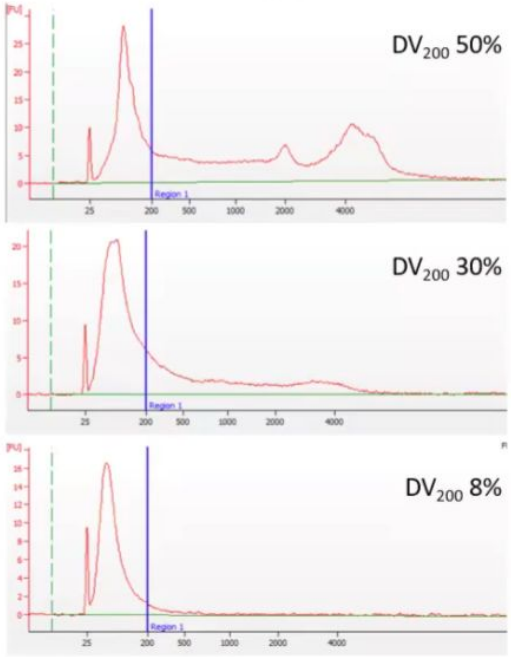


Fresh / Frozen: RIN > 7

RIN = RNA Integrity Number



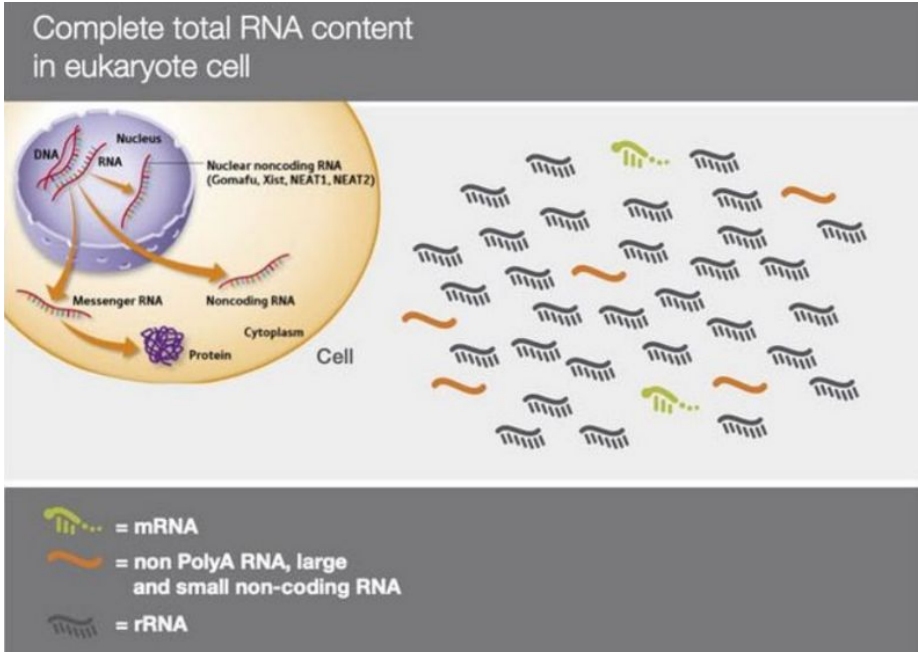
FFPE : DV_{200} > 30%



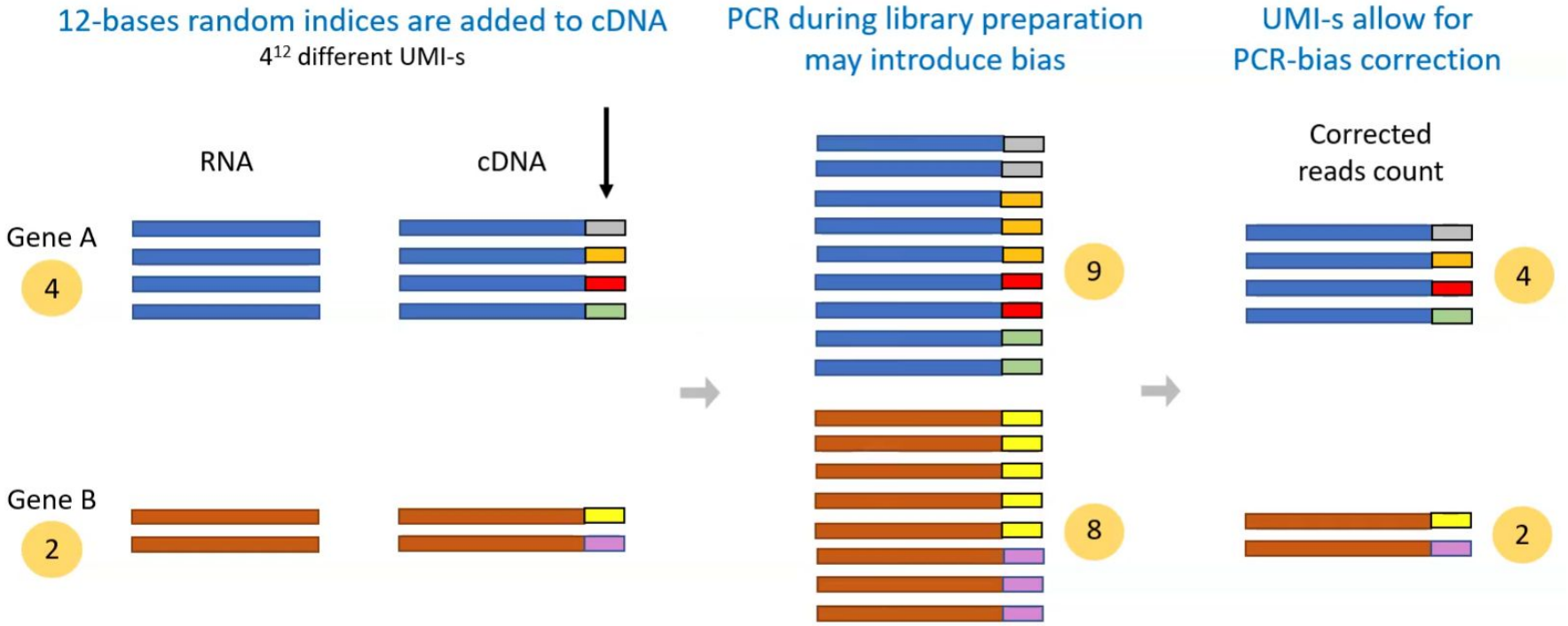
PLANNING: type of library

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

- total RNA with ribosomal depletion
- **polyA mRNA** library with oligo polyT capture
- unstranded vs **stranded** libraries
- specialized short RNA protocols and kits



PLANNING: Unique Molecular Identifiers (UMIs)

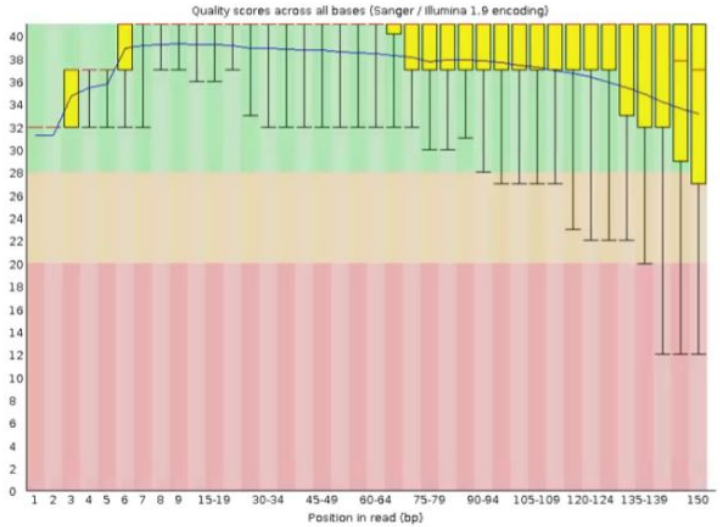


https://intelliseq.com/blog/unique-molecular-identifiers-umis-in-low-frequency-somatic-variant-detection/

QUALITY CONTROL (of sequencing reads)

FastQC

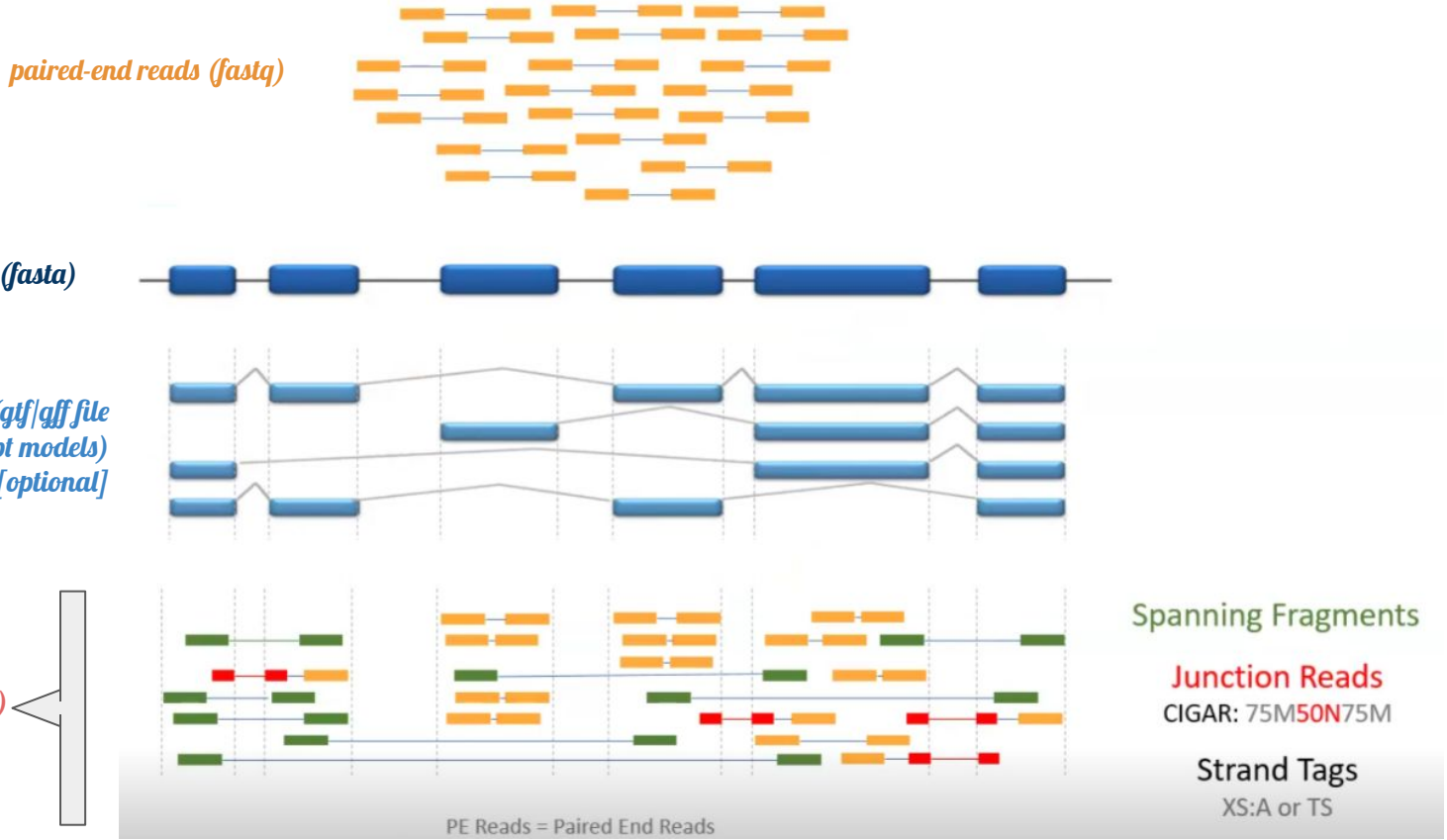
outputs results per one sample



MultiQC

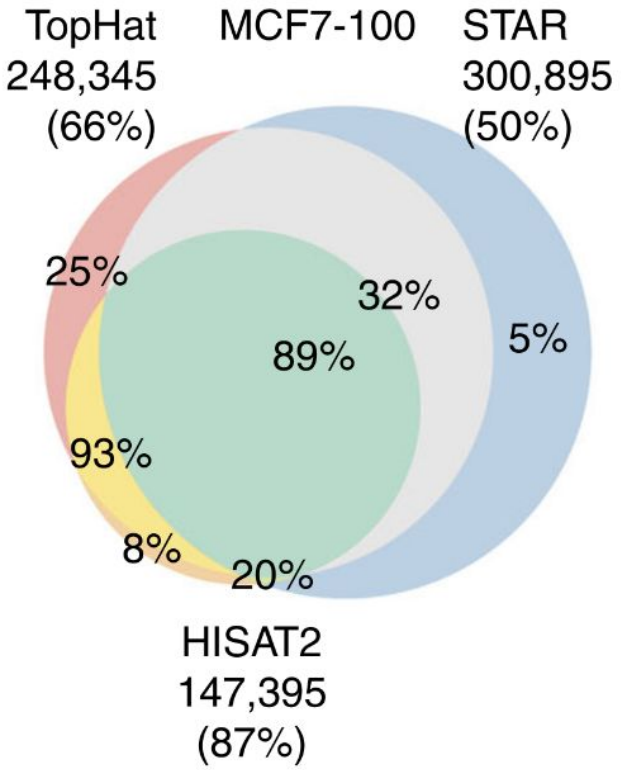
aggregates FastQC results from many samples

SPLICE-AWARE ALIGNMENT



SPLICE-AWARE ALIGNMENT

detected splice junctions by different aligners:



fraction of mapped reads by different aligners and techs:

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% **

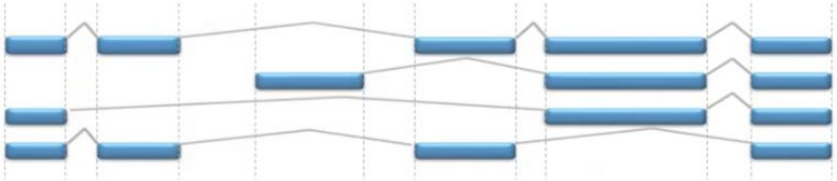
Krizanovic et al, 2018; Li 2018

Sahraeian et al, 2017

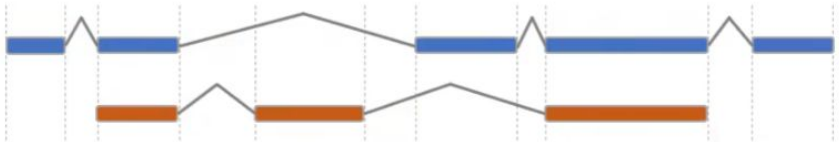
TRANSCRIPT ASSEMBLY (from genome alignment) [StringTie / Cufflinks]

spliced alignments (bam)

reference transcriptome (gtf/gff file with gene and transcript models) [optional]



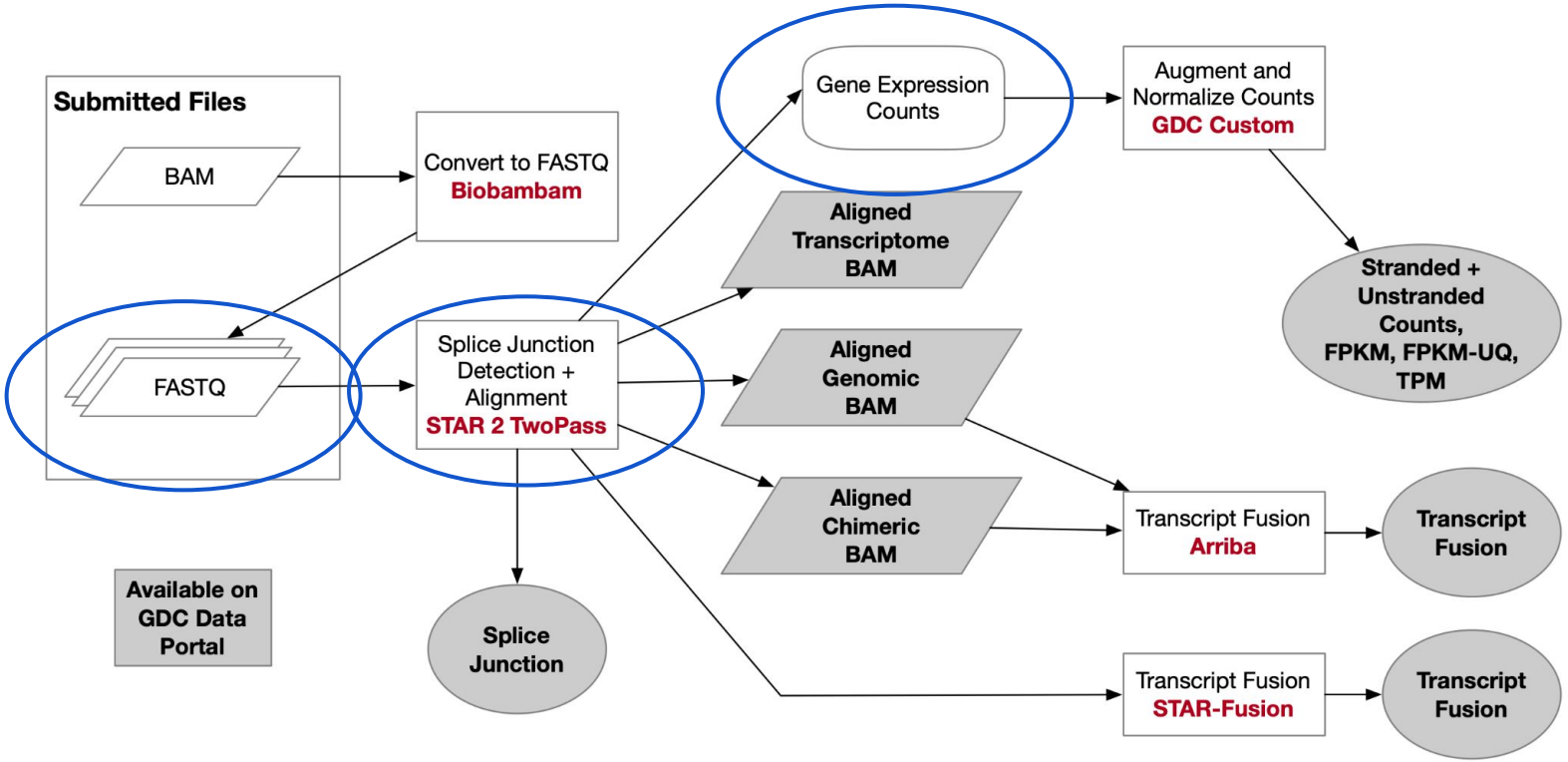
fasta with transcript sequences
GTF/GFF file with transcripts and junctions



A known transcript

A new transcript

FROM ALIGNMENT TO COUNTS [e.g. TCGA]



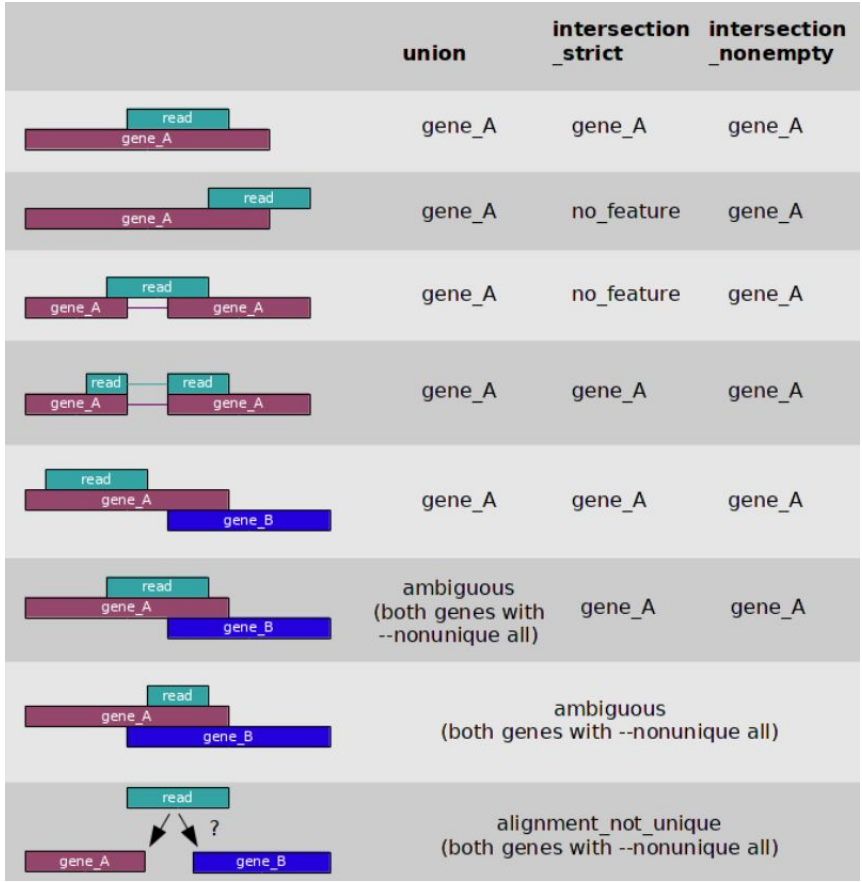
https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/

FROM ALIGNMENT TO COUNTS: HOW TO COUNT?

How to intersect?

What features to use?

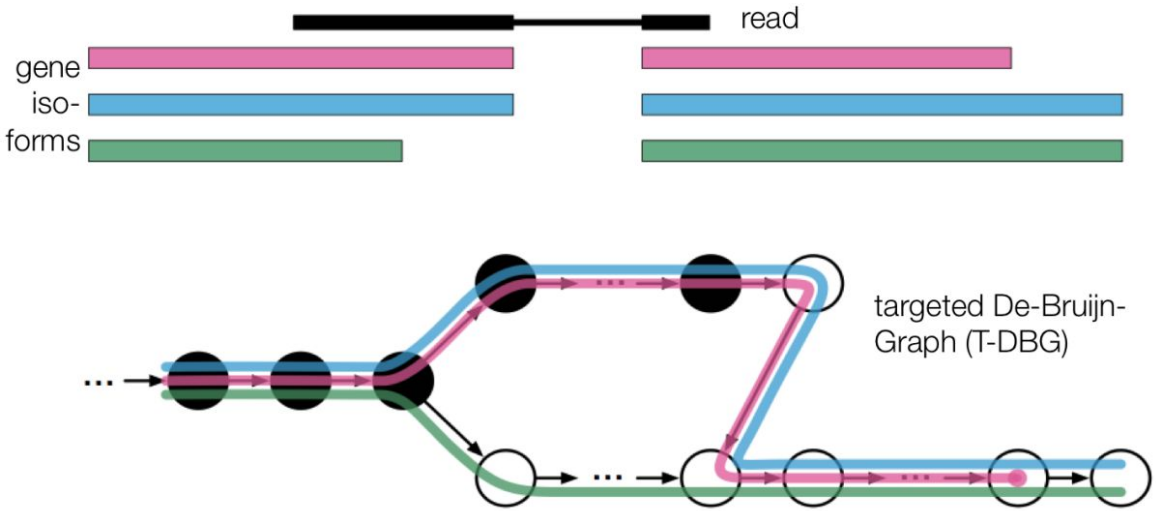
genes, transcripts, exons…



https://htseq.readthedocs.io/en/release_0.11.1/count.html#options

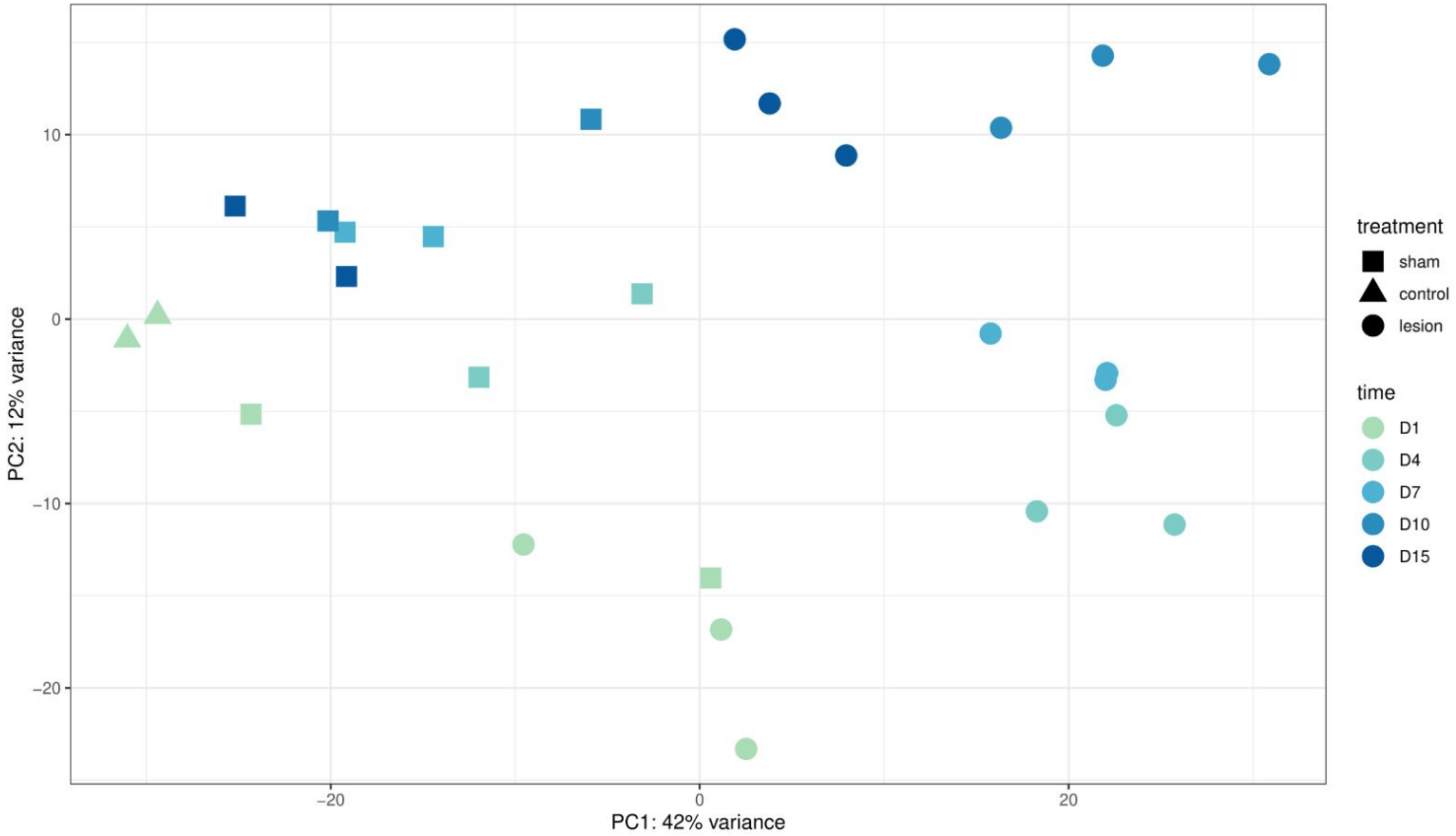
ALIGNMENT-FREE TOOLS (pseudo-alignment / quasi-mapping)[Kallisto / Salmon]

1. creates a k-mer index from all transcript sequences
2. each read is decomposed into k-mers
3. determines which transcripts are compatible with each read (pseudo-alignment)
4. using an expectation–maximization (EM) algorithm, Kallisto estimates the most likely number of reads coming from each transcript
5. output: transcript-level expression values

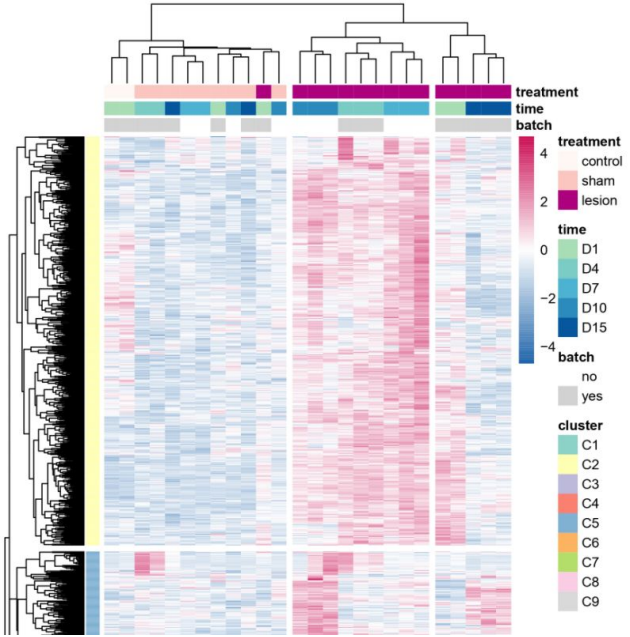


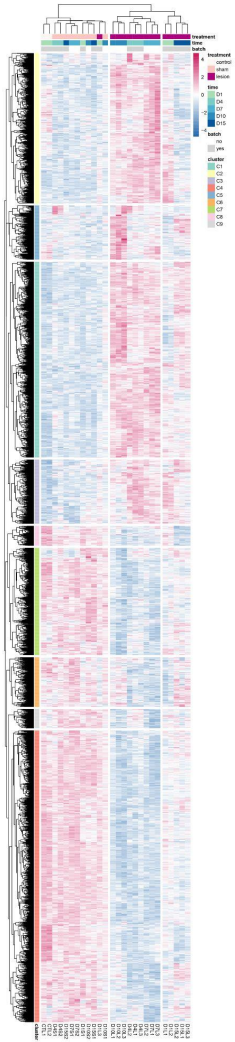
Bray et al, 2016

Visualization: PCA



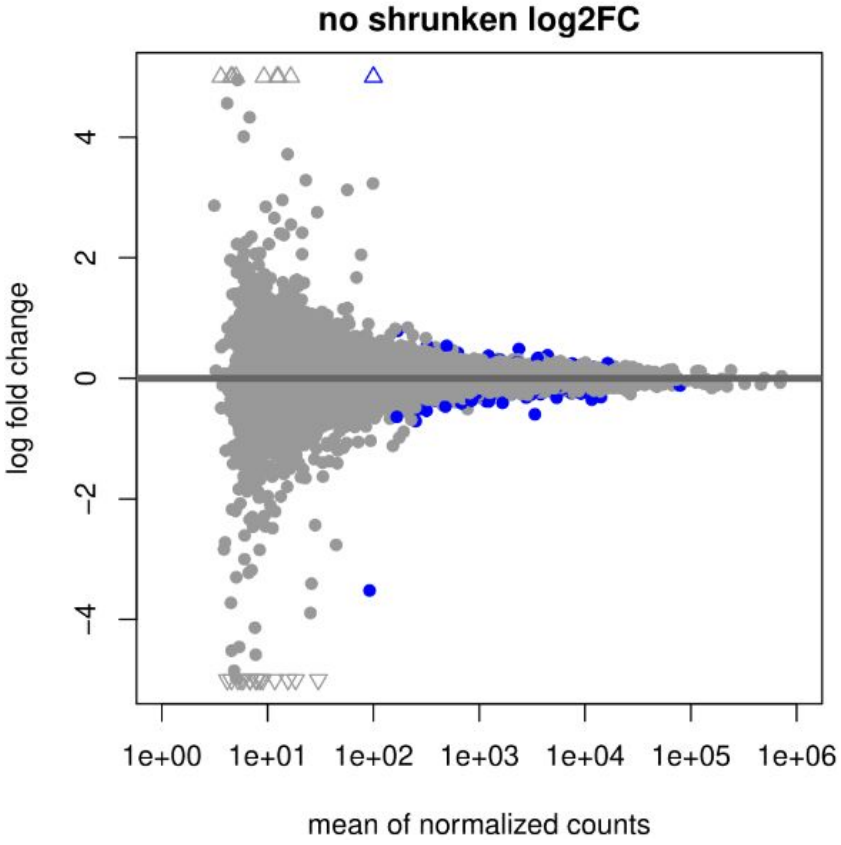
Visualization: HEATMAPs

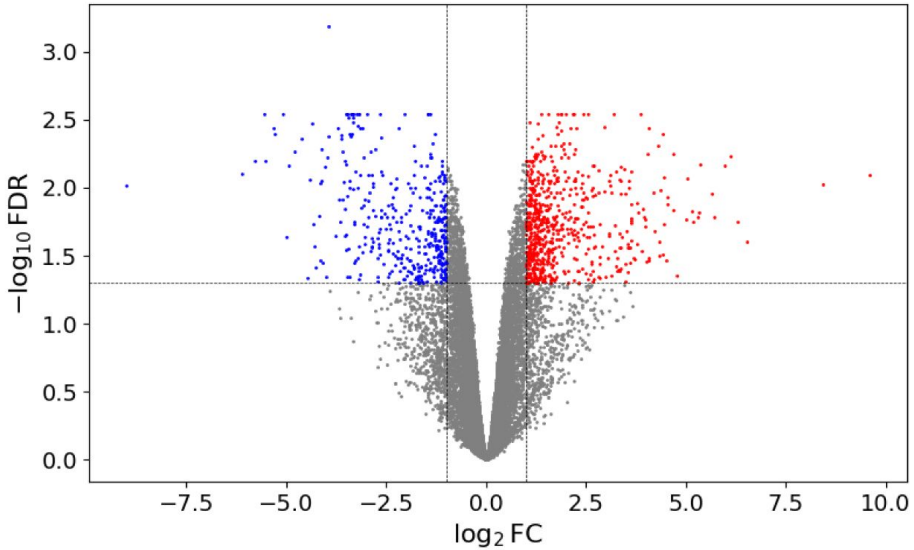




3418 genes with DE (padj>0.1 & abs(log2FC)>0.35) in at least one time point*

Visualization: MA-plot & Volcano plot

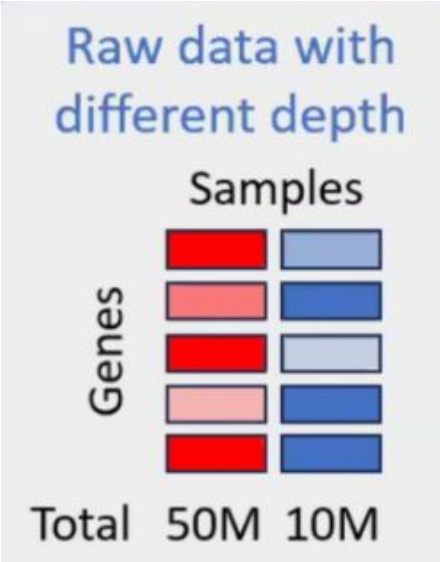




STATISTICS: normalization by sequencing depth and feature length

Accounts for:

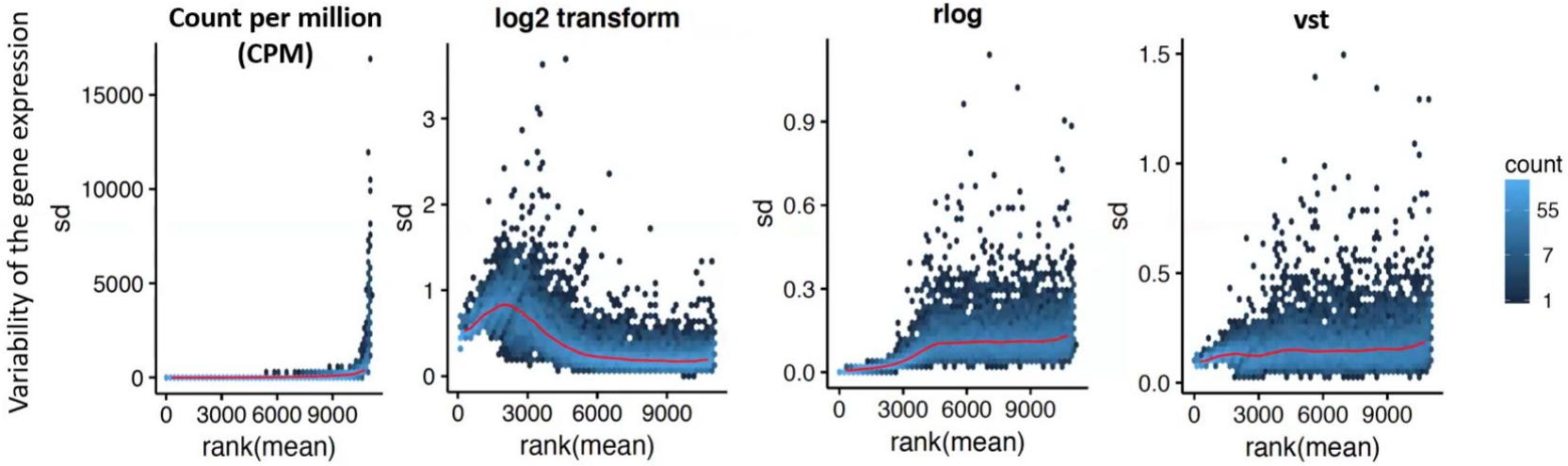
- different depth of sequencing (i.e. number of reads) per sample [between samples].
- different size of genes [for between-gene comparisons].
- RPKM, FPKM, **TPM**
  a. RPKM: Reads Per Kilobase of gene per Million reads
  b. FPKM: Fragments Per Kilobase of gene per Million reads
  c. **TPM: Transcripts per Million (Wagner et al, 2012)**



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

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

STATISTICS: transformation for visualization



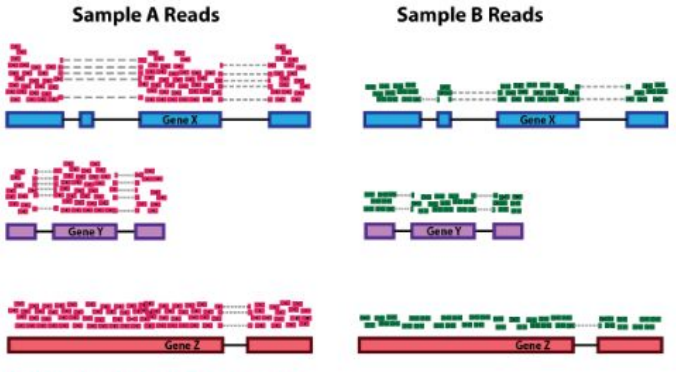
RNA-seq count data show variance that increases with mean expression, causing highly expressed genes to dominate analyses. Transformations like $\log_2$, rlog, or VST stabilize variance, making genes more comparable across the expression range.

STATISTICS: normalization for DE analysis (DESeq2, edgeR)

No need to normalize by gene length *

Make samples comparable by correcting for library size and composition bias.

Simply using *counts per million (CPM)* is not enough, since composition effects can bias totals — for instance, if one sample has many highly expressed genes, others may appear artificially downregulated.



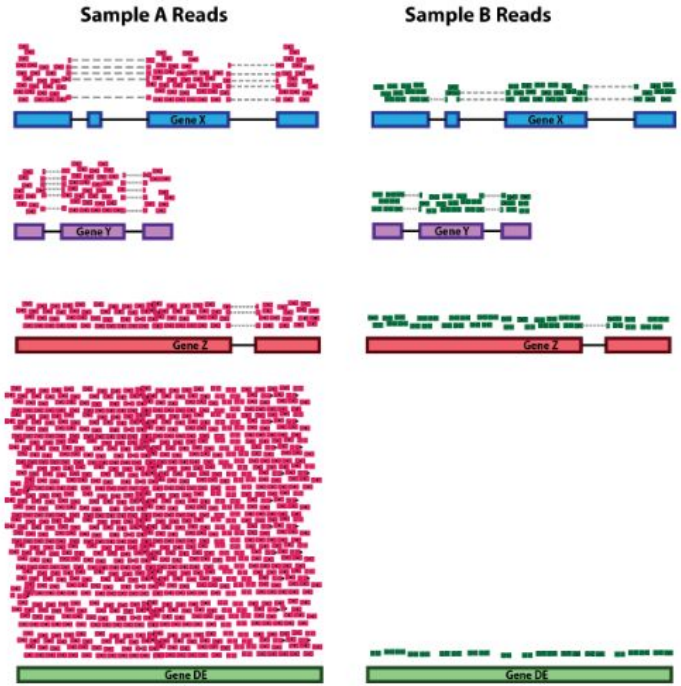
STATISTICS: normalization for DE analysis (DESeq2, edgeR)

No need to normalize by gene length *

Make samples comparable by correcting for library size and composition bias.

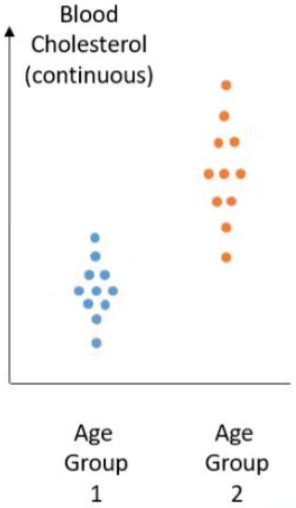
Simply using *counts per million (CPM)* is not enough, since composition effects can bias totals — for instance, if one sample has many highly expressed genes, others may appear artificially downregulated.

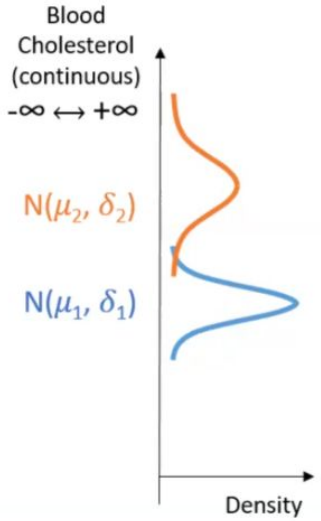
TMM (edgeR): Computes a trimmed mean of M-values versus a reference, removing extreme genes to estimate scaling factors.
Median of Ratios (DESeq2): Uses gene-wise geometric means and scales each sample by the median ratio, assuming most genes are not DE.

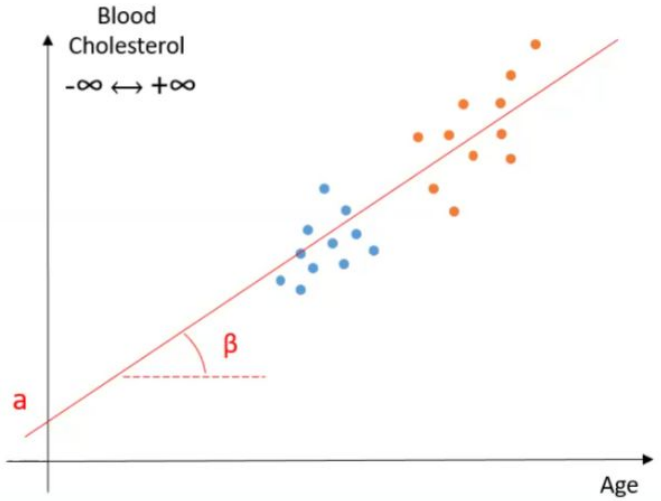


STATISTICS: detecting differences between groups

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





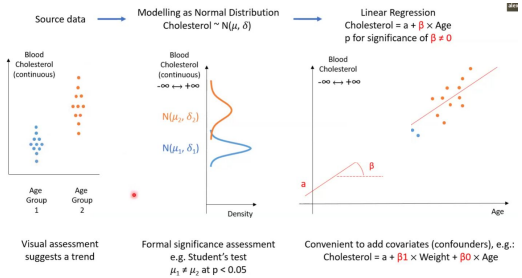


Visual assessment
suggests a trend

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

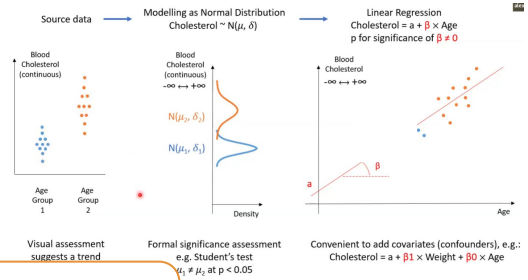
STATISTICS: detecting differences between groups



PROBLEMS:

1. raw counts in each sample depends on library size (sequencing depth)
2. counts don't go below 0
3. counts are discrete ⇒ better modeled by discrete distribution
4. small number of samples does not allow for accurate estimation of dispersion (variance)
5. testing for many genes at a time

STATISTICS: detecting differences between groups

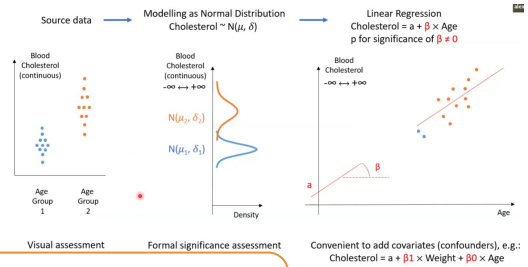


PROBLEMS:

1. raw counts in each s
2. counts don't go below 0
3. counts are discrete ⇒ better modeled by discrete distribution
4. small number of samples does not allow for accurate estimation of dispersion (variance)
5. testing for many genes at a time

normalizing by library size and composition bias (Median of Gene Ratios in DESeq2)

STATISTICS: detecting differences between groups



PROBLEMS:

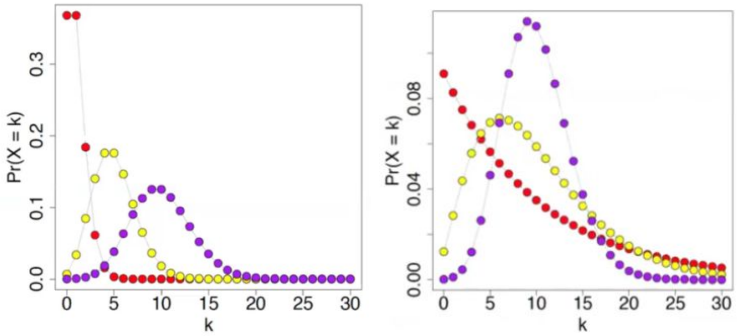
1. raw counts in each sample (
2. counts don't g
3. counts are dis
4. small number of samples does not allow for accurat

   dispersion (variance)
5. testing for many genes at a time

normalizing by library size and composition bias (Median of Gene Ratios in DESeq2)

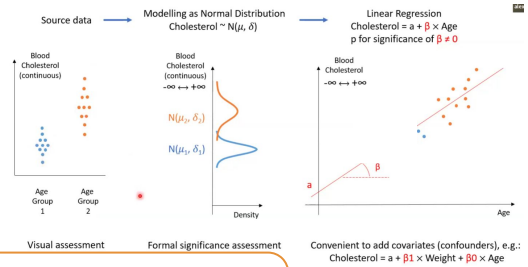
choosing an appropriate discrete distribution (negative binomial)

Marioni et al, 2008;
Robinson & Smyth, 2007



total variance in RNAseq = technical variance (Poisson) + biological variance
(overdispersion that can be modeled by negative binomial)

STATISTICS: detecting differences between groups



PROBLEMS:

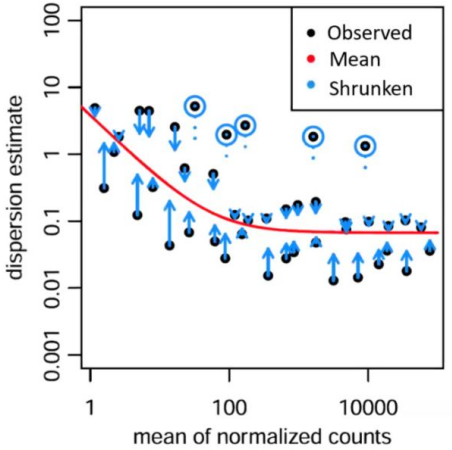
1. raw counts in each sample (
2. counts don't g
3. counts are dis
4. small numbe

   dispersion (v
5. testing for many genes at a time

normalizing by library size and composition bias (Median of Gene Ratios in DESeq2)

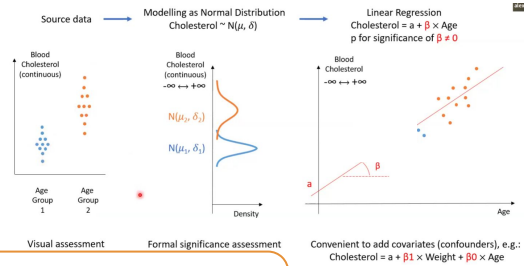
choosing an appropriate discrete distribution (negative binomial)

borrowing data between genes for estimation of dispersion



Love et al, 2014

STATISTICS: detecting differences between groups



PROBLEMS:

1. raw counts in each sample (
2. counts don't g
3. counts are dis
4. small numbe

   dispersion (v
5. tes

normalizing by library size and composition bias (Median of Gene Ratios in DESeq2)

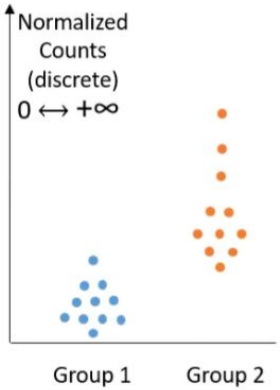
choosing an appropriate discrete distribution (negative binomial)

borrowing data between genes for estimation of dispersion

multiple testing correction (typically FDR)

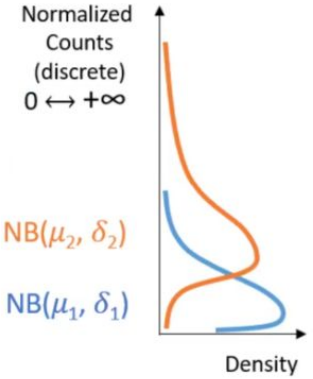
STATISTICS: detecting differences between groups

Source data
Normalized by library size
(median gene ratios)



Visual assessment
suggests a trend

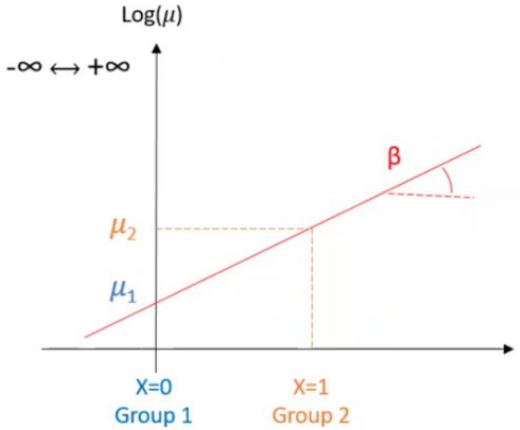
Modelling as Negative Binomial
Counts ~ NB(μ, δ)



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

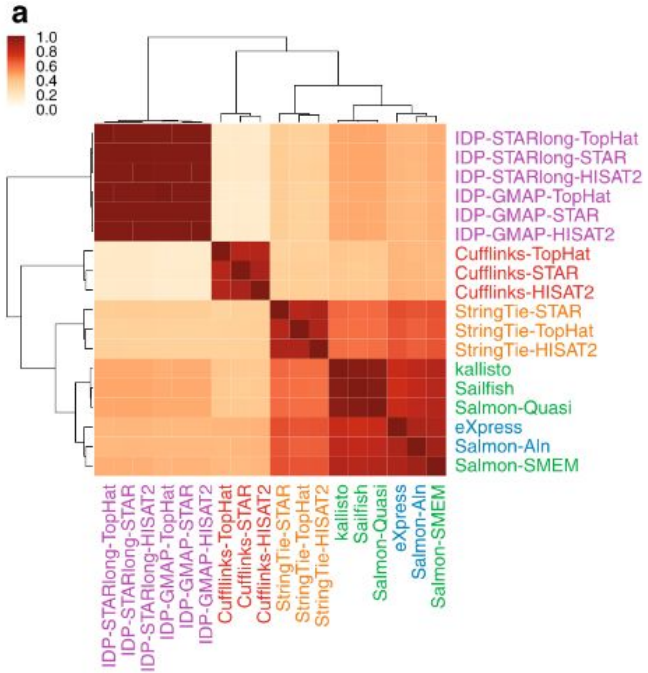
Log-link

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

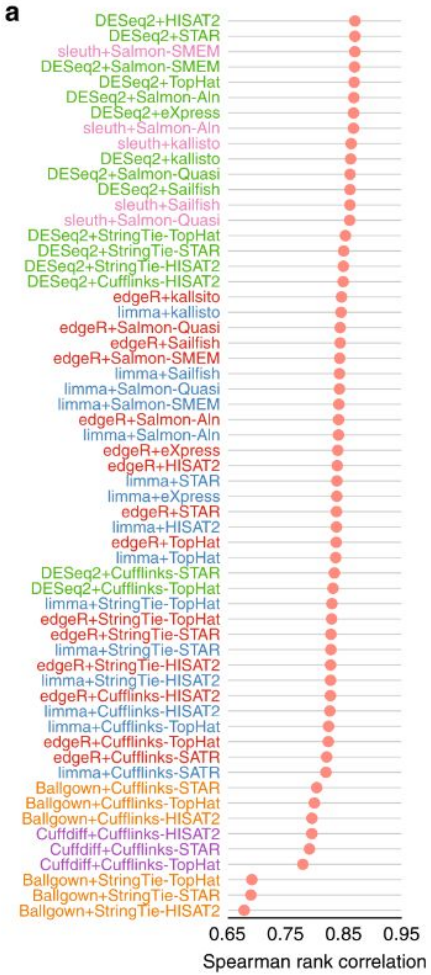


Significance of β ≠ 0
Convenient adding covariates (confounders):
Counts = a + β1 × Batch + β0 × Group

WHAT IS BETTER? DOES IT MATTER?



between-sample log2FC for 1000 genes evaluated by RNAseq and qRT-PCR



Sahraeian et al, 2017

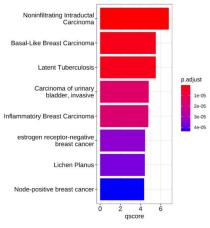
INTERPRETING RESULTS

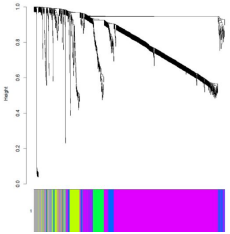
Differentially expressed (DE) genes are the starting point — but they are *not the end of the story*.

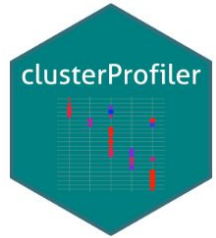
We can use them to:

- Identify enriched biological processes or pathways (GO, KEGG, Reactome, etc.).
- Discover potential biomarkers or signatures.
- Explore functional shifts using *all genes ranked* by expression change.

The goal: move from **individual genes → biological interpretation**.







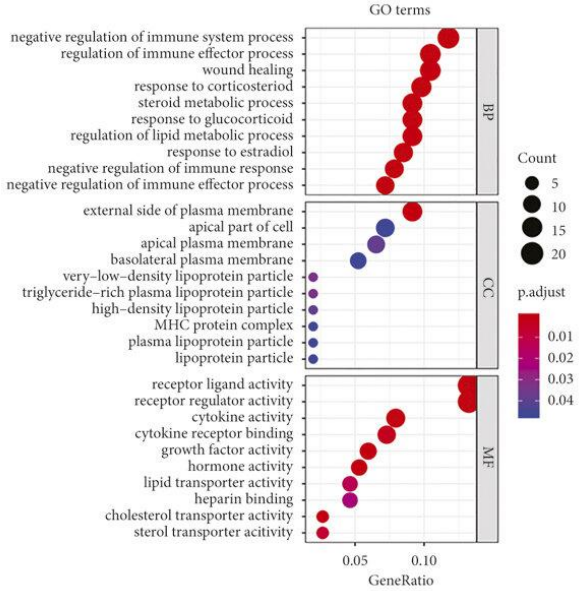
INTERPRETING RESULTS: Over-Representation Analysis (ORA)

Uses a list of DE genes (typically adjusted p < 0.05, |log2FC| > 1).

Asks: *"Are certain functional categories represented more often than expected by chance?"*

Works with predefined sets such as: GO, KEGG, REACTOME, MSigDB, custom gene sets.

Statistical test: *usually Fisher's exact test or hypergeometric test.*



100 DEG: 50 "proliferation"

50,000 background genes: 1,000 "proliferation"

Fisher exact test p < 10^{-16}

Pros: simple, intuitive
Cons: depends on arbitrary cutoff and ignores genes below threshold.

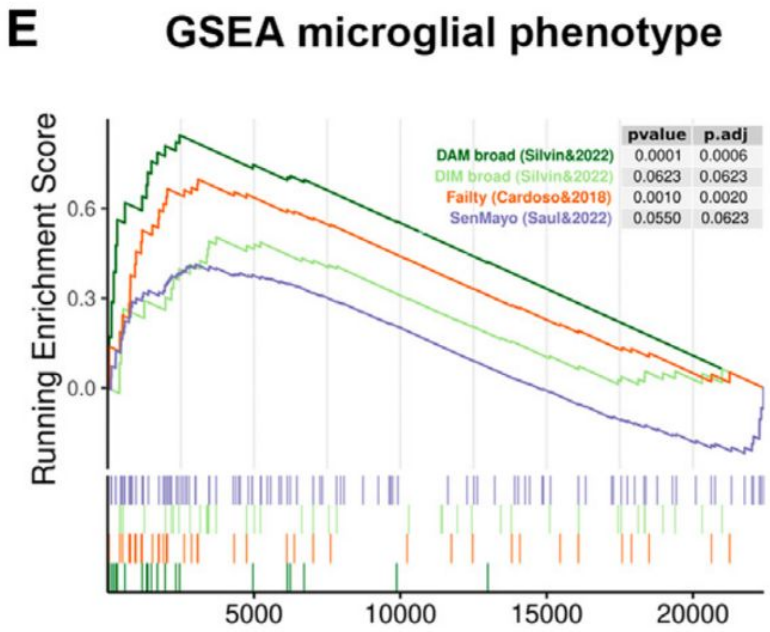
INTERPRETING RESULTS: Gene Set Enrichment Analysis (GSEA)

Uses all genes ranked by a continuous metric (e.g. *log2FC*, padj, signed(padj)).

Tests whether genes from a predefined set are concentrated at the top or bottom of the ranked list.

Output:

- Enrichment score (ES) and normalized ES (NES).
- *Leading-edge genes* that drive enrichment.

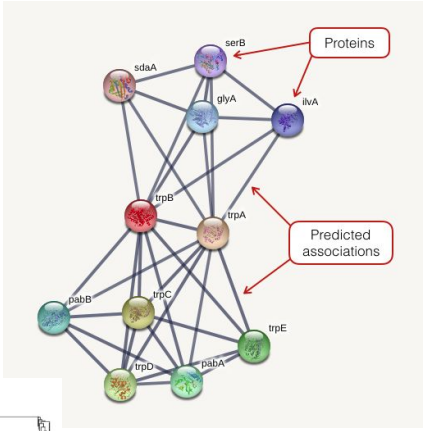


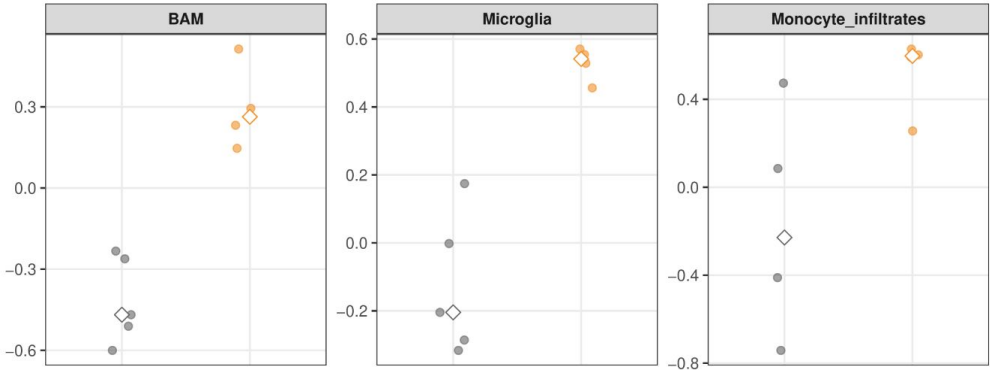
Pros: no arbitrary cutoff, sensitive to coordinated patterns.
Cons: requires ranking metric.

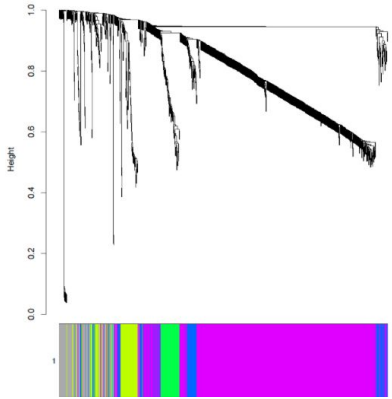
INTERPRETING RESULTS: beyond enrichment

Endless:

- network analysis (e.g. STRING, Cytoscape).
- co-expression modules (e.g. WGCNA)
- pathway activity score (e.g. GSVA)
- cell type deconvolution (e.g. CIBERSORT)







CIBERSORTX

SUMMARY

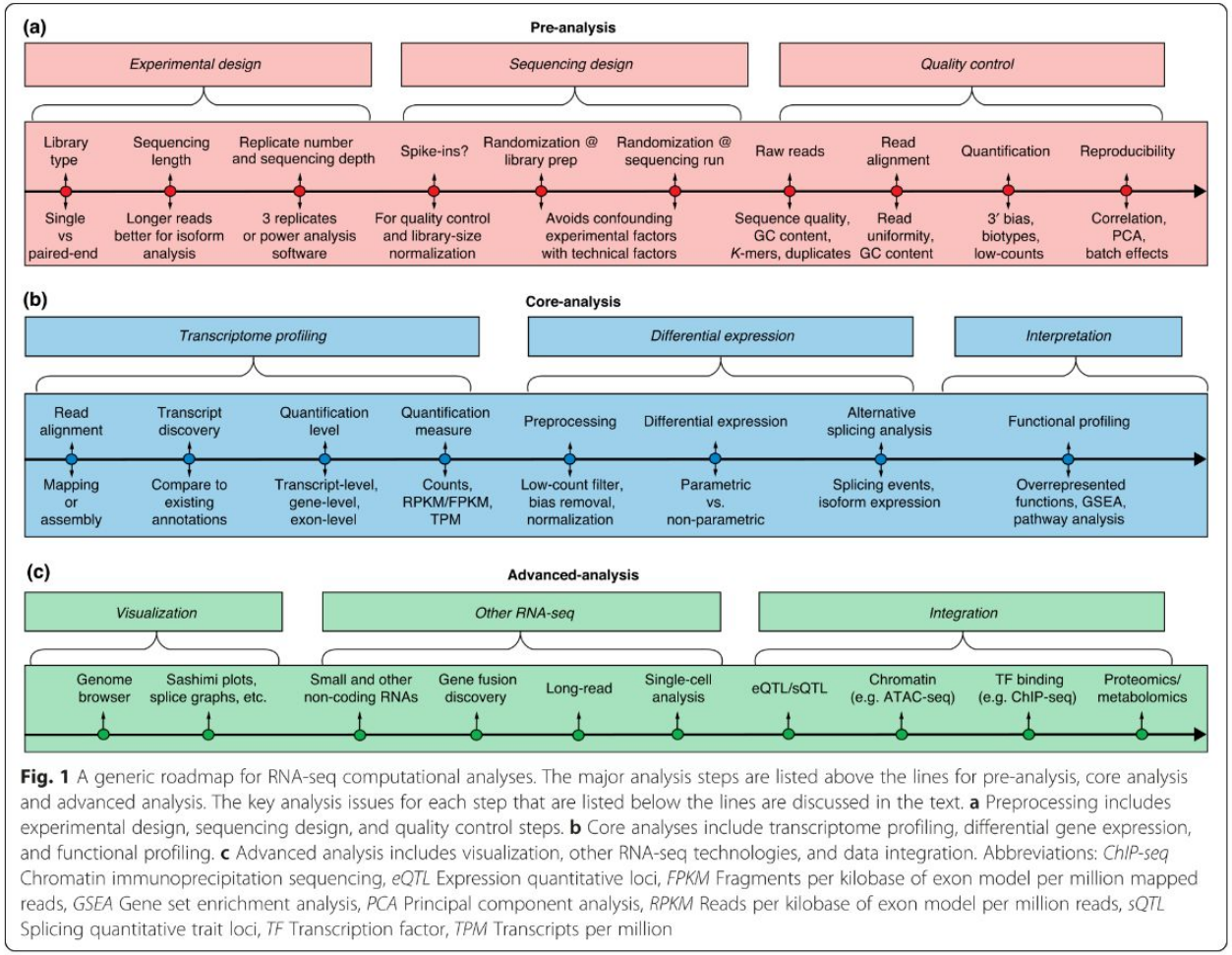


Fig. 1 A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

Conesa et al, 2016

A FEW REFERENCES:

Conesa et al 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology* 10.1186/s13059-016-0881-8

Sahraeian et al 2017. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature Communications* 10.1038/s41467-017-00050-4

Koch et al 2017. A beginner's guide to analysis of RNA sequencing data. *Translational Review* 10.1165/rcmb.2017-0430TR

Questions?