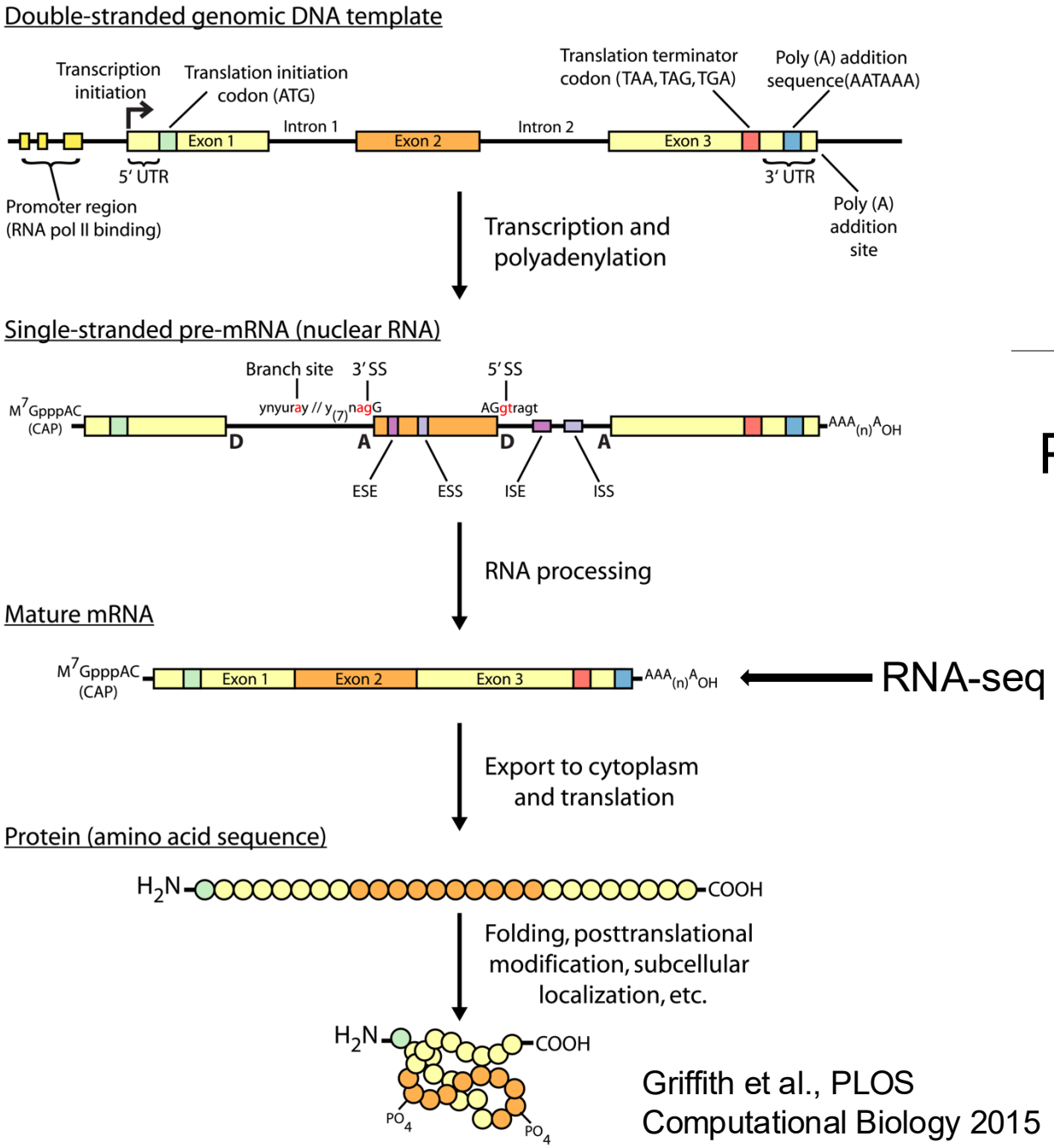
RNA-sequencing

Clinical Cancer Genomics
19 May 2025

Learning objectives and lecture agenda

- Learn about RNA-seq as a high-throughput method to profile the expression of genes in a sample.
- List the main steps involved in the analysis of RNA-seq data.
- Learn about bioinformatics tools used in preparing and analyzing RNA-seq data.
- Name and explain a few applications of RNA-seq data in the cancer research context.



Griffith et al., PLOS Computational Biology 2015

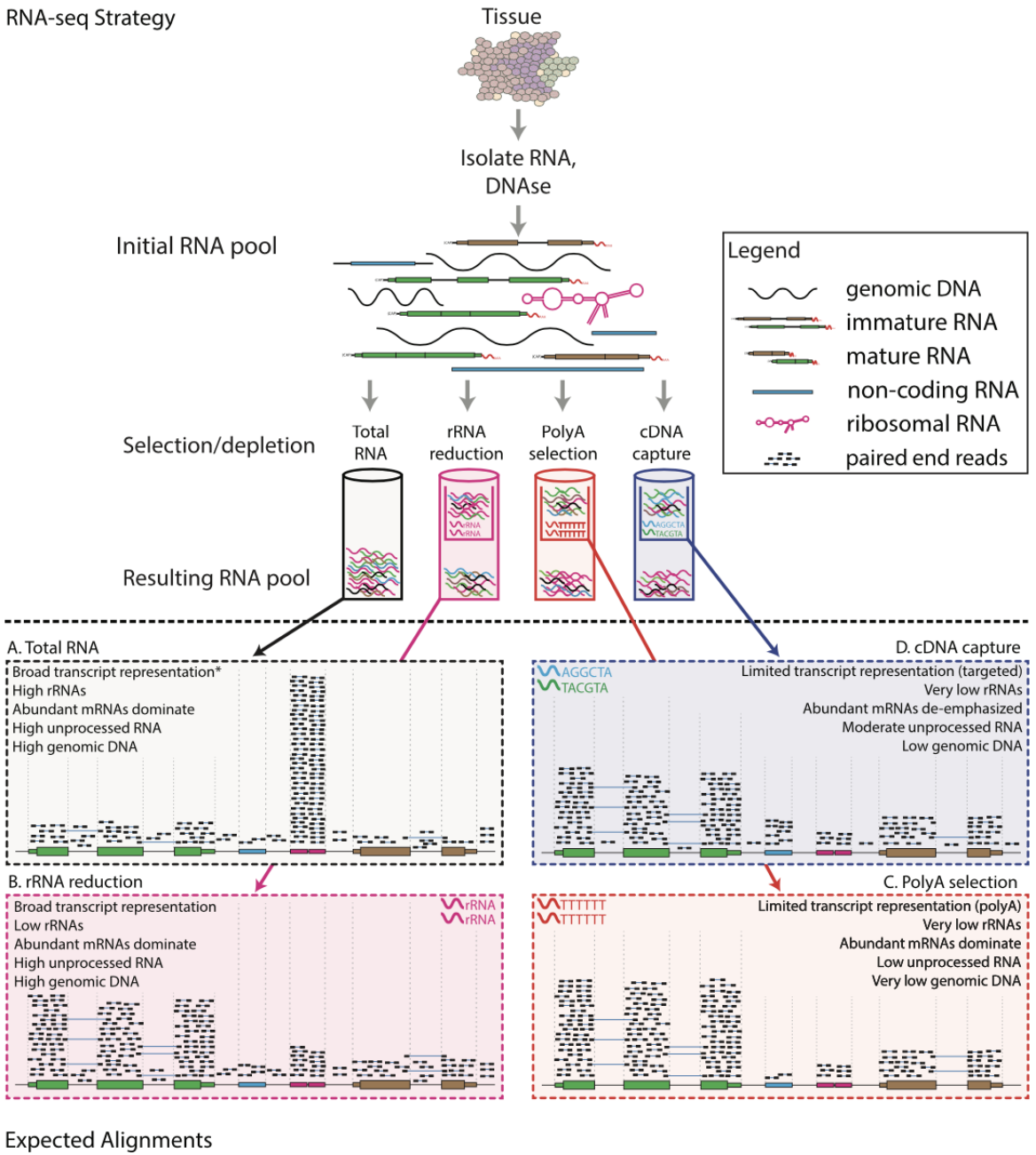
The central dogma of molecular biology

Phenotype = expression of genetic information modified by environmental factors

Cell identity and its activities are specified via the genes it transcribes

RNA-seq strategies

The analysis and interpretation of RNA-seq data is influenced by the protocols and methods used to generate it.

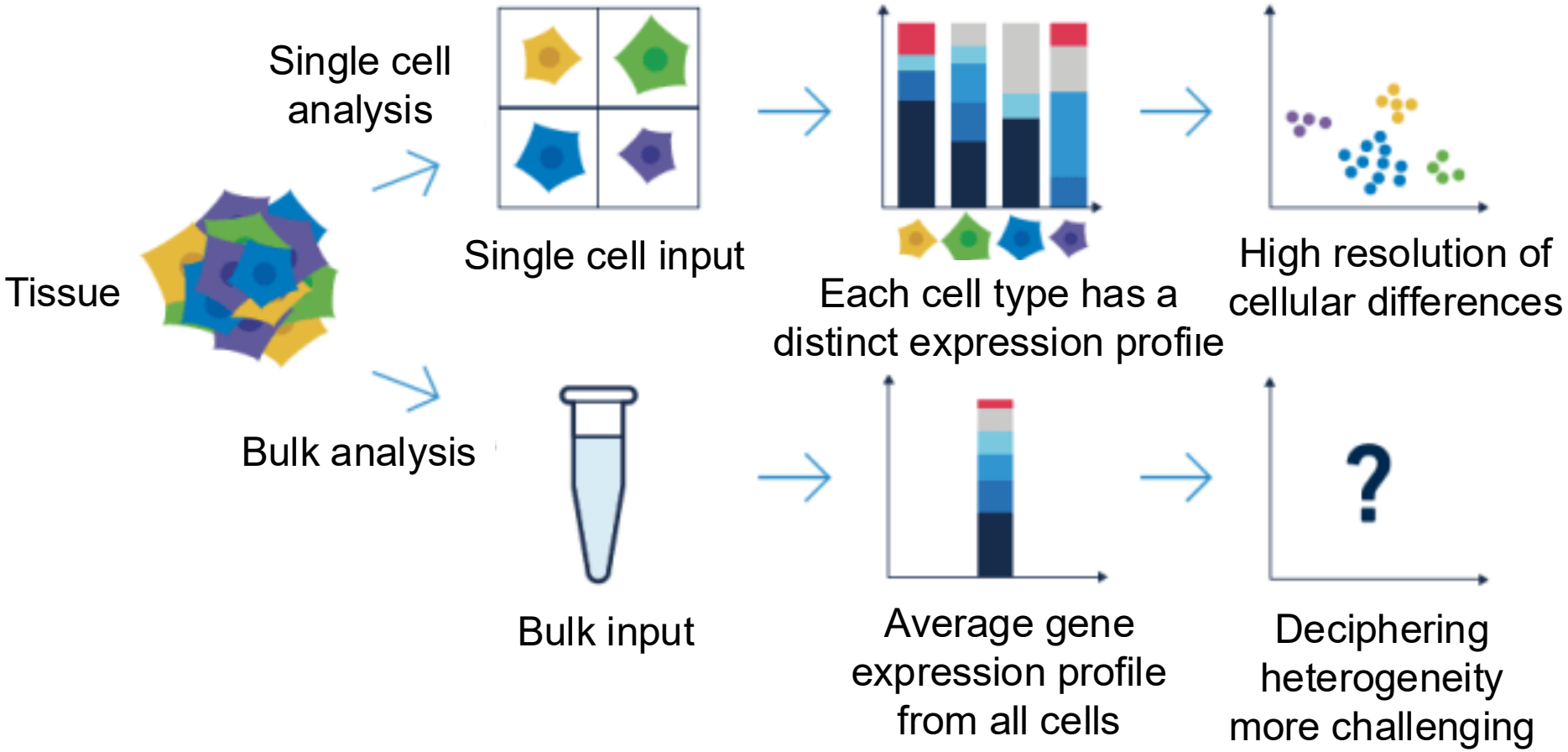


Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud, PLOS Computational Biology 2015

RNA-sequencing challenges

- Sample
  - Purity?, quantity?, quality?
- The relative abundance of RNAs vary a lot
  - 10^5 – 10^7 orders of magnitude
  - Since RNA-seq works by random sampling, a small fraction of highly expressed genes may consume majority of reads
- RNAs come in a wide range of sizes
  - Small RNAs have to be captured separately
- RNA is fragile compared to DNA (easily degraded)
- RNA consists of small exons that may be separated by large introns → mapping reads to the genome can be challenging

Bulk vs single-cell RNA-seq



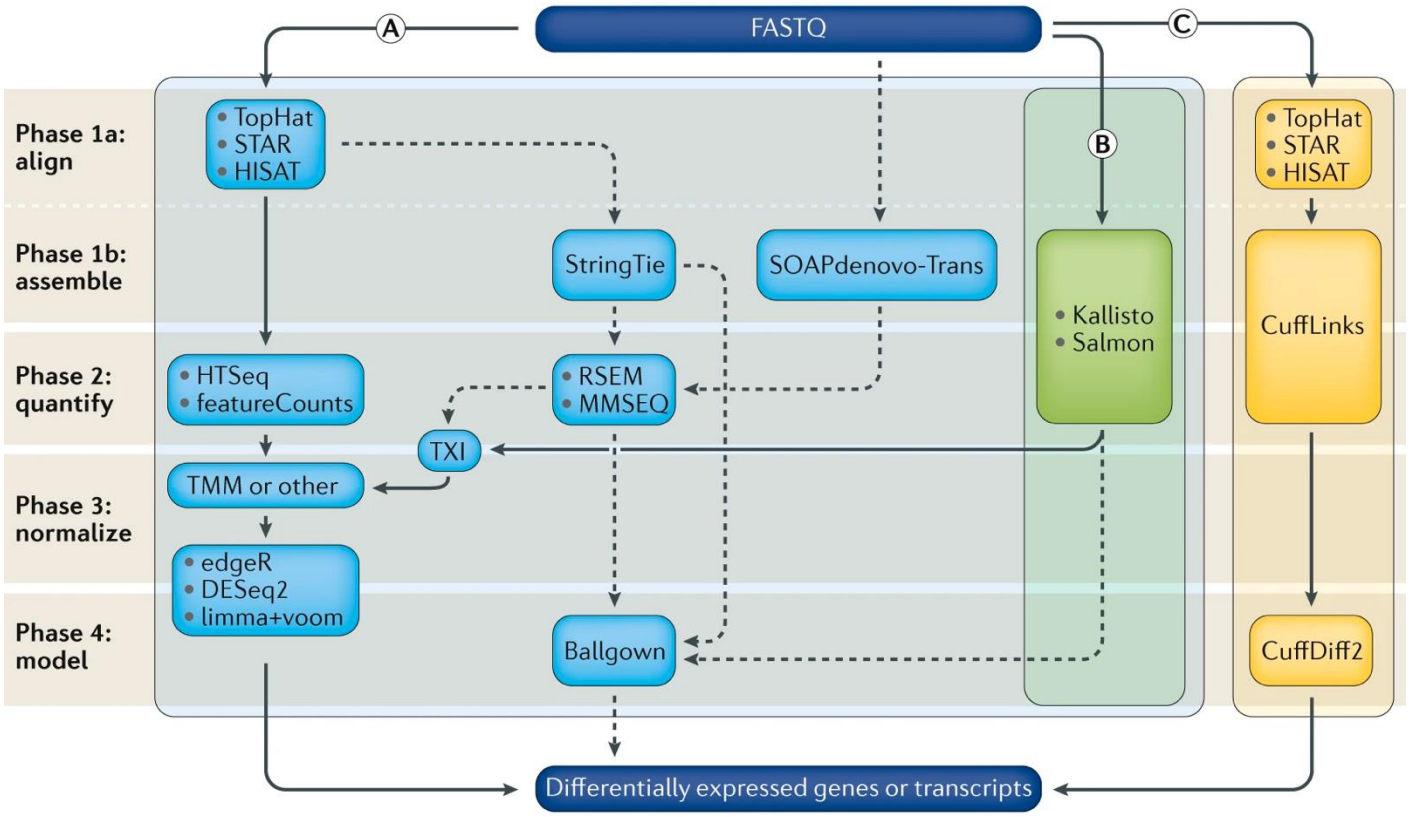
CHALLENGES:
Cost
Complexity of library prep
Data sparsity
Analysis tools

Source: 10x Genomics

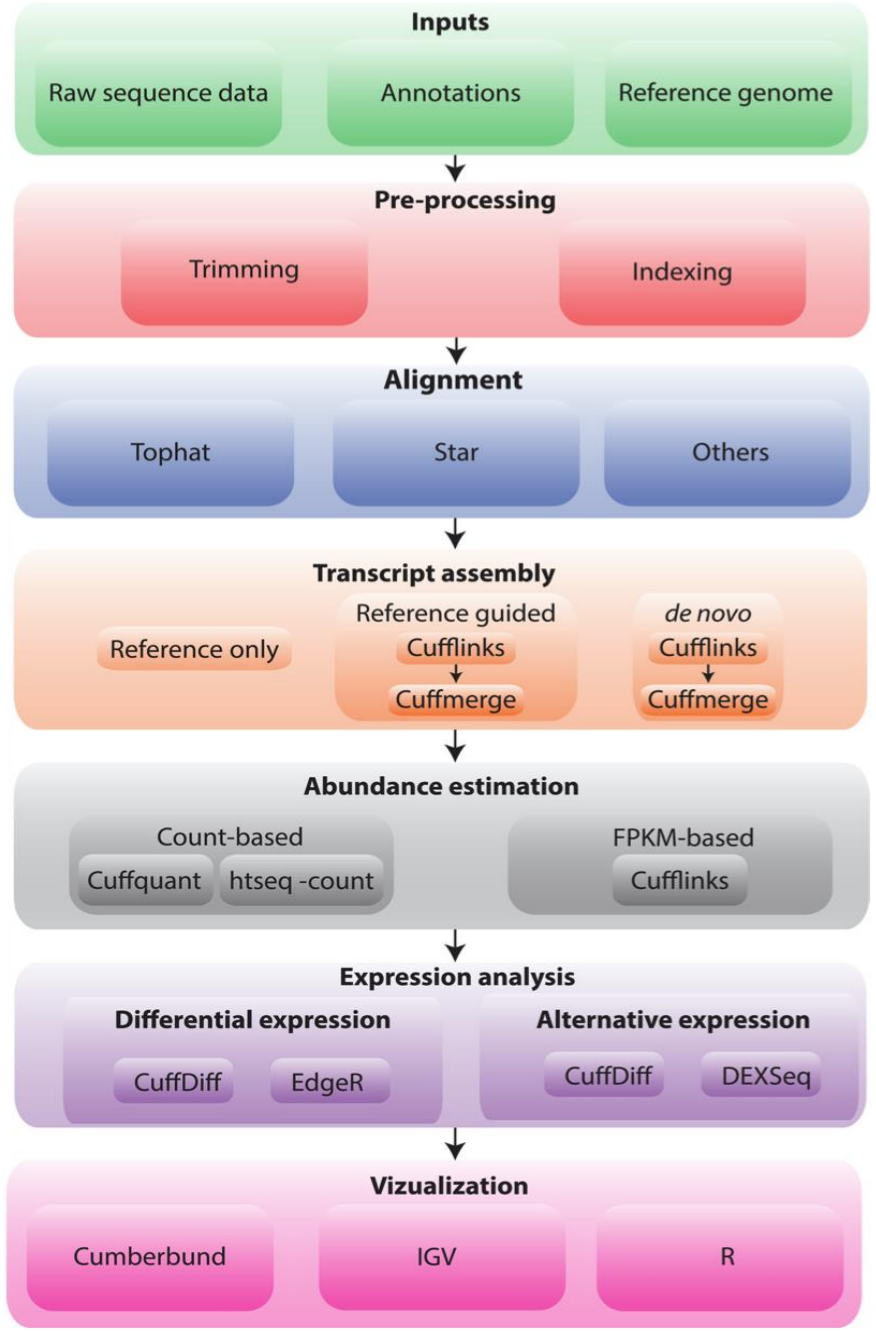
What are some common analysis goals of RNA-seq?

- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Fusion detection
- RNA editing
- …

RNA-seq workflows for gene expression and differential expression analysis



Stark et al., Nature Reviews Genetics 2019



Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud, PLOS Computational Biology 2015

Quality control (QC) of RNA-seq data

Karolinska Institutet

What are the concerns?

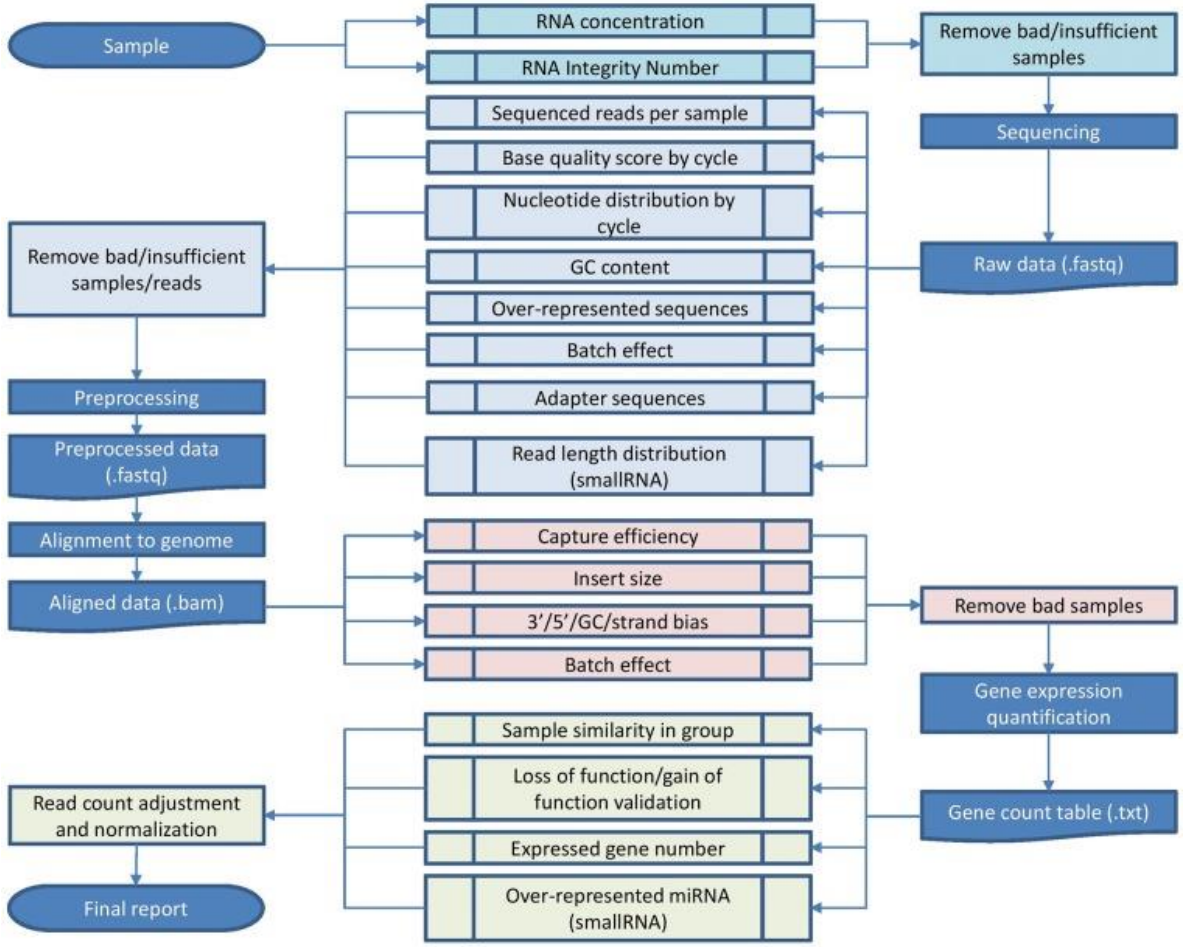
- Read quality and adapter contamination
- Ribosomal RNA fraction
- Fraction aligned reads
- Genomic origin of reads
- 5’-3’ bias in transcript coverage
- Batch effects

Tools for RNA-seq QC

- Some tools work for both DNA and RNA data, others are specific to RNA.
  - FastQC
  - Picard CollectRnaSeqMetrics
  - Qualimap
  - RNA-SeQC
  - RSeQC
  - RNA-QC-chain

Stages of RNA-seq QC

1. Checking RNA quality in samples
2. Checking quality of raw read data in FASTQ files
3. Checking RNA-seq alignment quality
4. Checking quality of gene expression estimates



Sheng et al., Briefings in Functional Genomics, 2017

Checking RNA quality in samples

- Is the RNA in the sample intact (and good for downstream analysis) or degraded?
- RNA can be digested by RNase enzymes also present in the sample
- Generally, RNA Integrity Number (RIN) > 8 is ideal
- Formalin-fixed samples may have lower RIN values (range ~2-5)
- Romero et al., *BMC Biol* 2014

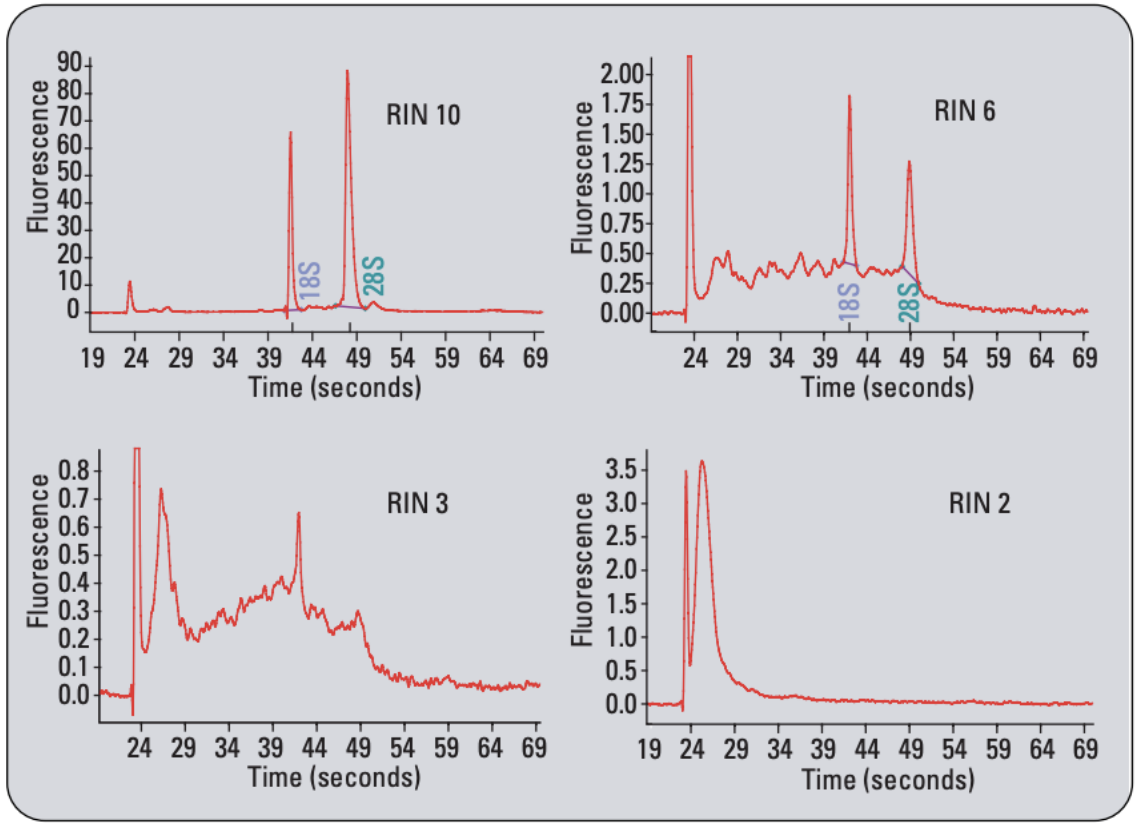
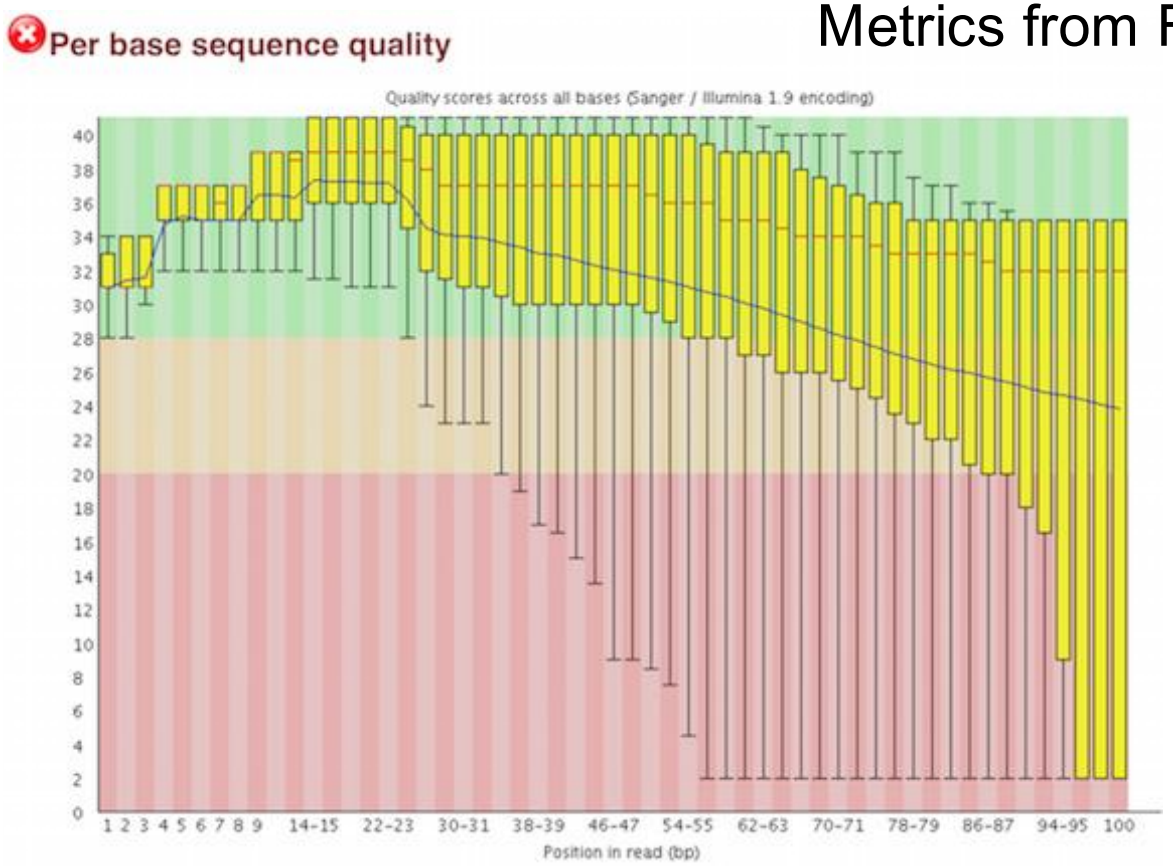


Figure 2
Sample electropherograms used to train the RNA Integrity Number (RIN) software. Samples range from intact (RIN 10), to degraded (RIN 2).

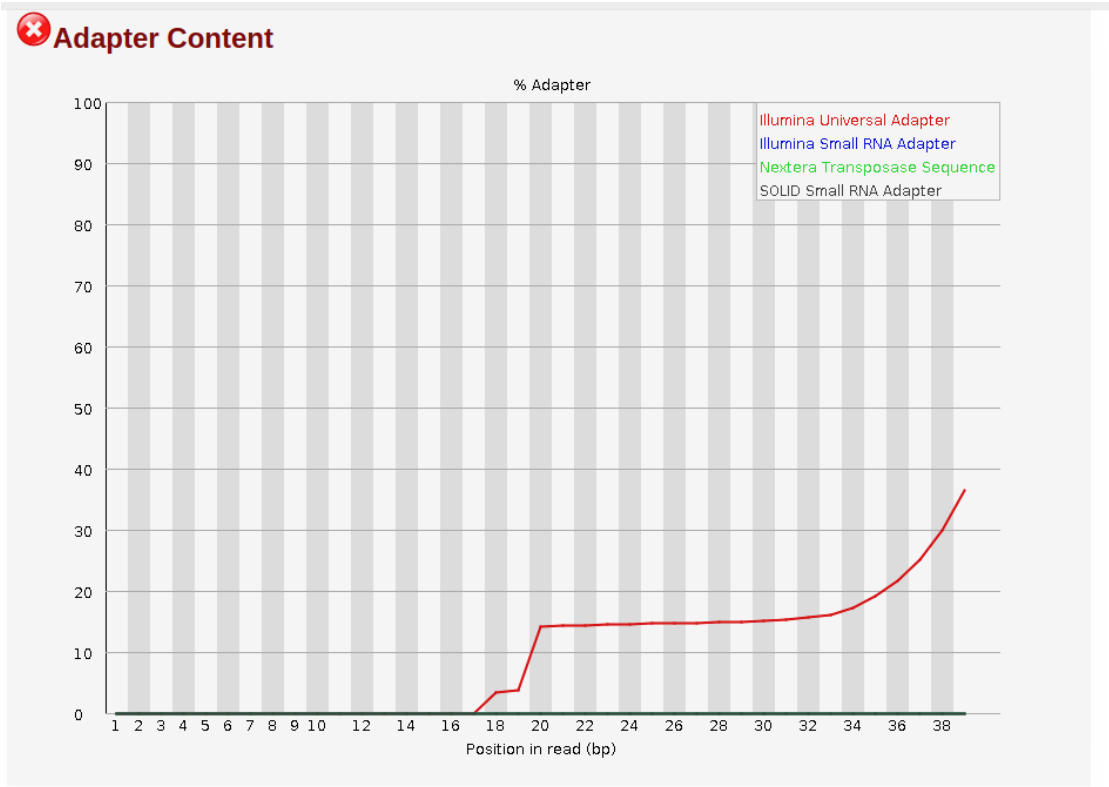
Source: Agilent

Checking quality of raw read data in FASTQ files

Metrics from FASTQC tool



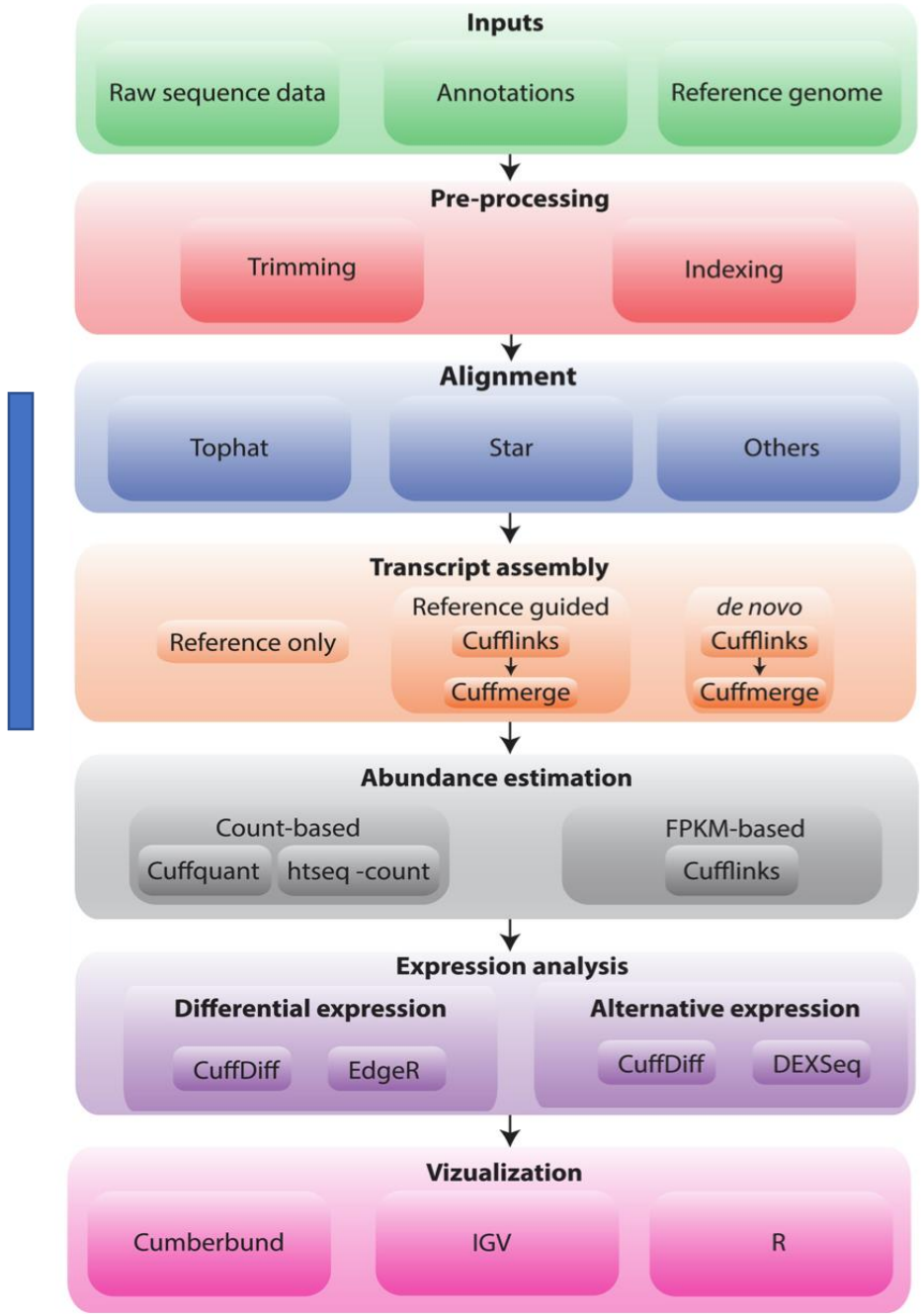
Are there low quality bases in the sequencing reads?



Is there adapter contamination?

RNA-seq analysis workflow

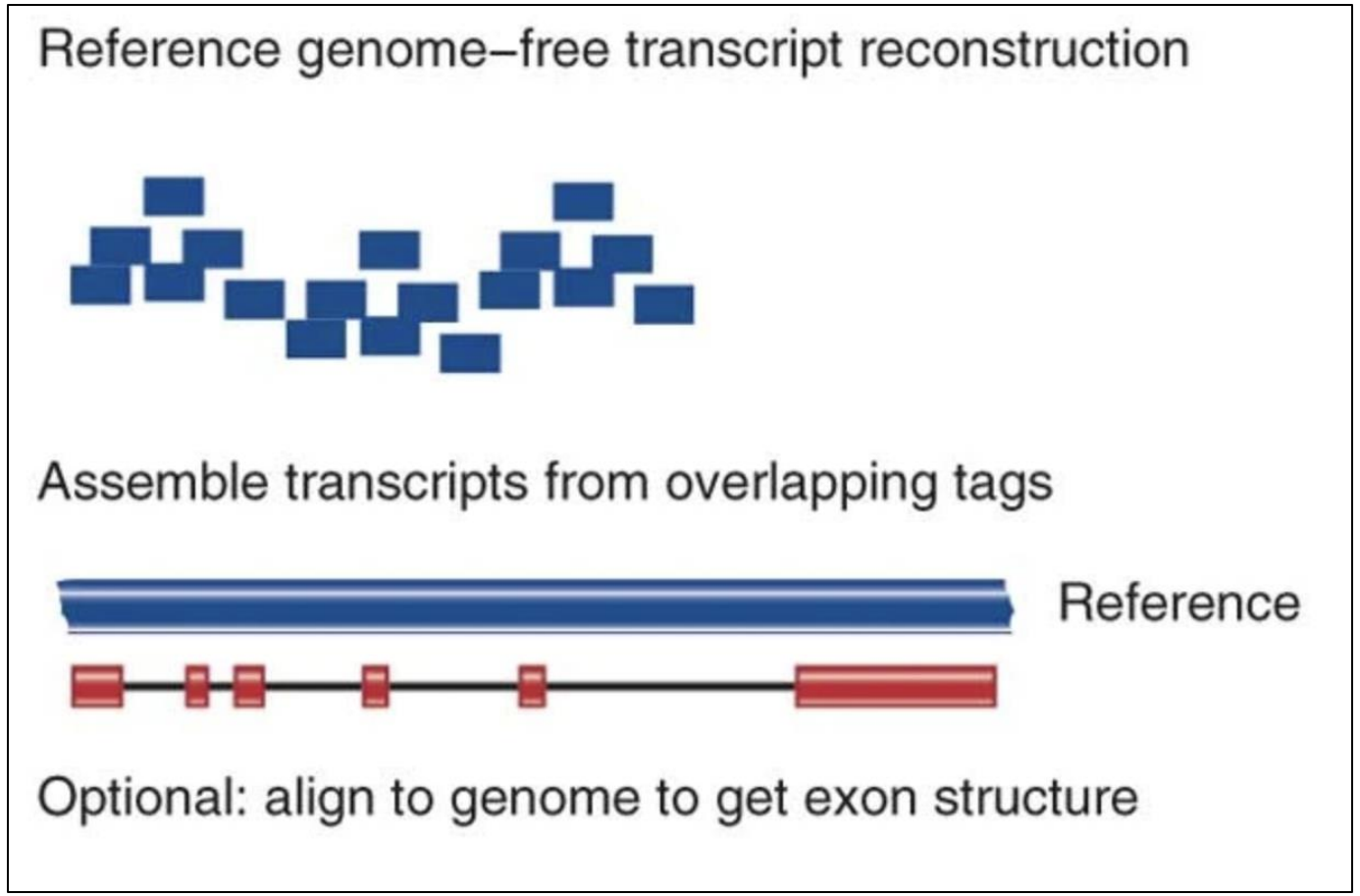
Alignment and assembly



Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud, PLOS Computational Biology 2015

Alignment strategies for RNA-seq data

De novo assembly

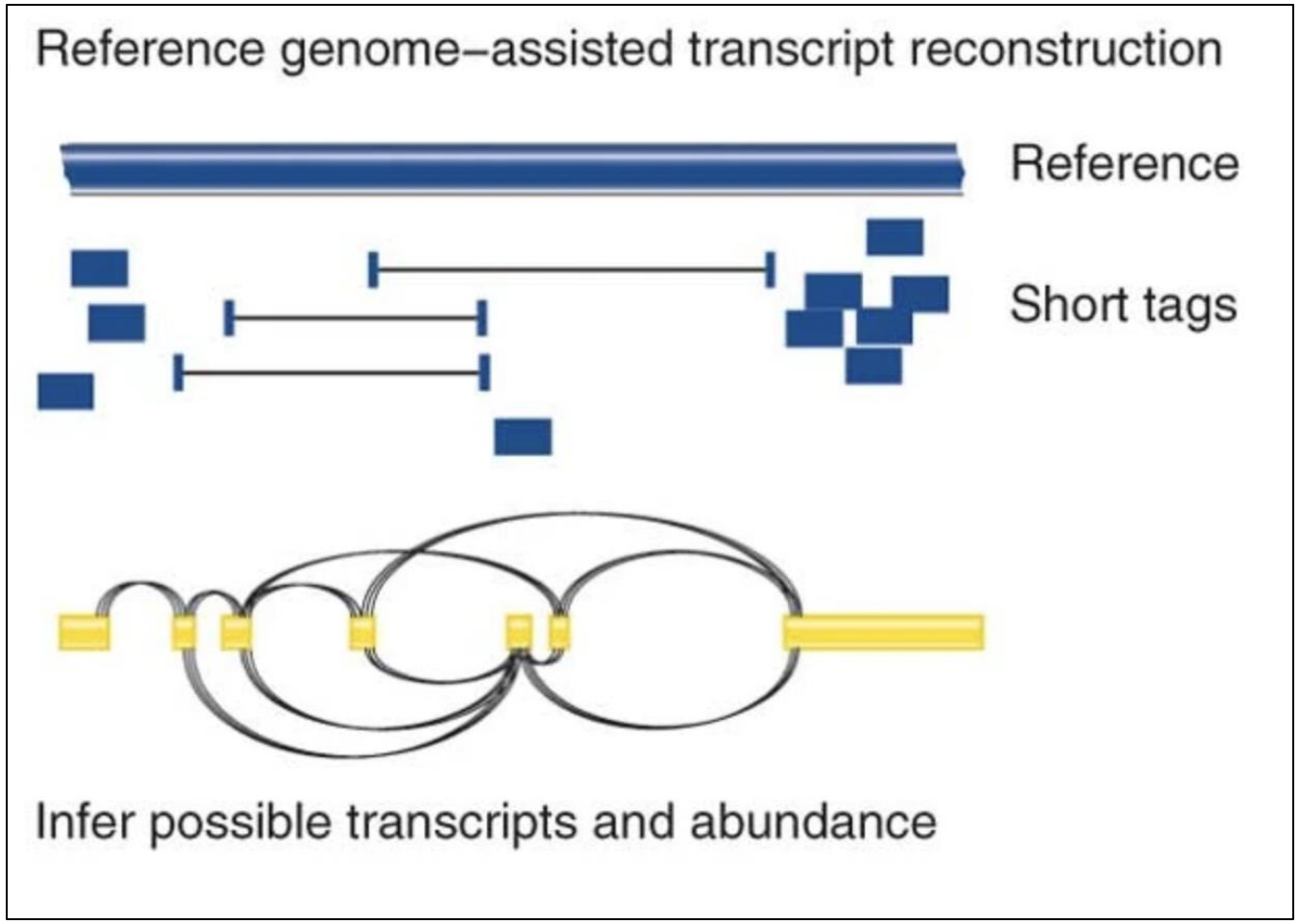


- Fusion transcripts
- Unannotated transcripts
- New splice variants

Cloonan and Grimmond, Nature Methods 2010

Alignment strategies for RNA-seq data

Align to reference genome

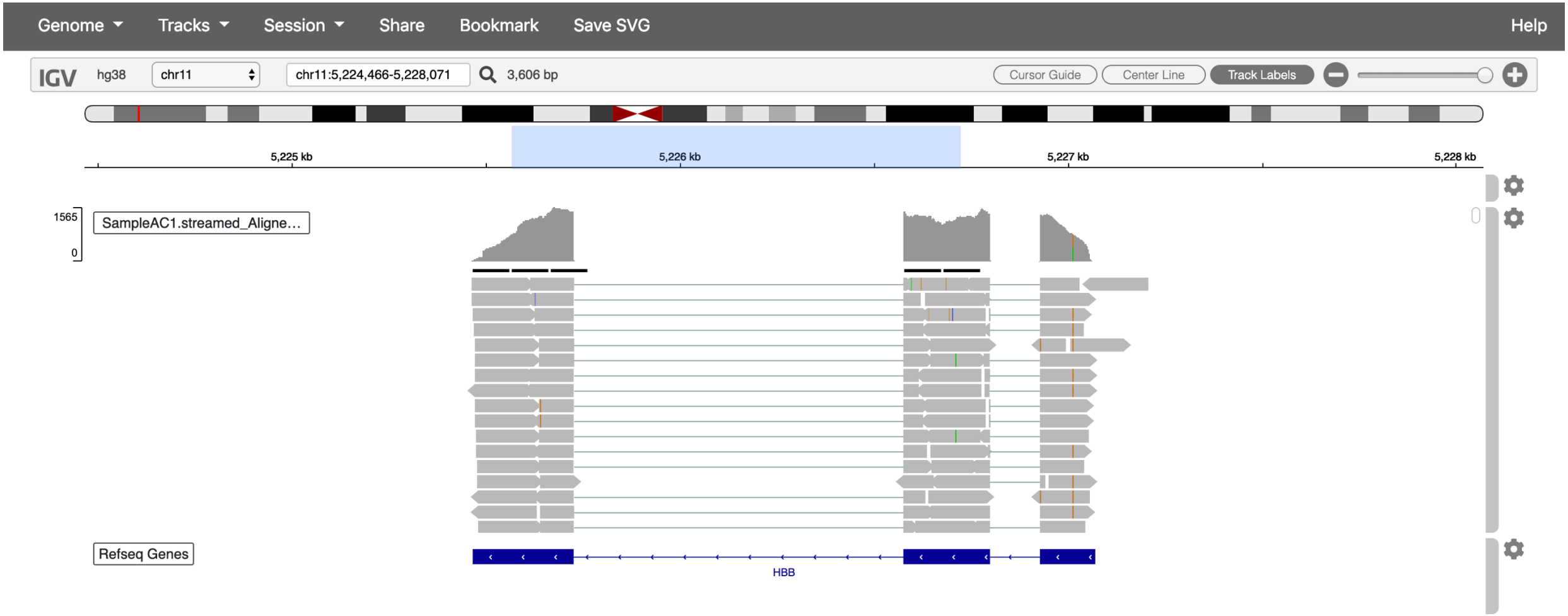


”Splice-aware” aligners like HISAT2

StringTie

Cloonan and Grimmond, Nature Methods 2010

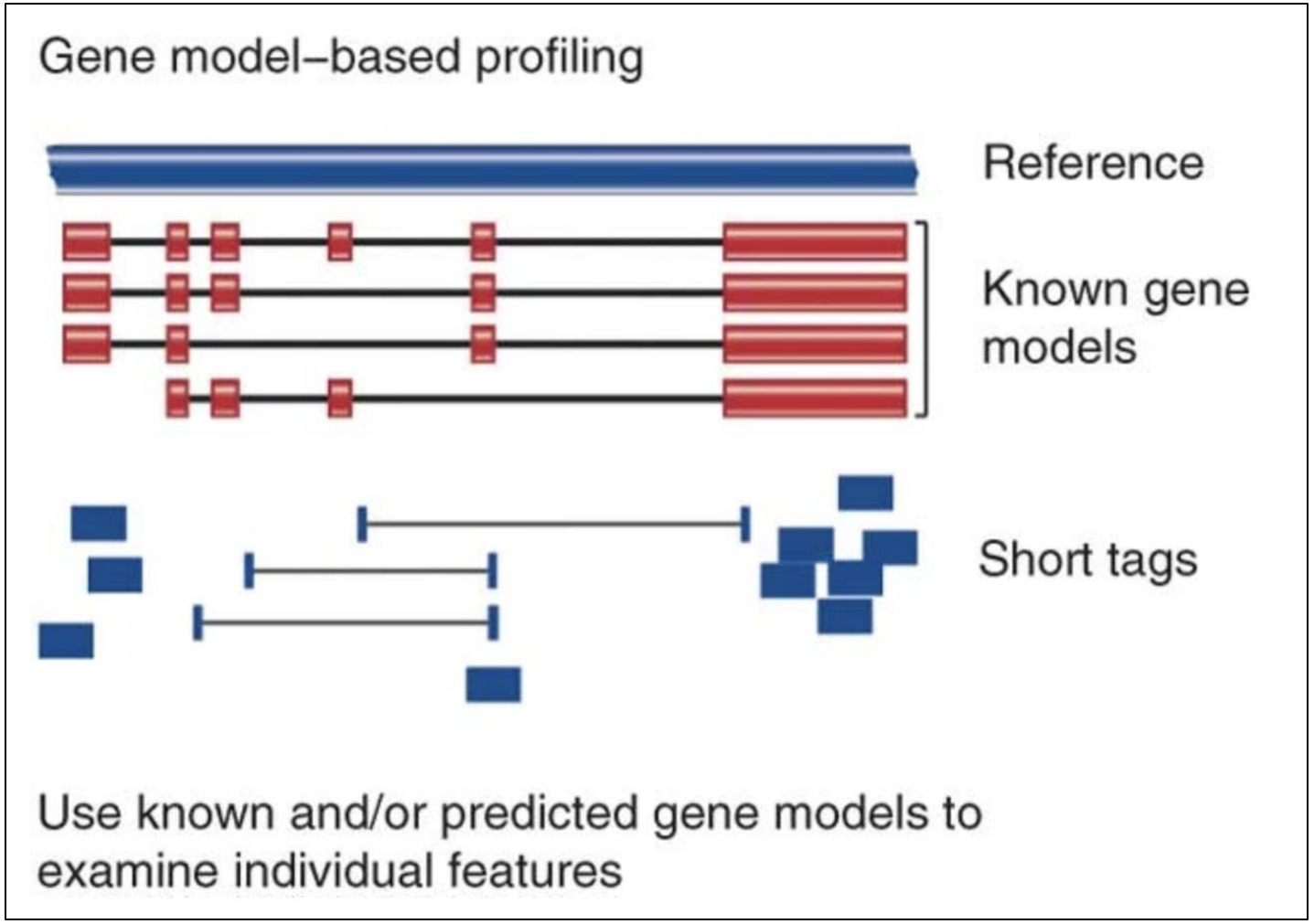
Alignment to reference genome



Source: UC Davis bioinformatics

Alignment strategies for RNA-seq data

“Align” to transcriptome



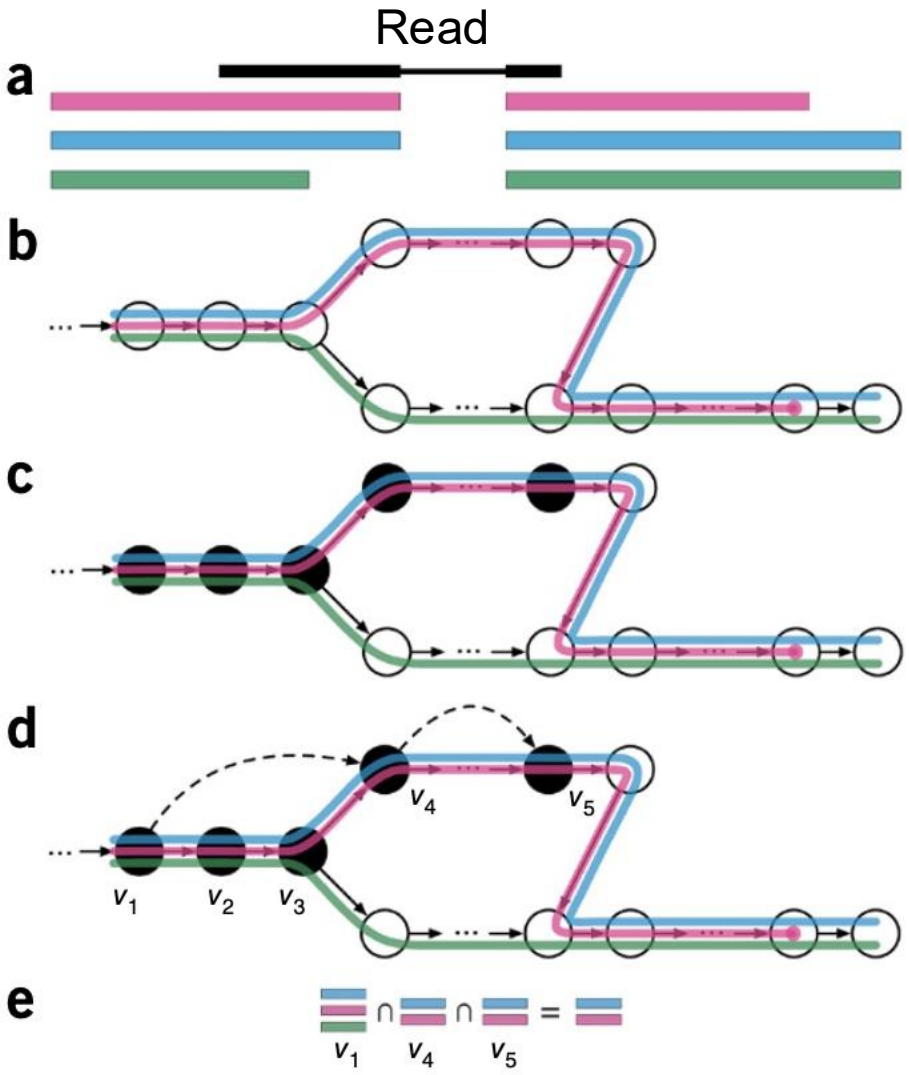
Kallisto

Cloonan and Grimmond, Nature Methods 2010

Kallisto ”pseudoalignment”

- A normal read alignment specifies where a read aligns and how (which nucleotides in read match which target sequences in reference)
- A pseudoalignment of a read is the set of transcript sequences that the read is compatible with

Which transcripts could this read have come from?

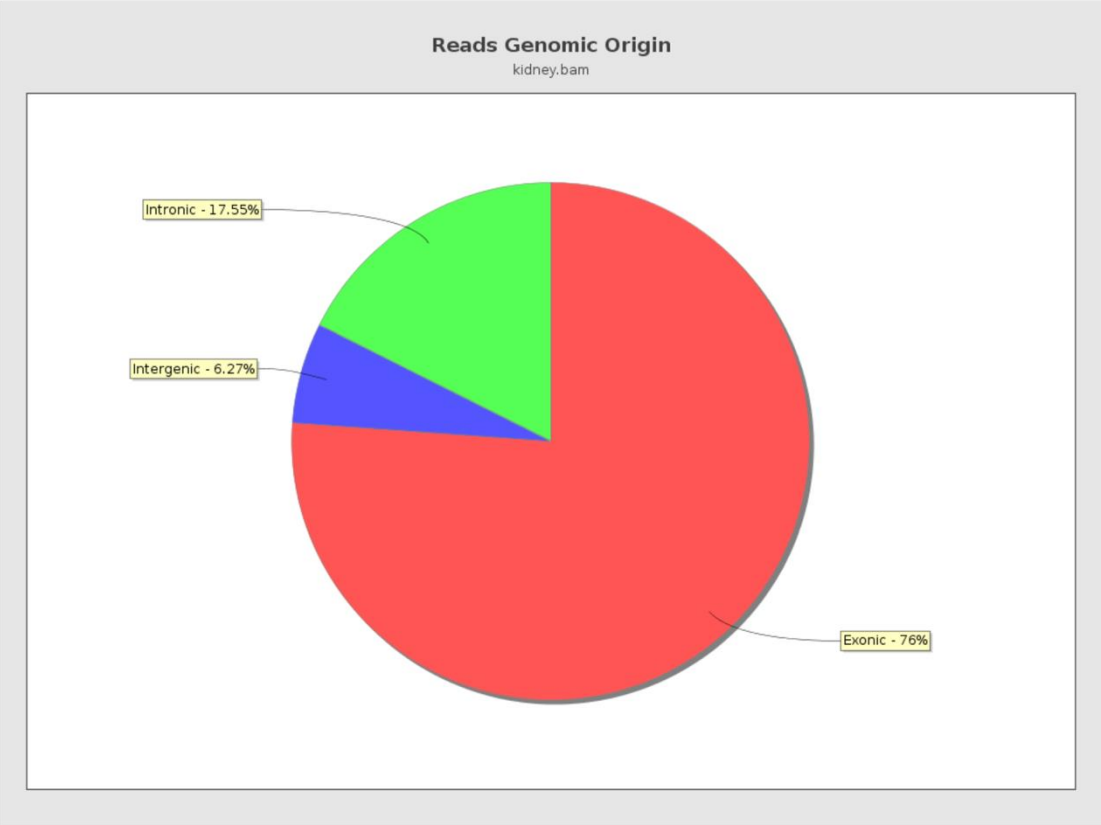


Transcripts from the reference transcriptome

Bray et al., Nature Biotechnology 2016

Checking RNA-seq alignment quality

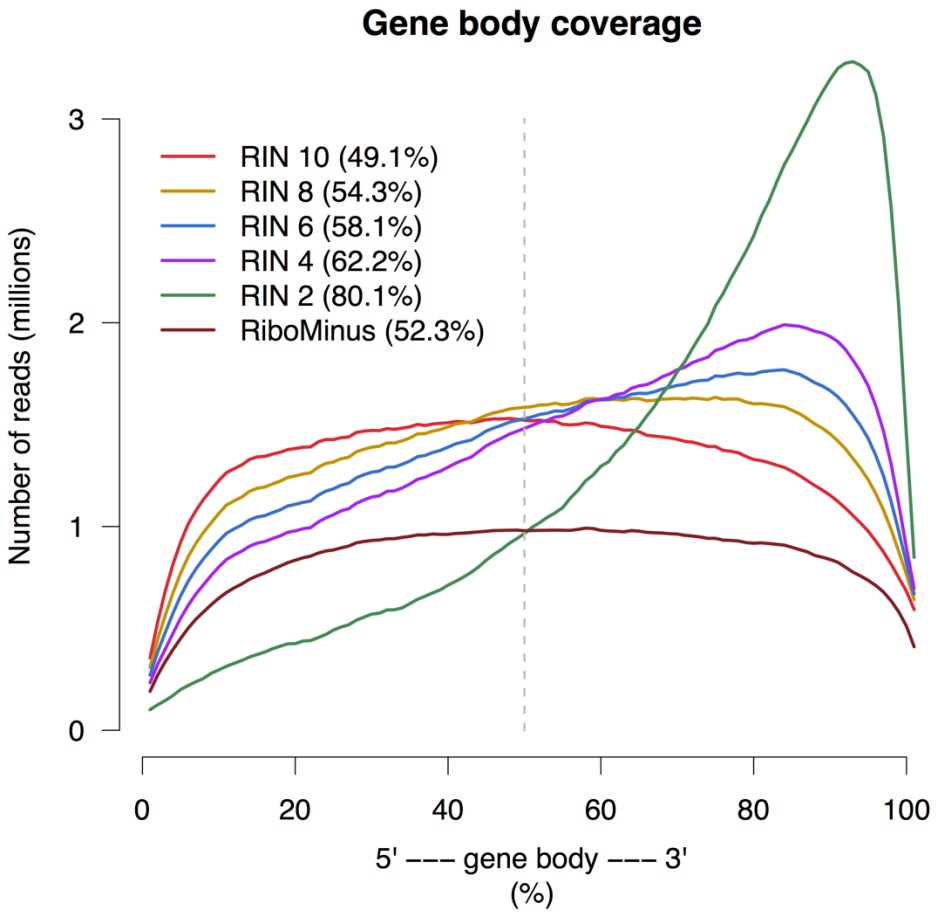
- Percentage of total sequenced reads mapped to the intended target region (e.g. coding regions).



From Qualimap tool, expect > 60% exonic reads

Checking RNA-seq alignment quality

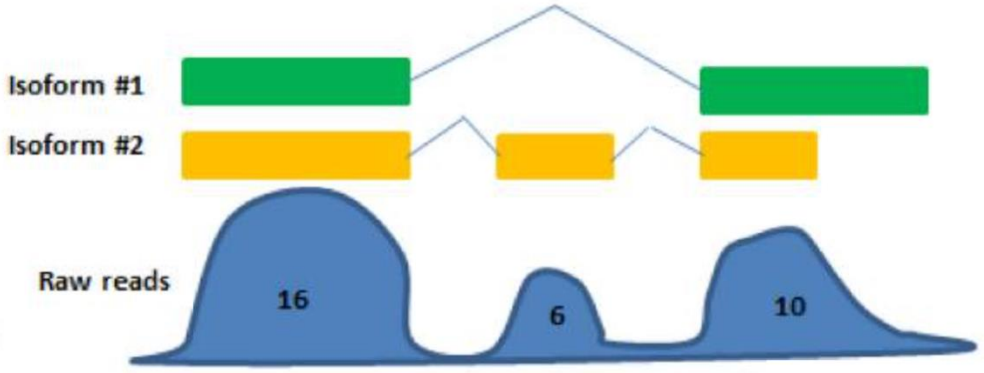
- **5’-3’ bias:** nonuniform coverage of genes in the samples
- If some bias is observed, important that it is similar across samples/groups.
- Can be reflective of sample quality, but bias can also be introduced by library preparation methods.



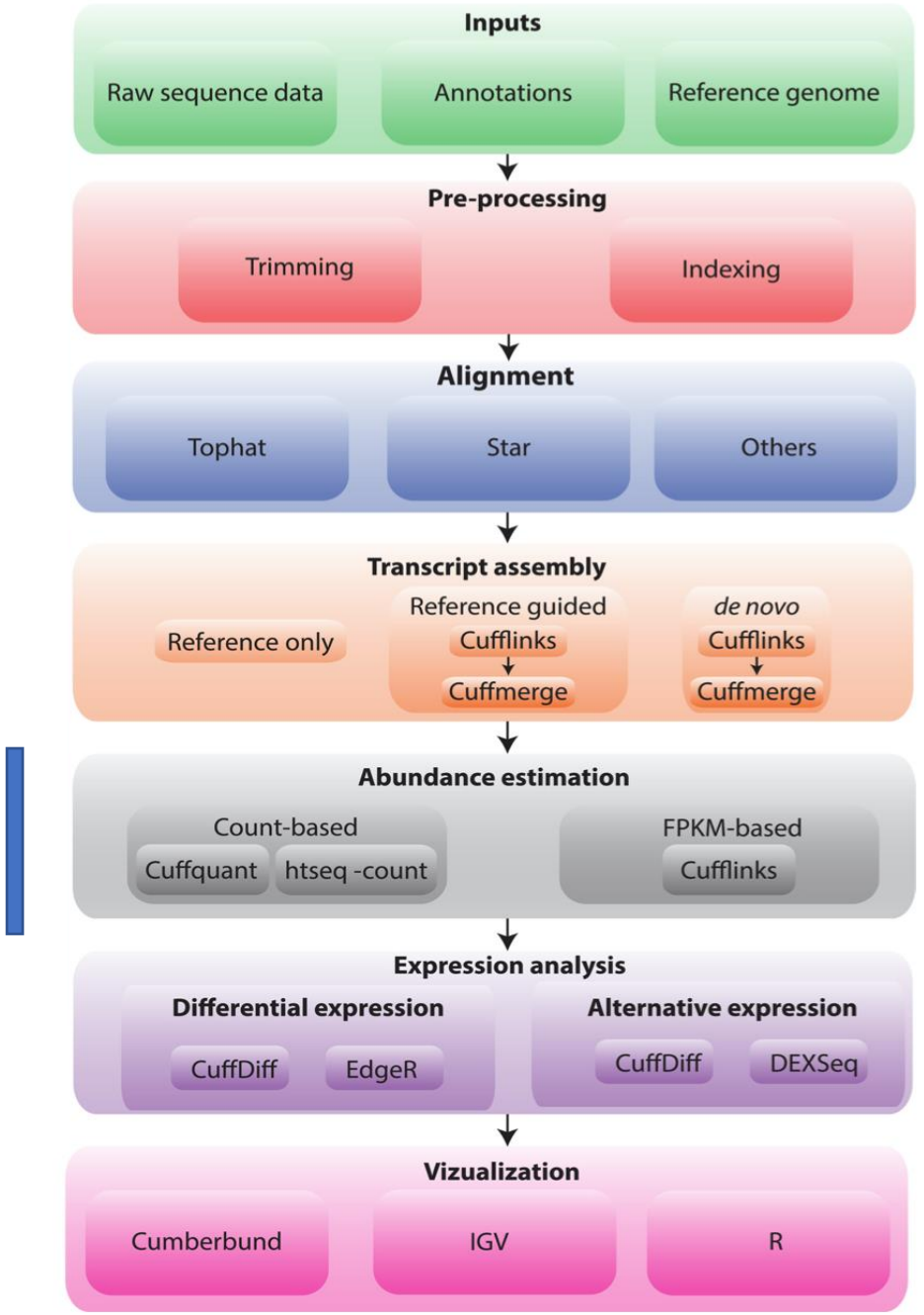
RIN = RNA integrity number

RNA-seq analysis workflow

- There are different approaches in quantifying the expression of genes from RNA-seq data.
- A simple approach for quantifying a gene would be to count the reads that fall in (or overlap with) each of the exons that are annotated to belong to that gene. Once we have the counts for each exon, we could sum them up to get a total count for that gene.
- Example tool: featureCounts



Zhao et al., PLOS One, 2015



Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud, PLOS Computational Biology 2015

Samples

Genes

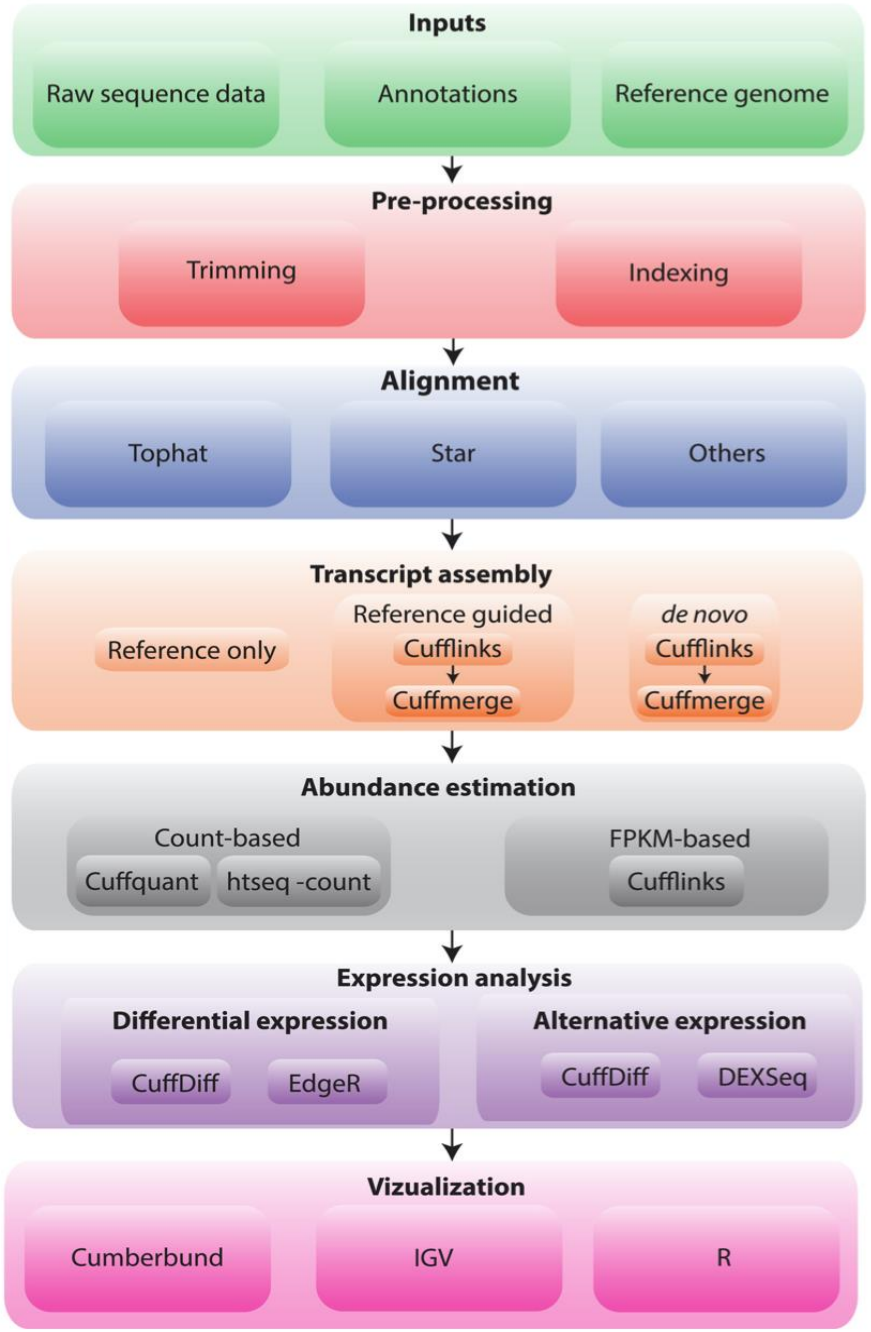
0	3	1	0	5
12	19	3	10	7
4	1	8	6	2
3	6	5	2	9
7	2	1	4	8

RNA-seq analysis workflow

RNA-seq data has several biases, e.g.

- Length bias
- Sequencing depth
- RNA composition

The number of reads from RNA-seq experiments therefore needs to be normalized to be comparable.



Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud, PLOS Computational Biology 2015

RNA-seq normalization methods

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

DE = differential expression

RNA-seq normalization methods

- **TPM** (Transcripts Per Million):

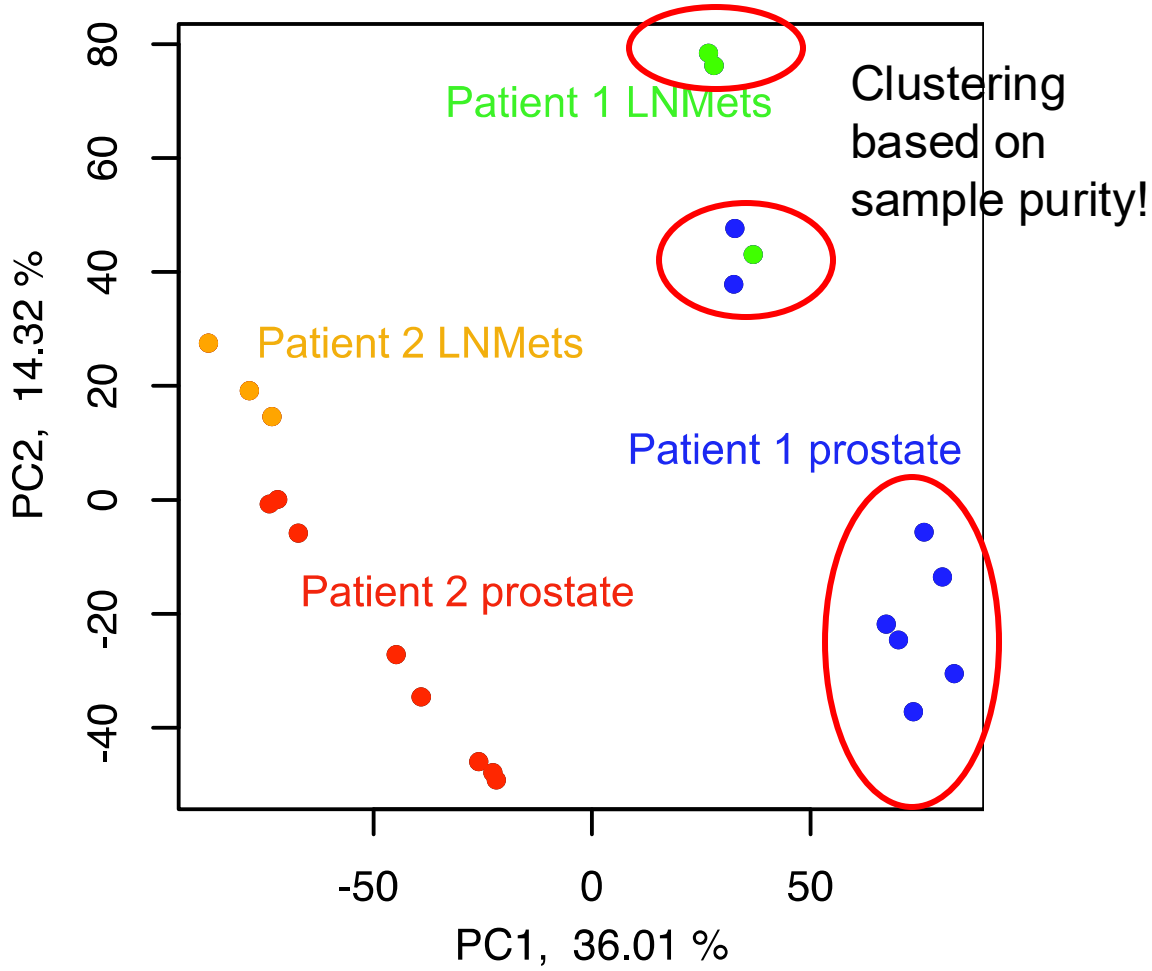
$$TPM = 10^6 \times \frac{Reads\ mapped\ to\ transcript / Transcript\ length}{Sum(Reads\ mapped\ to\ transcript / Transcript\ length)}$$

- **RPKM** (Reads Per Kilobase of transcript per Million mapped reads):

$$RPKM = 10^9 \times \frac{Reads\ mapped\ to\ transcript}{Total\ reads \times Transcript\ length}$$

Checking quality of gene expression estimates

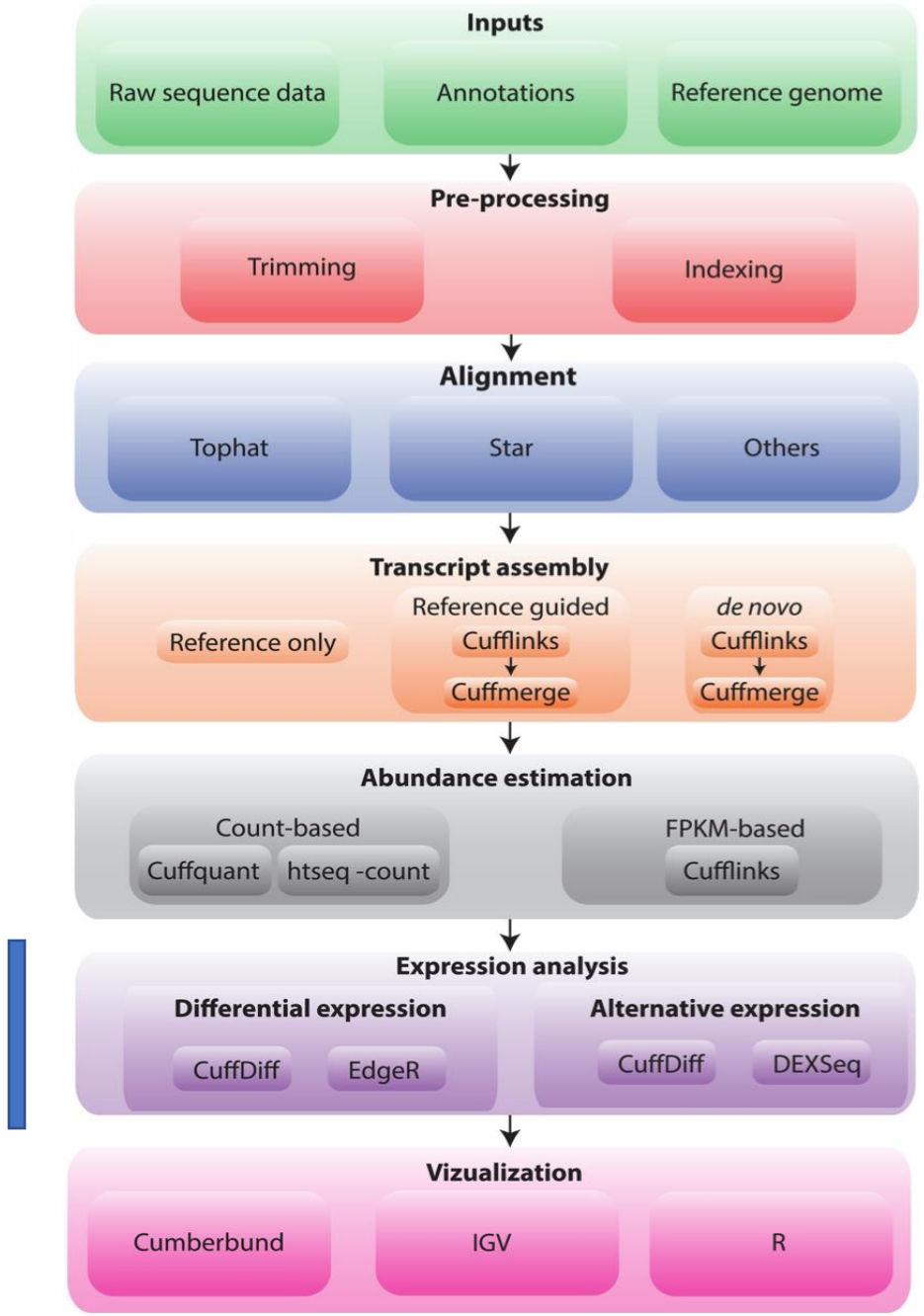
- Clustering can be used to identify which samples are closely related based on gene expression, but can also identify **outlier** samples or **batch effects** (e.g. sample purity, see Aran et al., Nature Communications, 2015).
- Samples might have to be excluded or re-sequenced.
- Can identify variables that have to be accounted for in downstream analyses.



RNA-seq analysis workflow

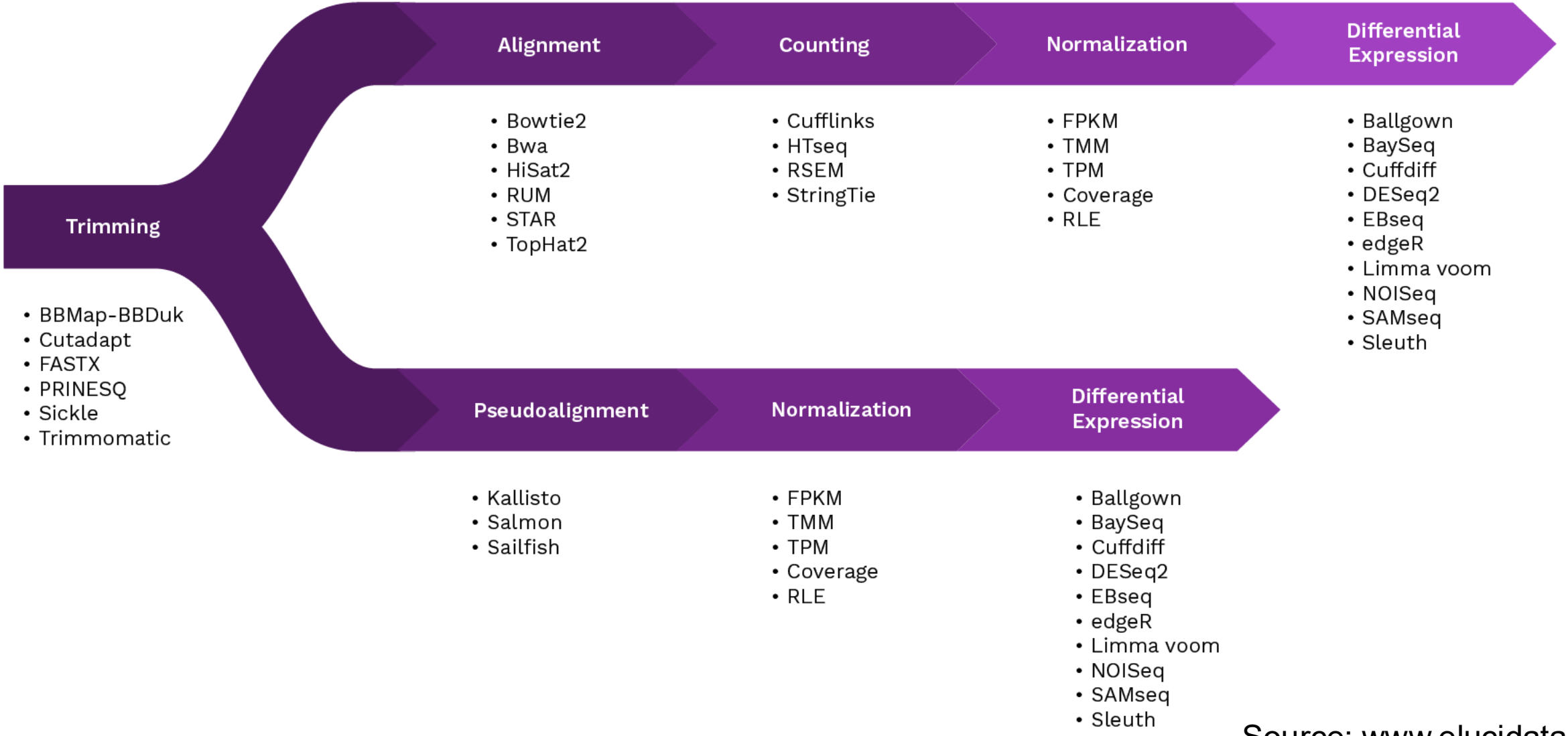
Differential expression

- Connecting gene expression back to genotype/phenotype
- What genes/transcripts are expressed at higher/lower levels in different sample groups?
  - Tumor vs normal samples
  - Are the differences statistically significant, accounting for variance and noise in the data?
- Tools: DESeq2, edgeR, Cuffdiff…
- [DESeq2 tutorial]



Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud, PLOS Computational Biology 2015

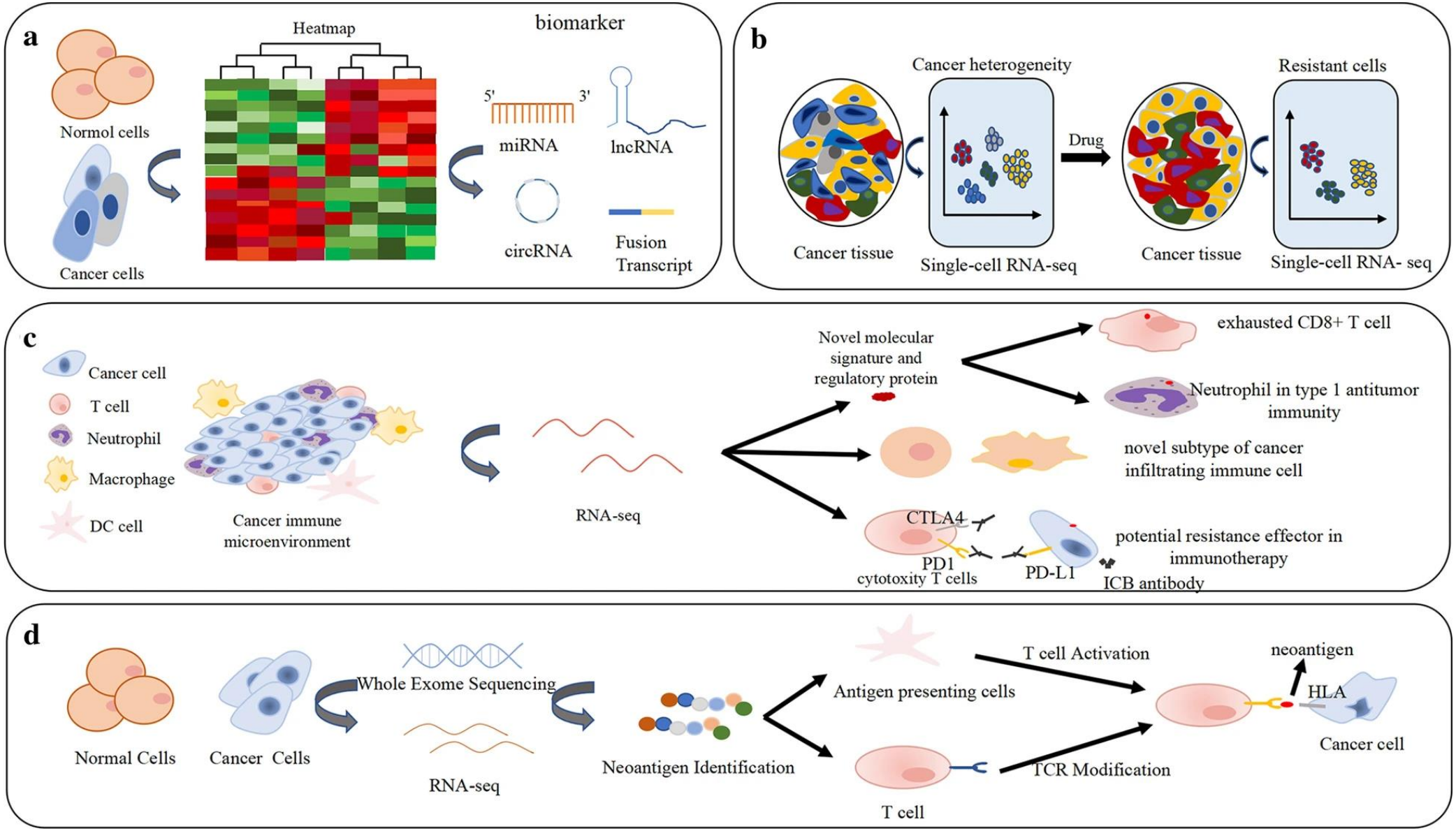
Popular tools and approaches for gene expression quantification and differential expression analysis



Source: www.elucidata.io

Differential gene expression, biomarkers, fusions

Diagnostic relevance- heterogeneity, evolution, and drug resistance in cancer

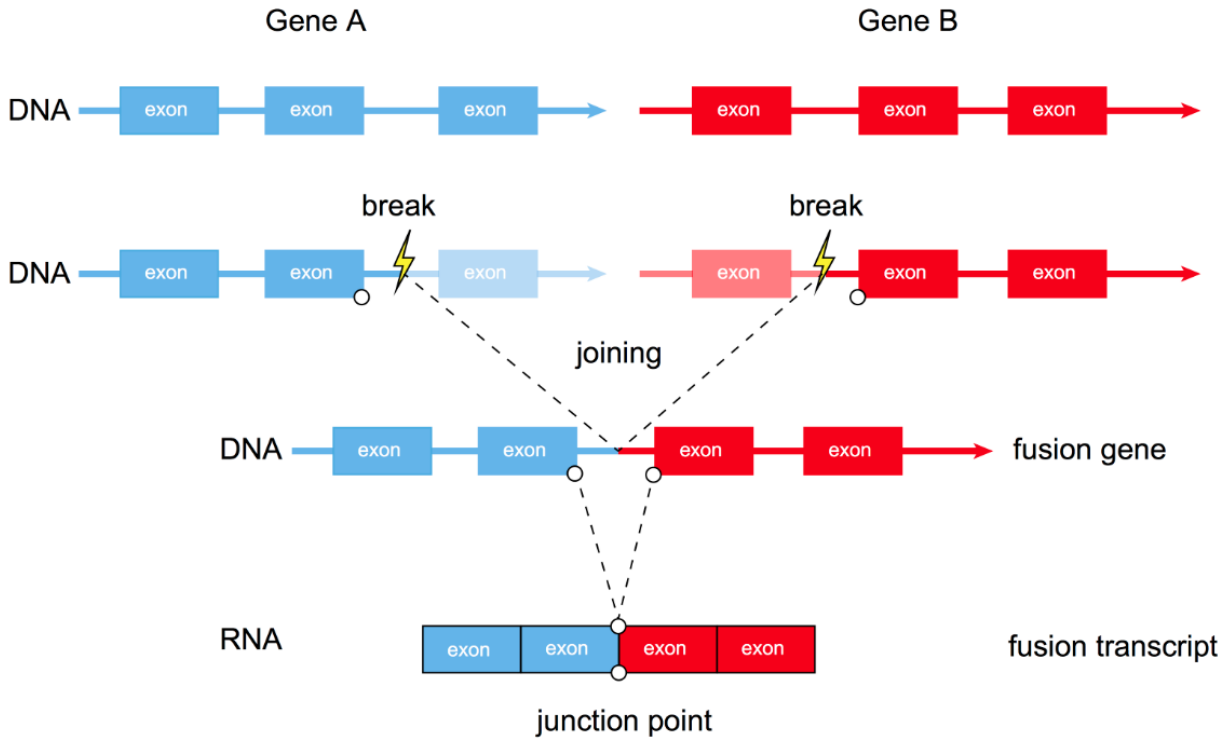


Molecular signatures, subtyping, estimation of cell type composition

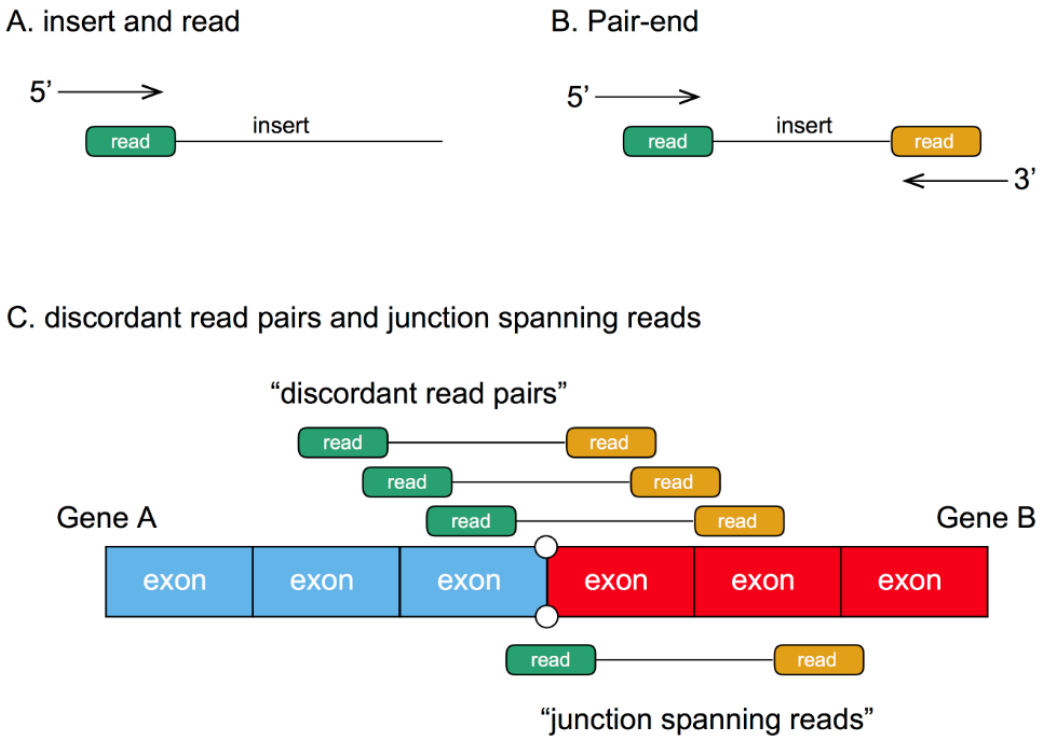
Neoantigen profiling and prediction

Hong et al., Journal of Hematology & Oncology, 2020

Application 1: Fusion detection

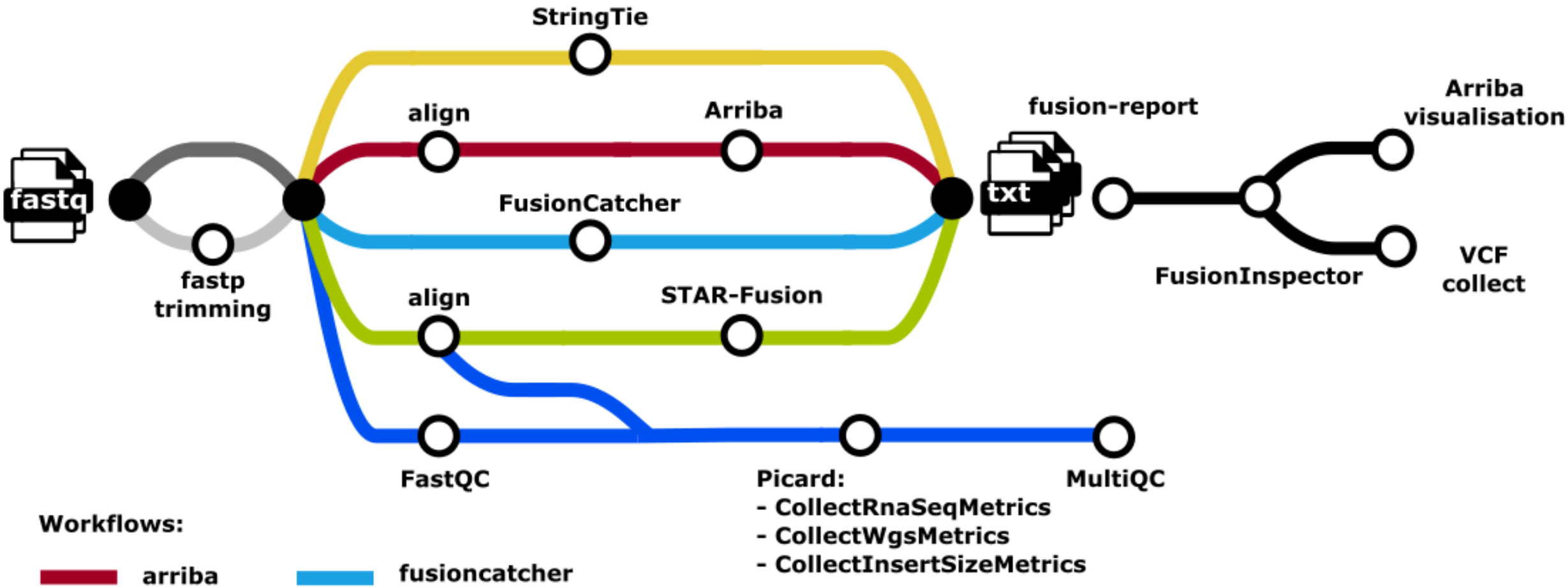


Fusion transcripts can arise due to genomic rearrangements.



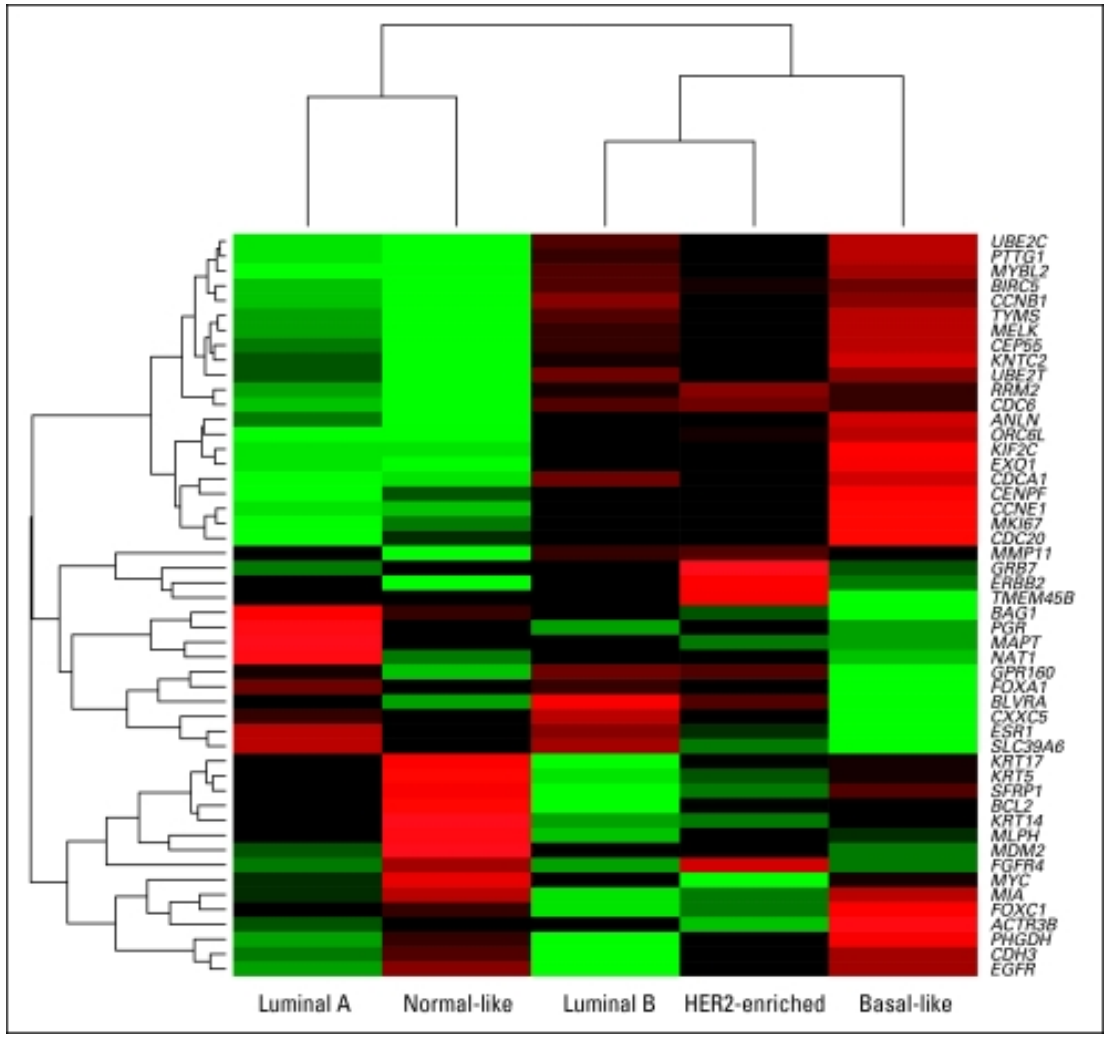
Paired-end RNA-seq can be used to detect fusion transcripts.

Source: The Jackson Laboratory



nf-core/rnafusion pipeline

Application 2: classification + subtyping



Parker et al., Journal of Clinical Oncology, 2009

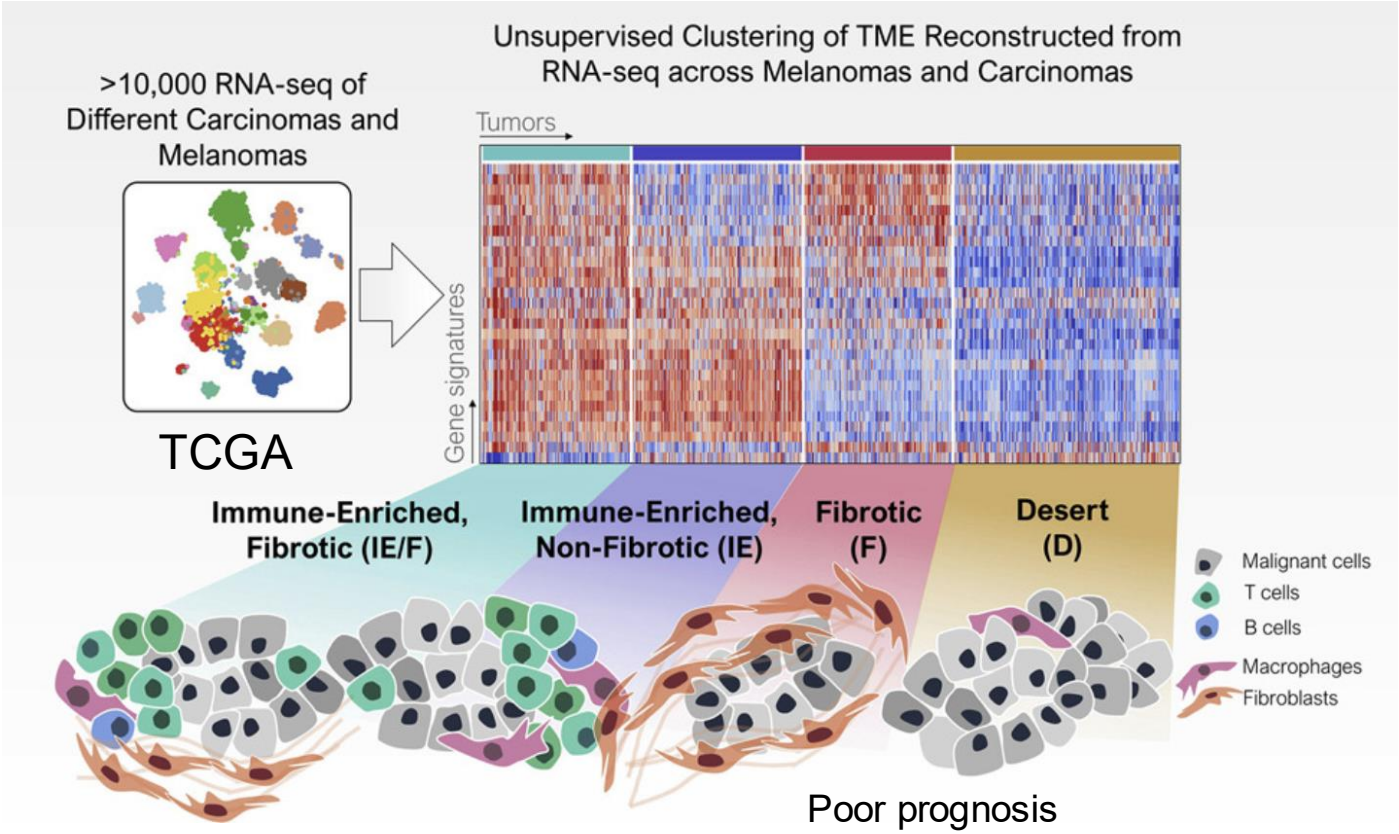
PAM50 in breast cancer

- List of 50 genes that classify breast cancers into one of five subtypes based on expression (Parker et al., Journal of Clinical Oncology, 2009).
- The different subtypes reflect different levels of aggressiveness and can benefit from different treatment strategies.
- For example, luminal subtypes have been shown to benefit more from hormonal therapy tamoxifen (Chia et al., Clinical Cancer Research, 2012).

Application 2: classification + subtyping

"A tumor personality test to guide therapeutic decision making"

Conserved pan-cancer microenvironment subtypes



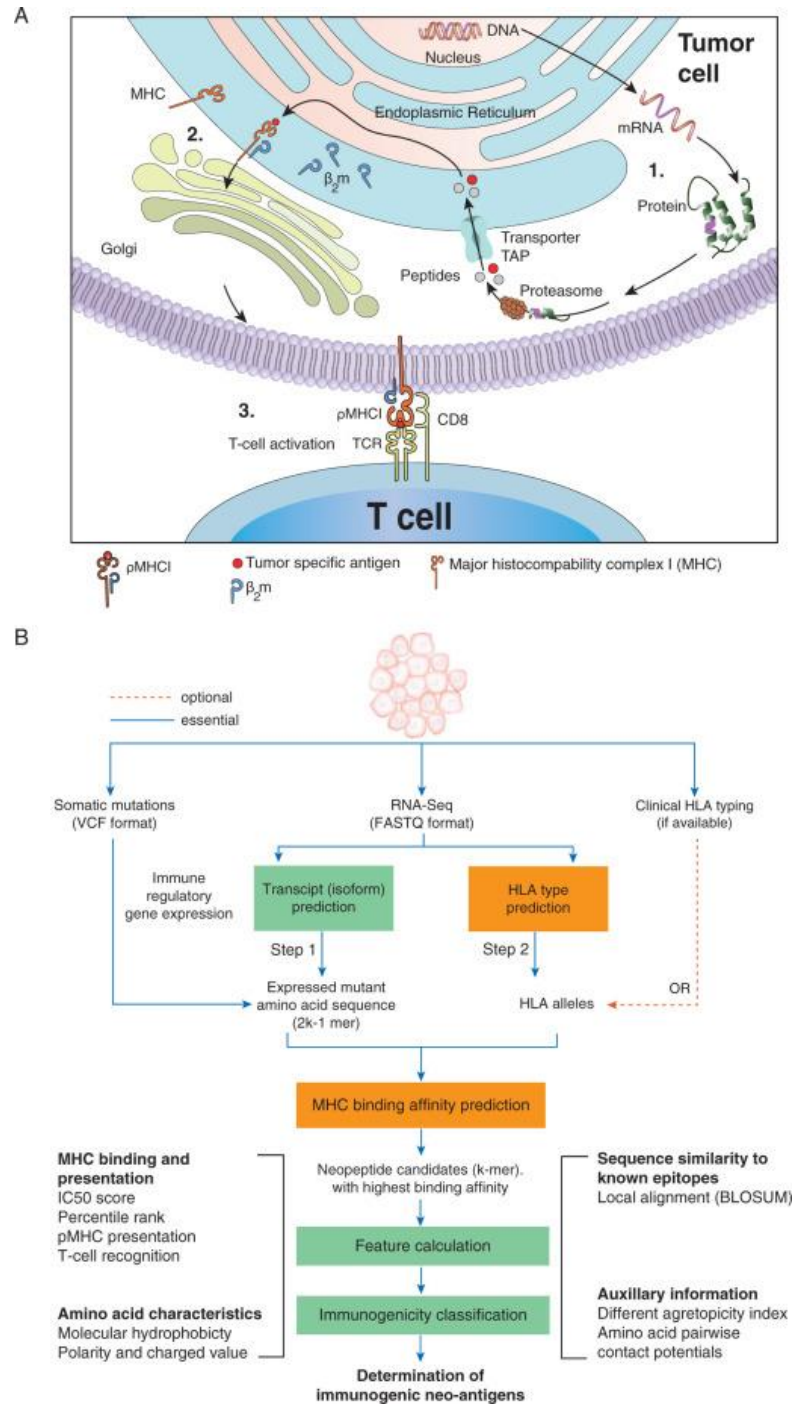
Better prognosis
Better immunotherapy response
Lower tumor proliferation

Poor prognosis
Poor response to immunotherapy

Bagaev et al, Cancer Cell, 2021

Application 3: neoantigens and peptides

- Tumor-specific mutations form novel immunogenic peptides called neoantigens, some of which can elicit T cell responses.
- Neoantigens can be used as a biomarker for predicting patient response to cancer immunotherapy.
- Combine information from somatic mutations (DNA) and RNA-seq data to identify candidate peptides.
- Tools for neoantigen discovery: Neopepsee, ScanNeo, ASNEO…



Kim et al., Annals of Oncology, 2018

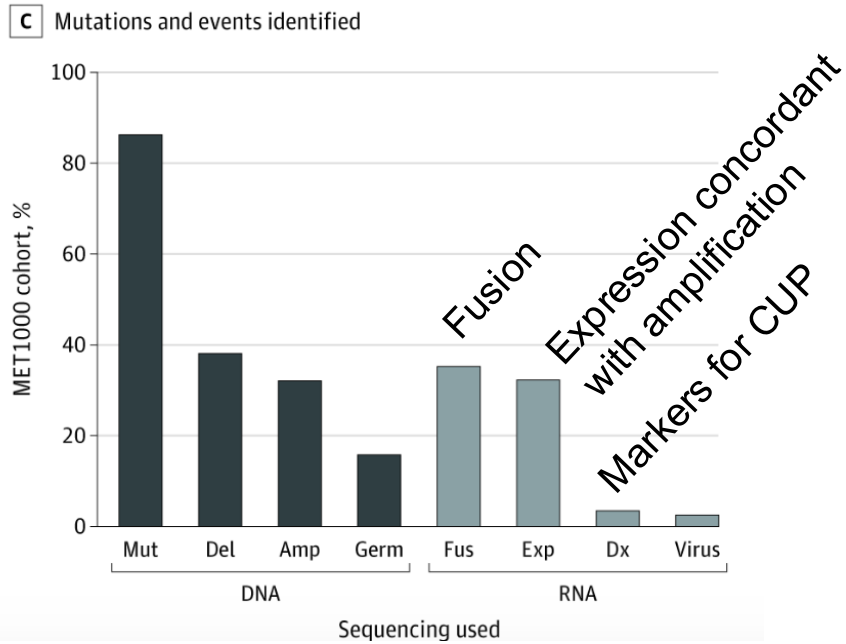
Clinical impact of comprehensive DNA and RNA sequencing

Research

JAMA Oncology | Original Investigation

Assessment of Clinical Benefit of Integrative Genomic Profiling in Advanced Solid Tumors

Erin F. Cobain, MD; Yi-Mi Wu, PhD; Pankaj Vats, PhD; Rashmi Chugh, MD; Francis Worden, MD; David C. Smith, MD; Scott M. Schuetze, MD, PhD; Mark M. Zalupski, MD; Vaibhav Sahai, MD; Ajjai Alva, MD; Anne F. Schott, MD; Megan E. V. Caram, MD; Daniel F. Hayes, MD; Elena M. Stoffel, MD; Michelle F. Jacobs, MS, CGC; Chandan Kumar-Sinha, PhD; Xuhong Cao, MS; Rui Wang, MS; David Lucas, MD; Yu Ning, MS; Erica Rabban, BS; Janice Bell, AS; Sandra Camelo-Piragua, MD; Aaron M. Udager, MD, PhD; Marcin Cieslik, PhD; Robert J. Lonigro, PhD; Lakshmi P. Kunju, MD; Dan R. Robinson, PhD; Moshe Talpaz, MD; Arul M. Chinnaiyan, MD, PhD



The Michigan Oncology Sequencing Program

- Inclusion of 1138 advanced/metastatic patients between 2011-2018 (MET1000 cohort)
- Tumor biopsy sequencing with paired normal
- Whole-exome, targeted capture
- RNA-sequencing
  - Fusion detection
  - Classification of Cancer Of Unknown Primary (CUP)
- Clinical benefit rate from NGS-directed therapy

Exercise Set 4 overview

- Analyze RNA-seq from the same samples we used during Exercise Set 2 for alignment and variant calling
- Look at QC metrics from RNA-seq reads and alignments
- Quantify gene and transcript expression using HISAT2 + StringTie and Kallisto
- Perform fusion calling from RNA-seq using Kallisto and pizzly

Important: HISAT2 alignment and initial QC steps have been run for you!


KAROLINSKA INSTITUTET
ANNO 1810
Karolinska
Institutet