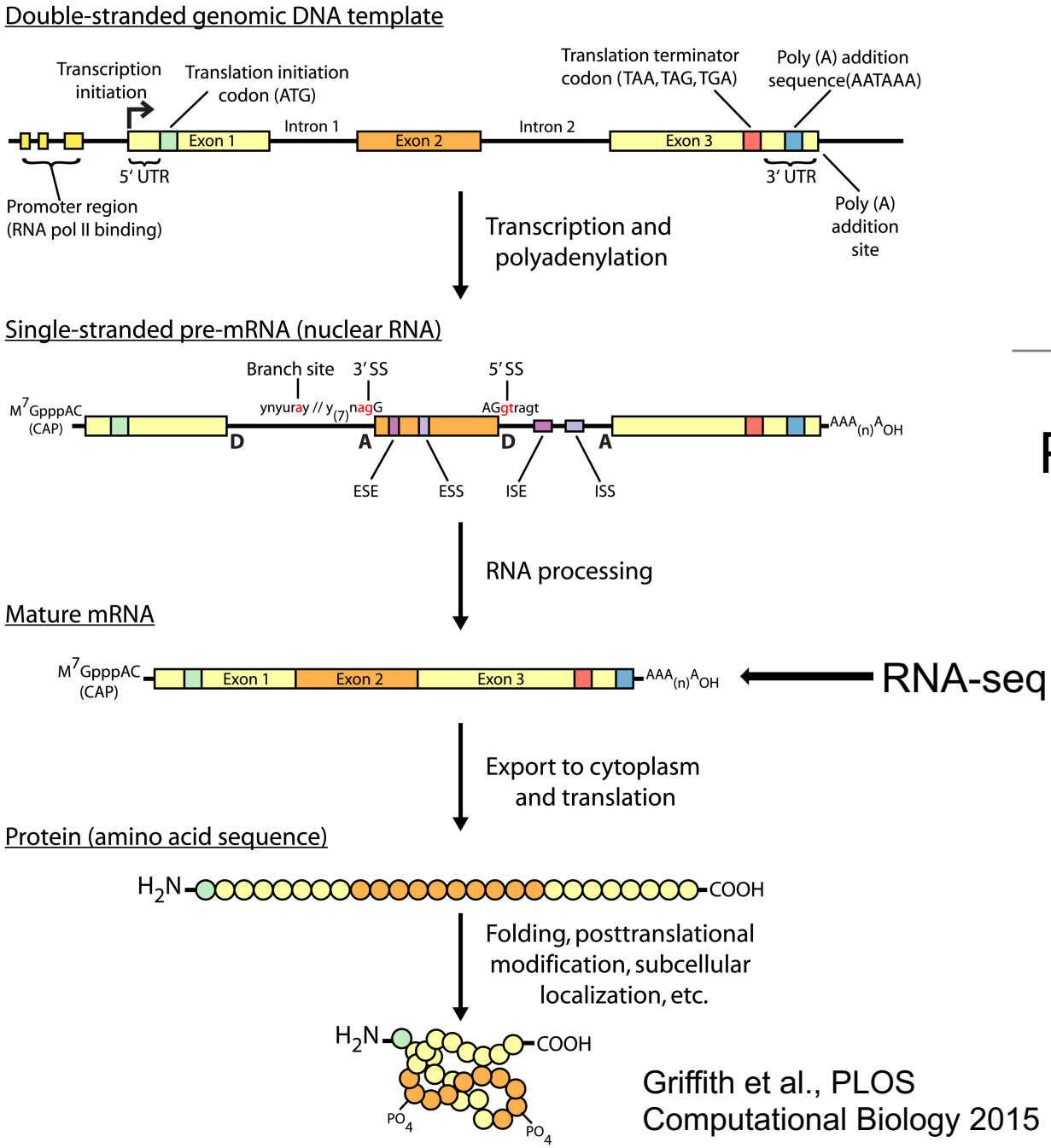
# RNA-sequencing

Clinical Cancer Genomics
26 April 2023

# 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

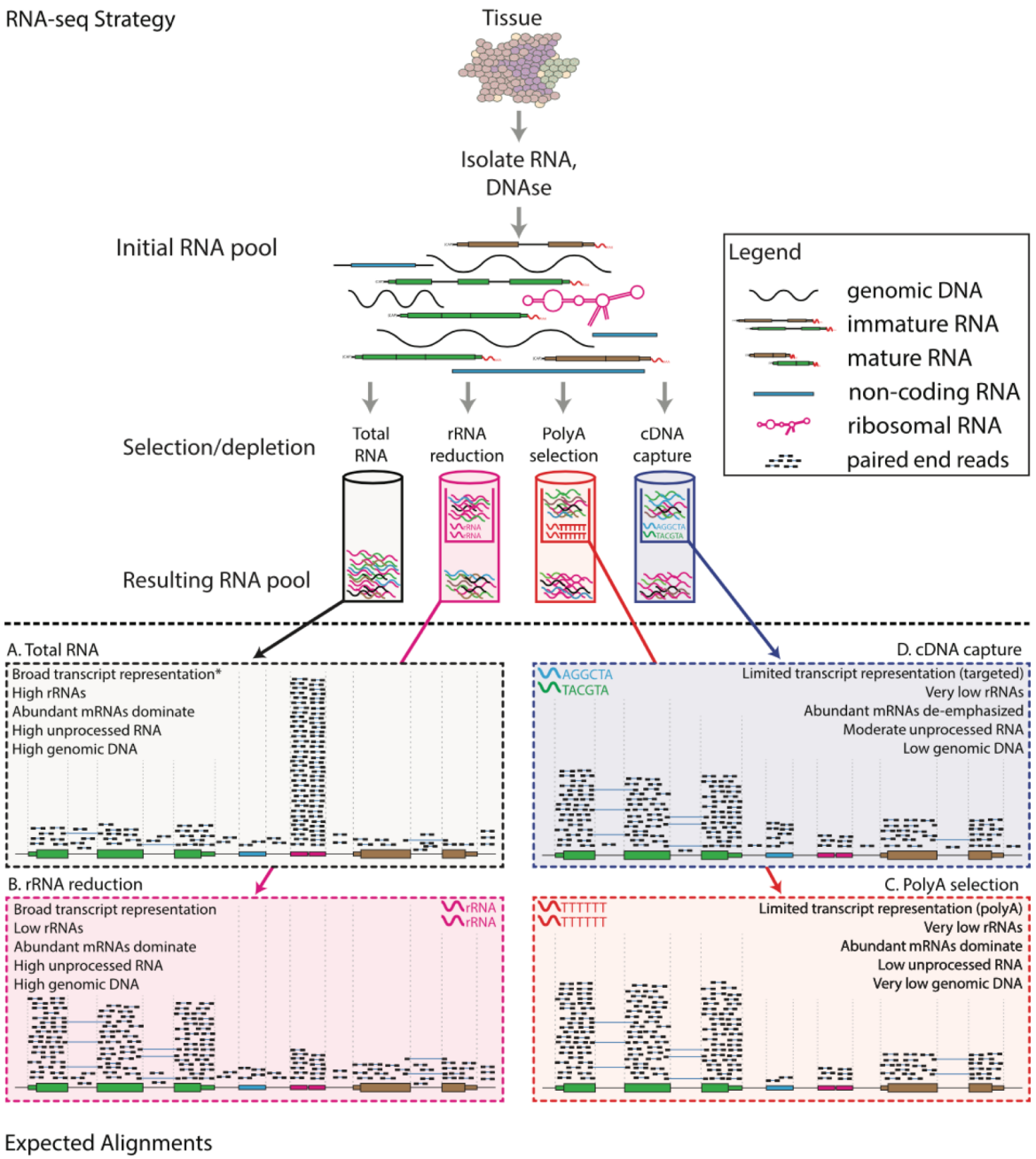
# The central dogma of molecular biology

Phenotype = expression of genetic information modified by environmental factors

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

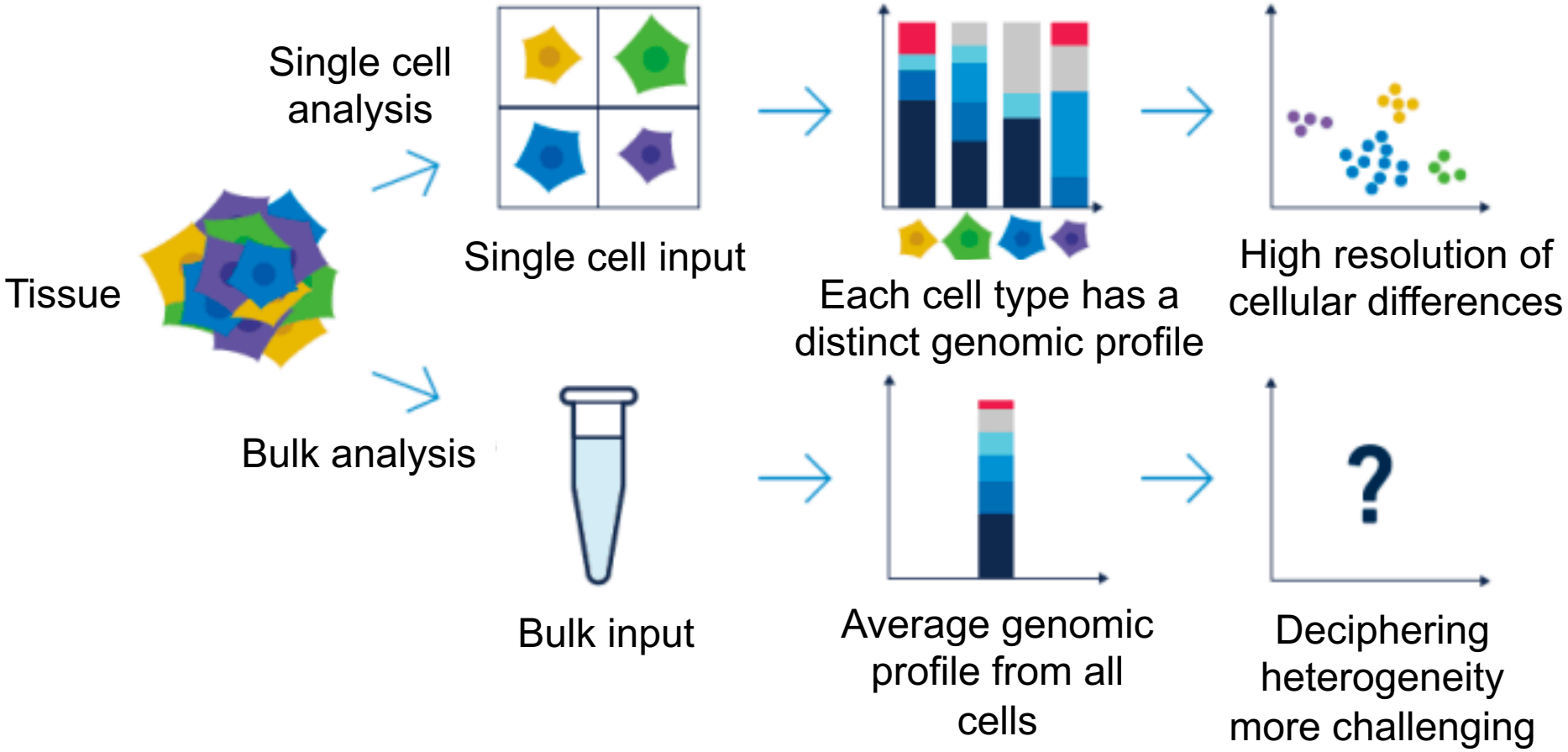
# RNA-seq strategies

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

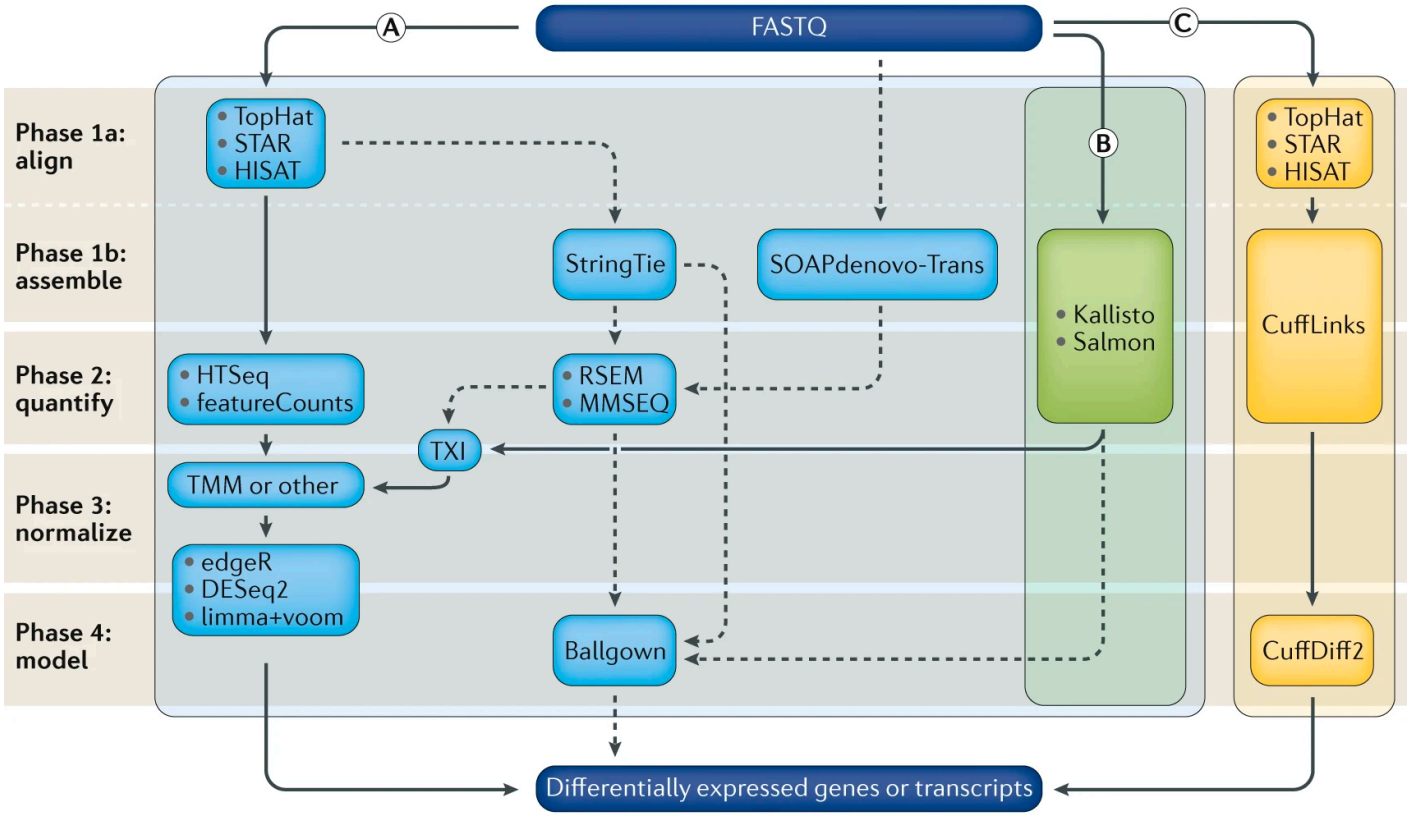
# Bulk vs single-cell RNA-seq



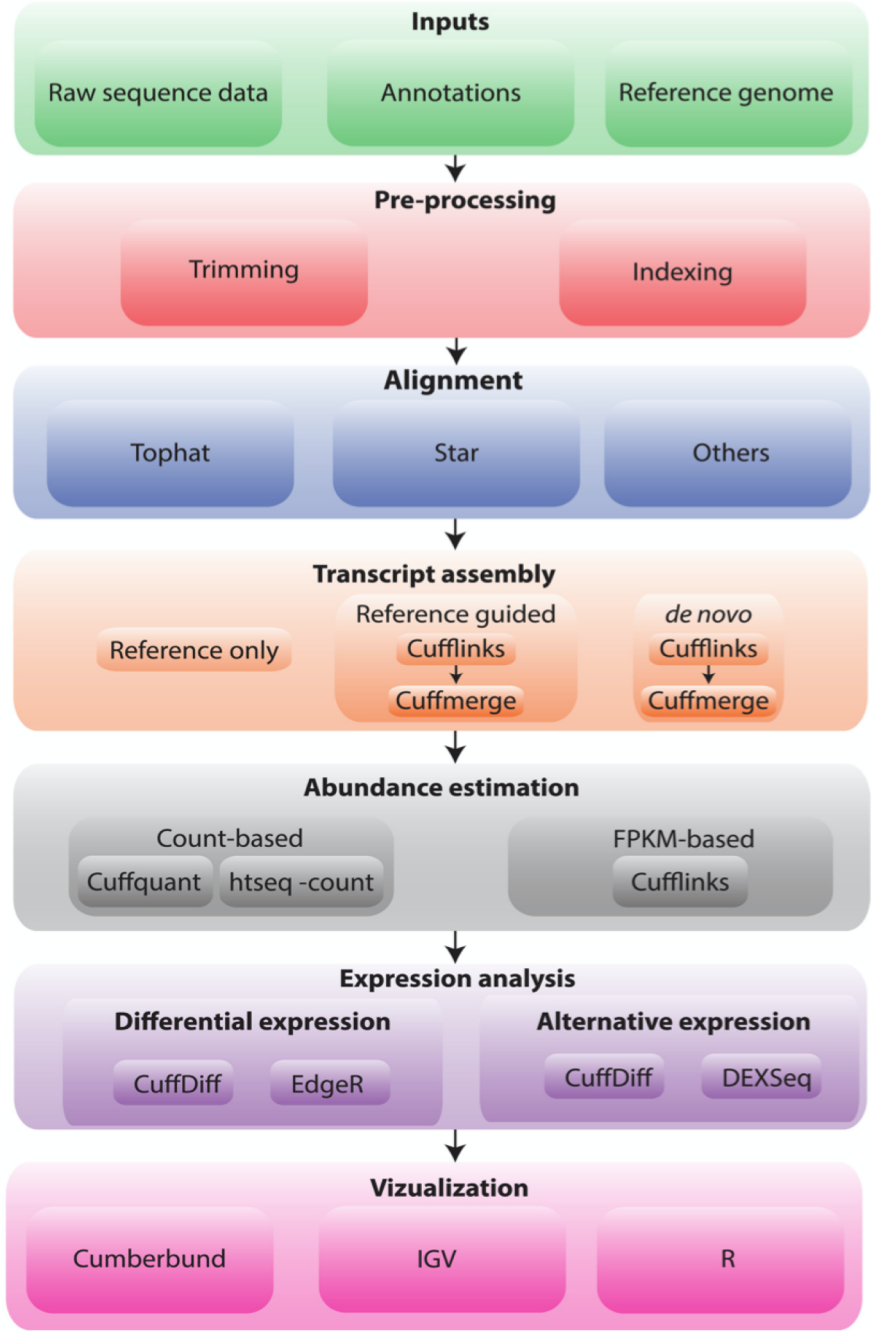

**CHALLENGES:**
Cost
Complexity of library prep
Data sparsity
Analysis tools

Source: 10x Genomics

# RNA-seq workflows for gene expression and differential expression analysis



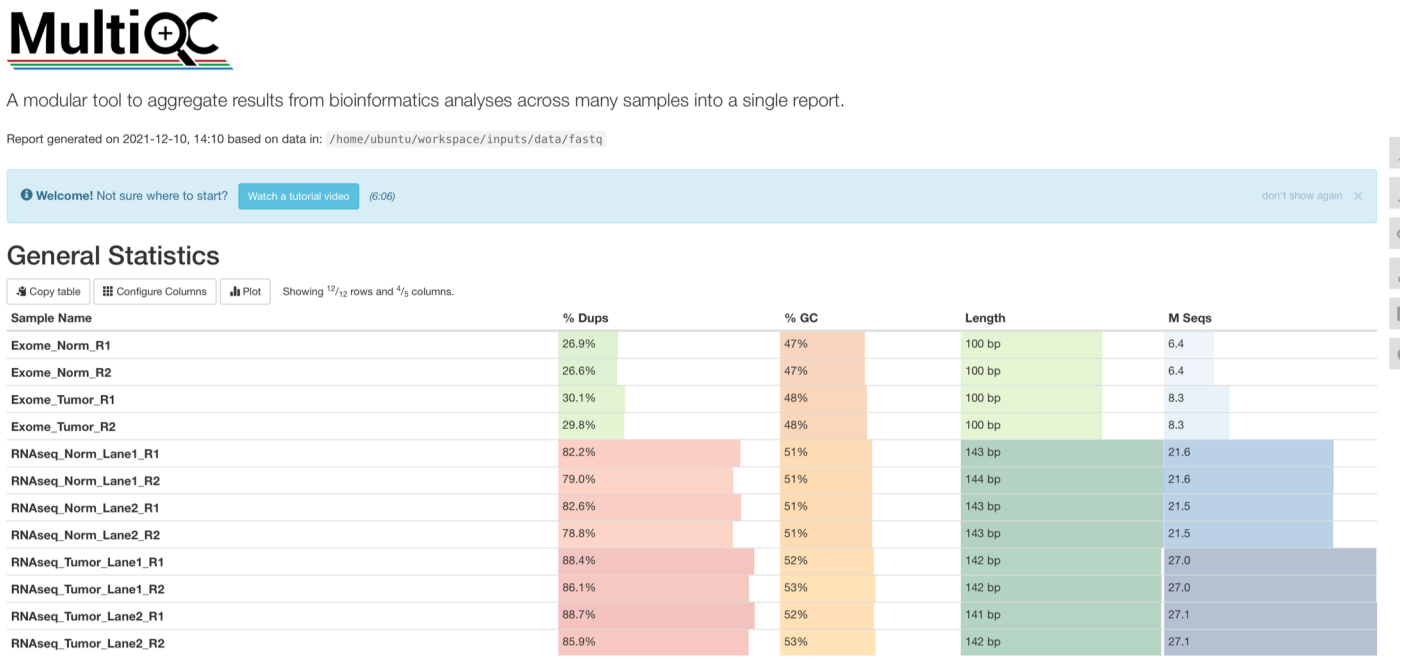

Stark et al., Nature Reviews Genetics 2019




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

# Quality control (QC) of RNA-seq data

- What are the concerns?
  - Adapter contamination
  - Ribosomal RNA fraction
  - Fraction aligned reads
  - Genomic origin of reads
  - 5’-3’ bias in transcript coverage
  - …
- Many tools exist for QC, e.g. FastQC, Qualimap, RSeQC

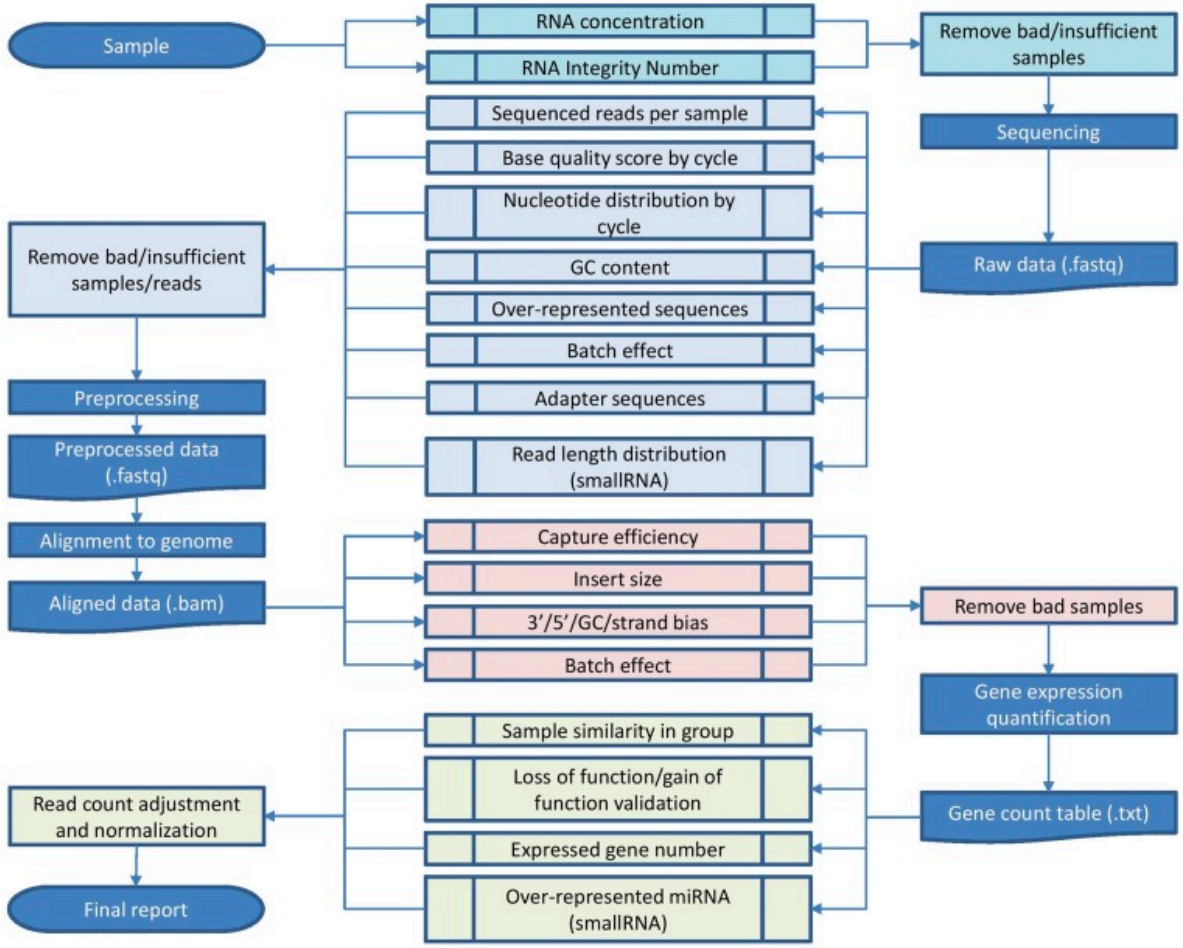



| Sample Name | % Dups | % GC | Length | M Seqs |
|---|---|---|---|---|
| Exome_Norm_R1 | 26.9% | 47% | 100 bp | 6.4 |
| Exome_Norm_R2 | 26.6% | 47% | 100 bp | 6.4 |
| Exome_Tumor_R1 | 30.1% | 48% | 100 bp | 8.3 |
| Exome_Tumor_R2 | 29.8% | 48% | 100 bp | 8.3 |
| RNAseq_Norm_Lane1_R1 | 82.2% | 51% | 143 bp | 21.6 |
| RNAseq_Norm_Lane1_R2 | 79.0% | 51% | 144 bp | 21.6 |
| RNAseq_Norm_Lane2_R1 | 82.6% | 51% | 143 bp | 21.5 |
| RNAseq_Norm_Lane2_R2 | 78.8% | 51% | 143 bp | 21.5 |
| RNAseq_Tumor_Lane1_R1 | 88.4% | 52% | 142 bp | 27.0 |
| RNAseq_Tumor_Lane1_R2 | 86.1% | 53% | 142 bp | 27.0 |
| RNAseq_Tumor_Lane2_R1 | 88.7% | 52% | 141 bp | 27.1 |
| RNAseq_Tumor_Lane2_R2 | 85.9% | 53% | 142 bp | 27.1 |

MultiQC can generate a joint summary report for all samples from the outputs of 128 different bioinformatics tools.

# Stages of RNA-seq QC

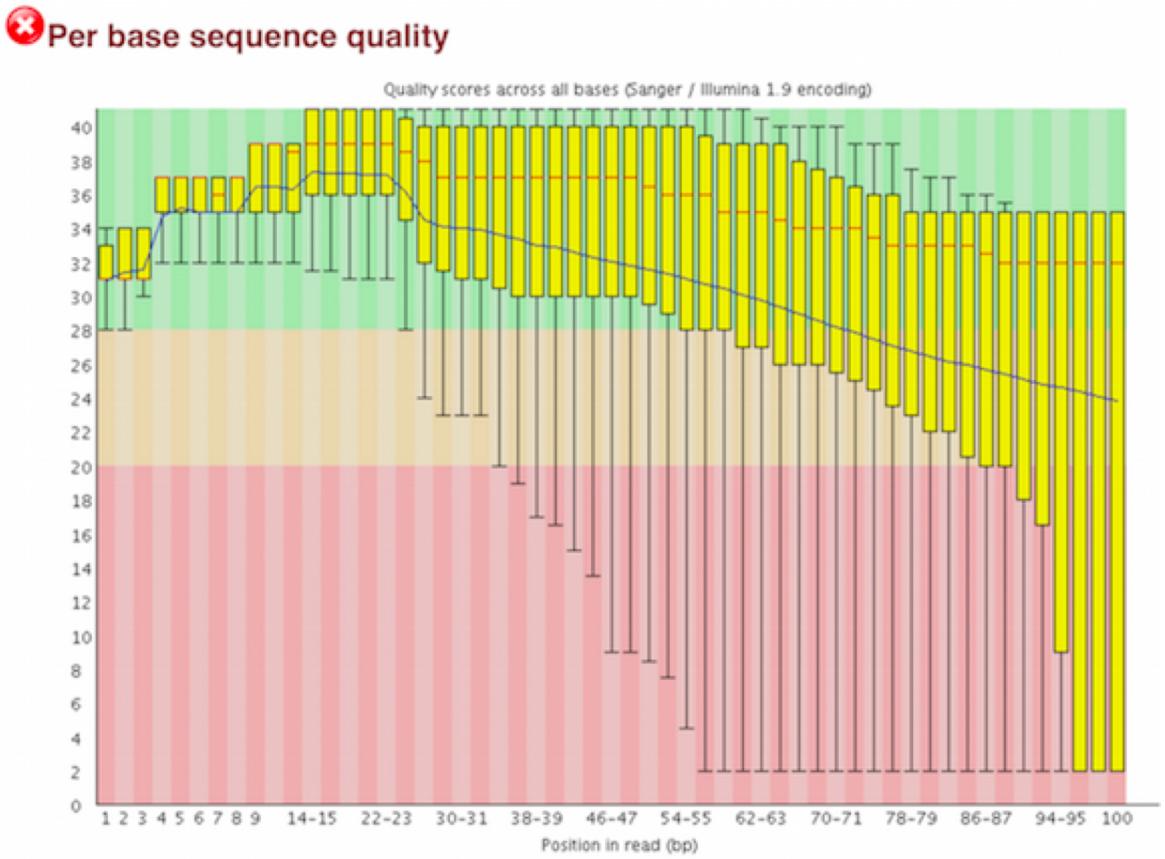
1. Checking RNA quality in samples
2. Checking quality of raw read data in FASTQ files
3. Checking RNA-seq alignment quality
4. Checking quality of gene expression estimates




Sheng et al., Briefings in Functional Genomics, 2017

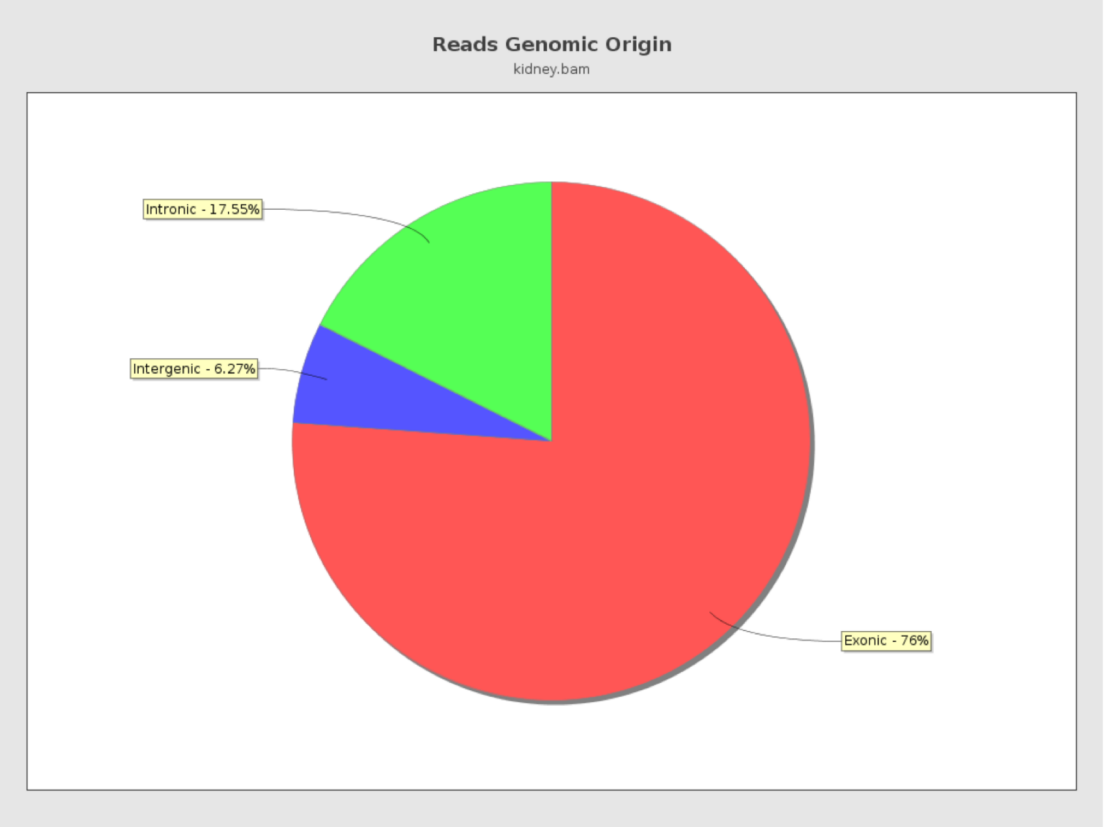
# Quality control (QC) of RNA-seq data

Raw read quality (FastQC)




Are there low quality bases in the sequencing reads?

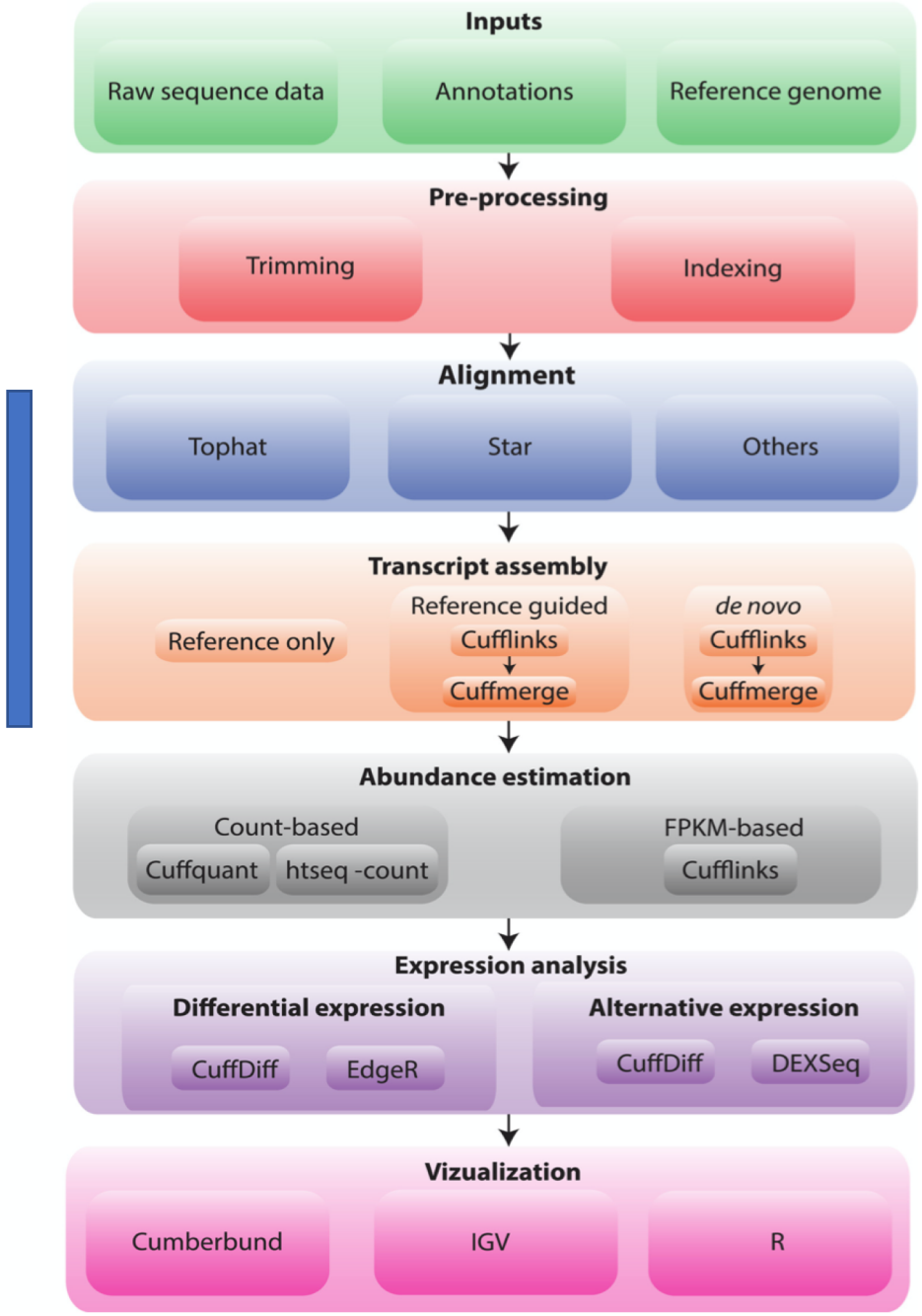
Alignment quality (Qualimap)




Expect > 60% exonic reads

# RNA-seq analysis workflow

## Alignment and assembly

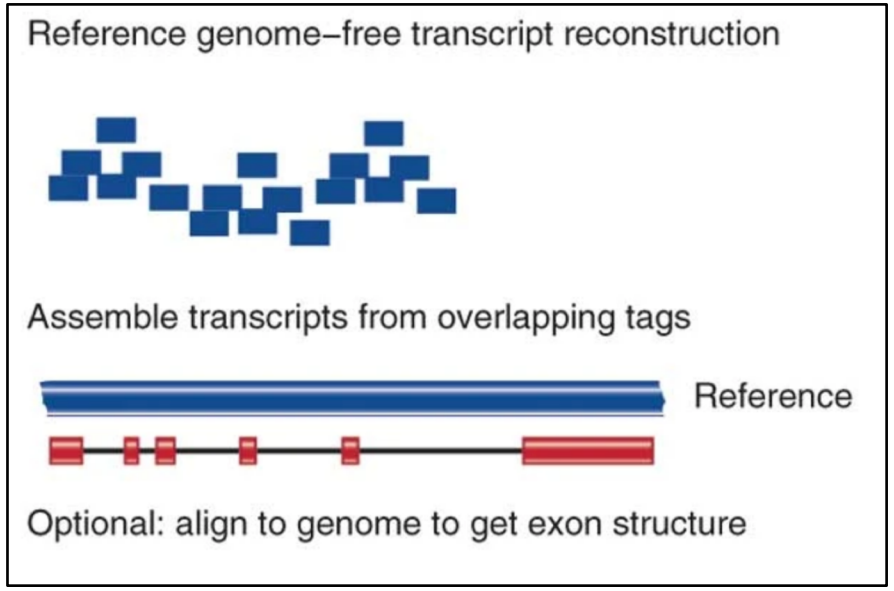



Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud, PLOS Computational Biology 2015

# Alignment strategies for RNA-seq data

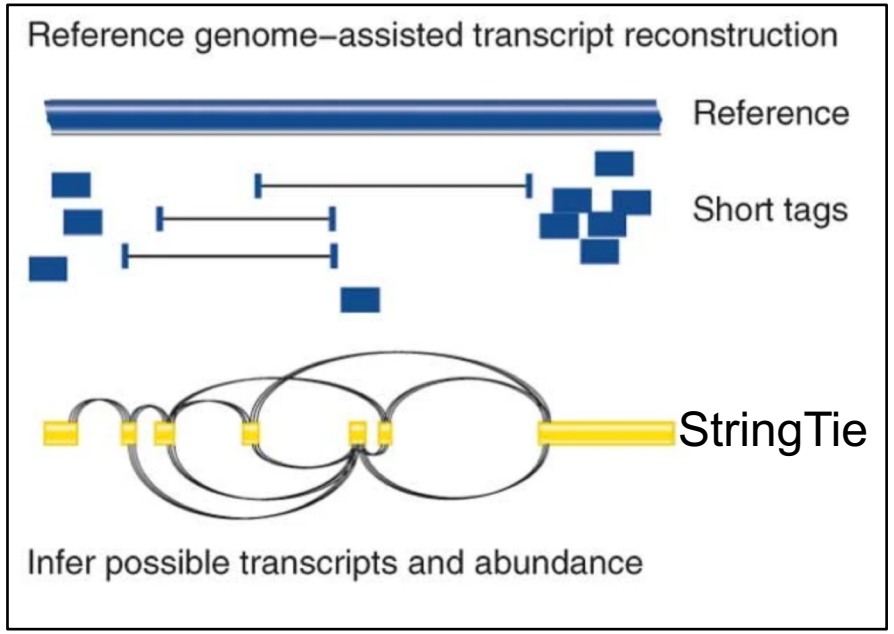
- Fusions
- Unannotated transcripts
- New splice variants

## De novo assembly

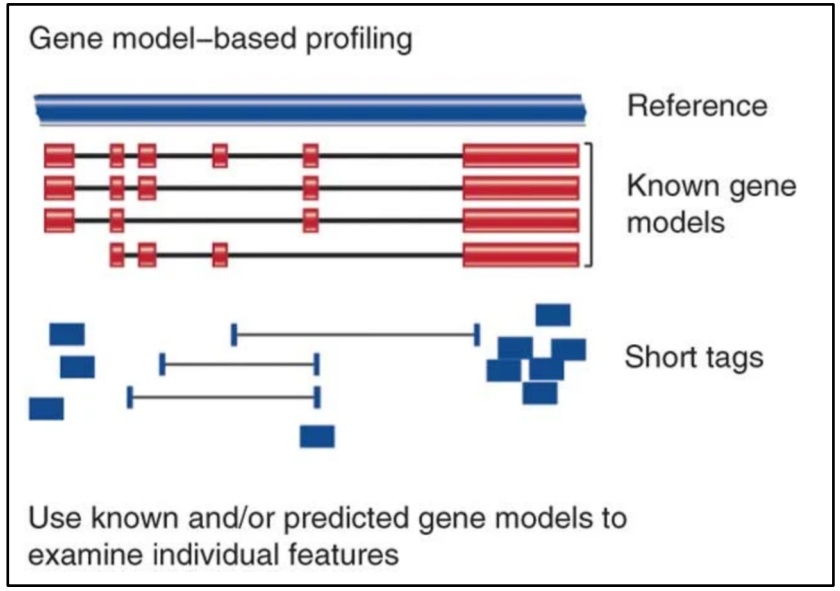



HISAT2

## Align to reference genome




## Align to transcriptome

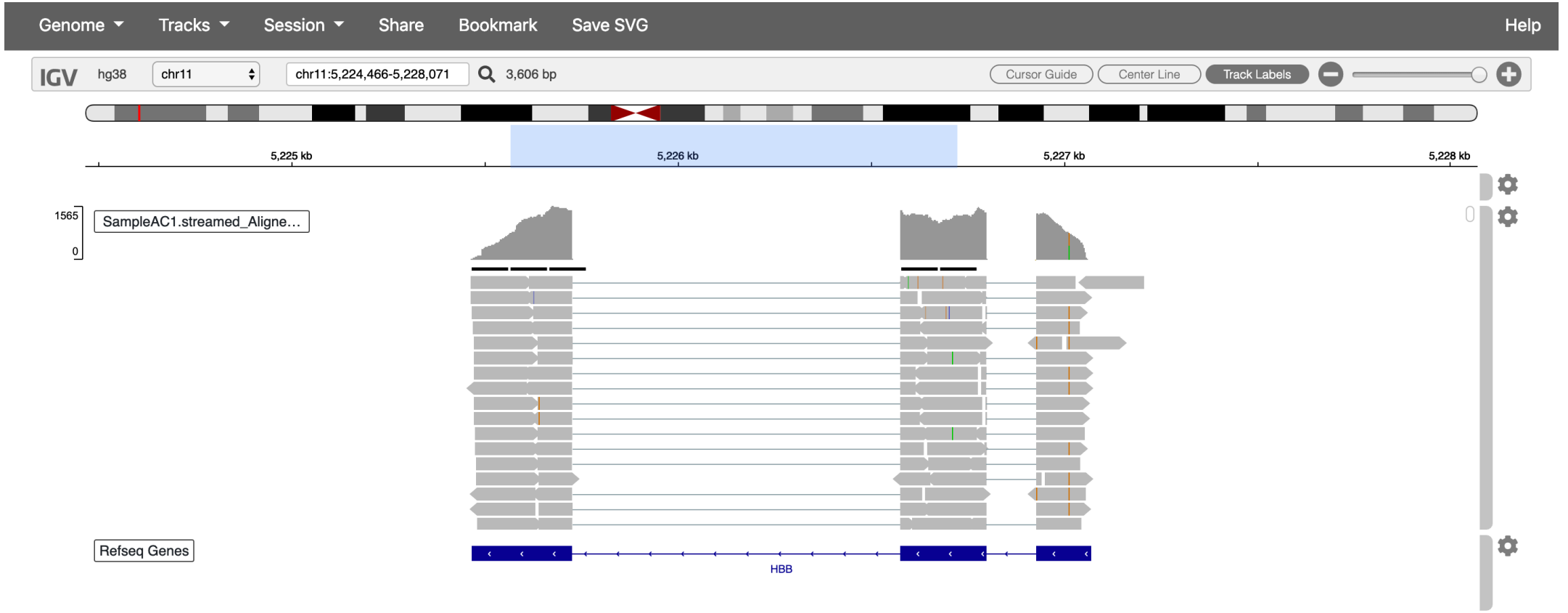



Cloonan and Grimmond, Nature Methods 2010

Kallisto

Is the transcriptome annotation complete?

# Alignment to reference genome



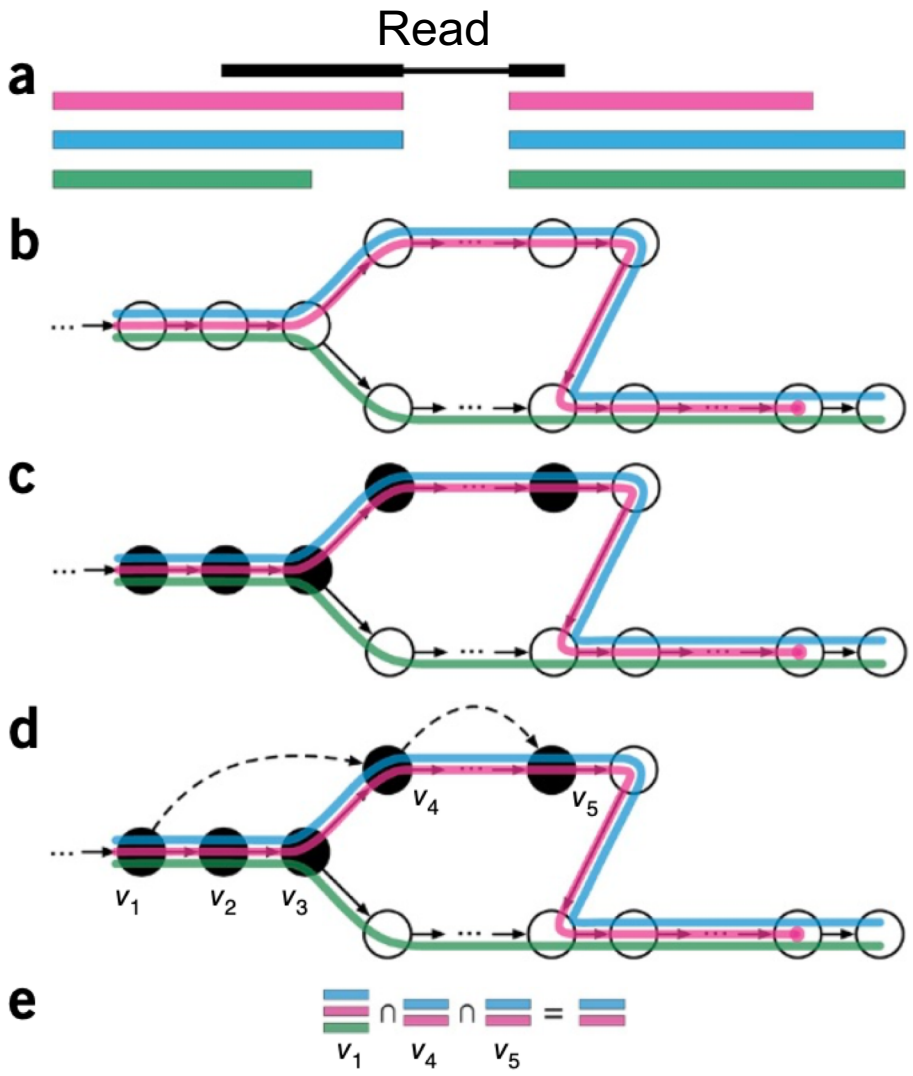

igv.org

UC San Diego BROAD INSTITUTE

Source: UC Davis bioinformatics

## 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 target sequences that the read is compatible with

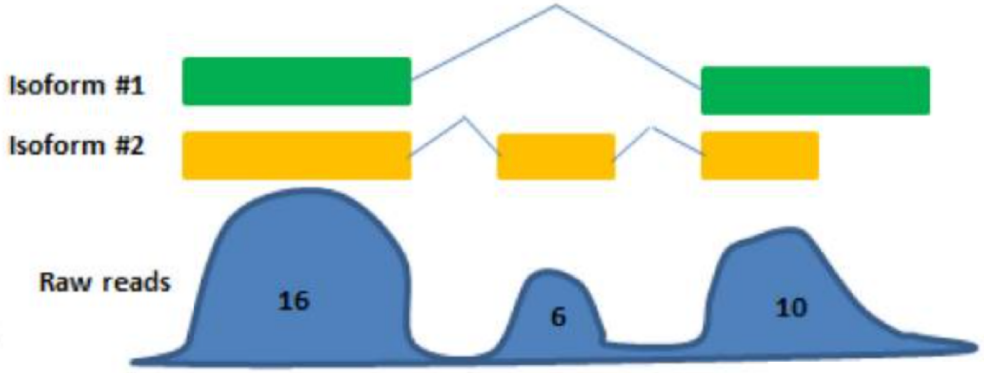



Overlapping transcripts from reference transcriptome

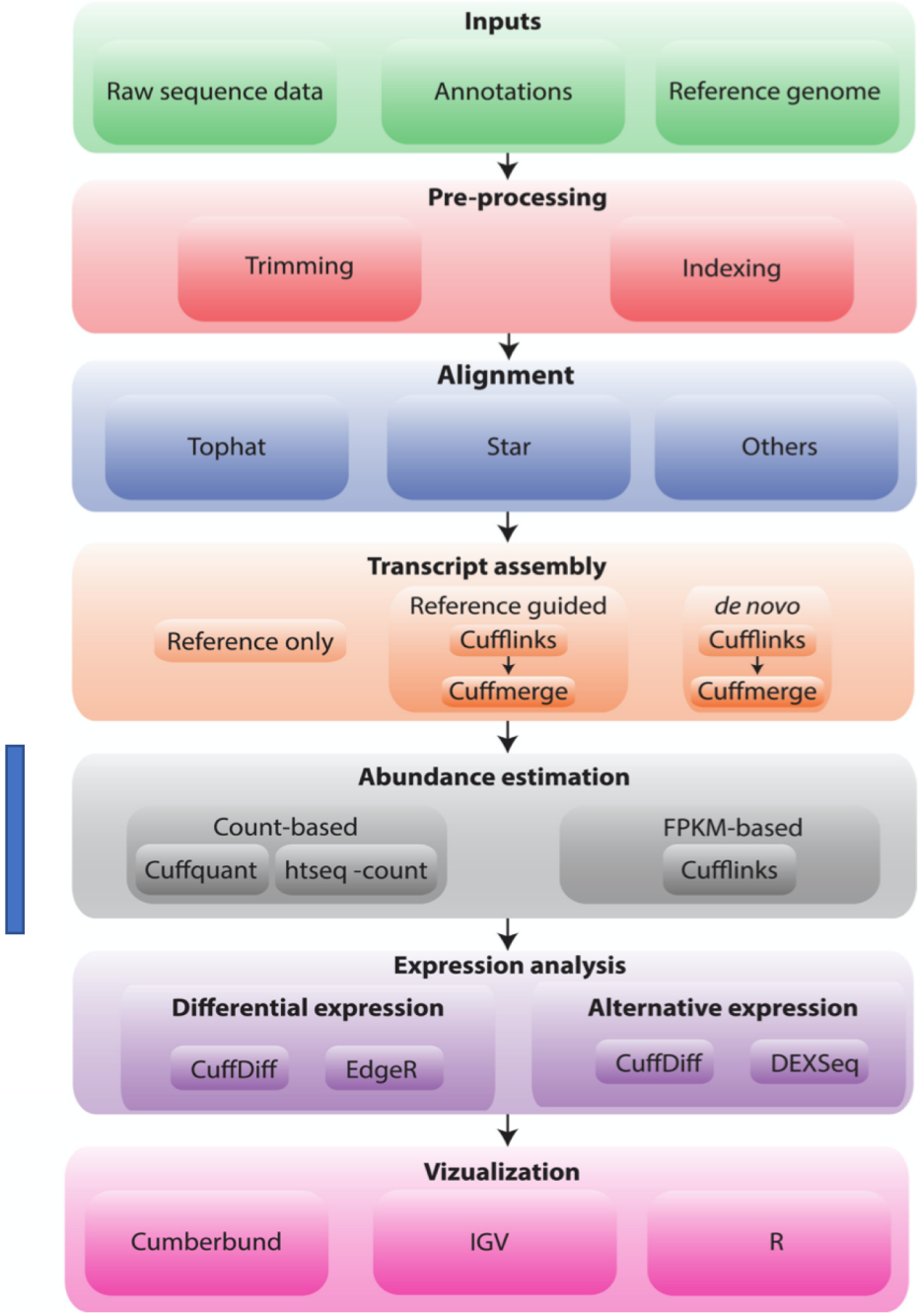
Bray et al., Nature Biotechnology 2016

# 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.




Zhao et al., PLOS One, 2015




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

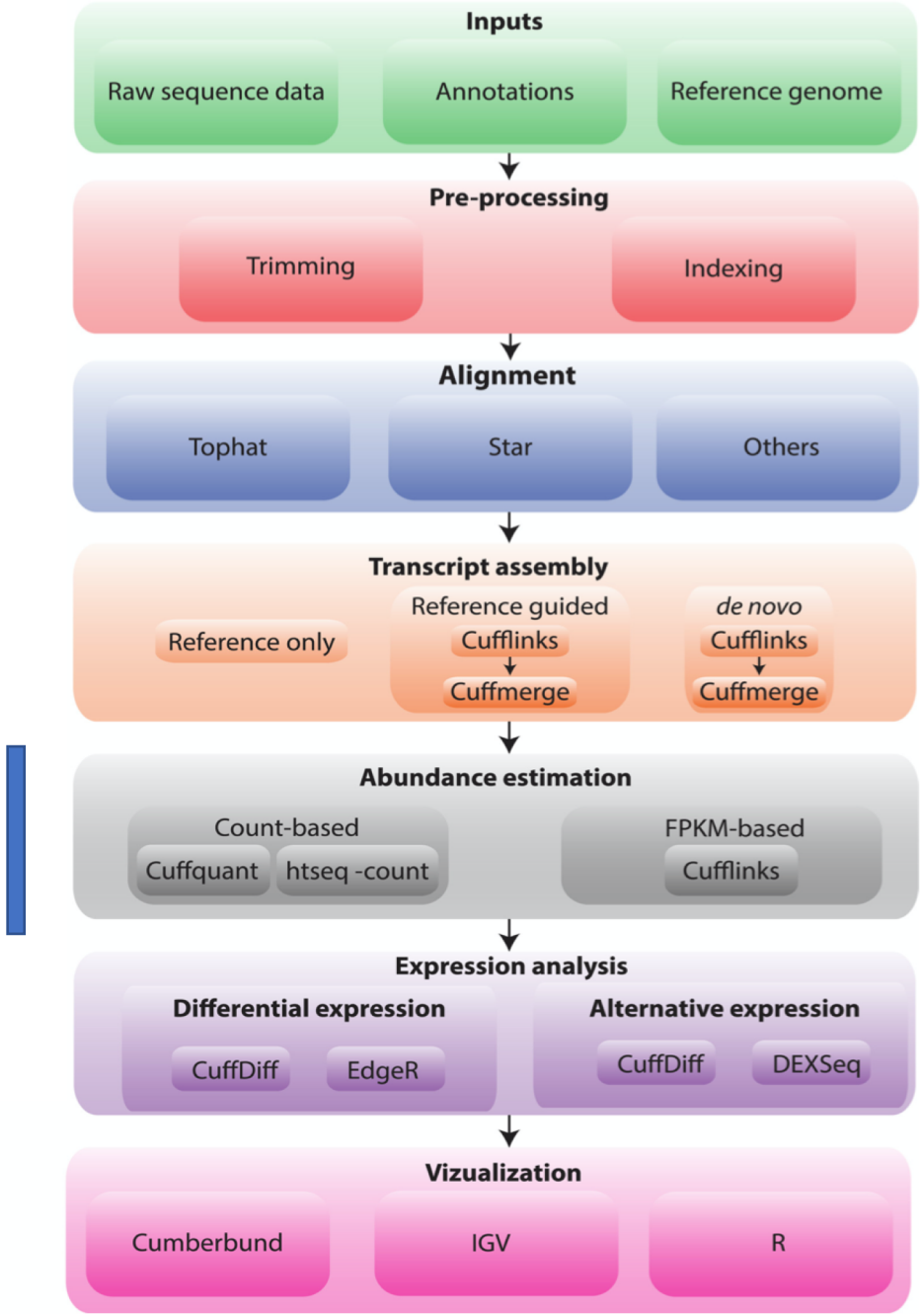
# RNA-seq analysis workflow

- RNA-seq data has several biases, e.g.
  - Length bias
  - GC content bias (sequence-specific bias)
  - Sequencing depth
- The number of reads from RNA-seq experiments therefore needs to be normalized to be comparable.

We will use **TPM** (Transcript Per Million) to assess expression:

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

The sum of all TPMs in each sample is the same.



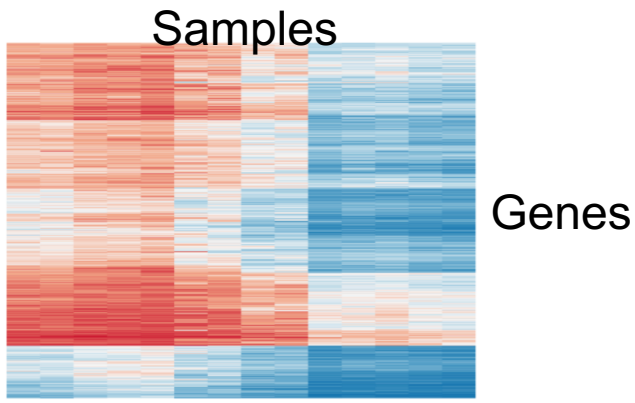

Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud, PLOS Computational Biology 2015

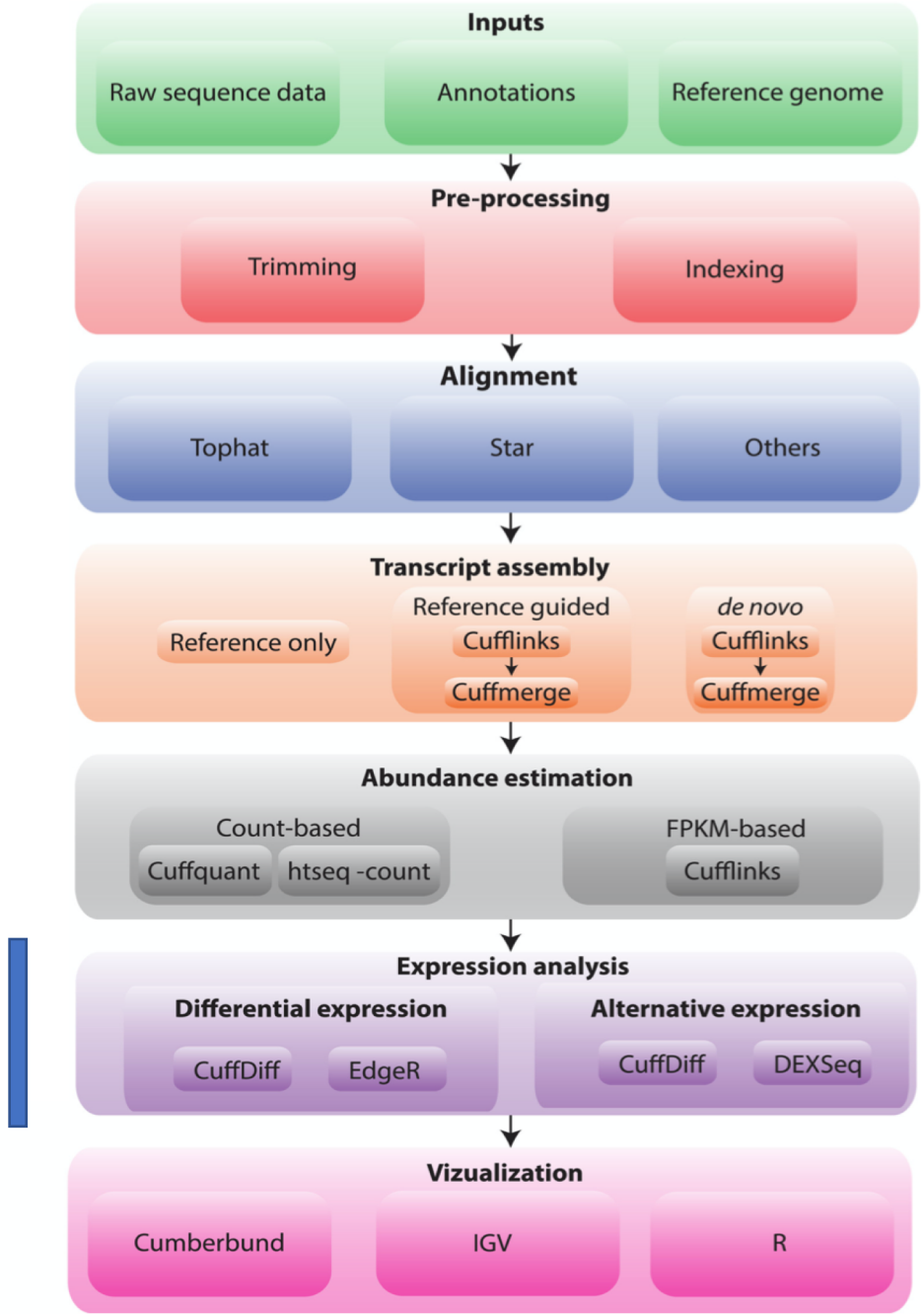
# RNA-seq analysis workflow

## 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
…

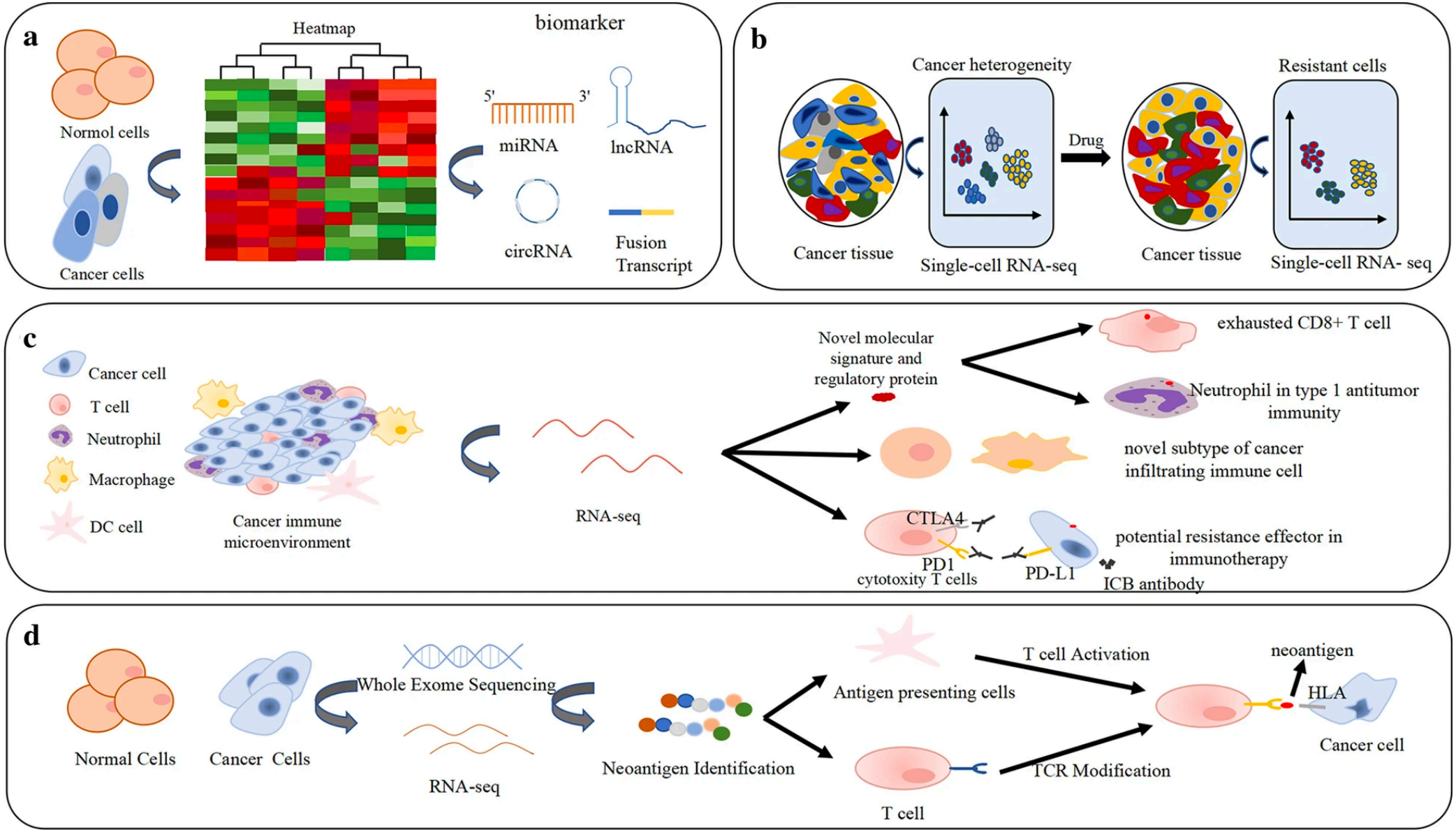



Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud, PLOS Computational Biology 2015

Differential gene expression, biomarkers, fusions

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

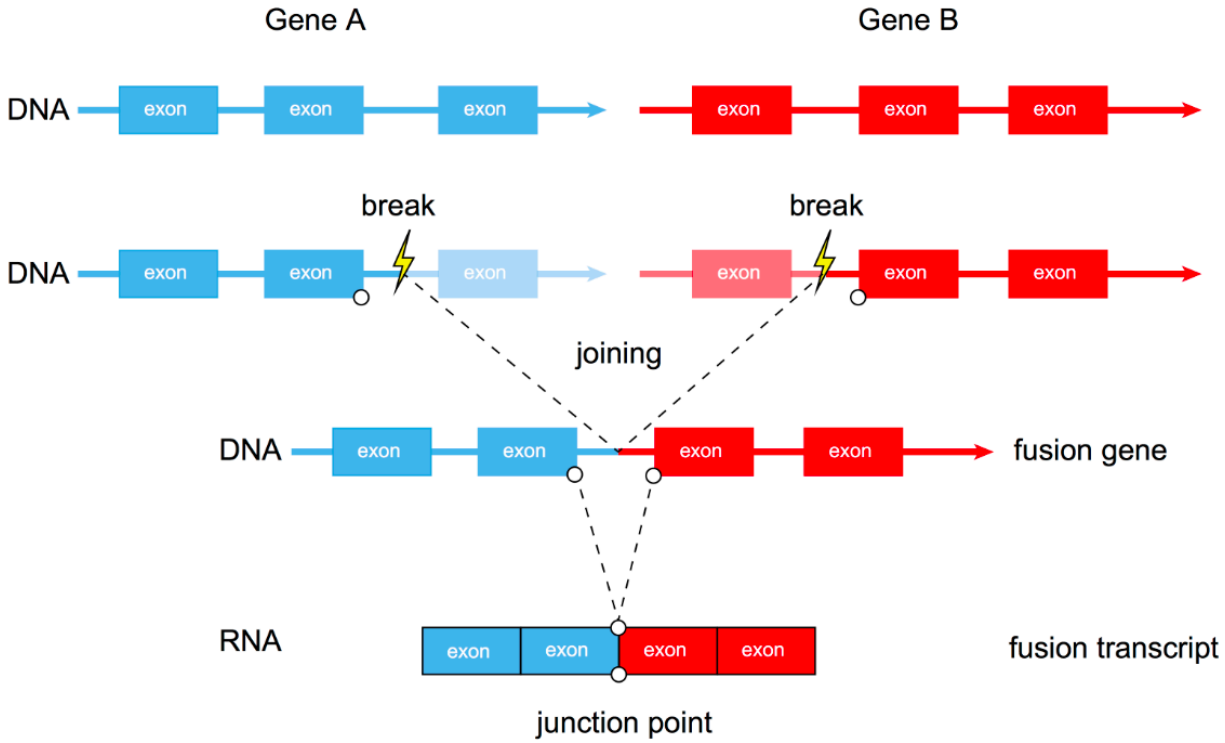



Molecular signatures, subtyping, estimation of cell type composition

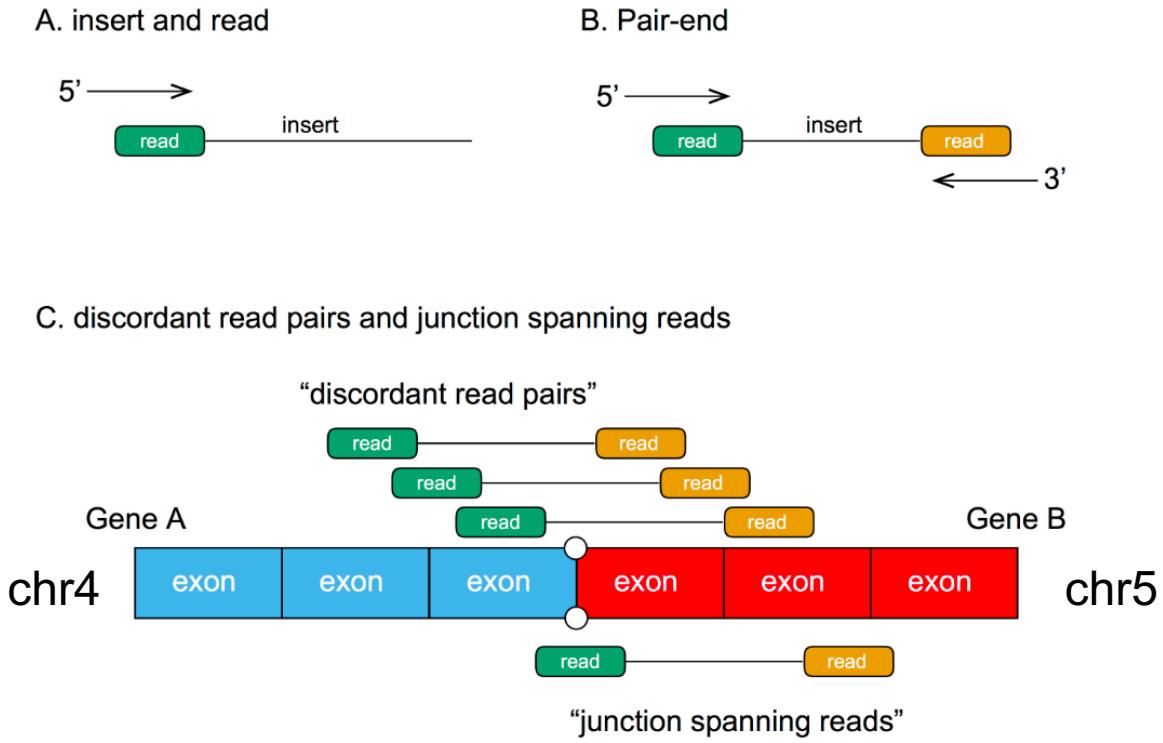
Neoantigen profiling and prediction

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

# Application 1: fusion detection




Fusion transcripts can arise due to genomic rearrangements.

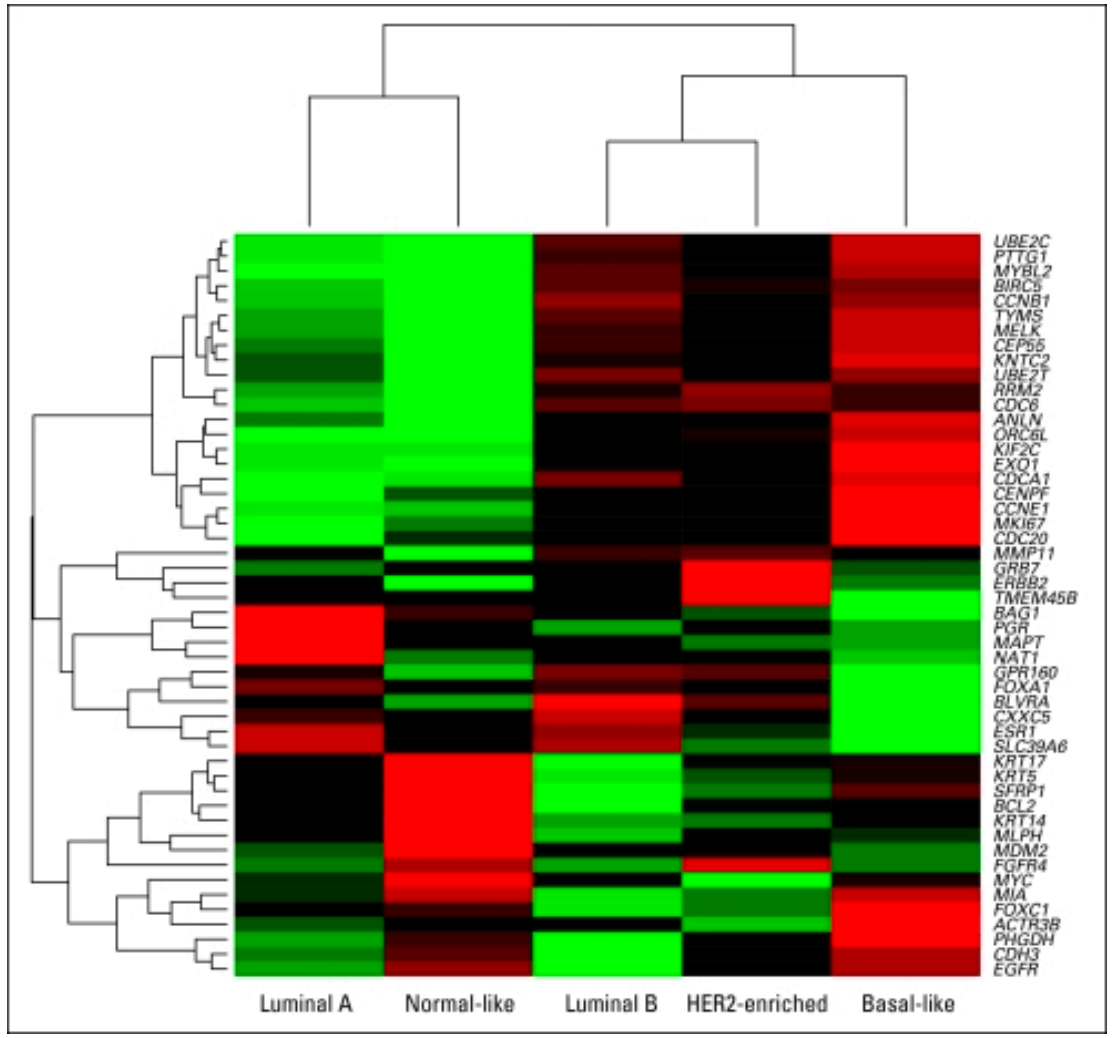



Paired-end RNA-seq can be used to detect fusion transcripts.

Tools: Fusion-Inspector, fusion-report, Pizzly, Squid, Star-Fusion…

Source: The Jackson Laboratory

# Application 2: classification + subtyping




Parker et al., Journal of Clinical Oncology, 2009

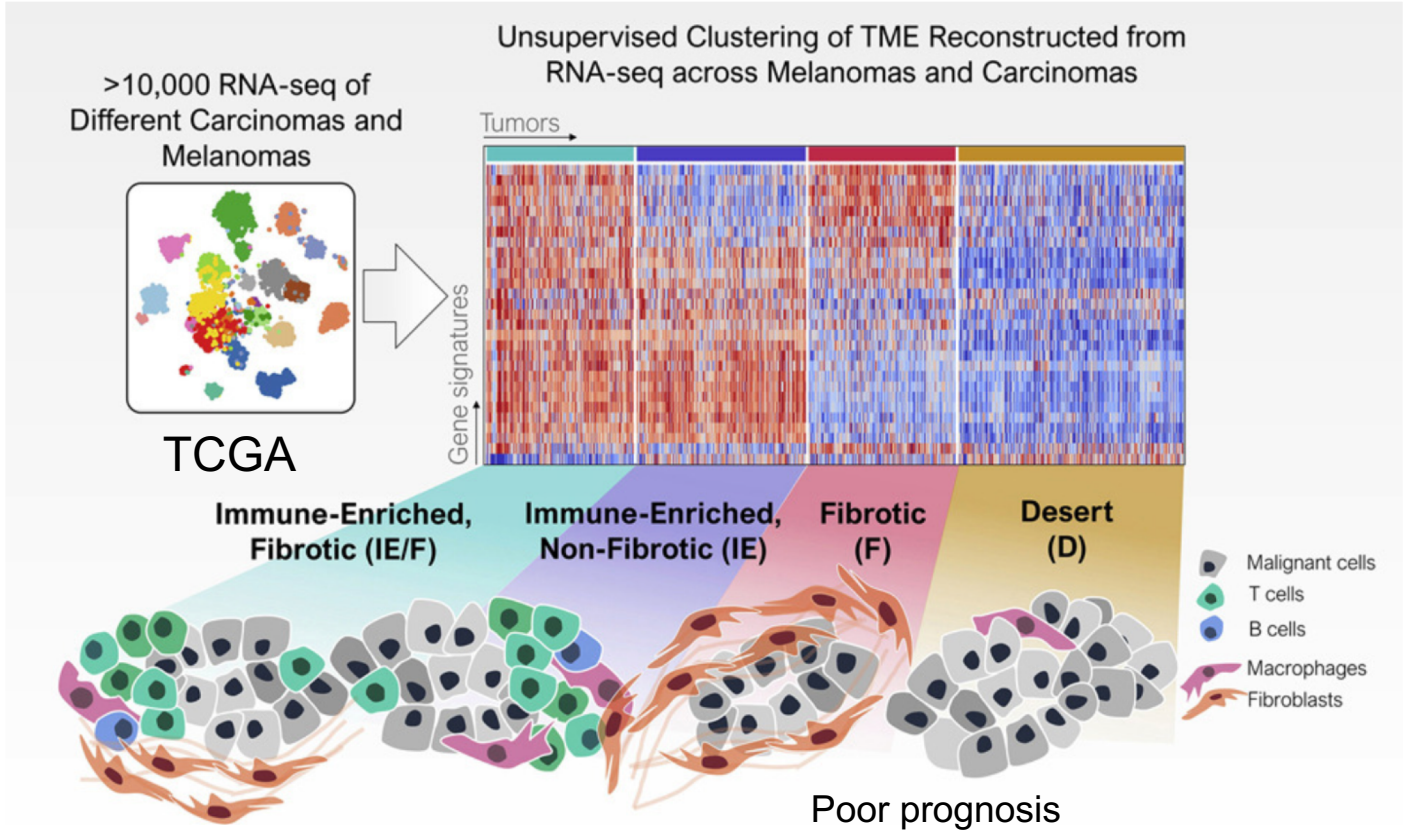
**PAM50 in breast cancer**

- 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



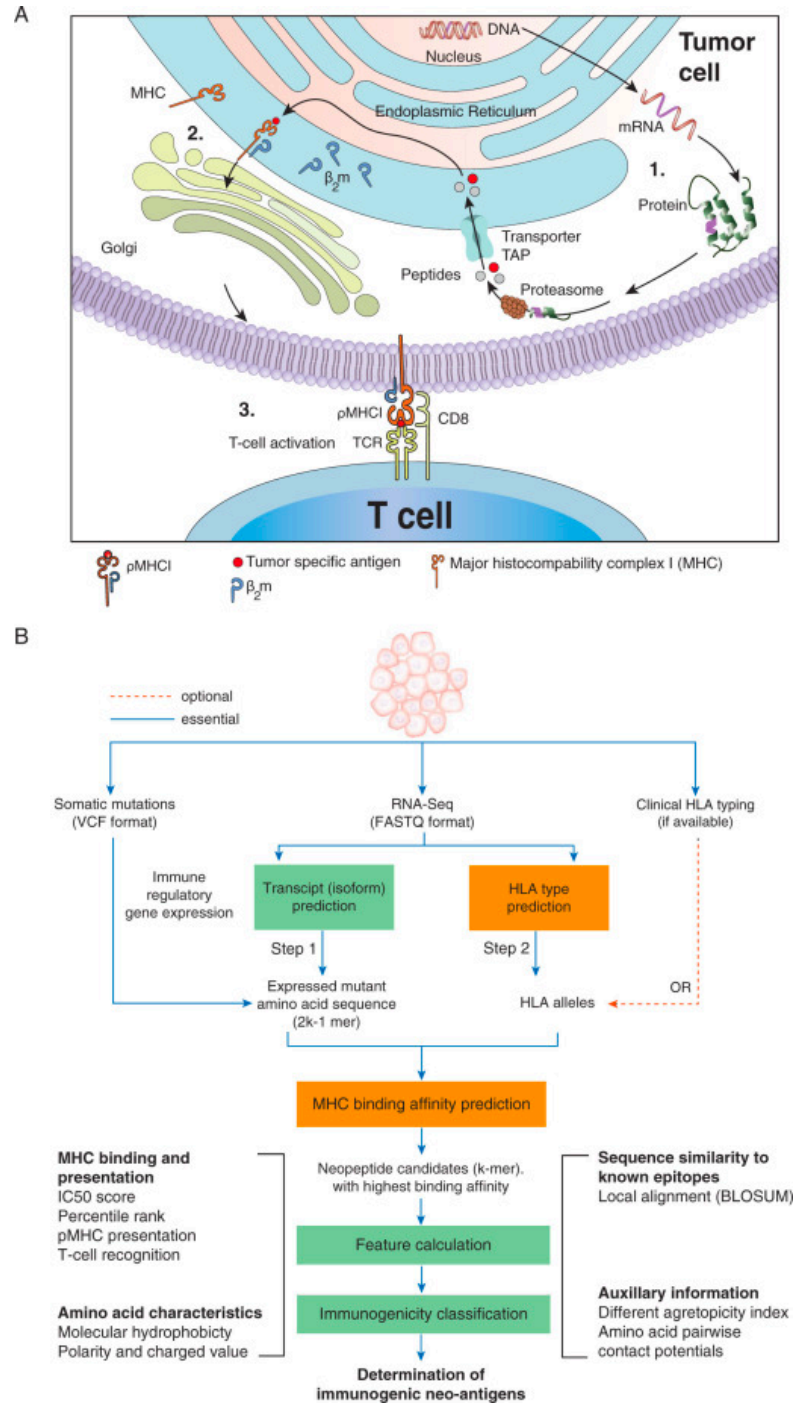

Better prognosis
Better immunotherapy response
Lower tumor proliferation

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

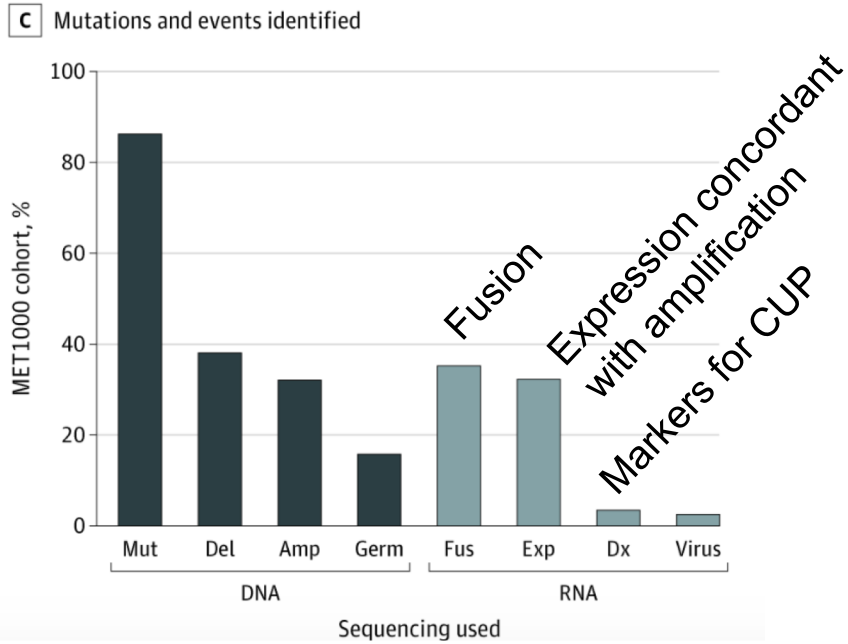
# Clinical impact of comprehensive DNA and RNA sequencing

Research

JAMA Oncology | Original Investigation

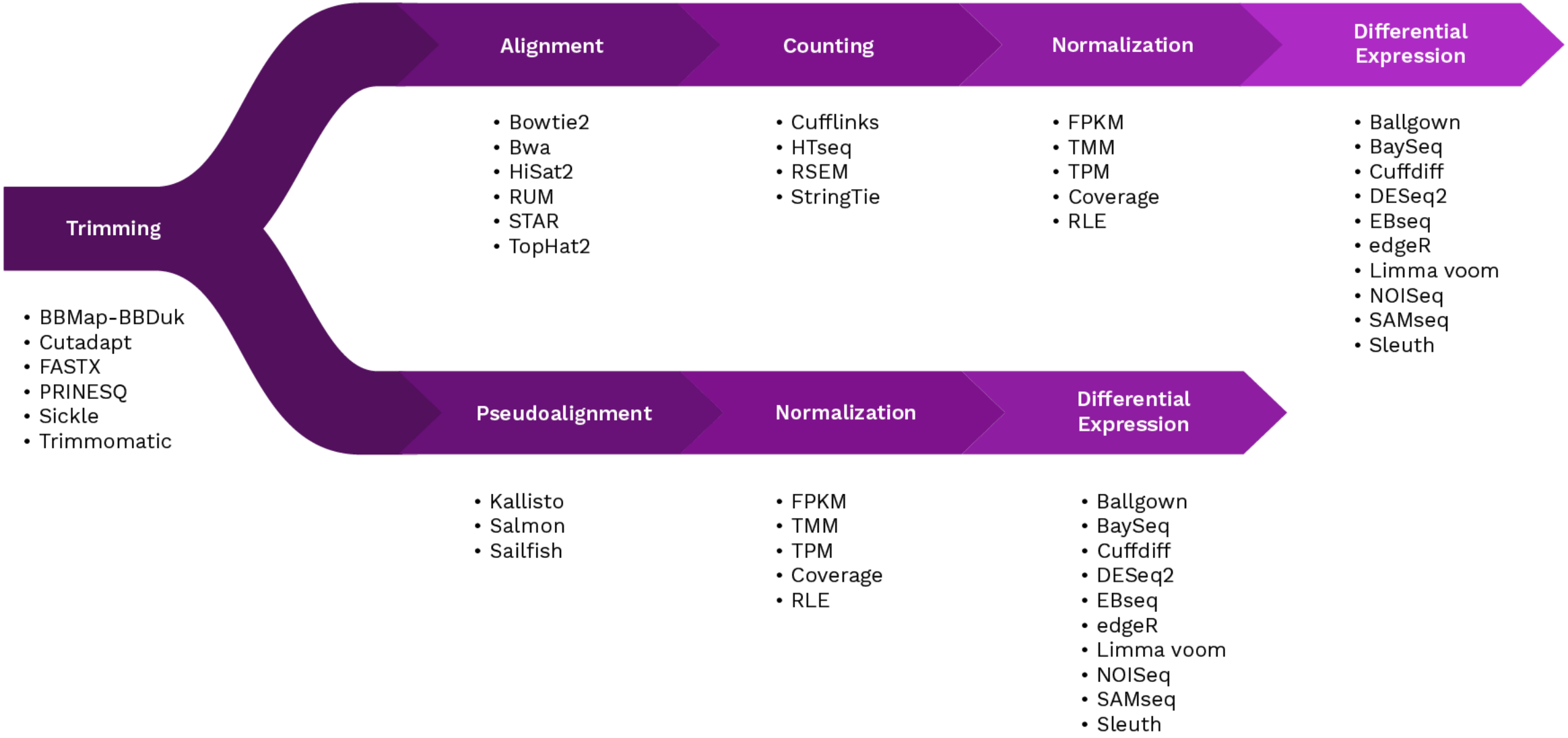
## Assessment of Clinical Benefit of Integrative Genomic Profiling in Advanced Solid Tumors

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 gDNA
  - Whole-exome or targeted capture
  - RNA-sequencing
    - Fusion detection
    - Classification of Cancer Of Unknown Primary (CUP)
  - Clinical benefit rate from NGS-directed therapy

# Popular tools and approaches…




Source: www.elucidata.io


Karolinska
Institutet