

RNA-sequencing

Clinical Cancer Genomics
26 April 2023

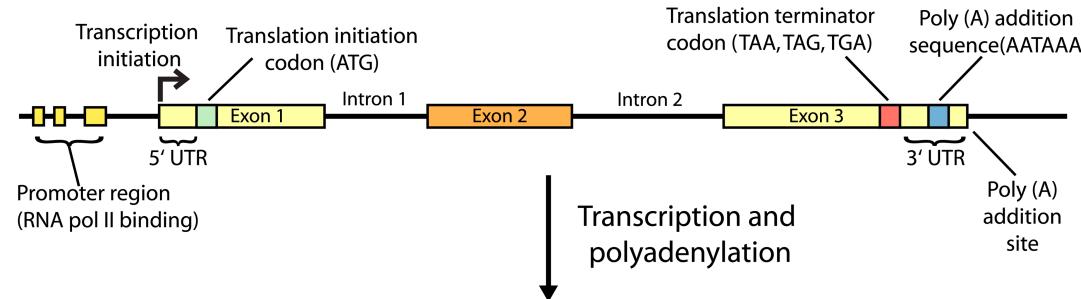


**Karolinska
Institutet**

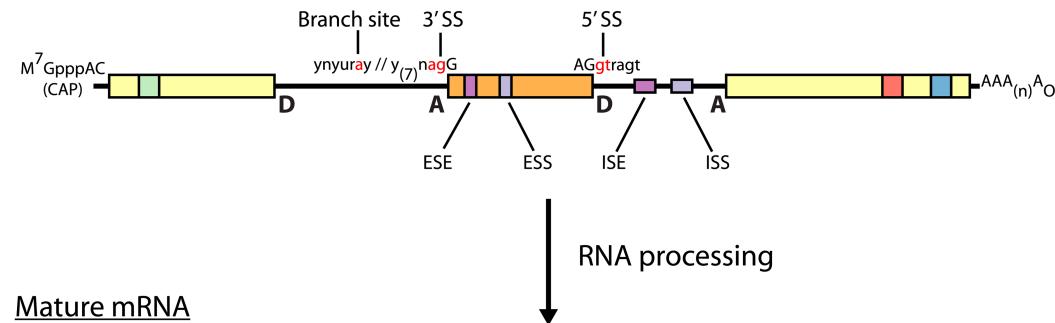
Learning objectives and lecture agenda

- 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.

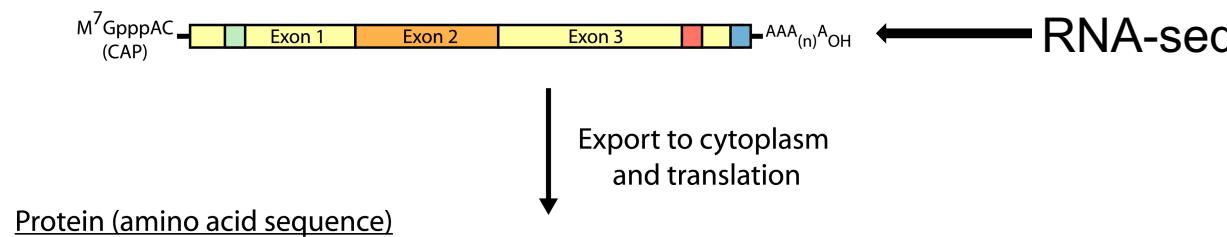
Double-stranded genomic DNA template



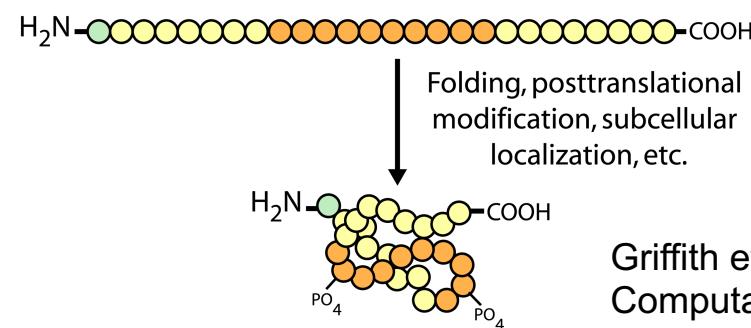
Single-stranded pre-mRNA (nuclear RNA)



Mature mRNA



Protein (amino acid sequence)



Griffith et al., PLOS
Computational Biology 2015

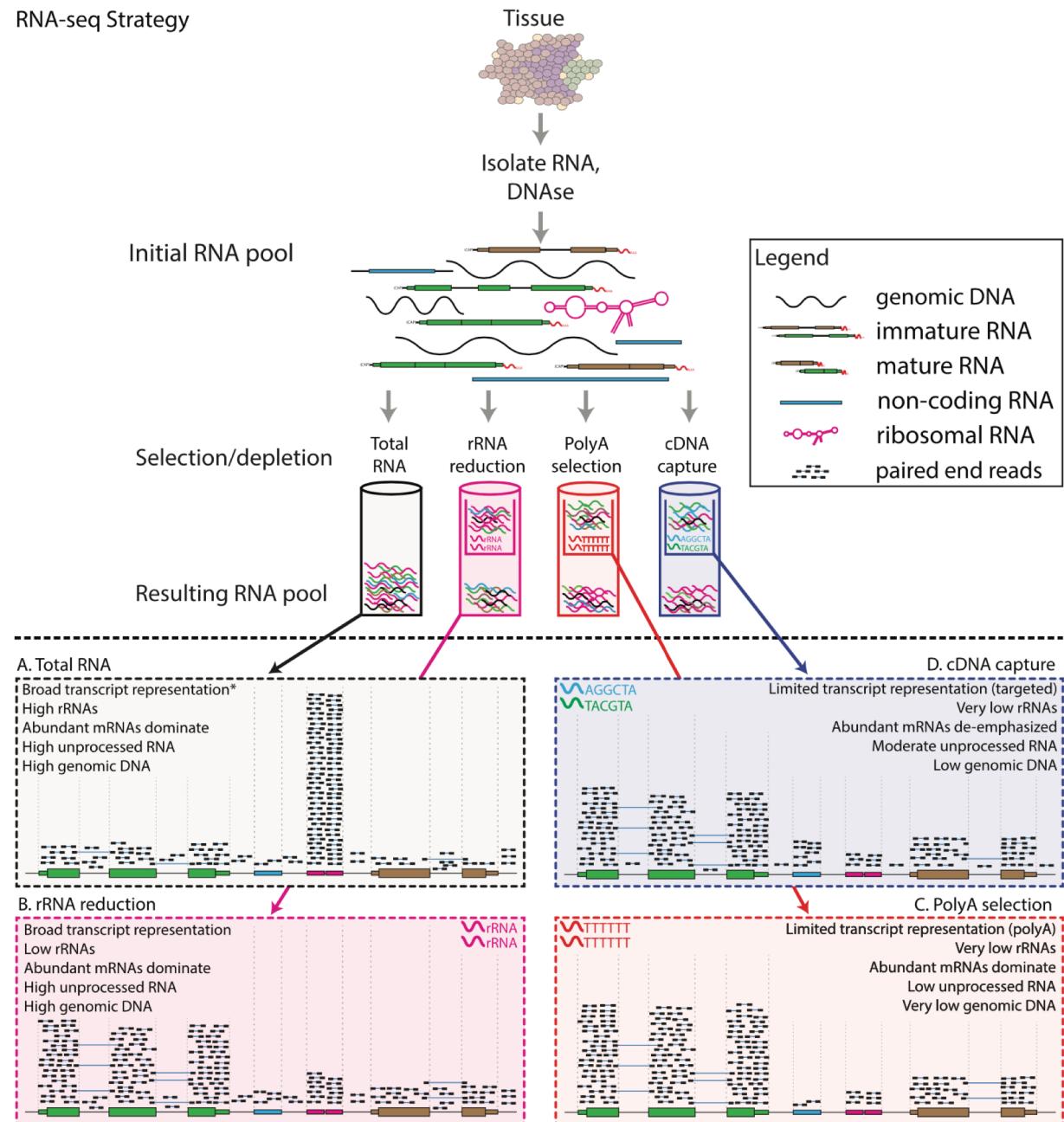
The central dogma of molecular biology

Phenotype = expression of genetic information modified by environmental factors

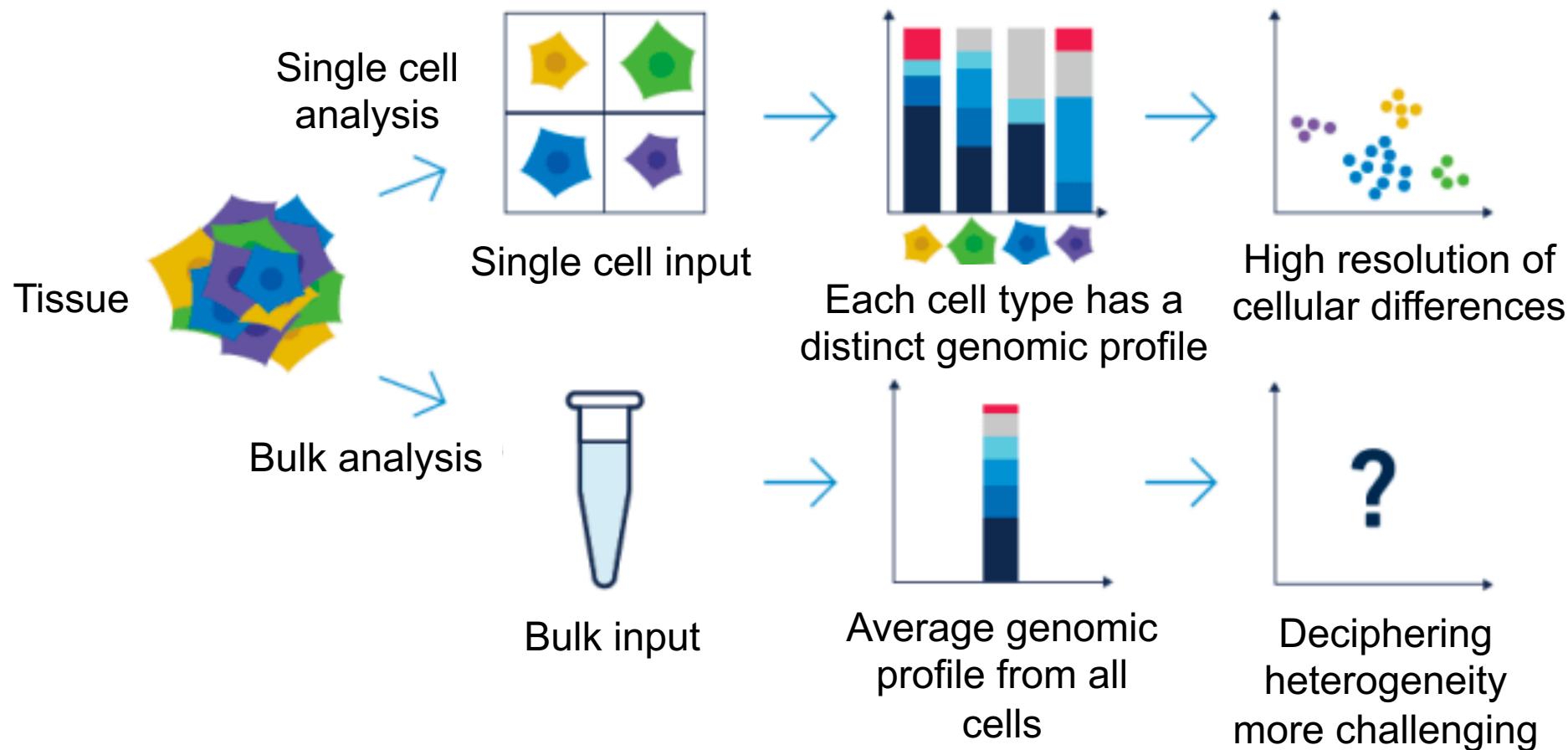
Cell identity and its activities are specified via the genes it transcribes

RNA-seq strategies

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



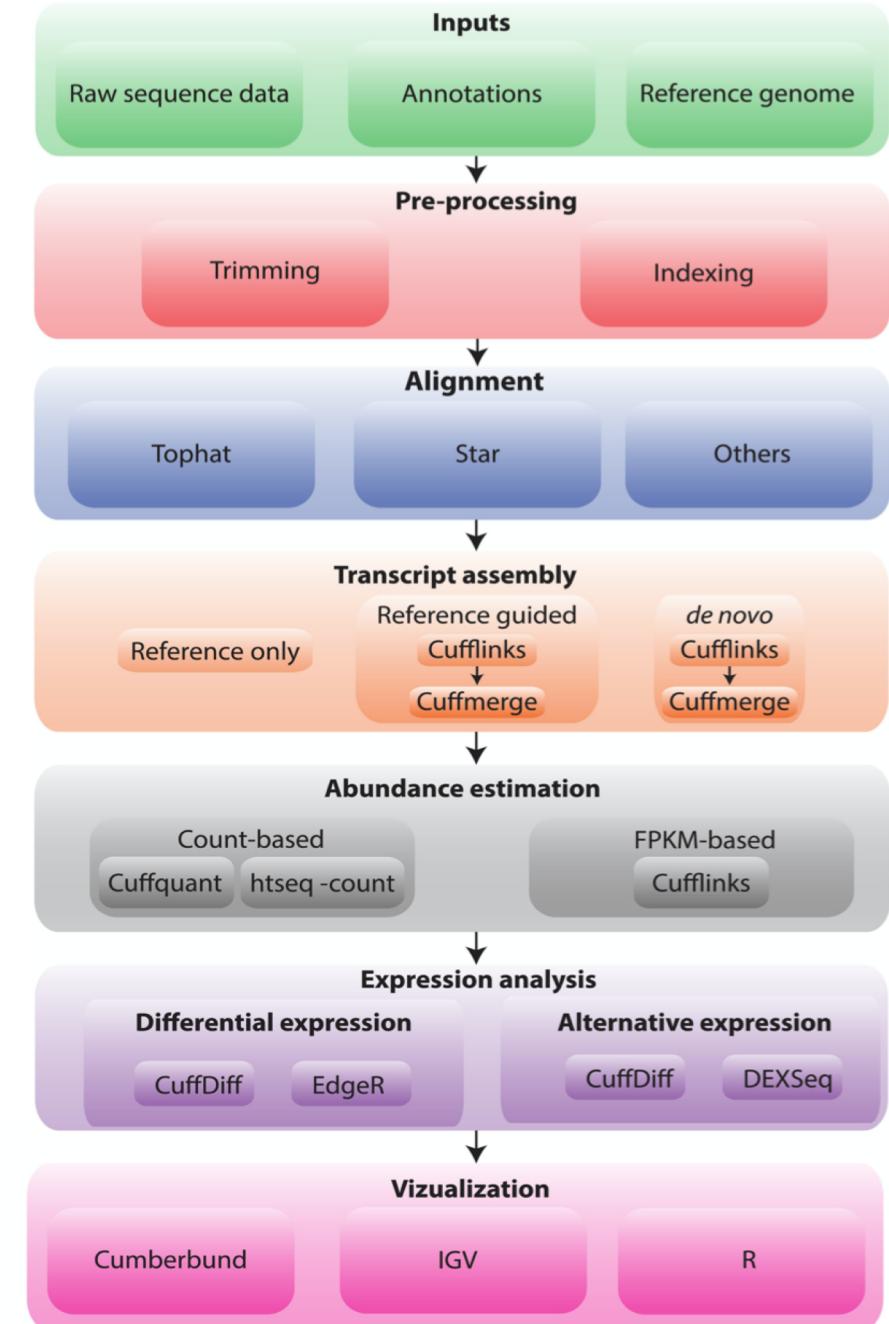
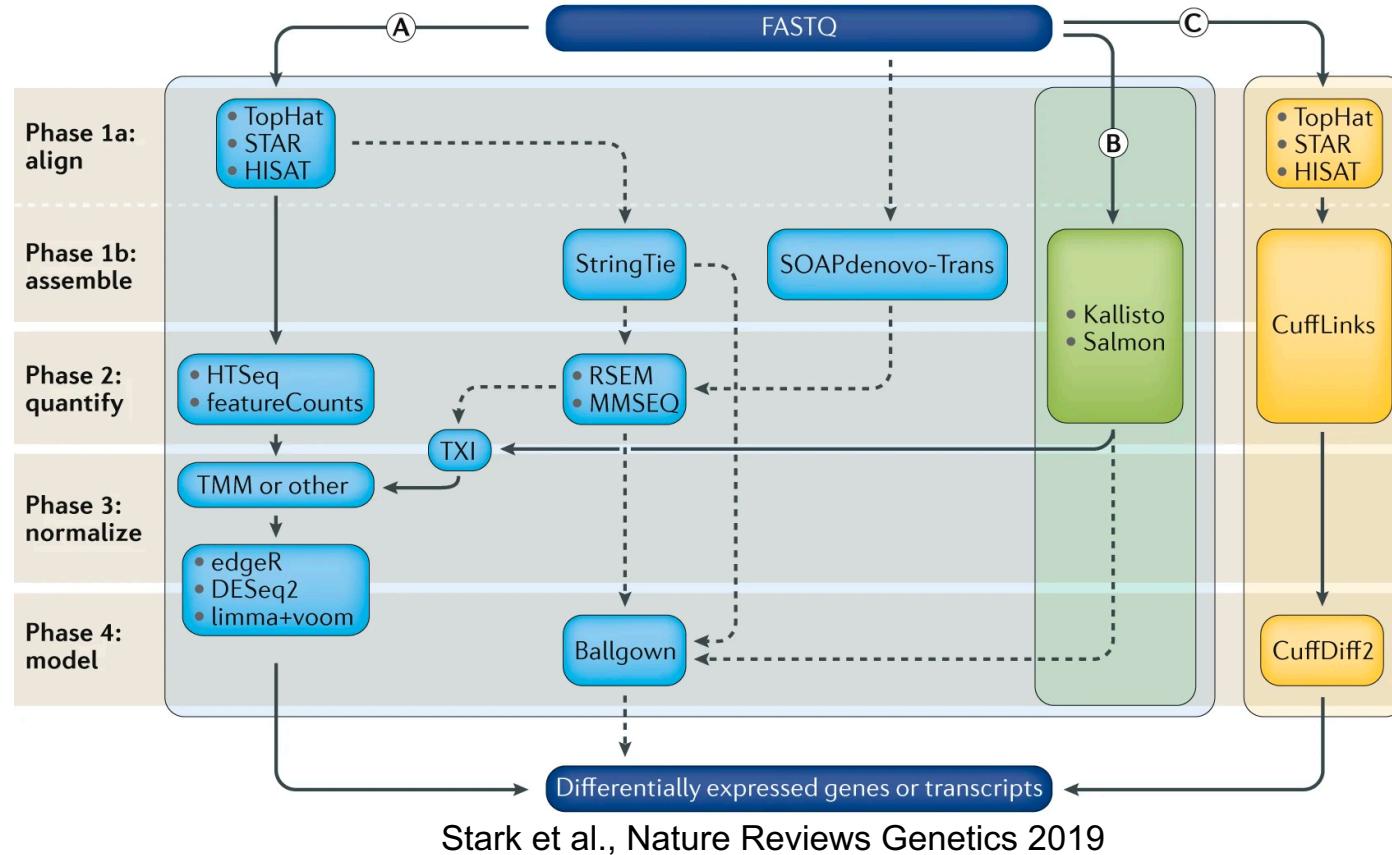
Bulk vs single-cell RNA-seq



CHALLENGES:

- Cost
- Complexity of library prep
- Data sparsity
- Analysis tools

RNA-seq pipelines for gene expression and differential expression analysis



Quality control (QC) of RNA-seq data

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

MultiQC

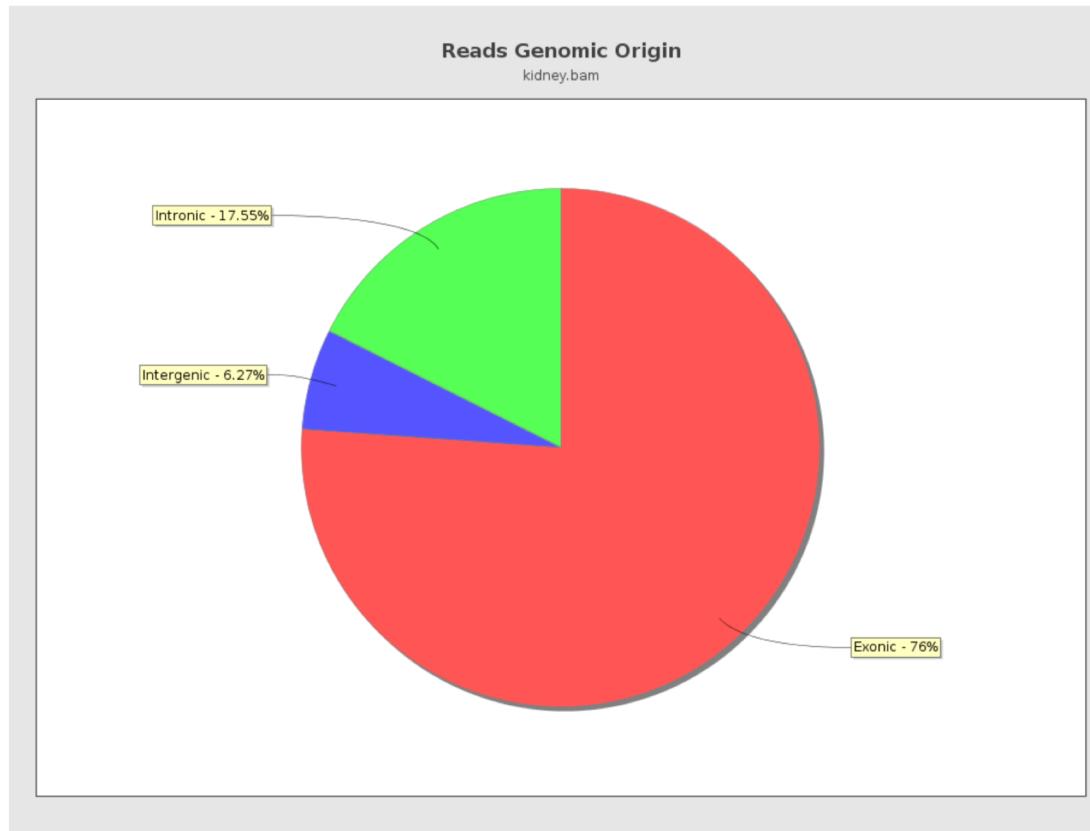
A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2021-12-10, 14:10 based on data in: /home/ubuntu/workspace/inputs/data/fastq

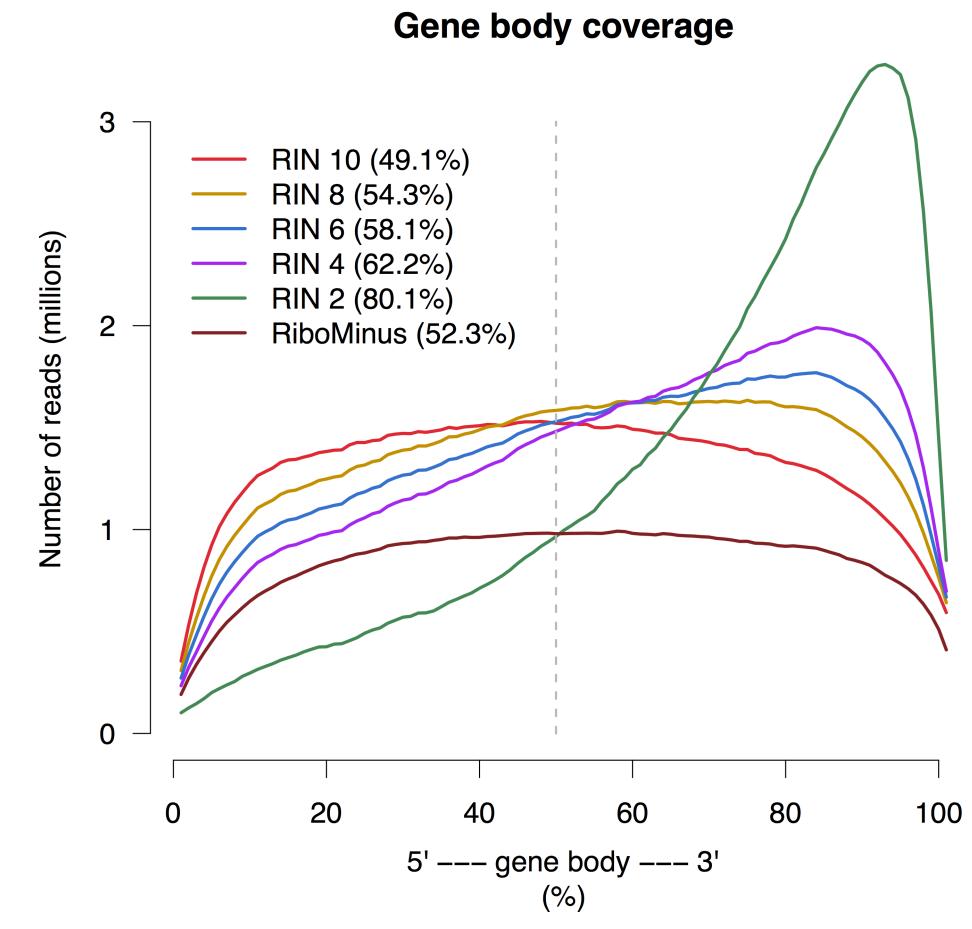


MultiQC can generate a joint summary report for all samples from the outputs of 128 different bioinformatics tools.

Quality control (QC) of RNA-seq data

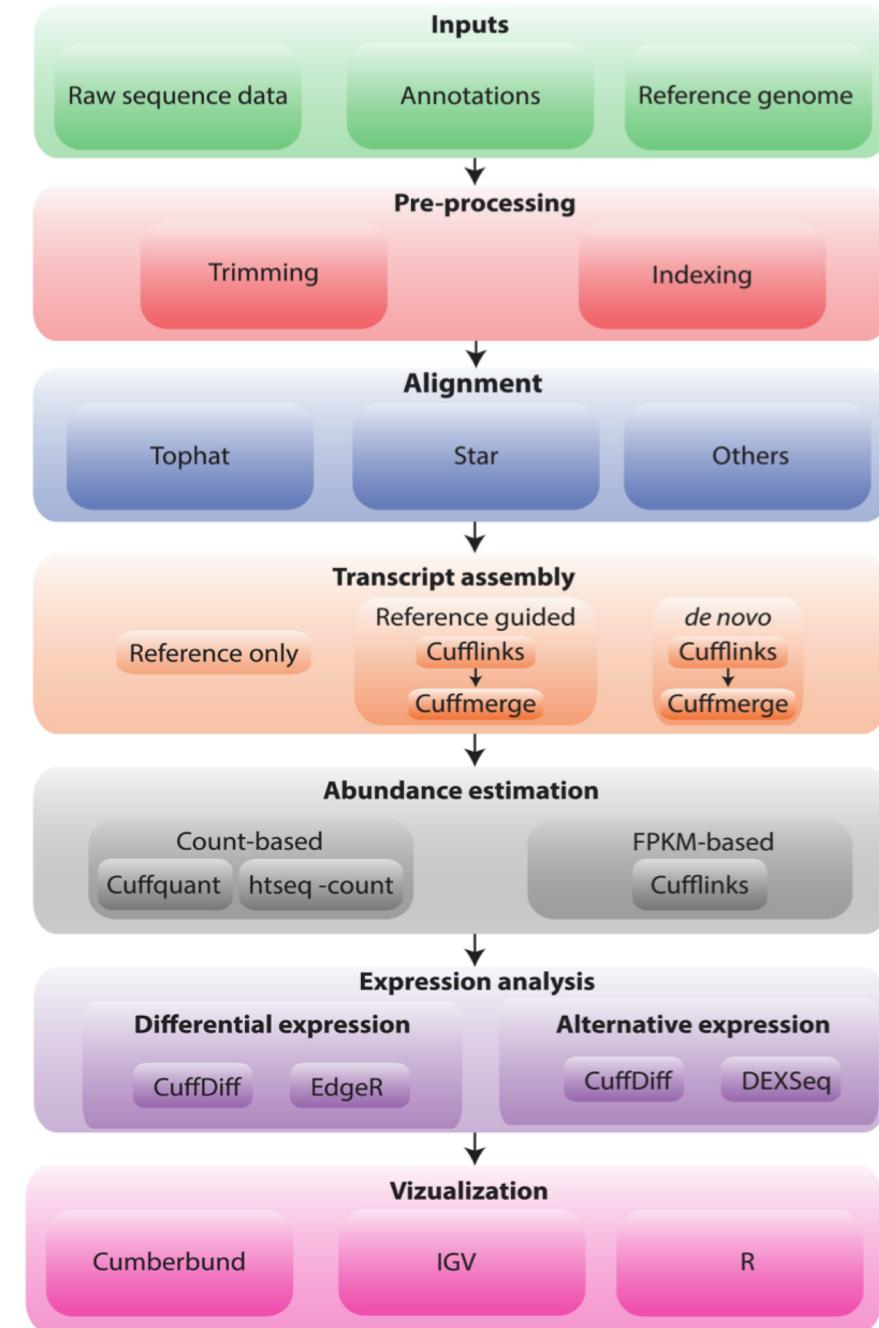


From Qualimap, expect > 60% exonic reads



RIN = RNA integrity number

RNA-seq analysis pipeline



De novo assembly

Reference genome-free transcript reconstruction



Assemble transcripts from overlapping tags



Optional: align to genome to get exon structure

Align to reference genome

Reference genome-assisted transcript reconstruction



Reference

Short tags

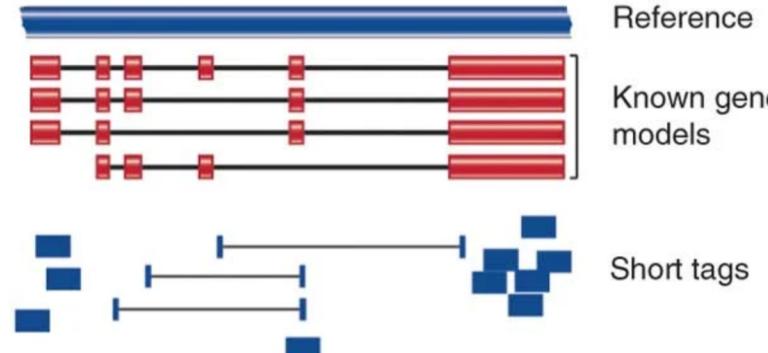


Infer possible transcripts and abundance

HISAT

Align to transcriptome

Gene model-based profiling



Use known and/or predicted gene models to examine individual features

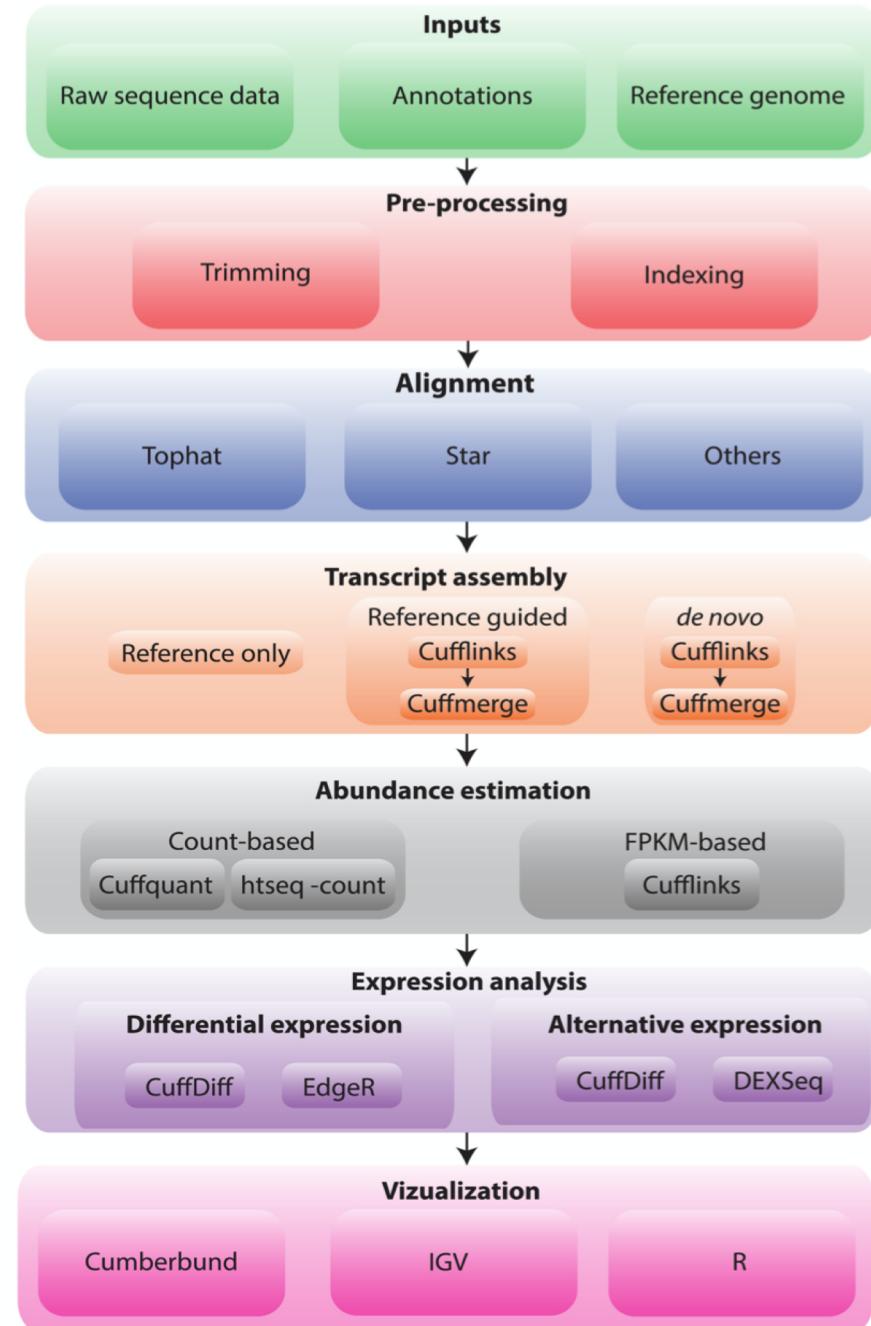
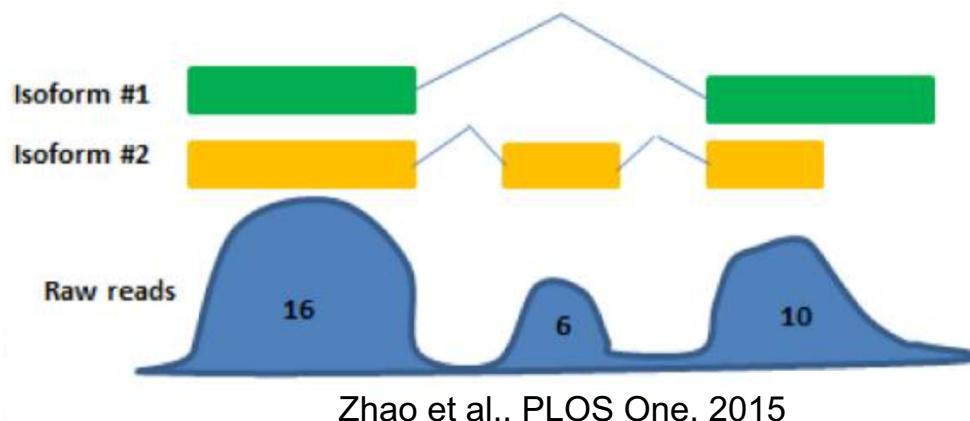
Kallisto

Cloonan and Grimmond, Nature Methods 2010

Alignment strategies
for RNA-seq data

RNA-seq analysis pipeline

- 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.



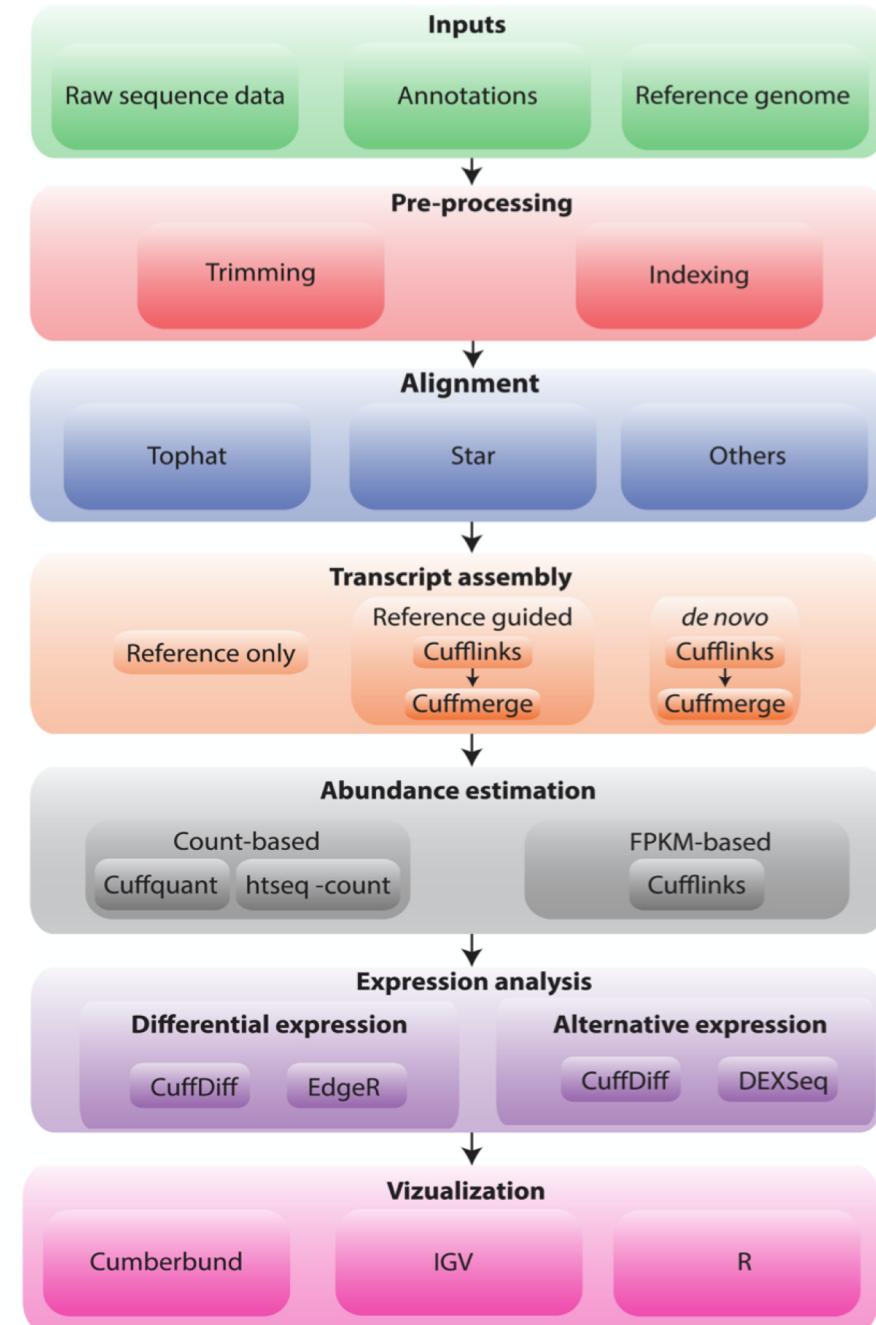
RNA-seq analysis pipeline

- RNA-seq data has several biases, e.g.
 - Length bias
 - GC content bias (sequence-specific bias)
 - PCR artefacts (duplicates)
- 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:

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

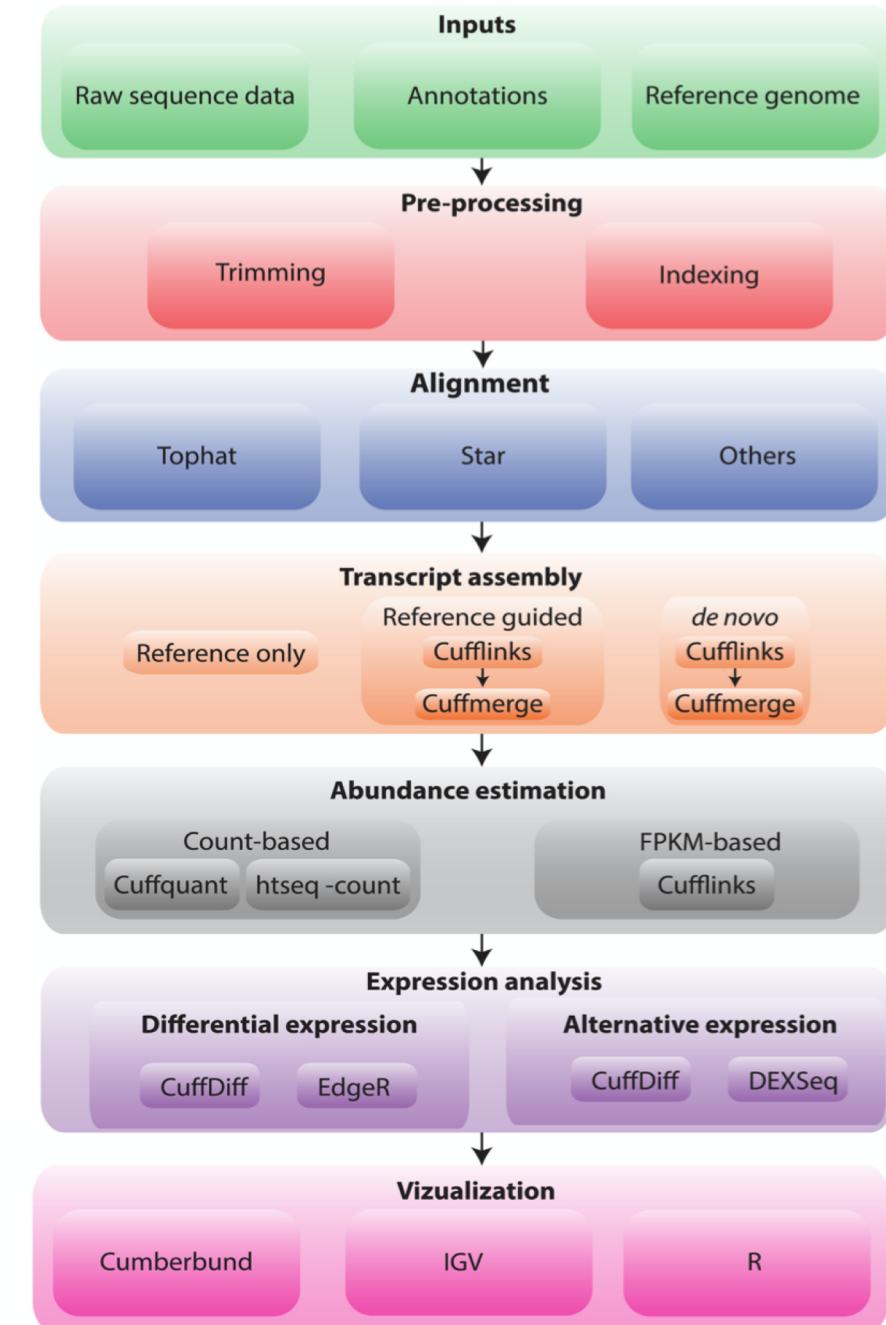
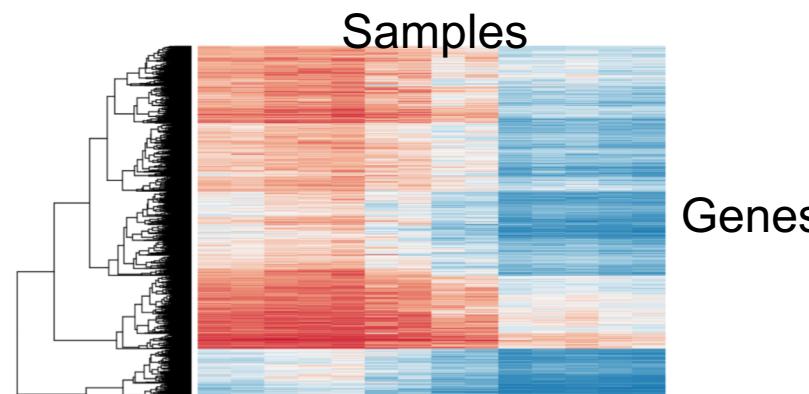
The sum of all TPMs in each sample is the same.



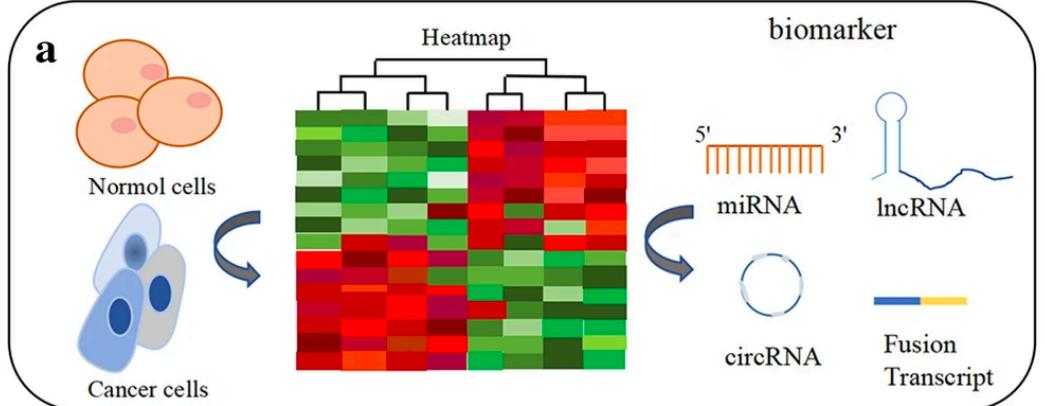
RNA-seq analysis pipeline

Differential expression

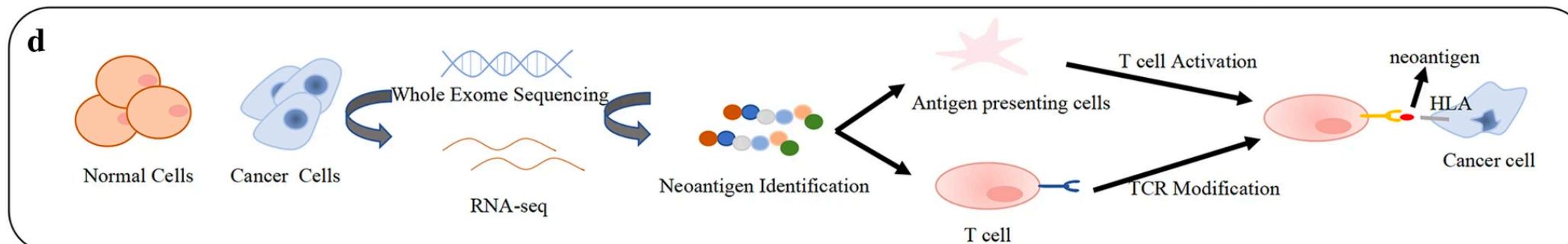
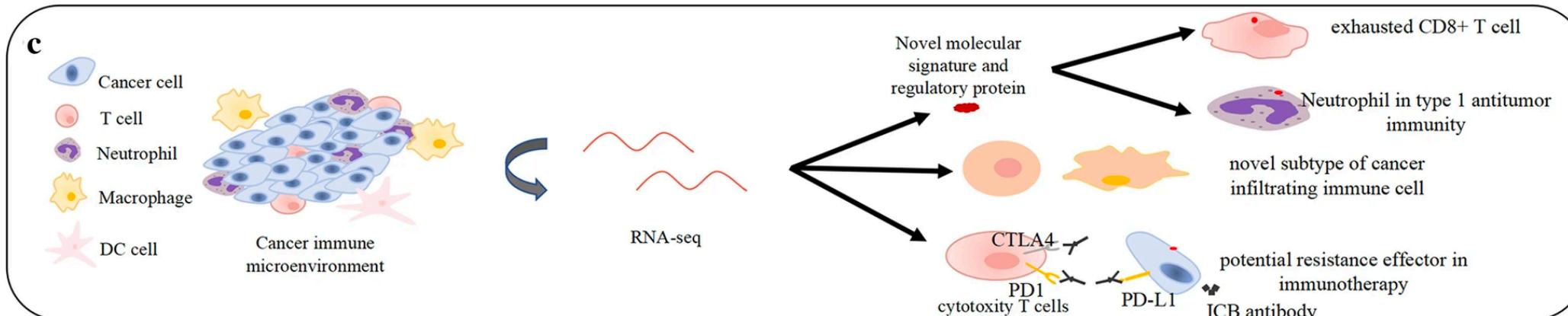
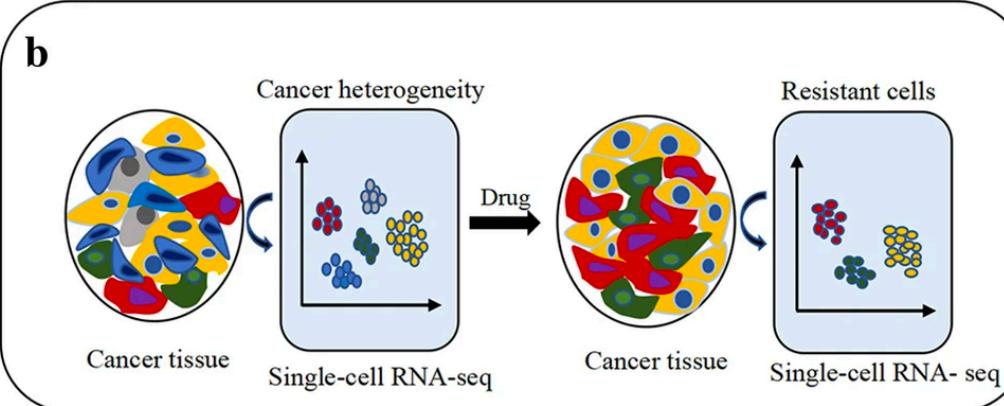
- Connecting gene expression back to genotype/phenotype
- What genes/transcripts are expressed at higher/lower levels in different sample groups?
 - Tumor vs normal samples
 - Are the differences statistically significant, accounting for variance and noise in the data?



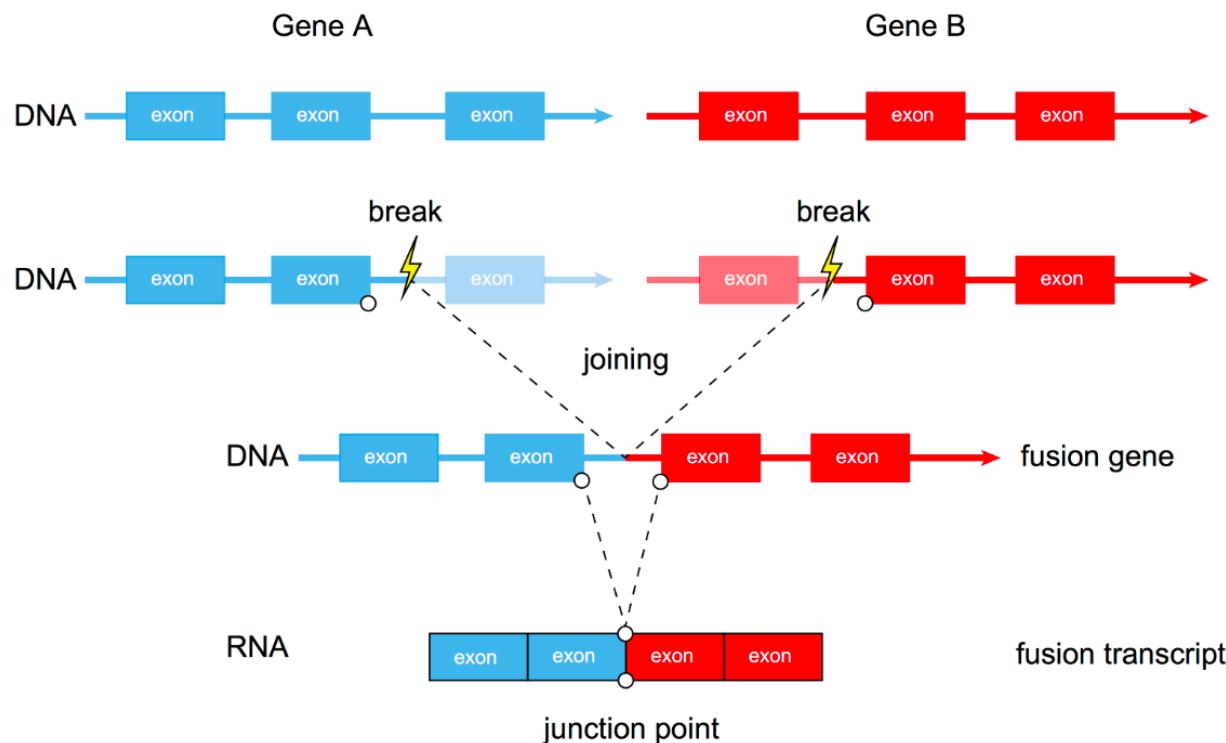
Differential gene expression, biomarkers, fusions



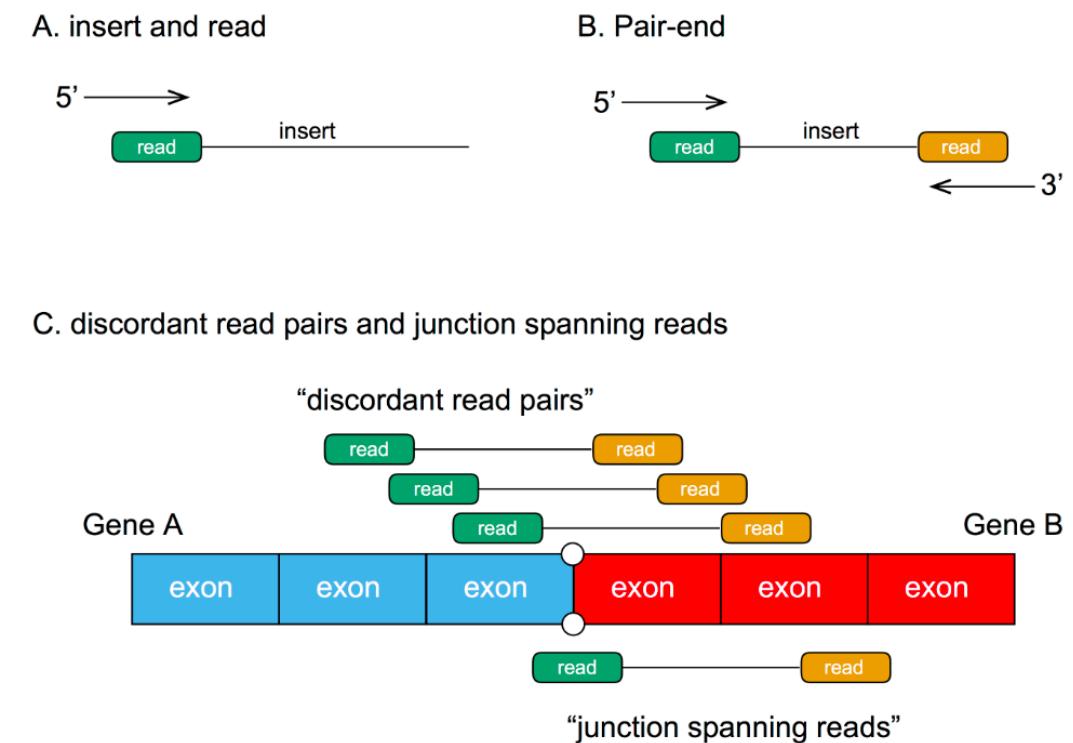
Diagnostic relevance- heterogeneity, evolution, and drug resistance in cancer



Application 1: fusion detection



Fusion transcripts can arise due to genomic rearrangements.

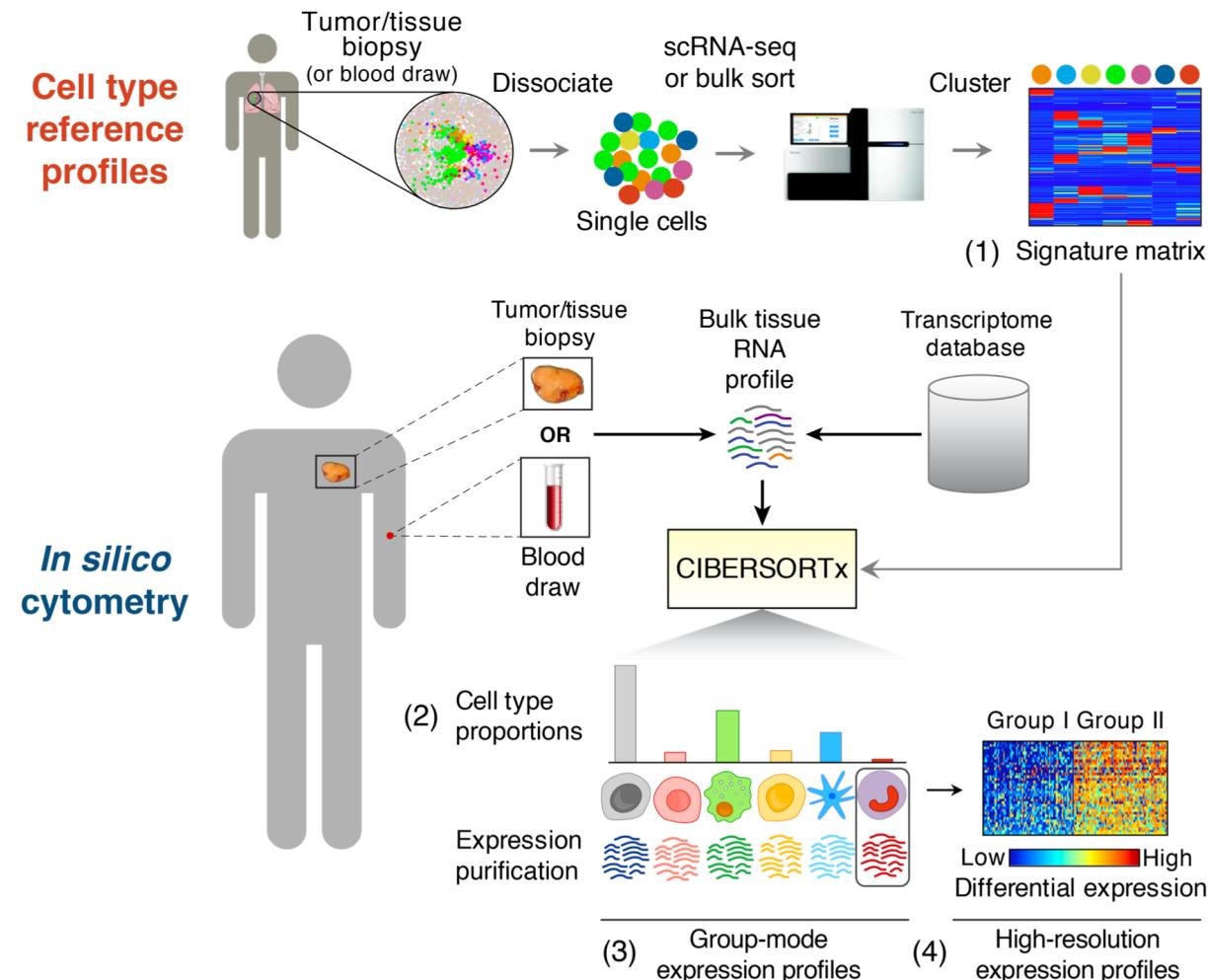


Paired-end RNA-seq can be used to detect fusion transcripts.

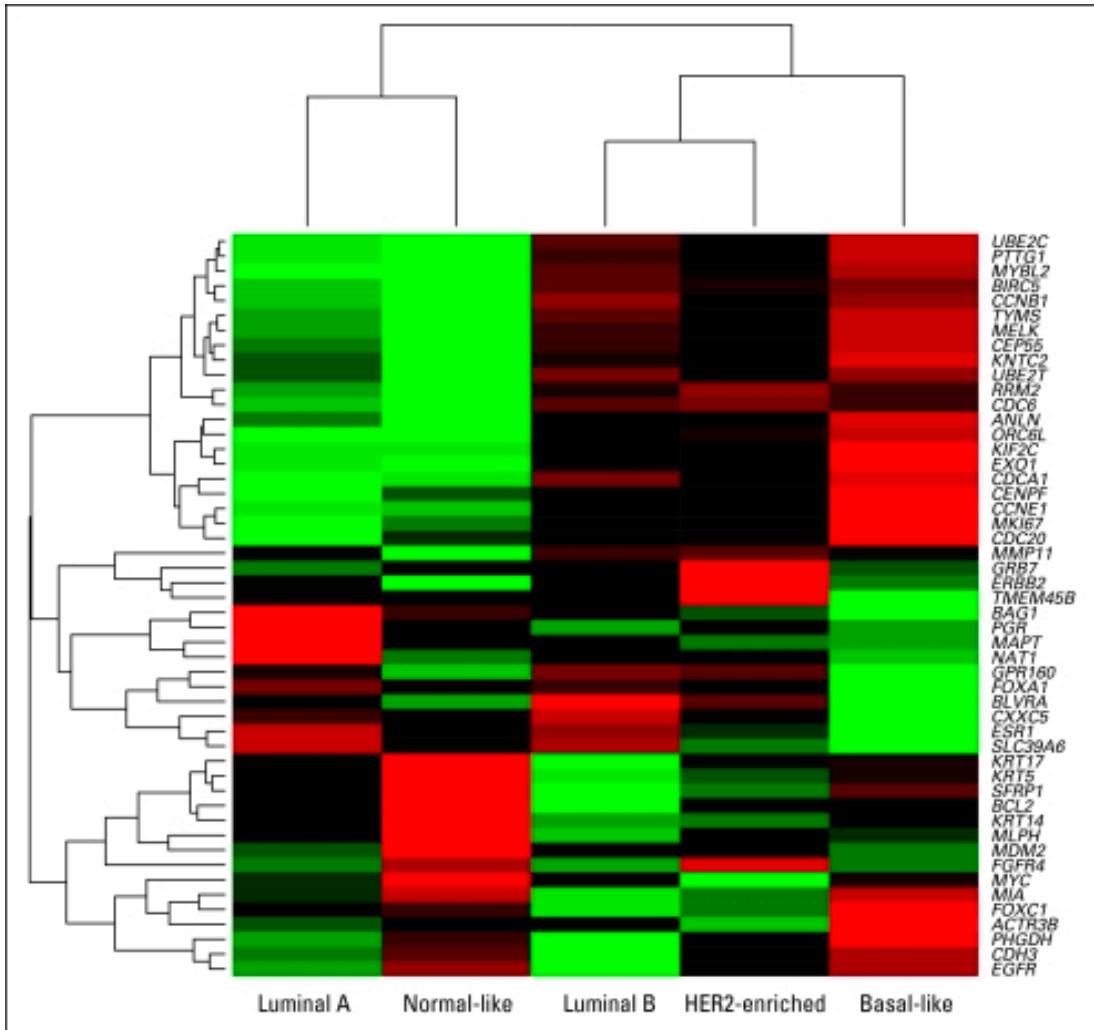
Application 2: estimation of cell type composition

Deconvolution methods can be used to break data up into its various composing elements. In RNA-seq analysis, this usually refers to the estimation of the relative fractions of individual cell types in a sample.

For example, **CIBERSORTx** and **EcoTyper** can estimate the cell types, their transcriptional states, and their co-associations from RNA-seq samples.



Application 3: classification + subtyping



Parker et al., Journal of Clinical Oncology, 2009

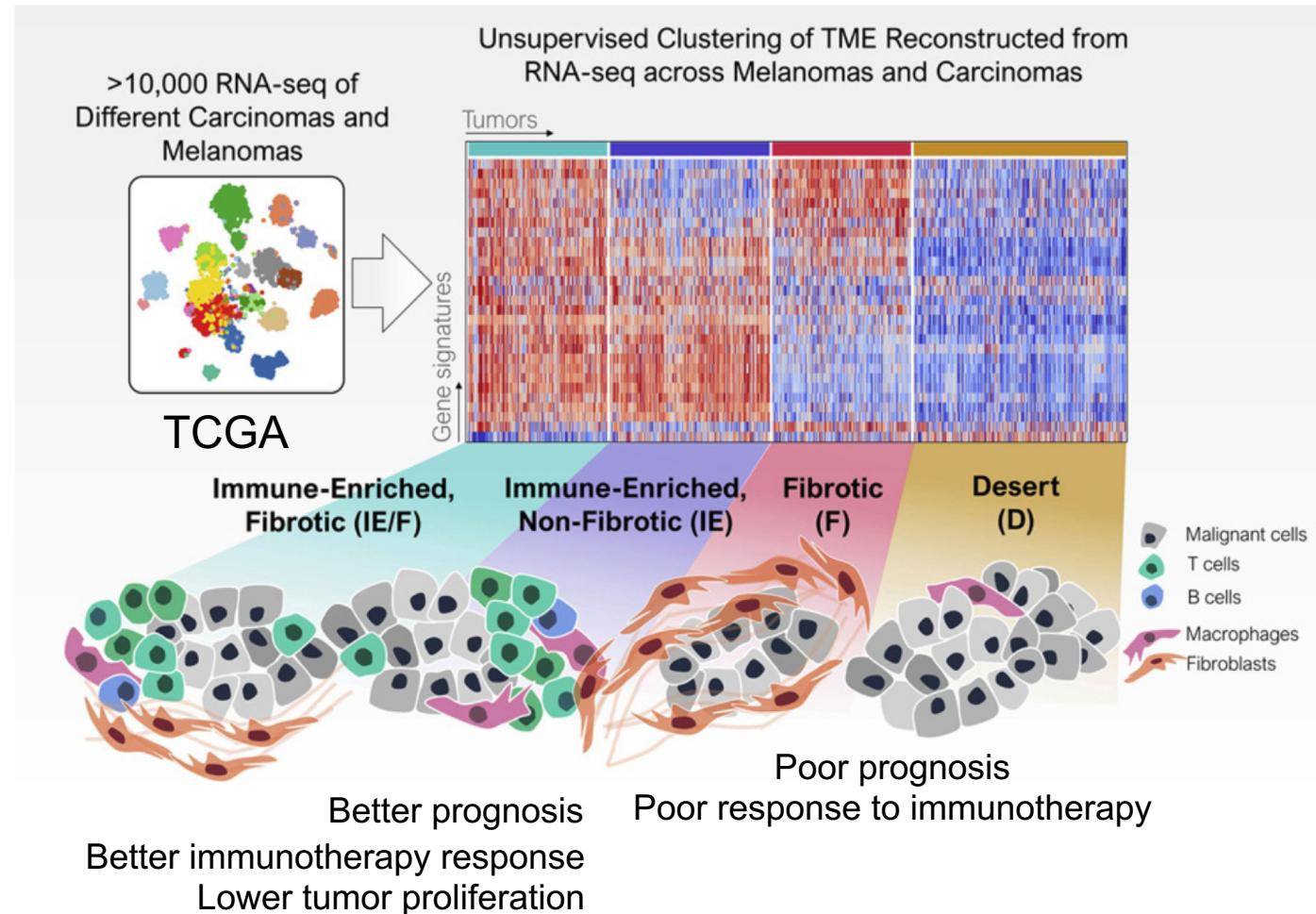
PAM50 in breast cancer

- 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 3: classification + subtyping

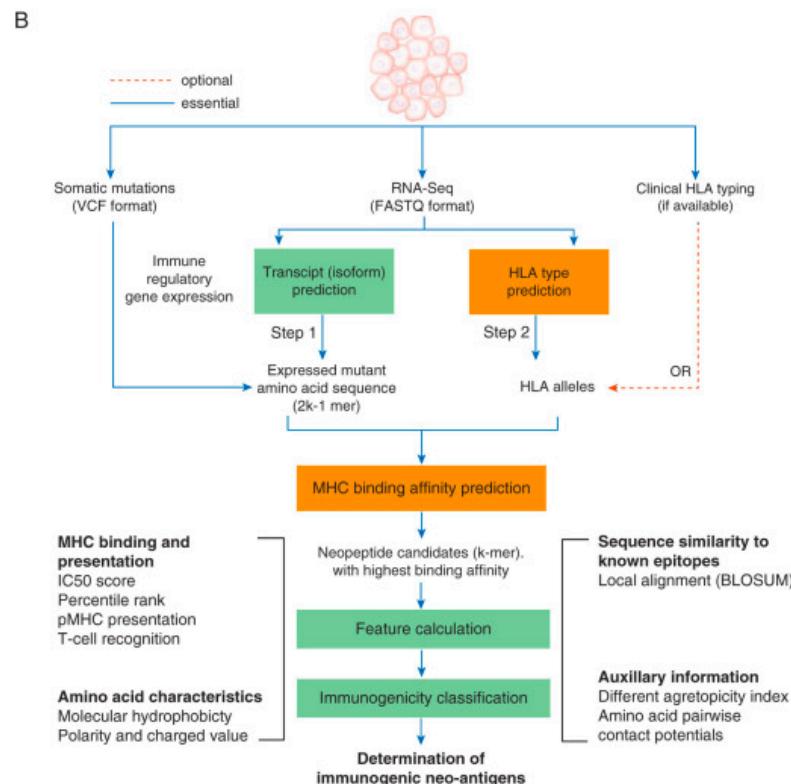
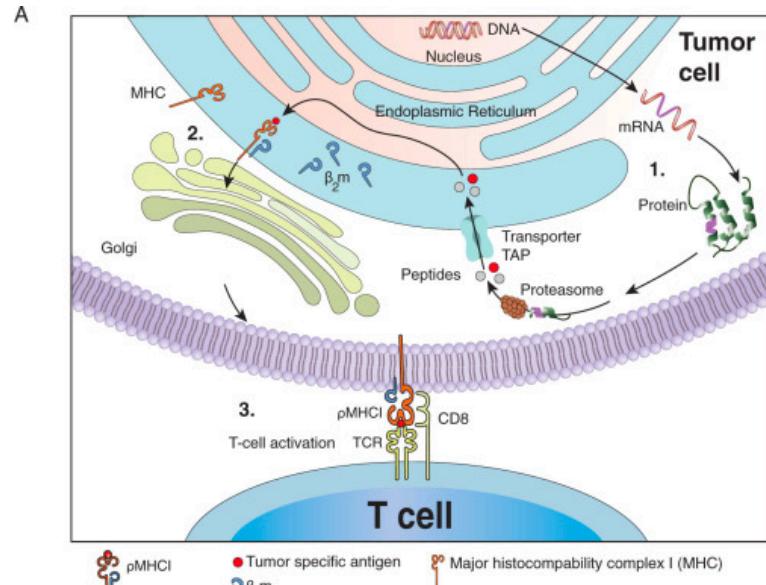
“A tumor personality test to guide therapeutic decision making”

Conserved
pan-cancer
microenvironment
subtypes



Application 4: 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 and RNA-seq data to identify candidate peptides.
- Tools for neoantigen discovery: Neopepsee, ScanNeo, ASNEO...



Clinical impact of comprehensive DNA and RNA sequencing

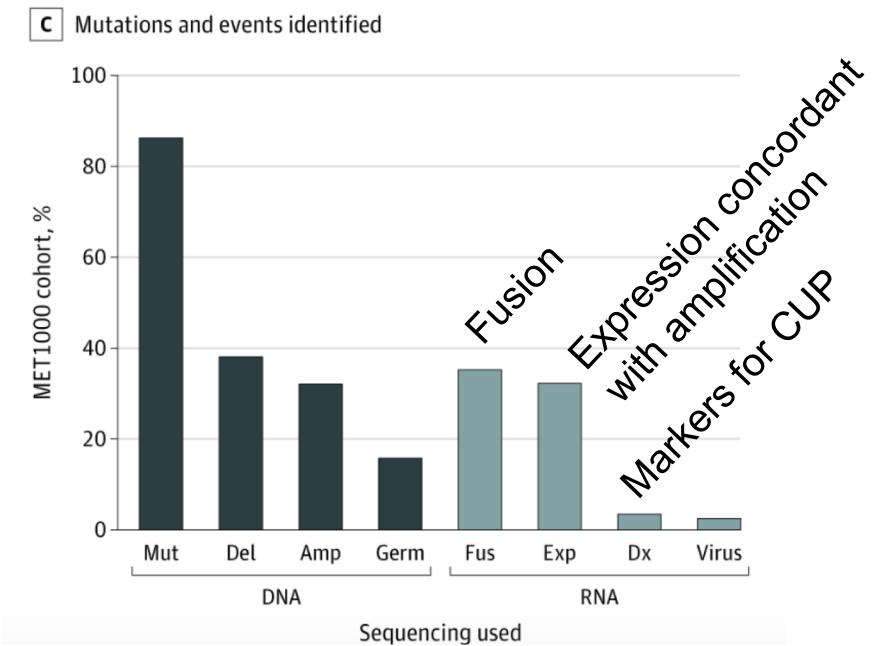
Research

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





**Karolinska
Institutet**