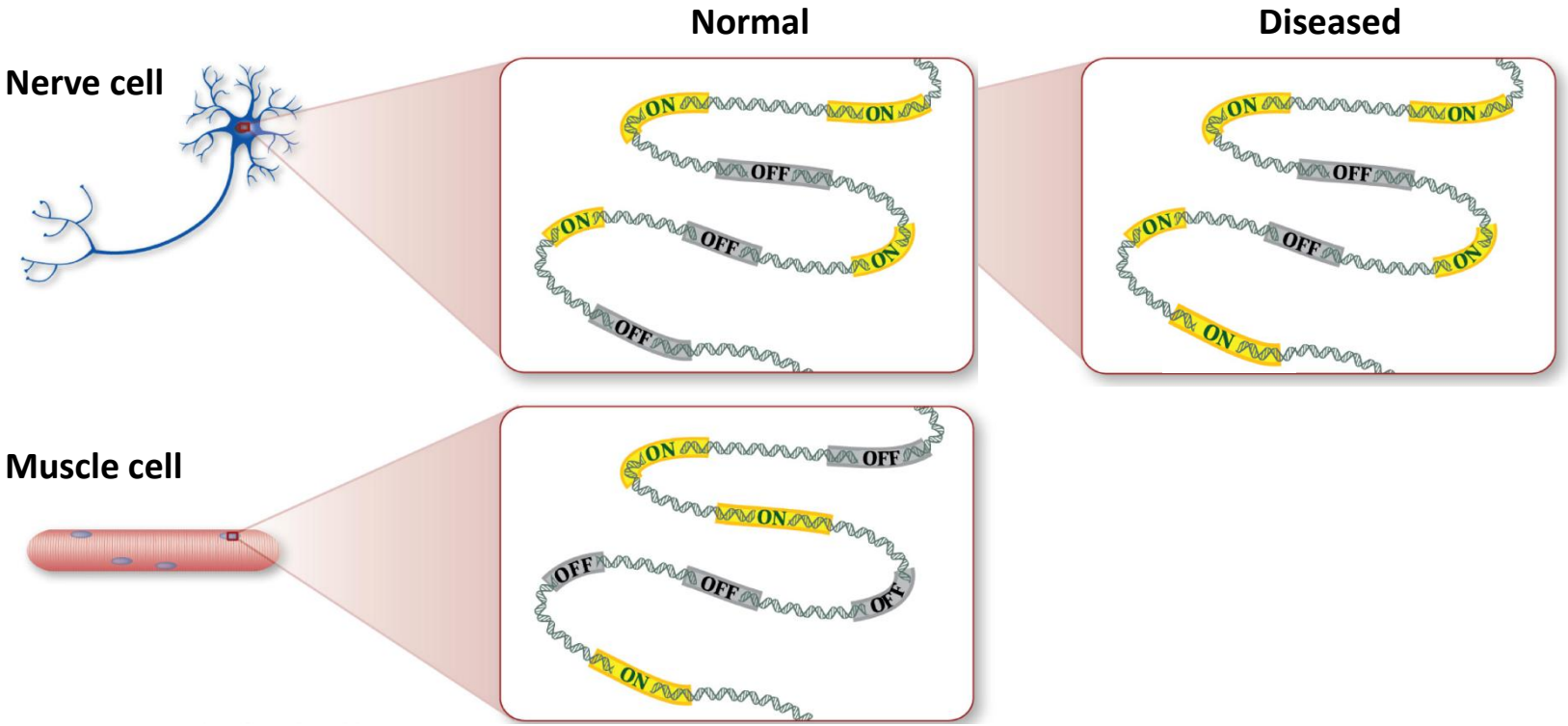
Biomarkers for disease identification/outcome

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

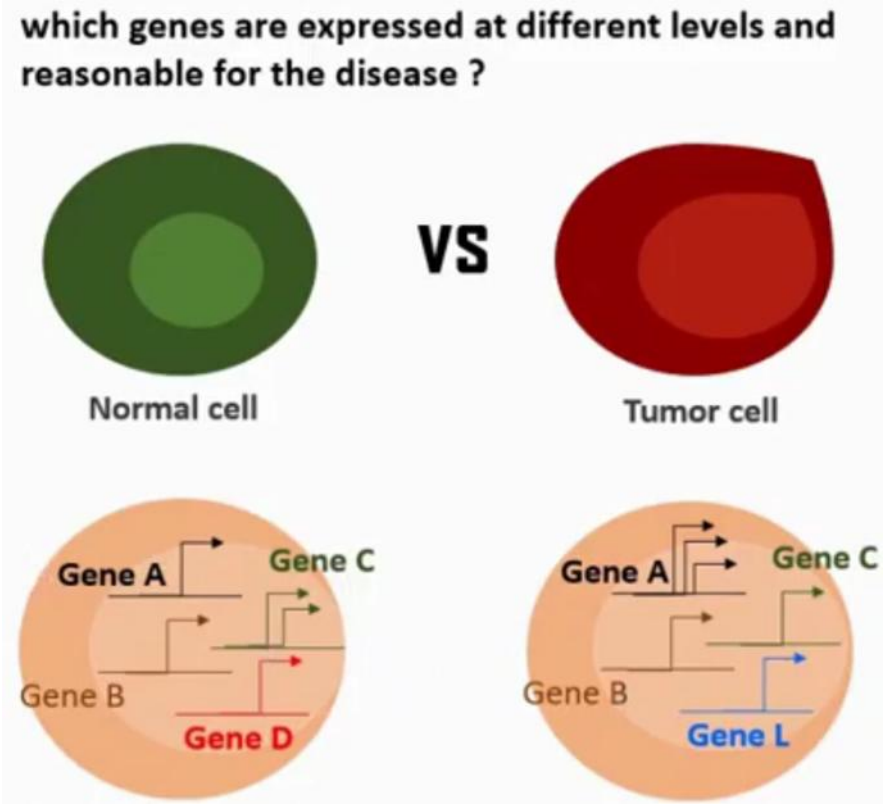
Dr. Jaspreet Kaur Dhanjal
Assistant Professor, Department of Computational Biology
Email ID: jaspreet@iiitd.ac.in

October 14, 2025

Why Transcriptome?



Differential gene expression analysis



What are the differentially expressed genes?

Gene A is up regulated
Gene c is down regulated
Gene D is turned off
Gene L is turned on

Why sequence RNA (Versus DNA)?

1. *Functional studies*

Genome may be constant but an experimental condition has a profound effect on the gene expression (differential expression)

Eg. Drug *vs.* untreated cells

Eg. Wild type *vs.* knock out mice cells

2. Predicting transcript sequence from genome sequence is difficult

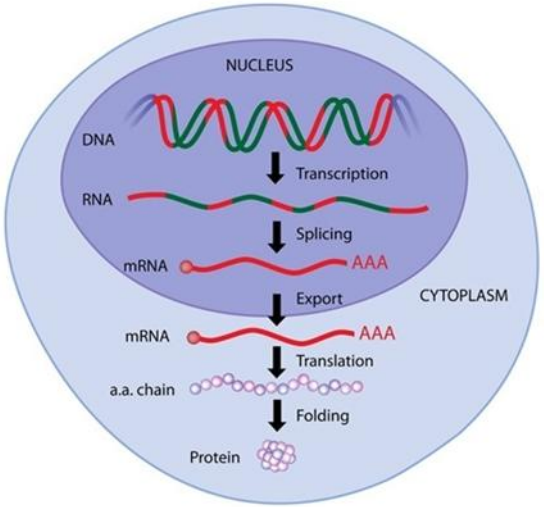
3. Some molecular features can only be observed at the RNA level

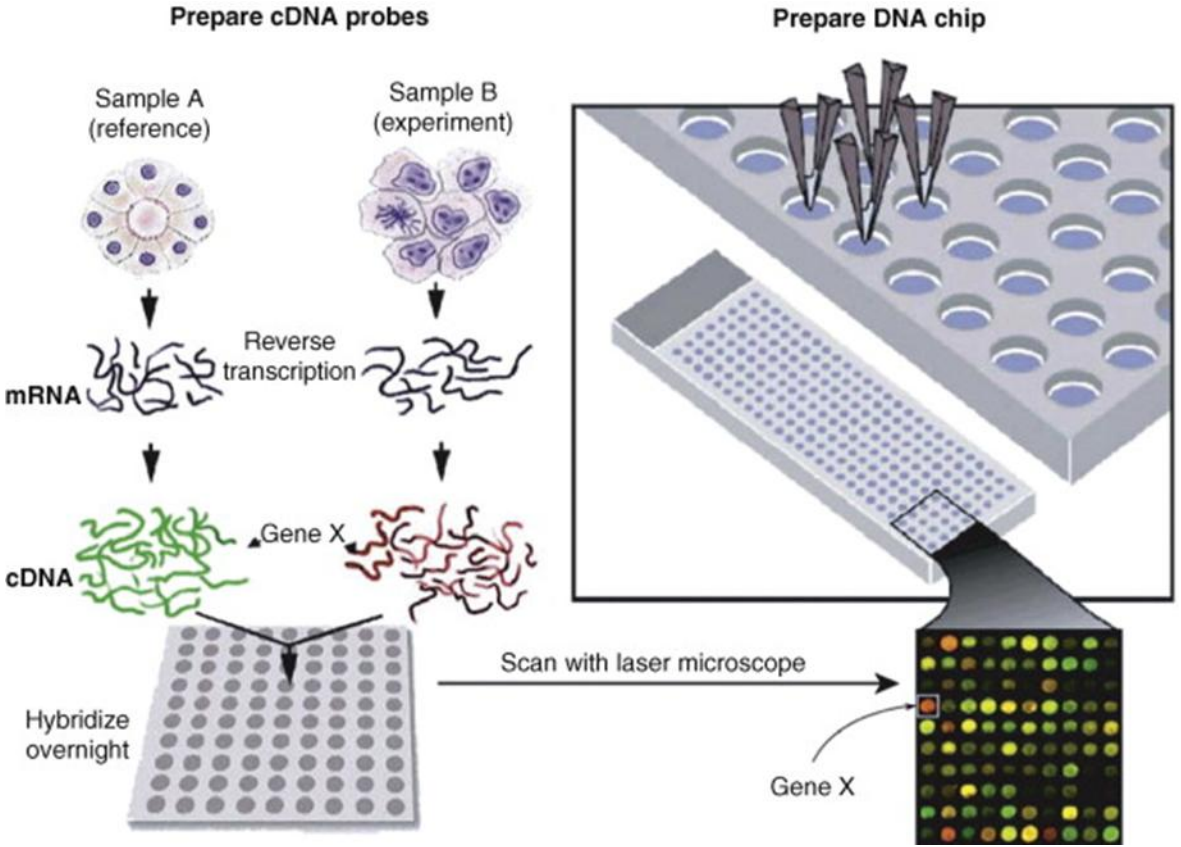
Alternative isoforms, fusion transcripts, RNA editing

4. Understand allele specific expression

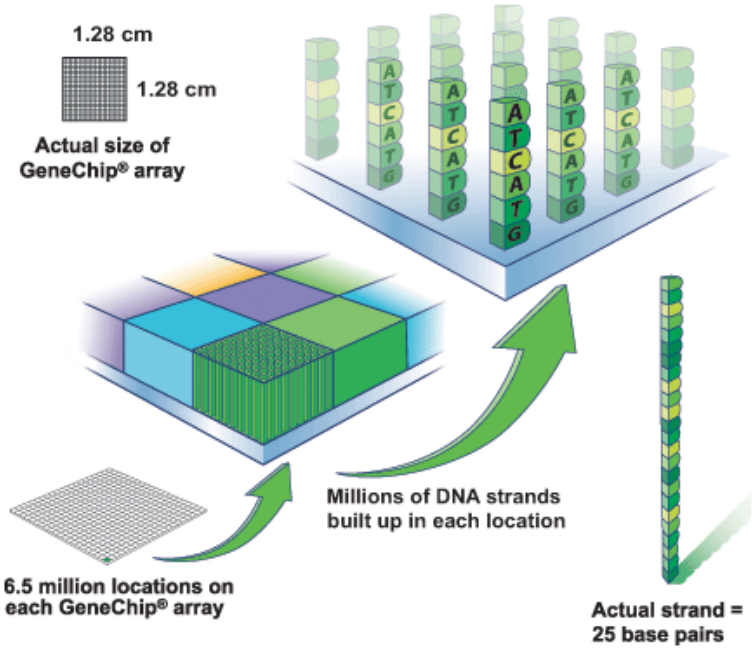
Gene expression-based biomarker identification

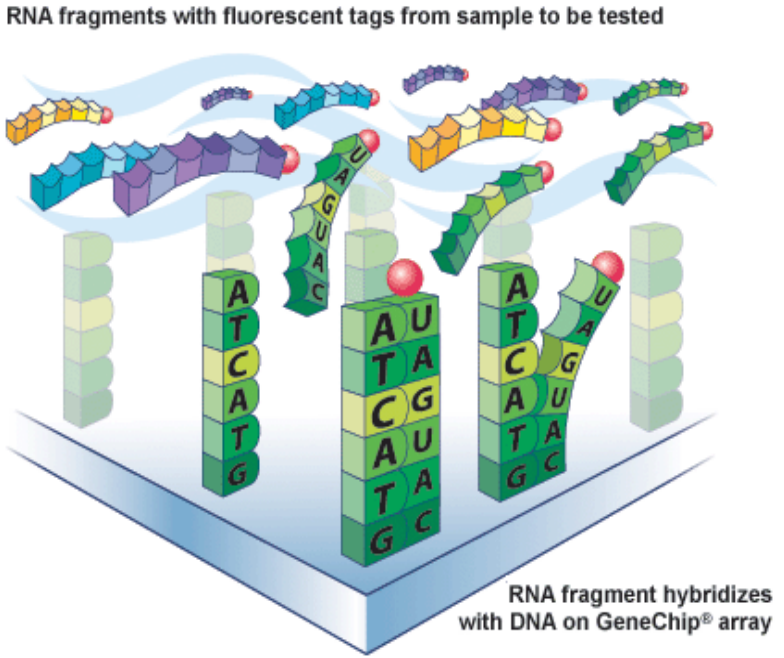
DNA Microarray





Microarray analysis

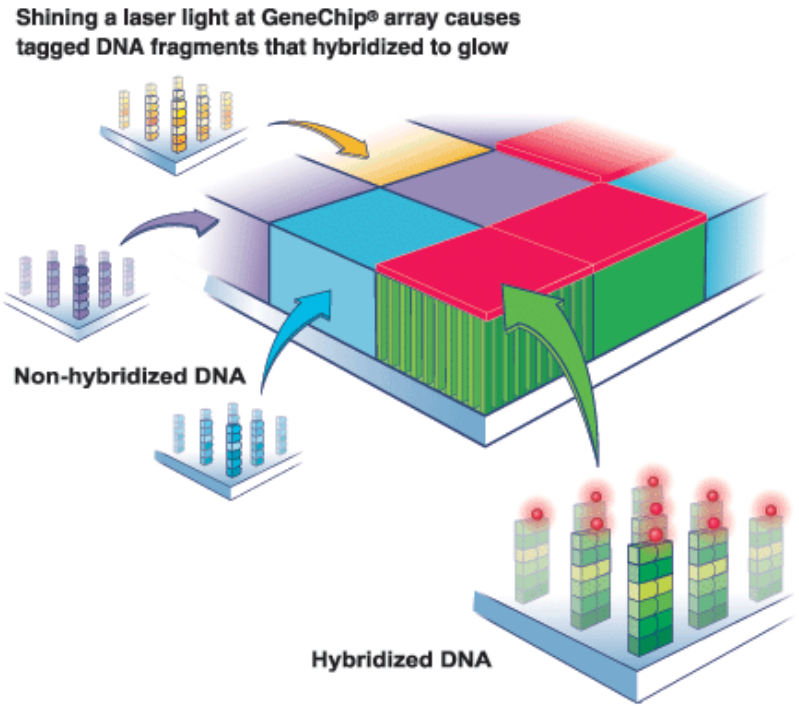




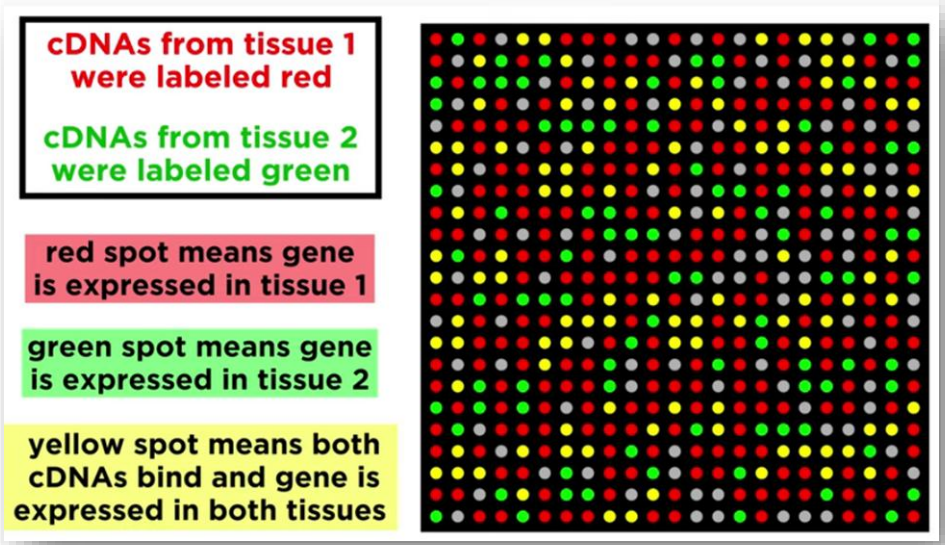
1. Micro Array Features

2. Hybridization (Pairing)

Microarray analysis



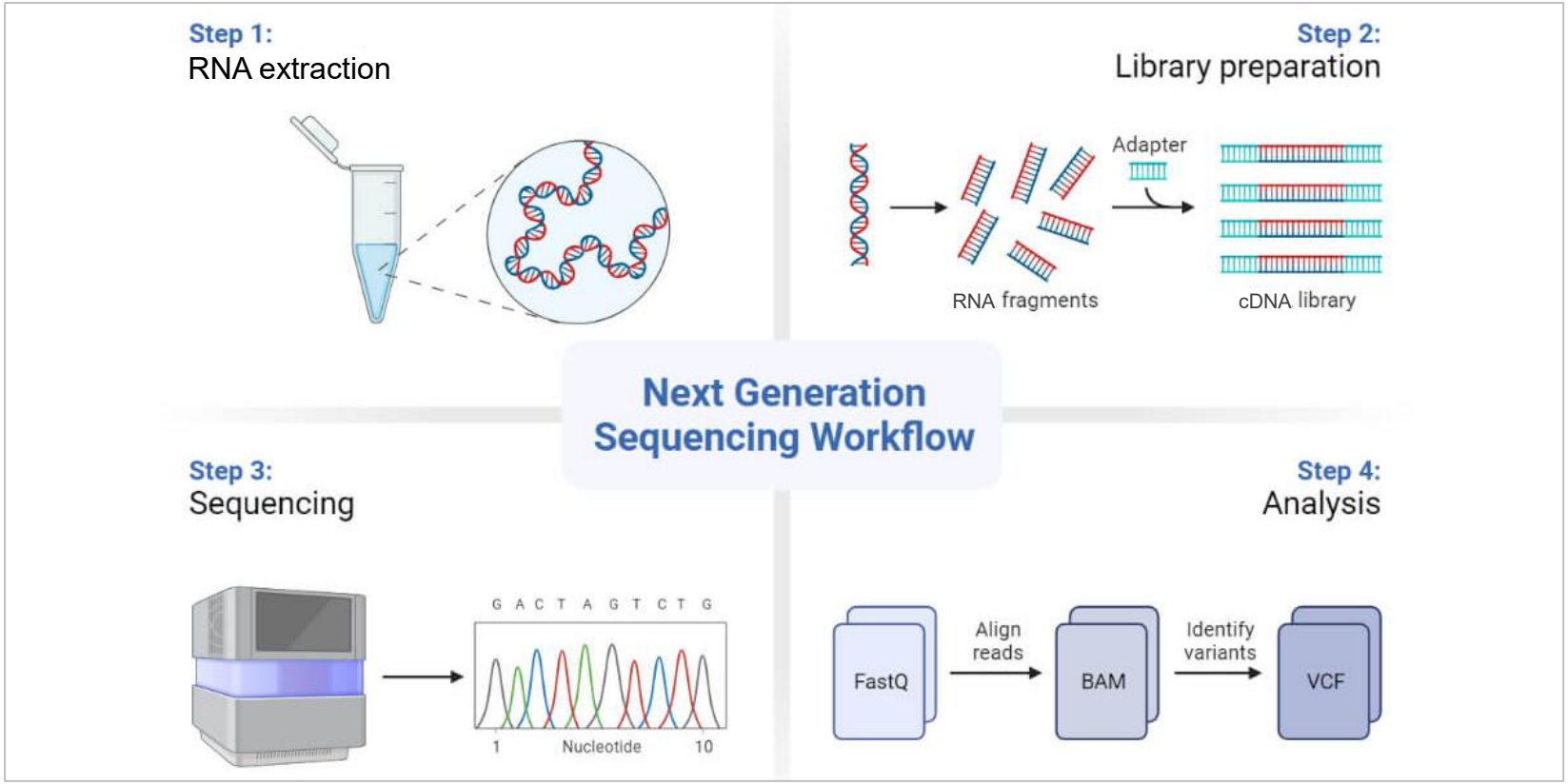
3. Detection



Limitations:

1. **Relies on existing knowledge about genome sequence.**
2. **Technical problems like high background levels owing to cross-hybridization**
3. **Comparison of expression across different samples/experiments is often complicated.**

Next generation RNA-sequencing



RNA-seq experiment workflow

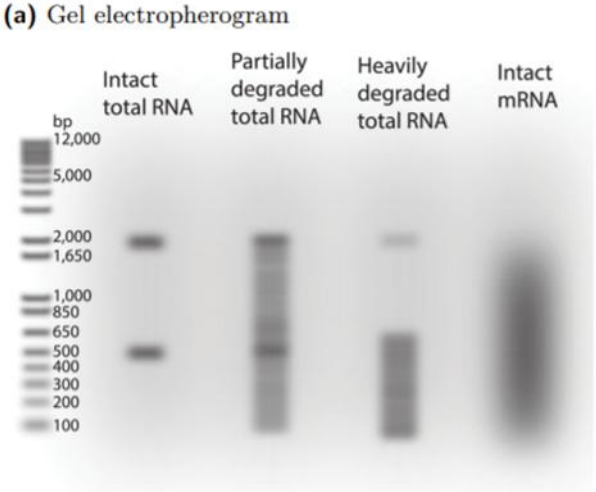
Types of RNA

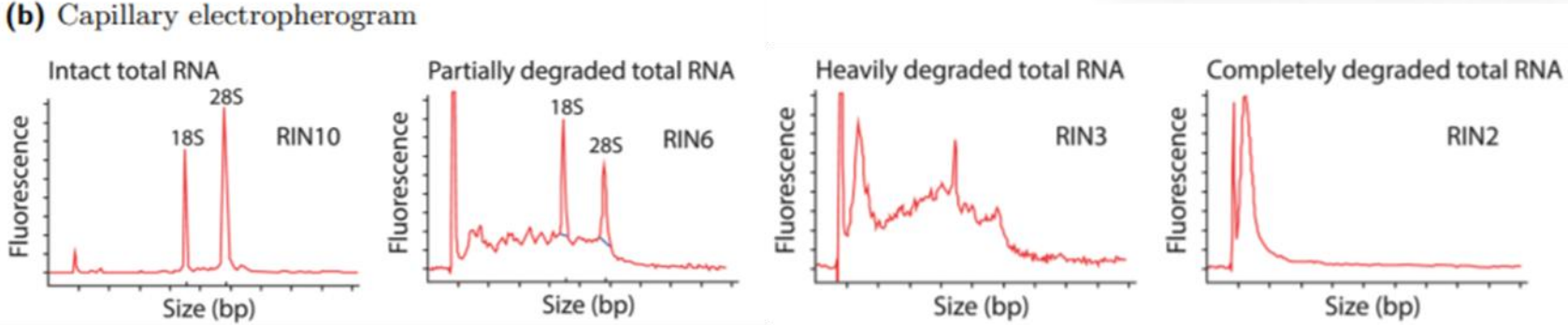
poly-adenylated (coding) RNAs, "genes"
short non-coding RNAs (ncRNA), "microRNA"
long non-coding RNAs
} Total RNA

ribosomal RNA

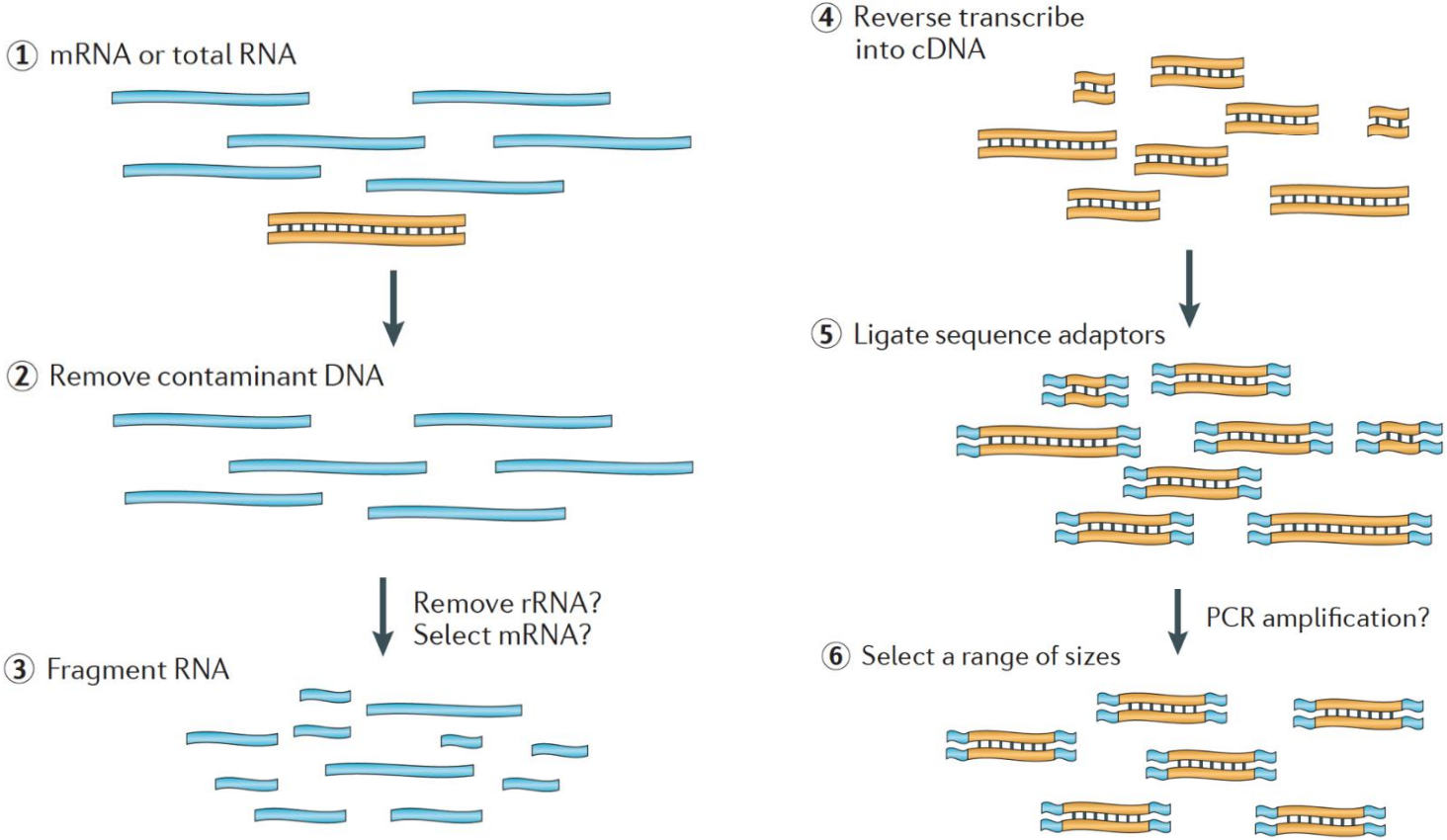
"Enrichment"

Most of the RNA in the cell { polyA capture
ribominus



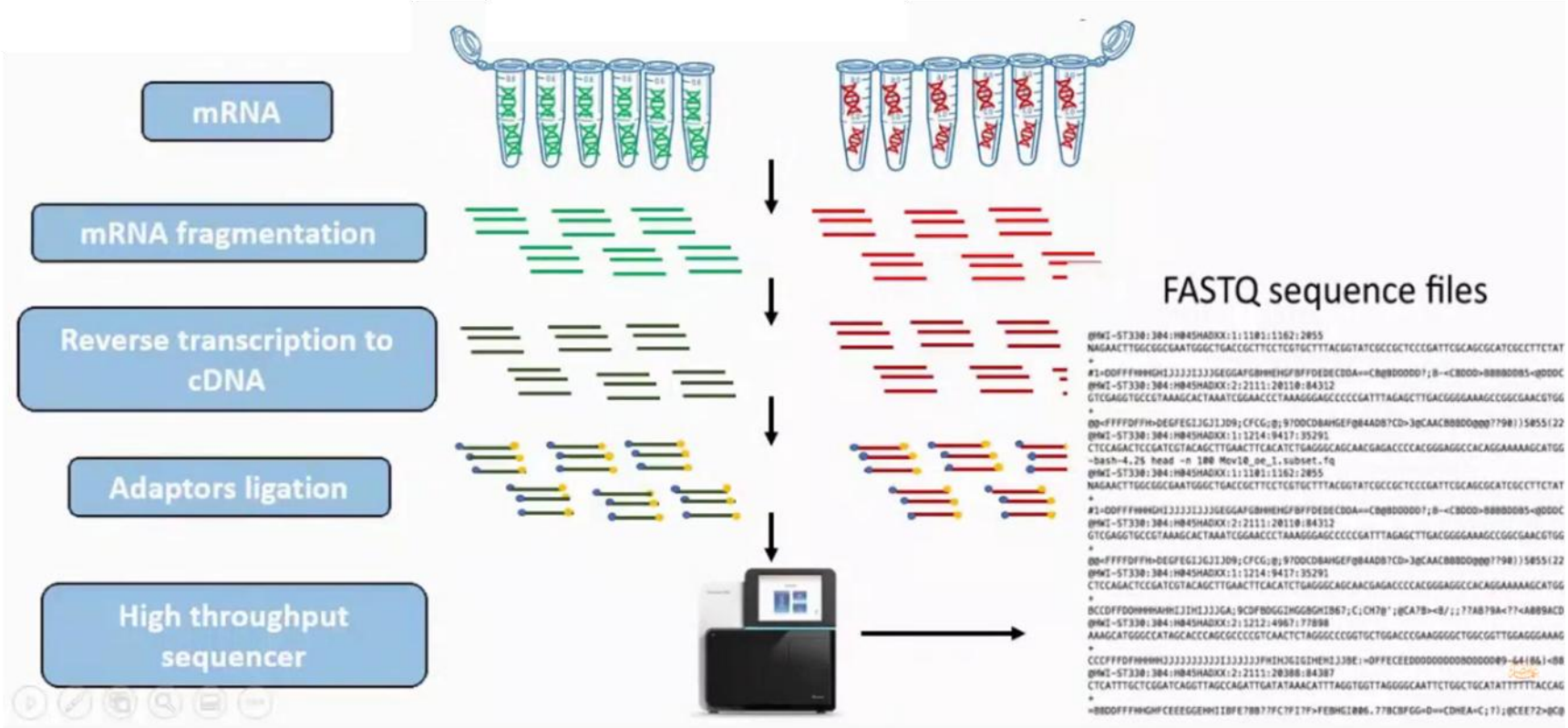


Library Preparation



RNA-seq experiment workflow

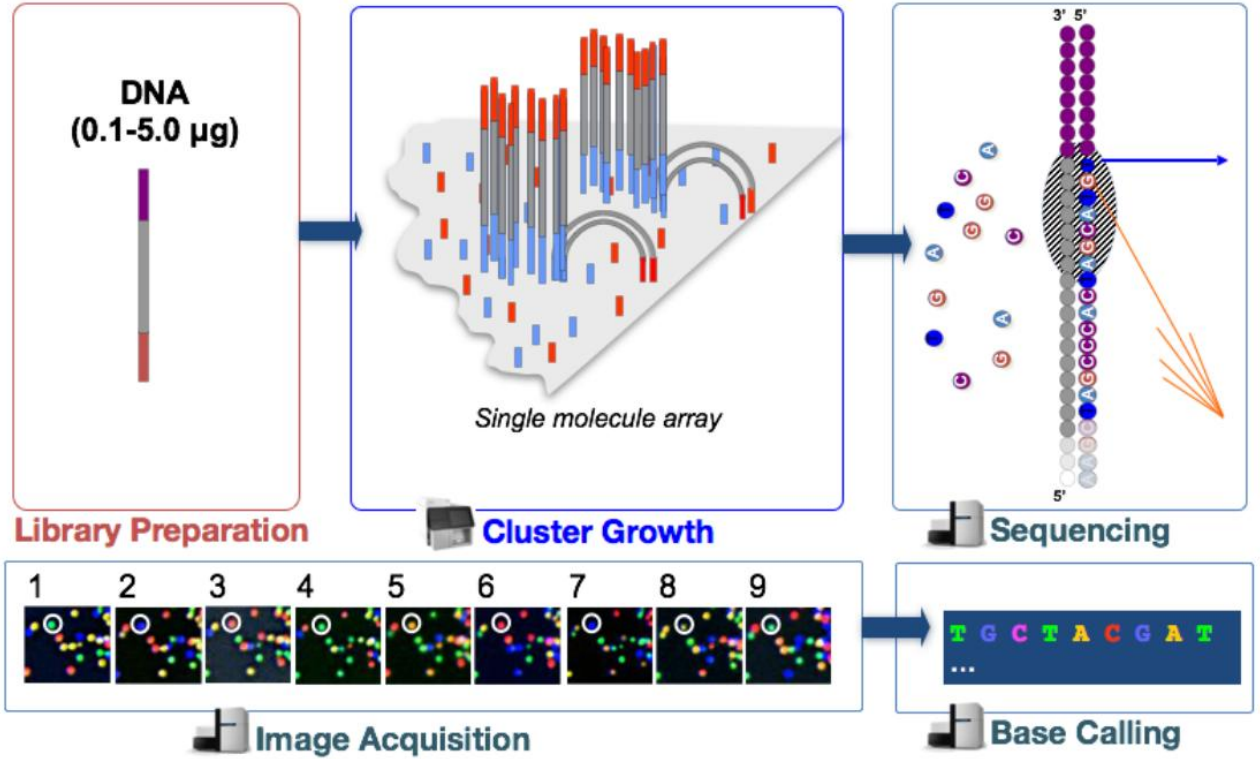
Library preparation



Illumina sequencing platforms

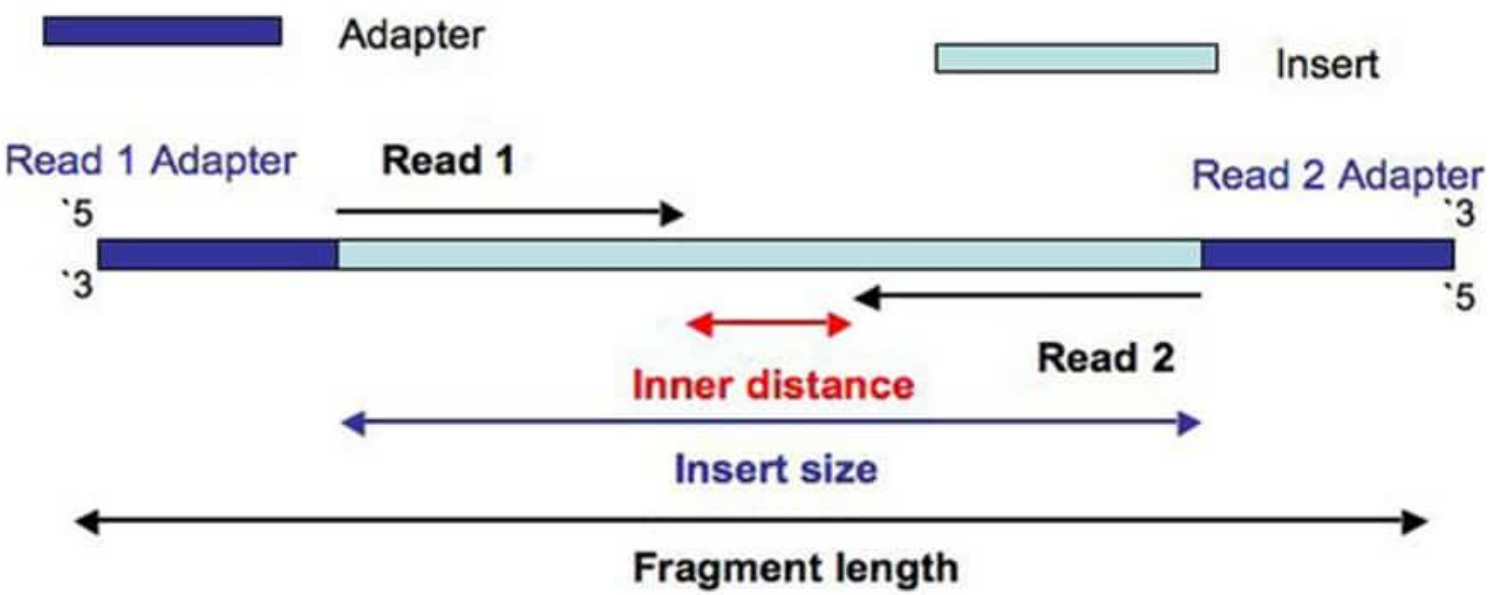
Other sequencing platforms: Pacific Bioscience, Oxford Nanopore, 10X Genomics

Sequencing by synthesis



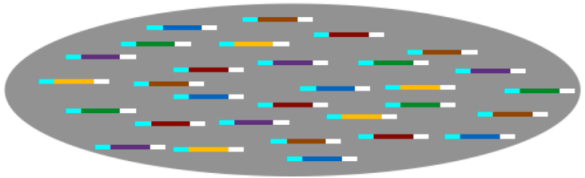
Illumina sequencing video URL: https://www.youtube.com/watch?v=womKfikWlxM

Single- and Paired-end sequencing



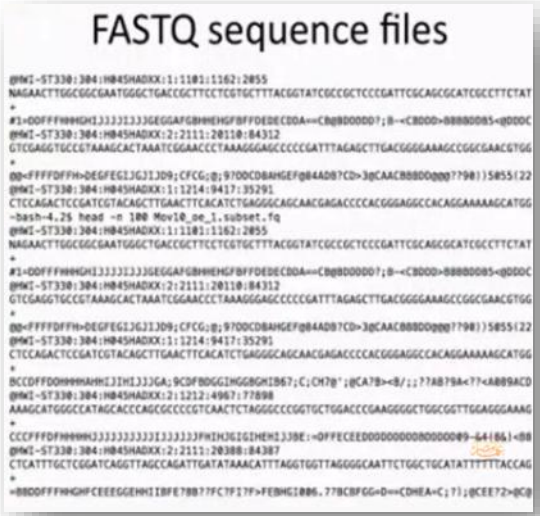
Multiplexing

Generate & pool indexed cDNA libraries



Sequence pooled libraries on a single lane

FASTQ sequence files



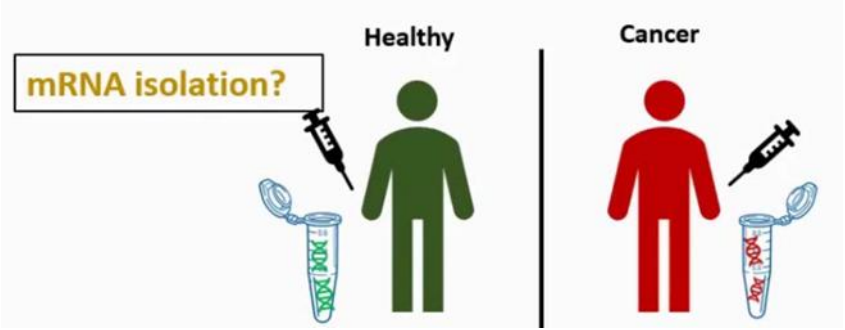
in silico: Demultiplex the data on index

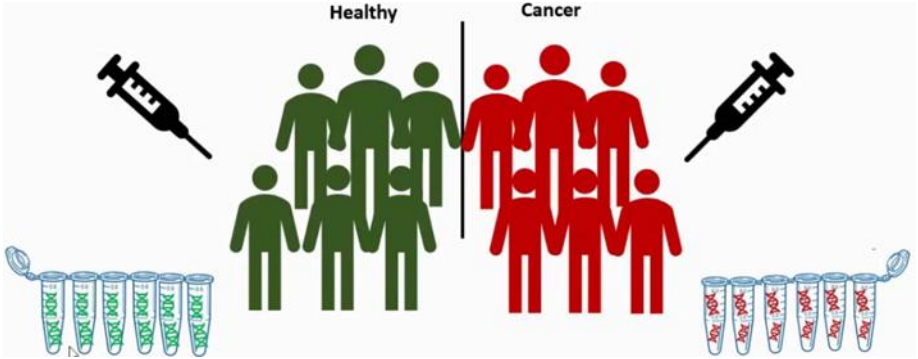
Microarray *vs.* RNA-seq

Microarray	RNA-seq
• Limited probe-set based on prior knowledge of the transcriptome	• Comprehensive overview of the transcriptome
• Higher throughput	• Best dynamic range
• Analysis is more user-friendly than RNA-seq currently	• Gene fusion, isoform, SNPs detection

RNA-seq experiment workflow

Sample preparation



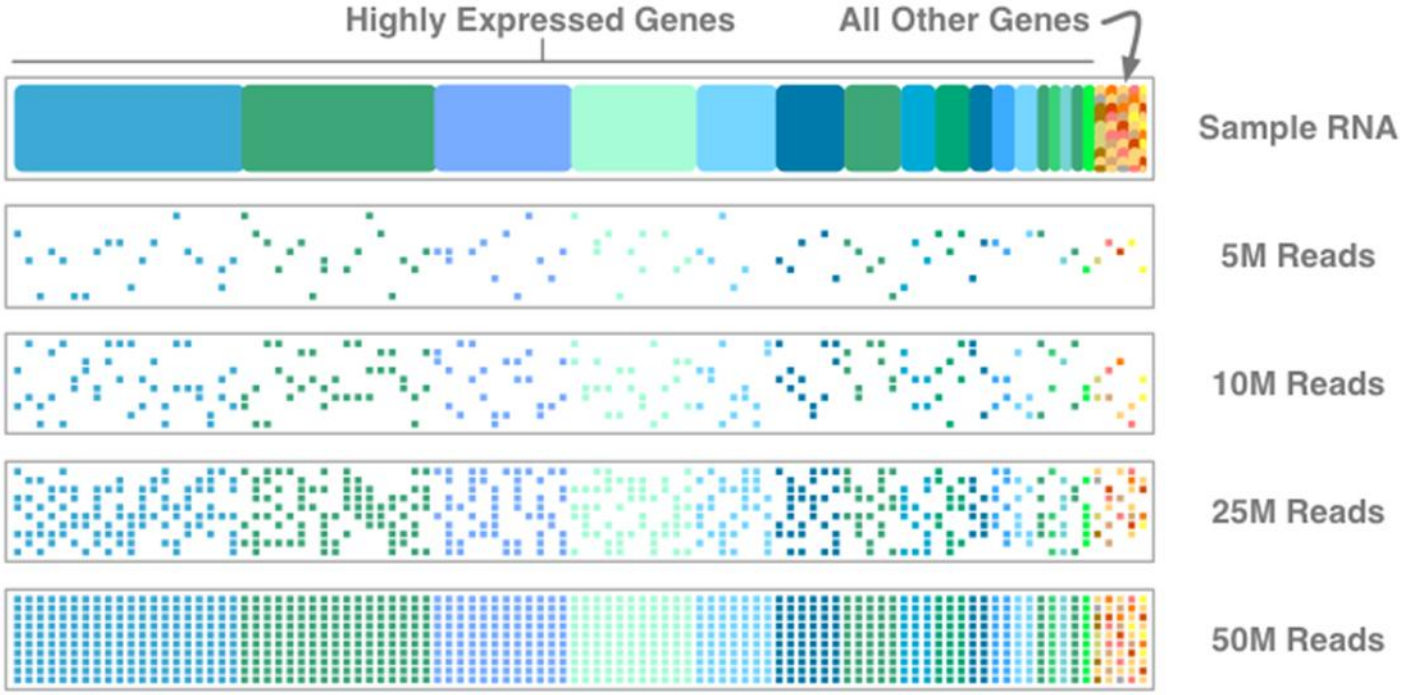


1. *Biological replicates* : Include multiple sampling within the population
2. *Technical replicates* : Include multiple preparation and re-sequencing of the same sample

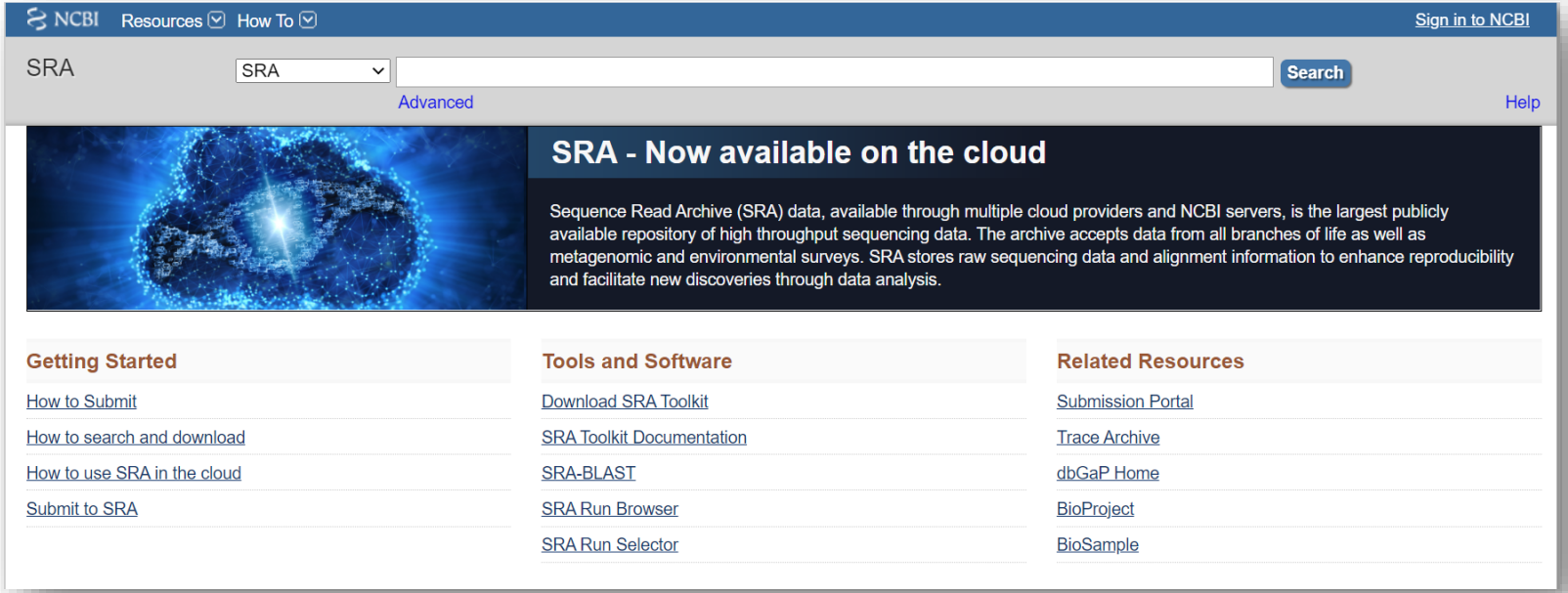
Biological replicates generally increase statistical power more than technical replicates

- Biological variability is generally greater than technical variability
- Biological replicates contain both biological and technical variability

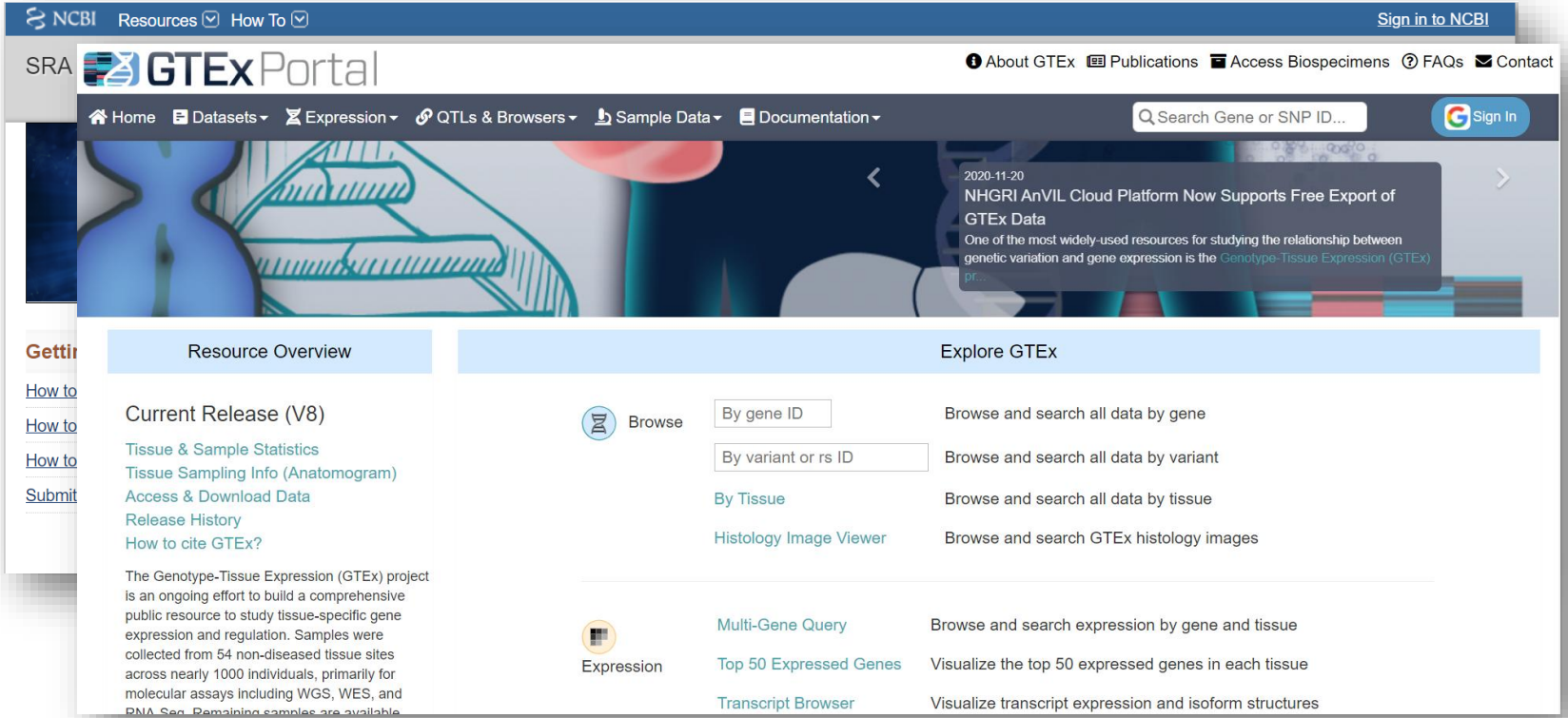
Sequencing depth



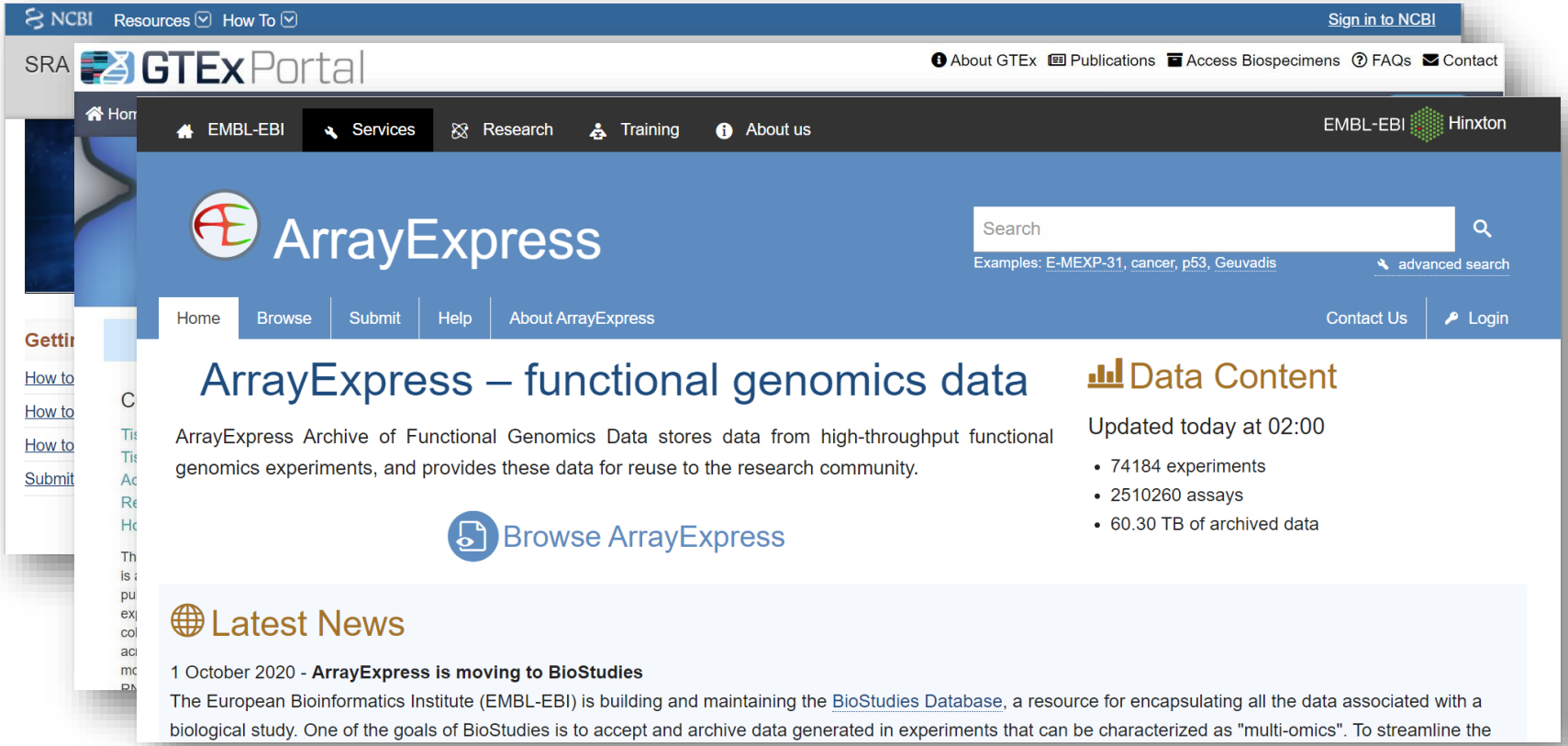
Sources for RNA-seq datasets



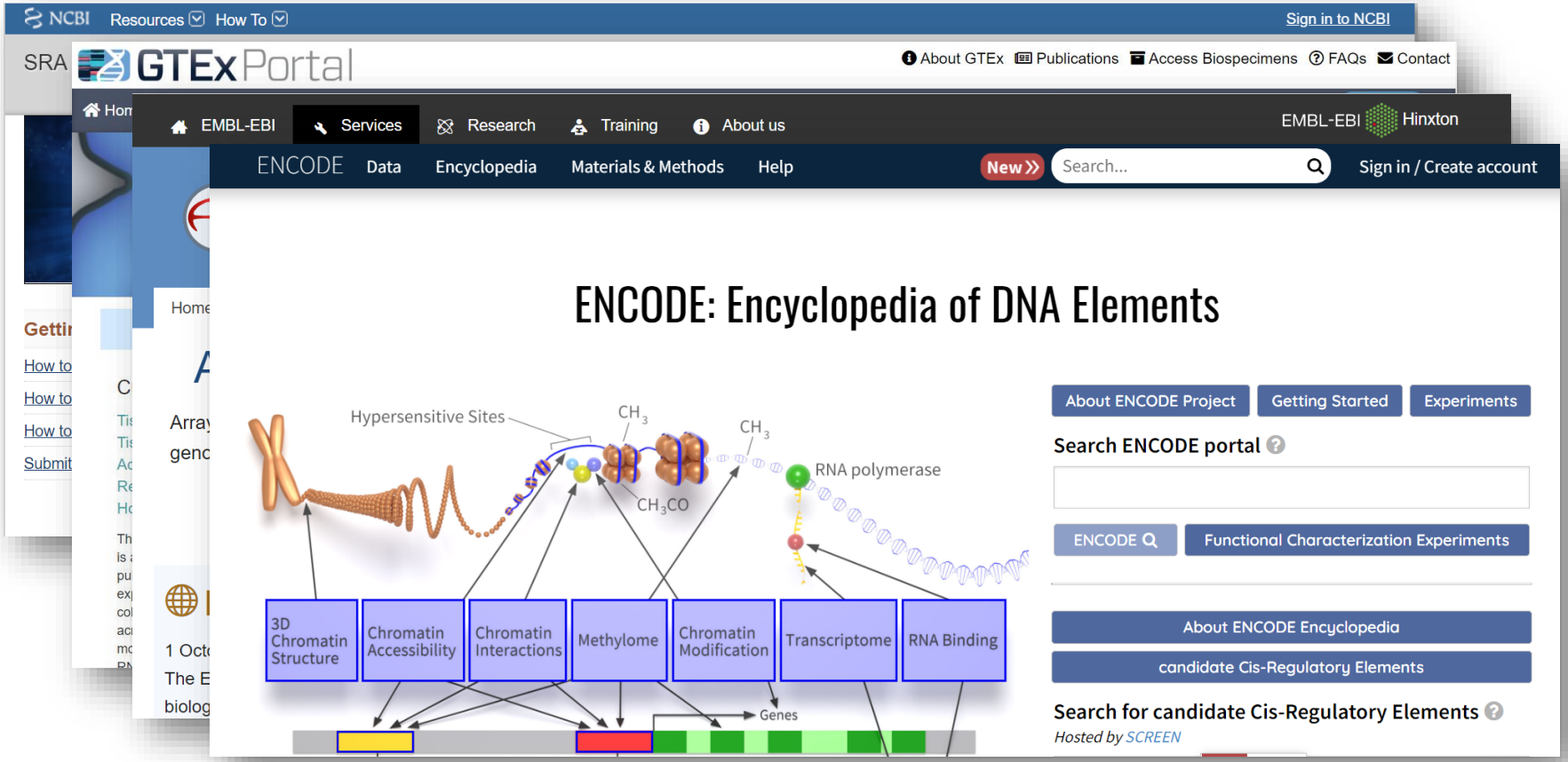
Sources for RNA-seq datasets



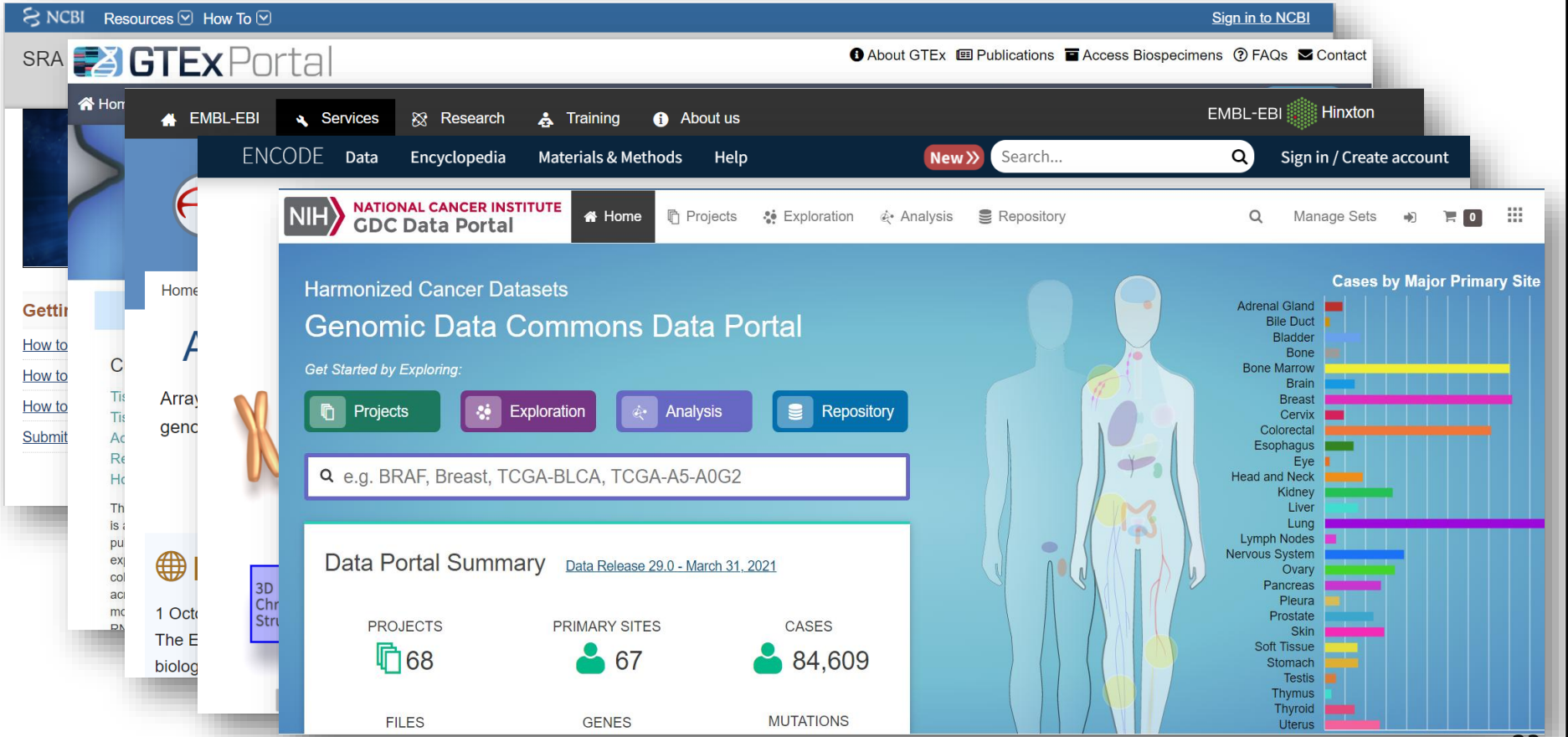
Sources for RNA-seq datasets



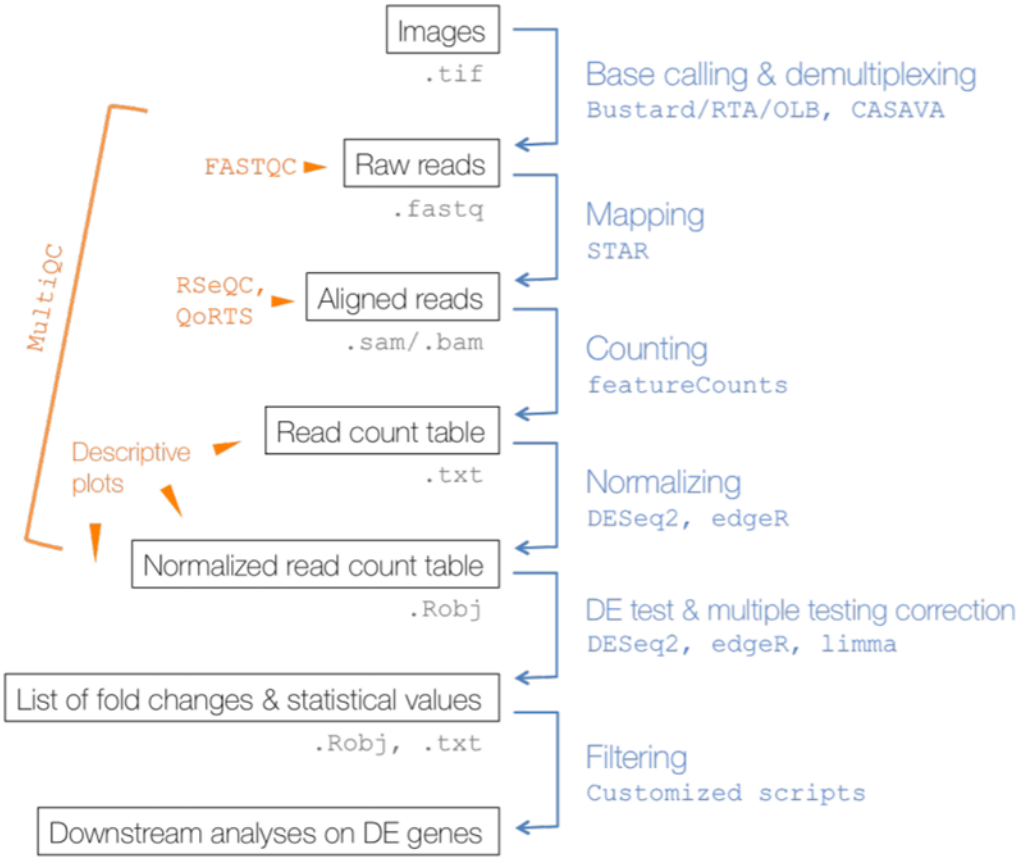
Sources for RNA-seq datasets



Sources for RNA-seq datasets



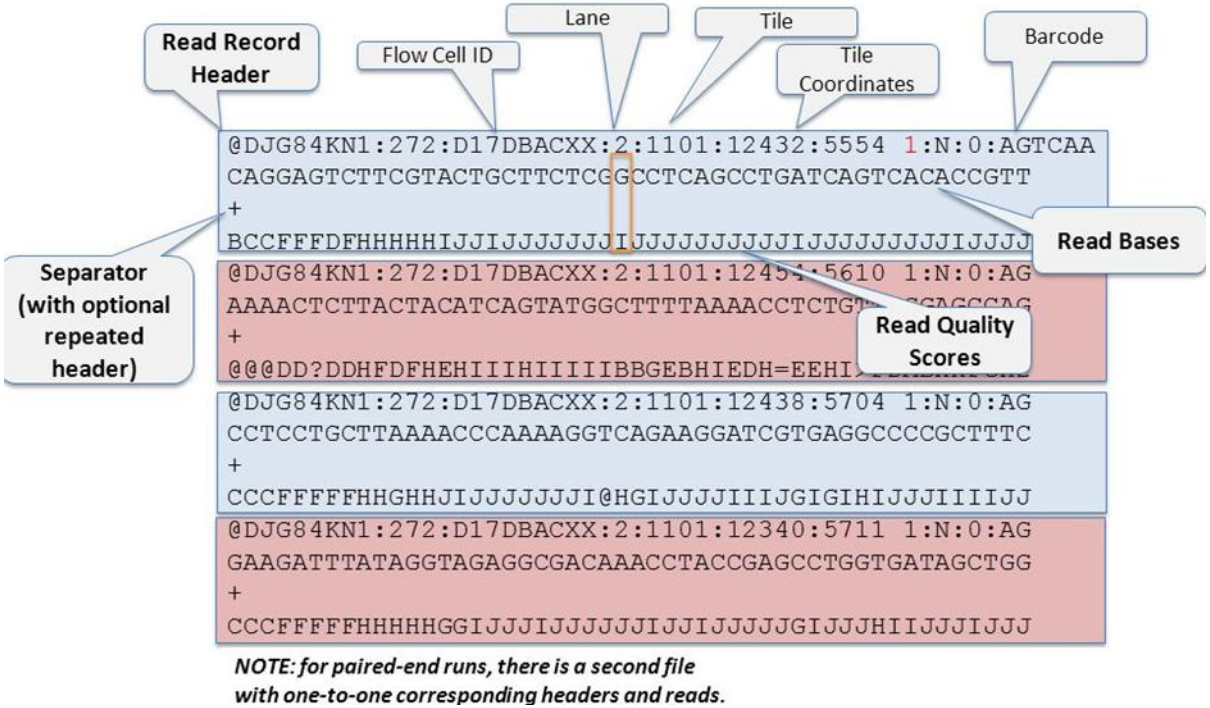
Workflow of differential gene expression analysis



Gene expression-based biomarker identification

Problems in sequencing

1. Low confidence bases, Ns
2. Specific sequence bias, GC bias
3. Adaptors
4. Sequence contamination

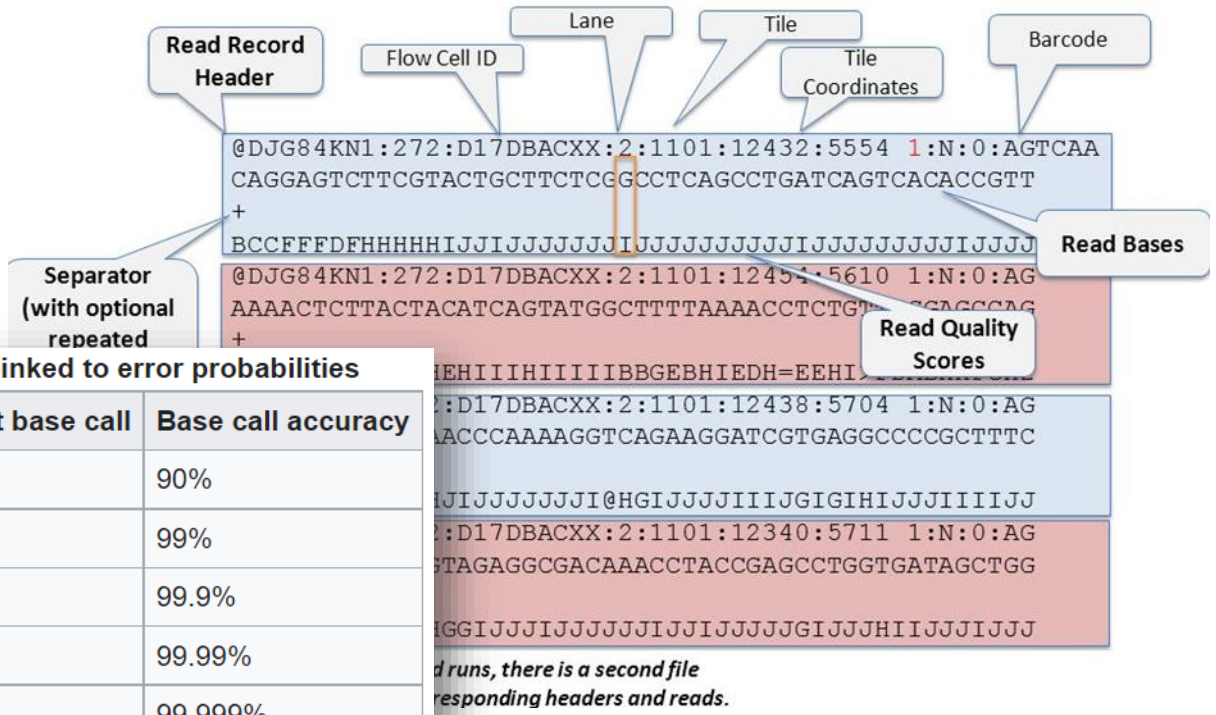


NOTE: for paired-end runs, there is a second file with one-to-one corresponding headers and reads.

Gene expression-based biomarker identification

Problems in sequencing

1. Low confidence bases, Ns
2. Specific sequence bias, GC bias
3. Adaptors
4. Sequence contamination



d runs, there is a second file
responding headers and reads.

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%