

Biomarkers for disease identification/outcome



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

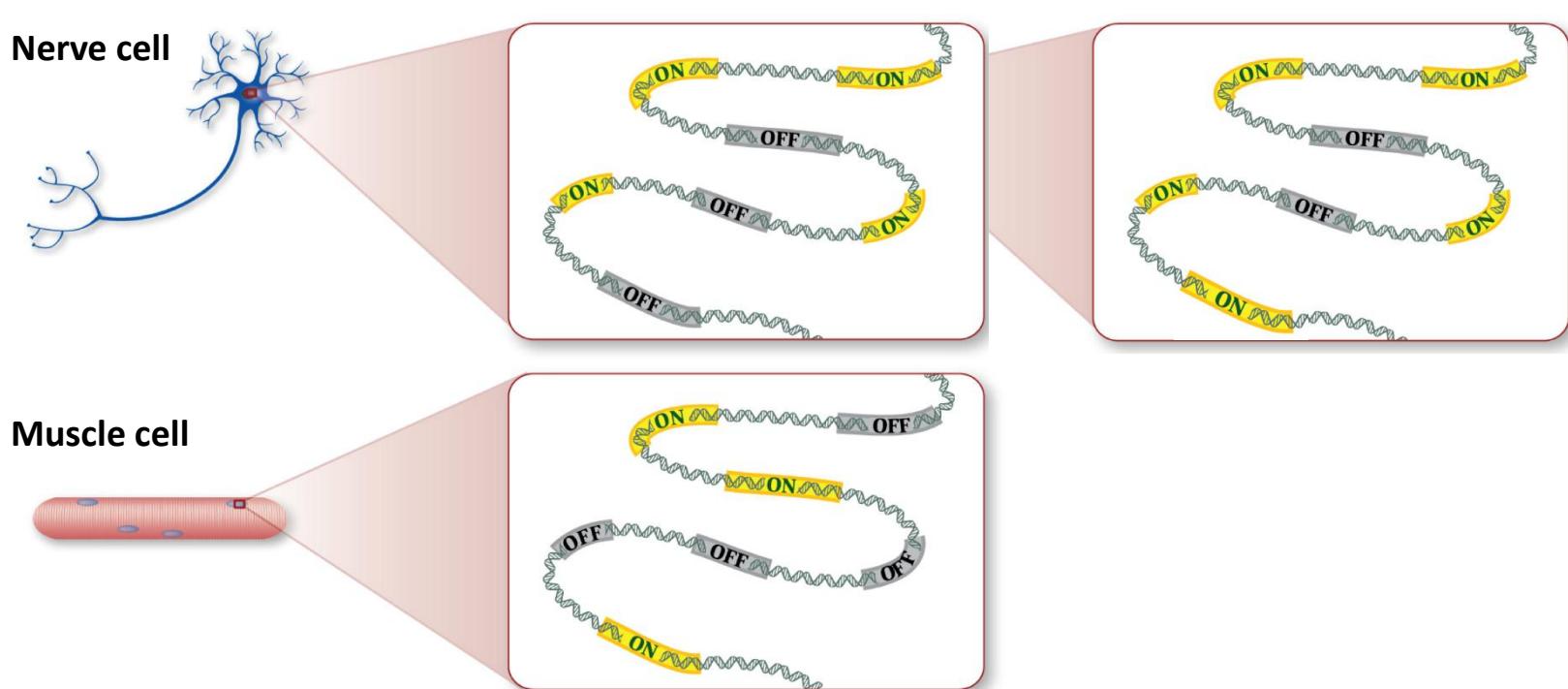
Dr. Jaspreet Kaur Dhanjal

Assistant Professor, Department of Computational Biology

Email ID: jaspreet@iiitd.ac.in

October 14, 2025

Why Transcriptome?

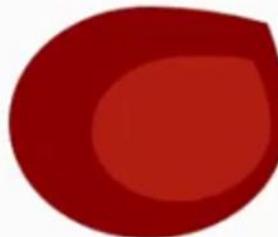


Differential gene expression analysis

which genes are expressed at different levels and reasonable for the disease ?



VS



Normal cell

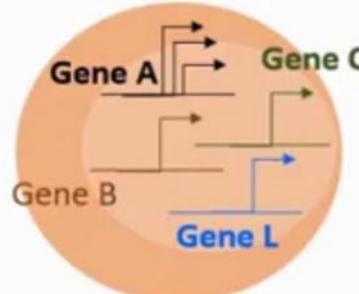
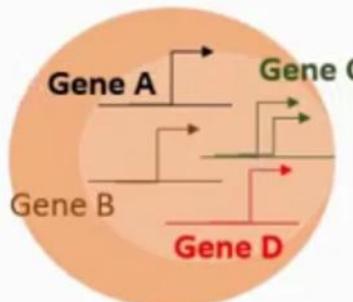
What are the differentially expressed genes?

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*

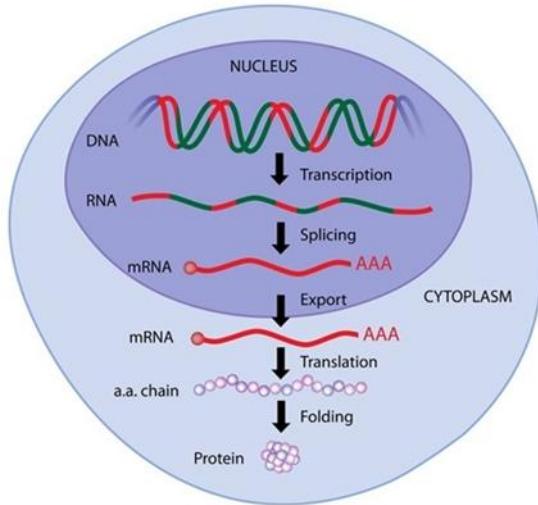
Alternative isoforms, fusion transcripts, RNA editing

‘

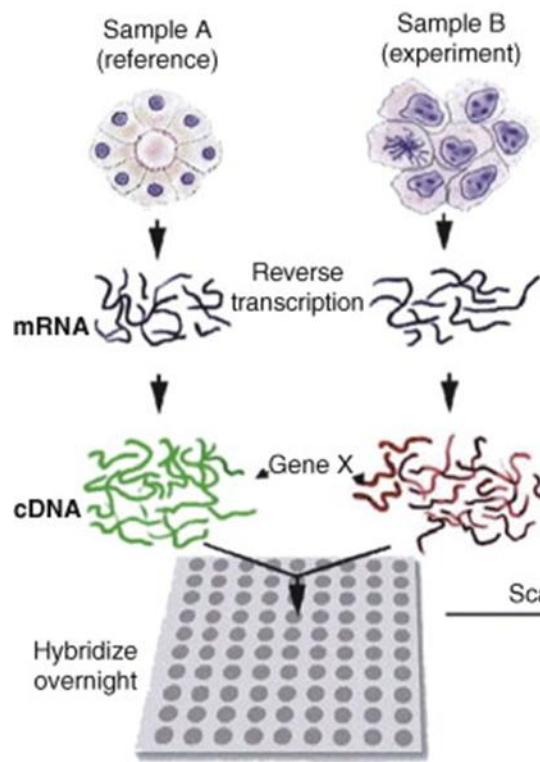
4. *Understand allele specific expression*

Gene expression-based biomarker identification

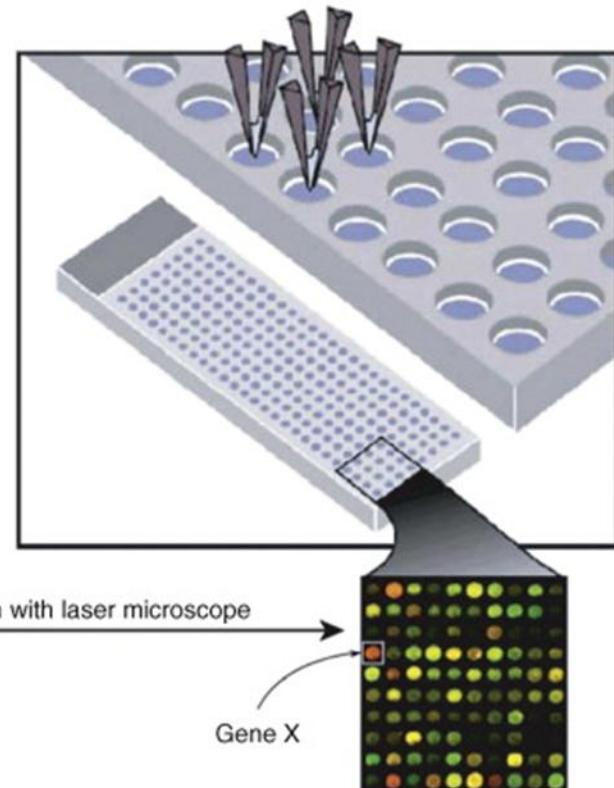
DNA Microarray



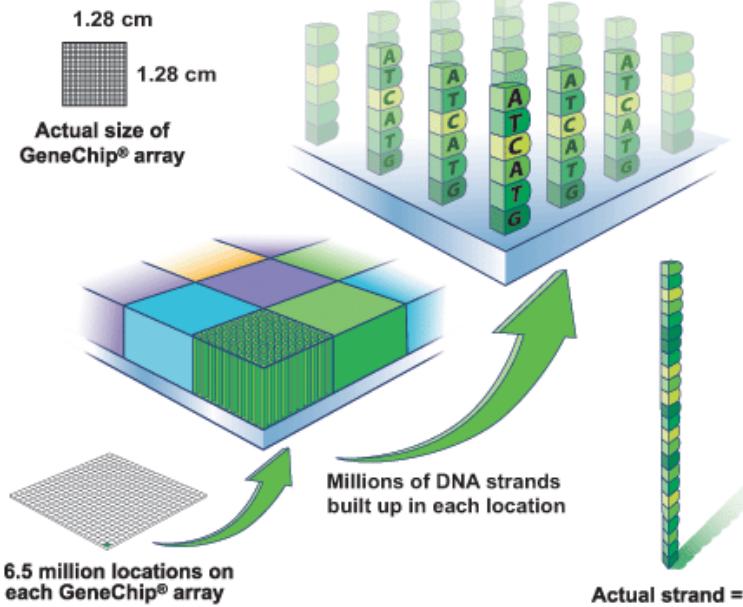
Prepare cDNA probes



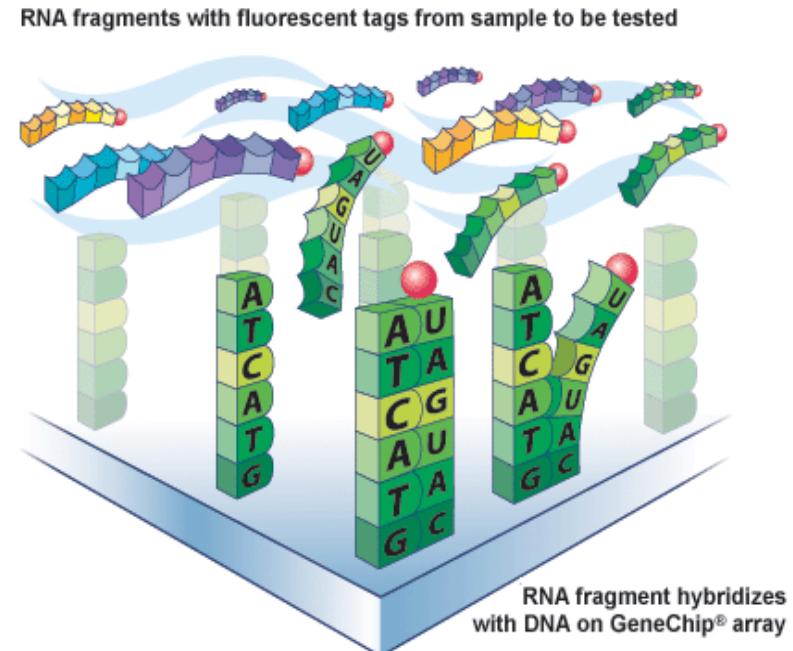
Prepare DNA chip



Microarray analysis



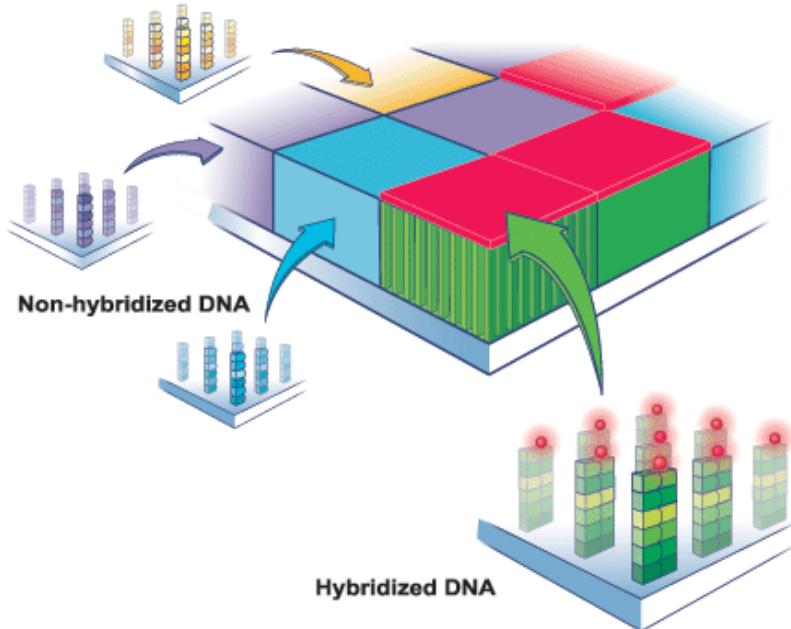
1. Micro Array Features



2. Hybridization (Pairing)

Microarray analysis

Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow



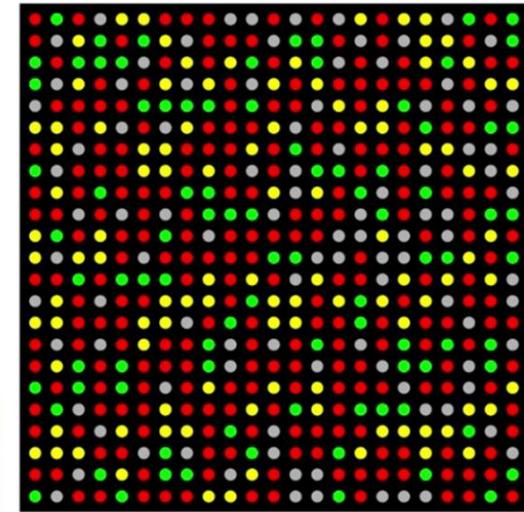
cDNAs from tissue 1 were labeled red

cDNAs from tissue 2 were labeled green

red spot means gene is expressed in tissue 1

green spot means gene is expressed in tissue 2

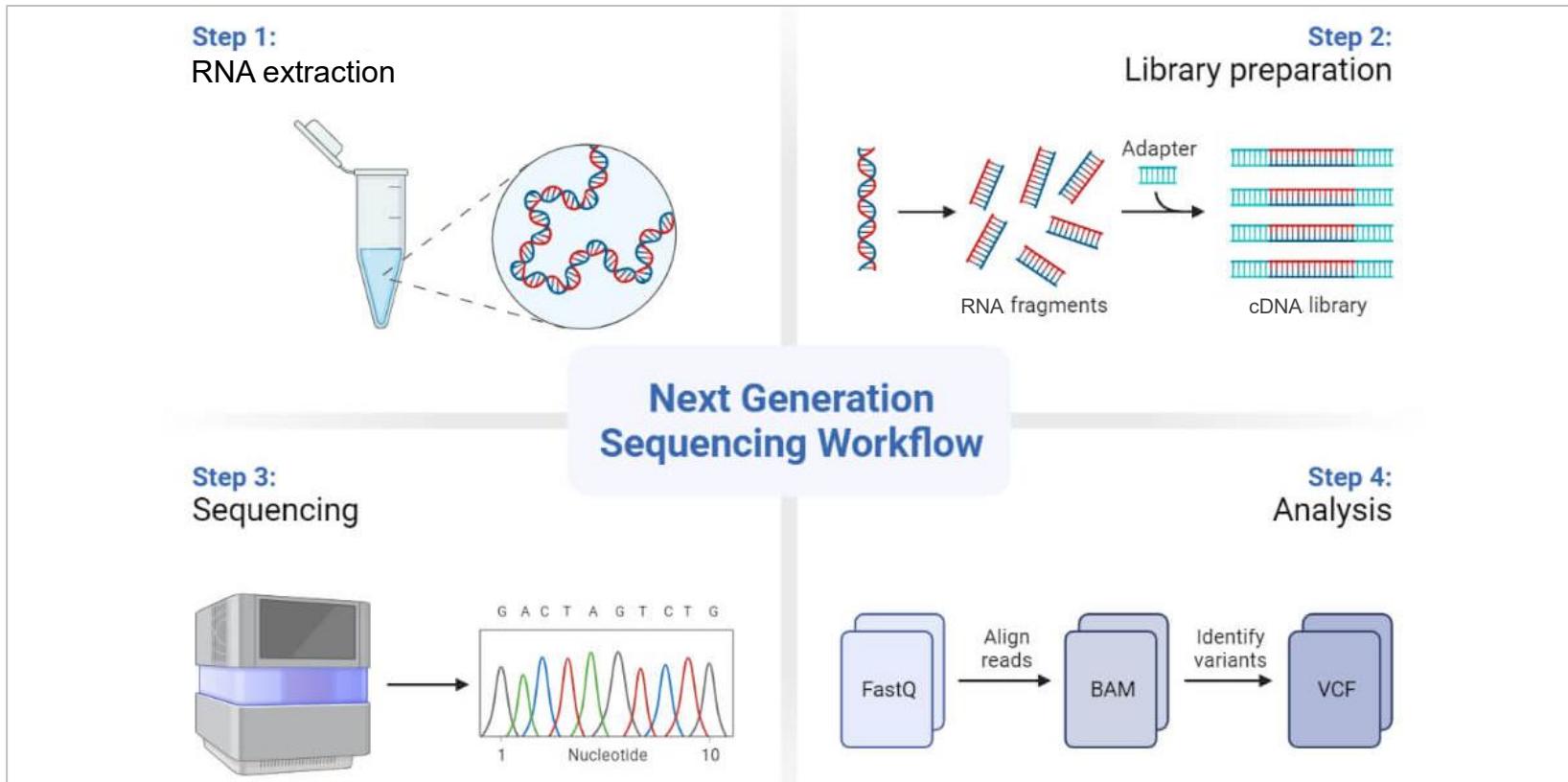
yellow spot means both cDNAs bind and gene is expressed in both tissues



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
ribosomal RNA

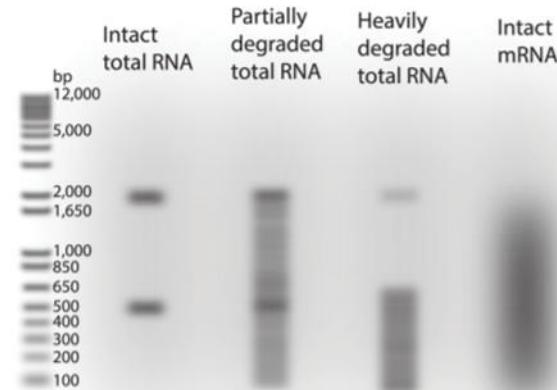
Most of the RNA in the cell

Total RNA

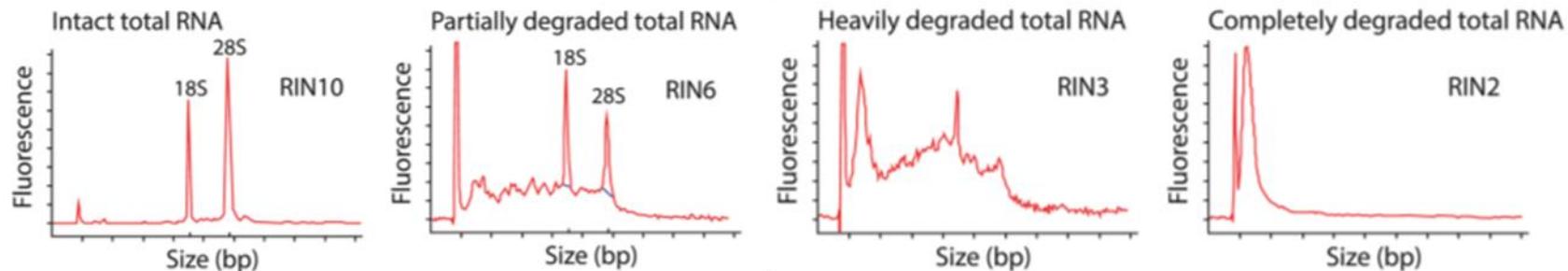
Enrichment

polyA capture
ribominus

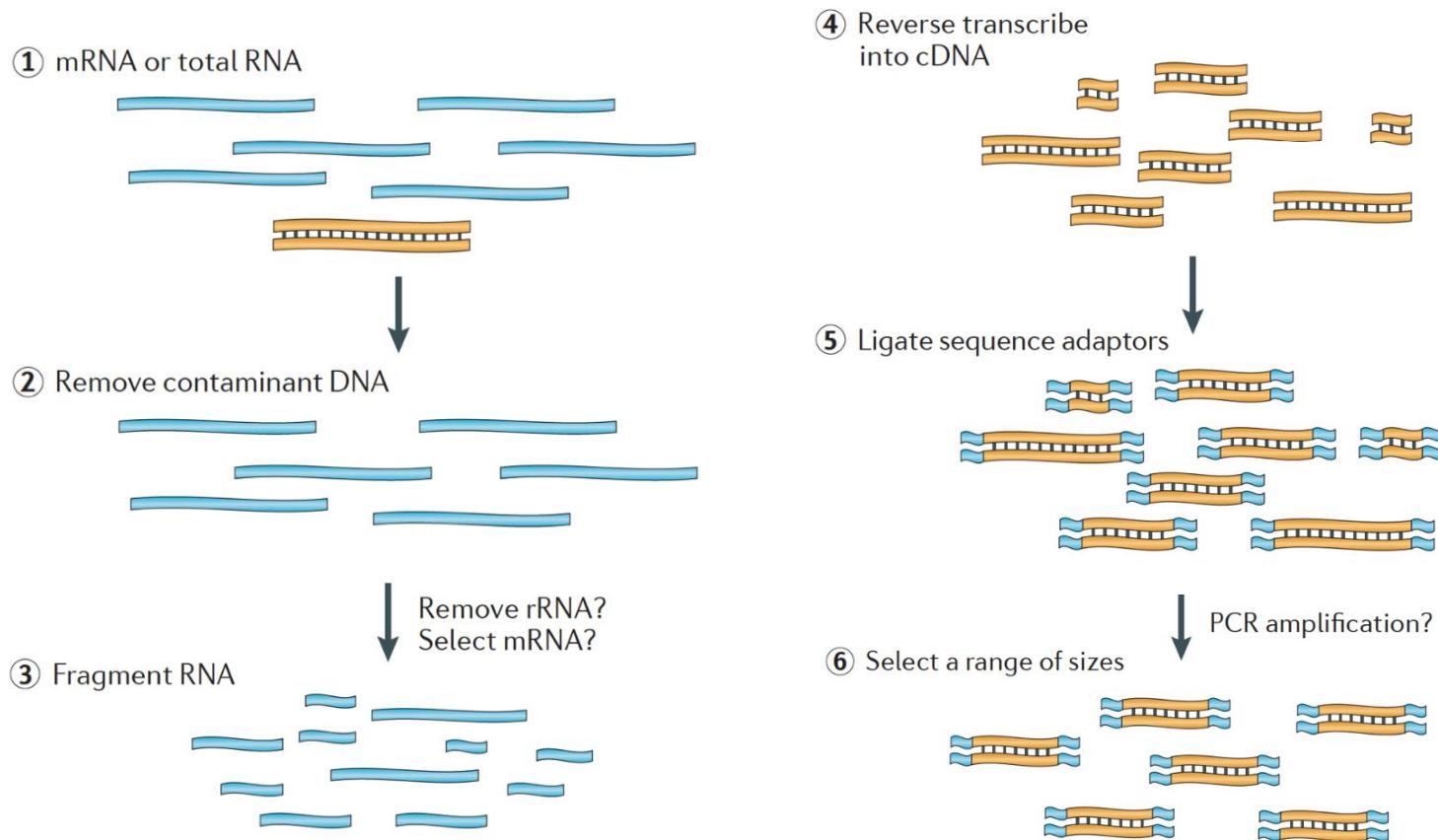
(a) Gel electropherogram



(b) Capillary electropherogram

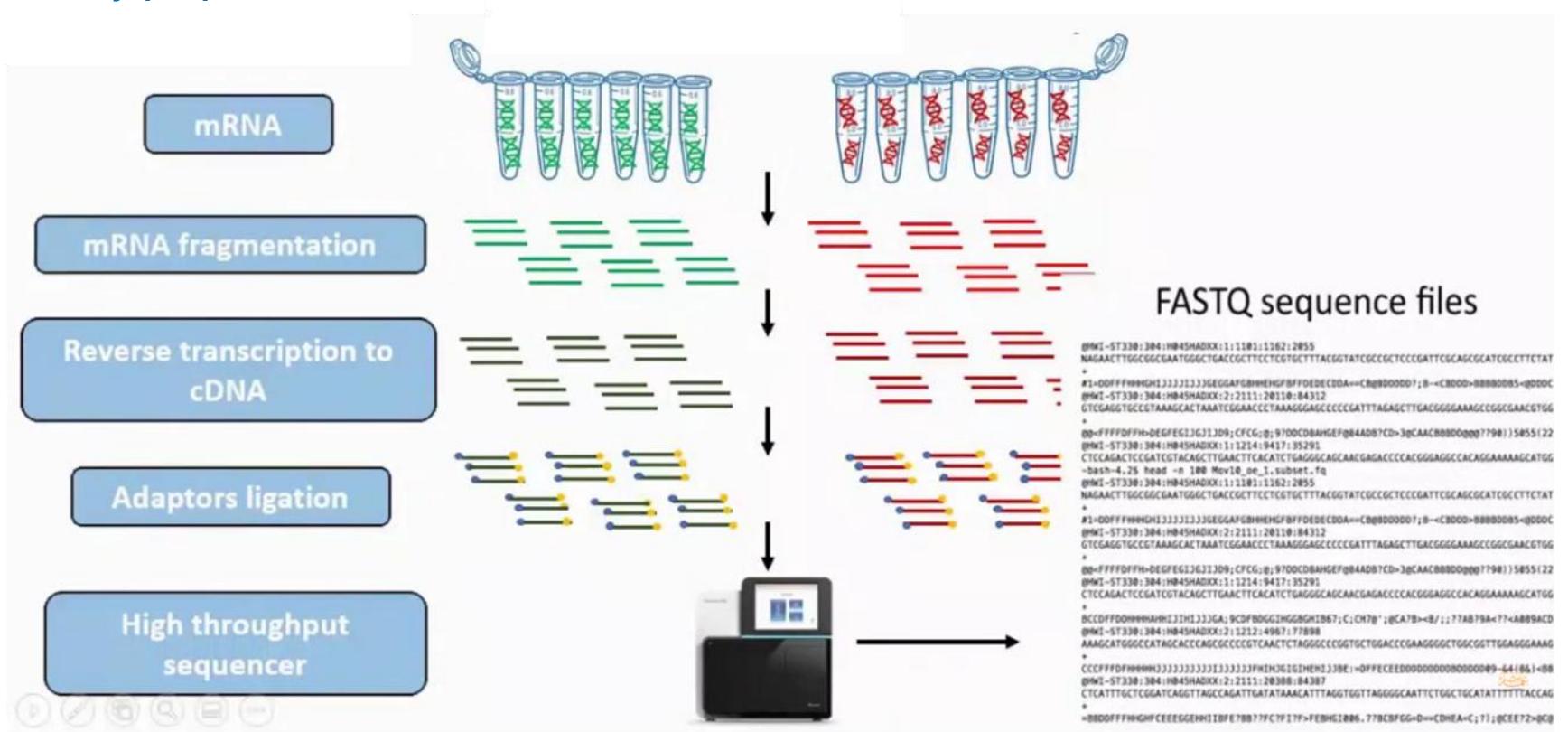


Library Preparation

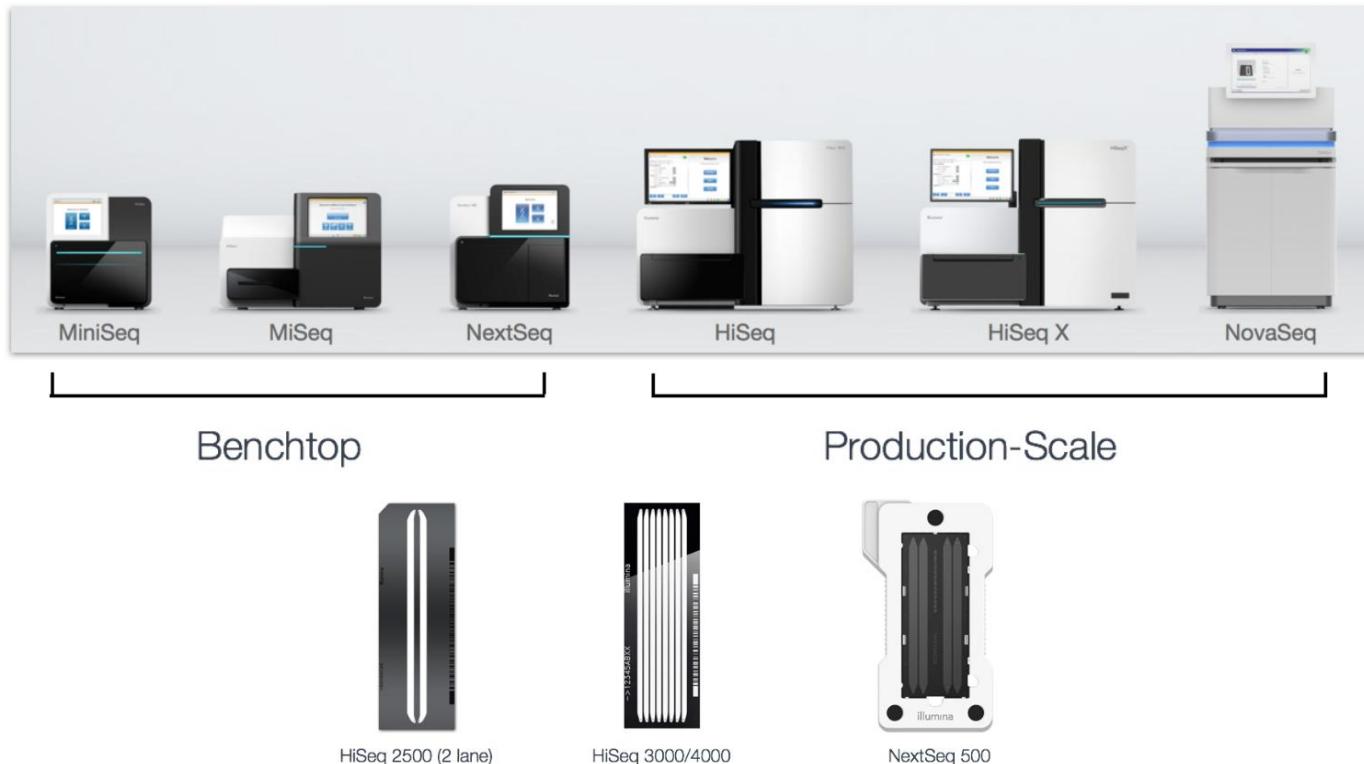


RNA-seq experiment workflow

Library preparation

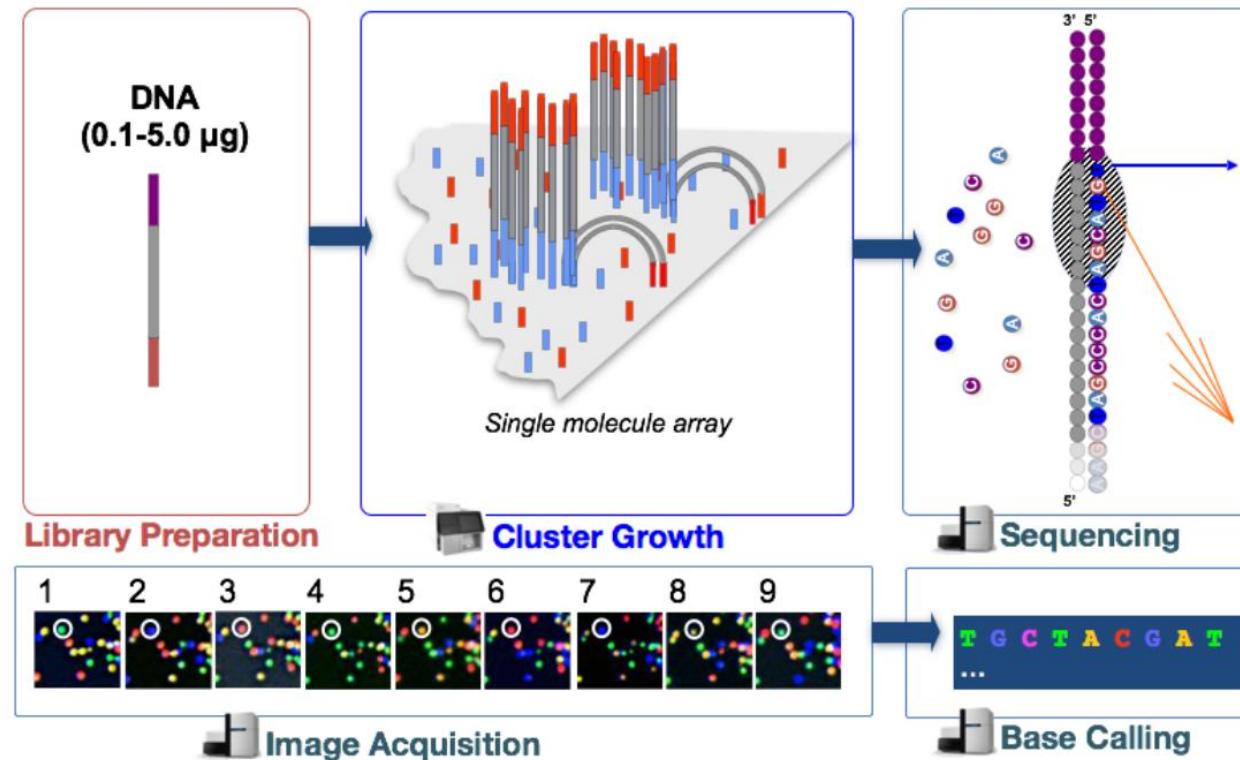


Illumina sequencing platforms



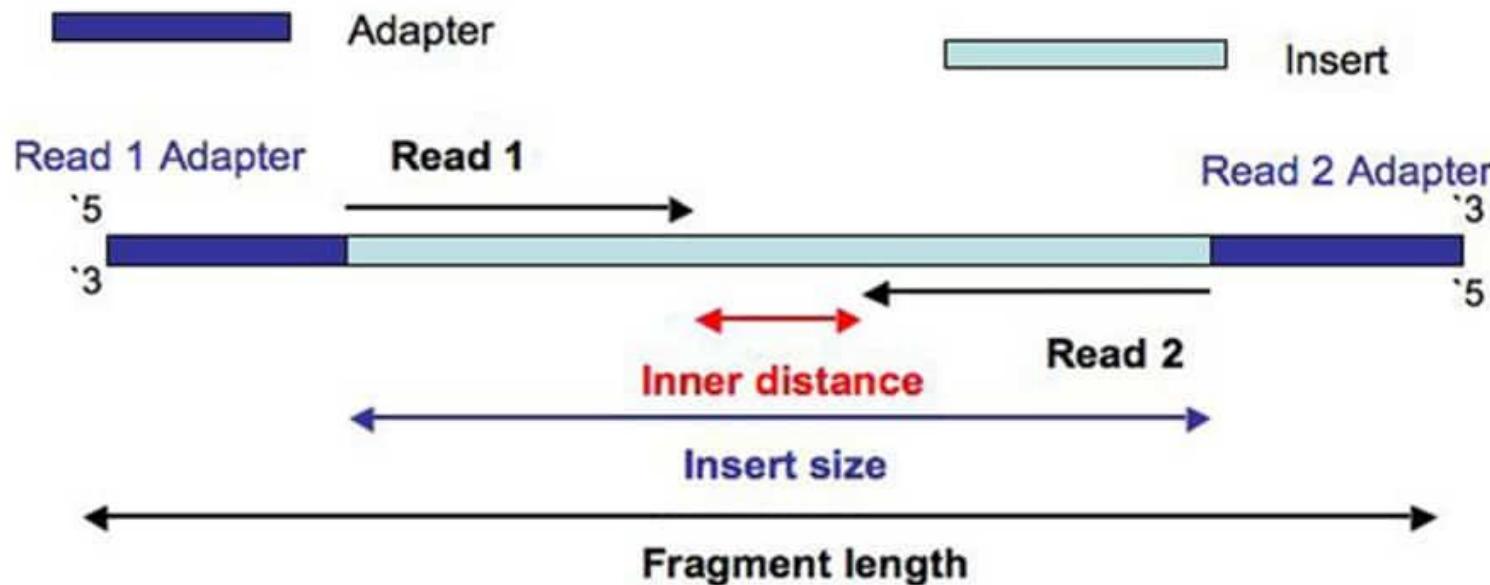
Other sequencing platforms: Pacific Bioscience, Oxford Nanopore, 10X Genomics

Sequencing by synthesis

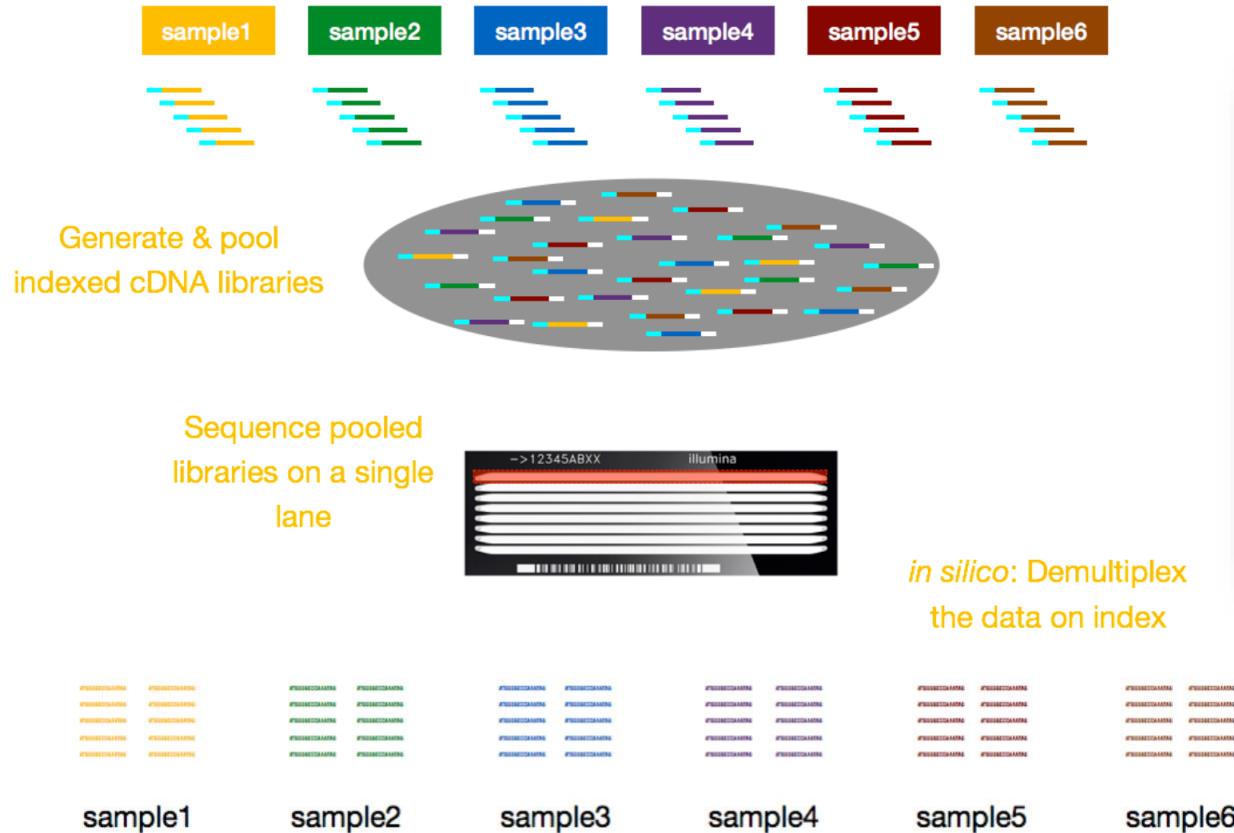


Illumina sequencing video URL: <https://www.youtube.com/watch?v=womKfikWlxM>

Single- and Paired-end sequencing



Multiplexing



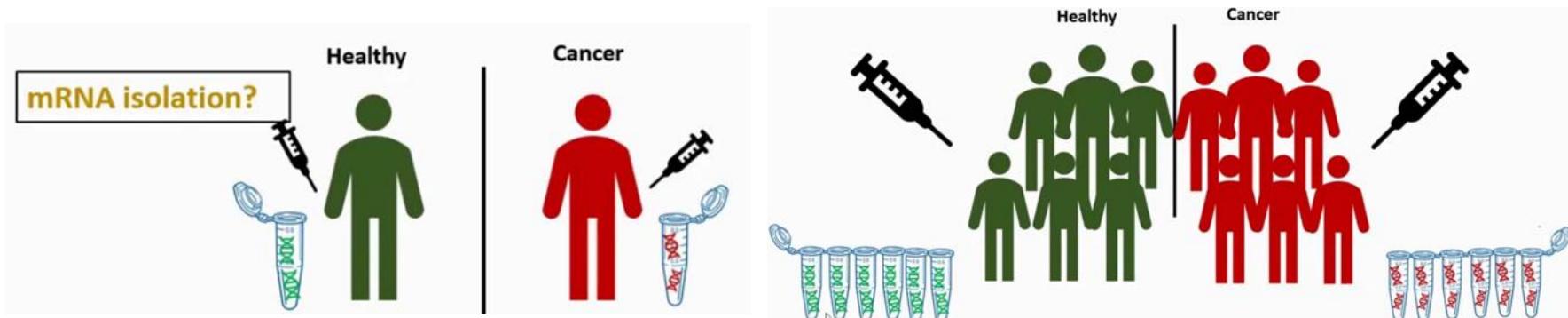
in silico: Demultiplex the data on index

Microarray vs. RNA-seq

Microarray	RNA-seq
<ul style="list-style-type: none">Limited probe-set based on prior knowledge of the transcriptome	<ul style="list-style-type: none">Comprehensive overview of the transcriptome
<ul style="list-style-type: none">Higher throughput	<ul style="list-style-type: none">Best dynamic range
<ul style="list-style-type: none">Analysis is more user-friendly than RNA-seq currently	<ul style="list-style-type: none">Gene fusion, isoform, SNPs detection

RNA-seq experiment workflow

Sample preparation

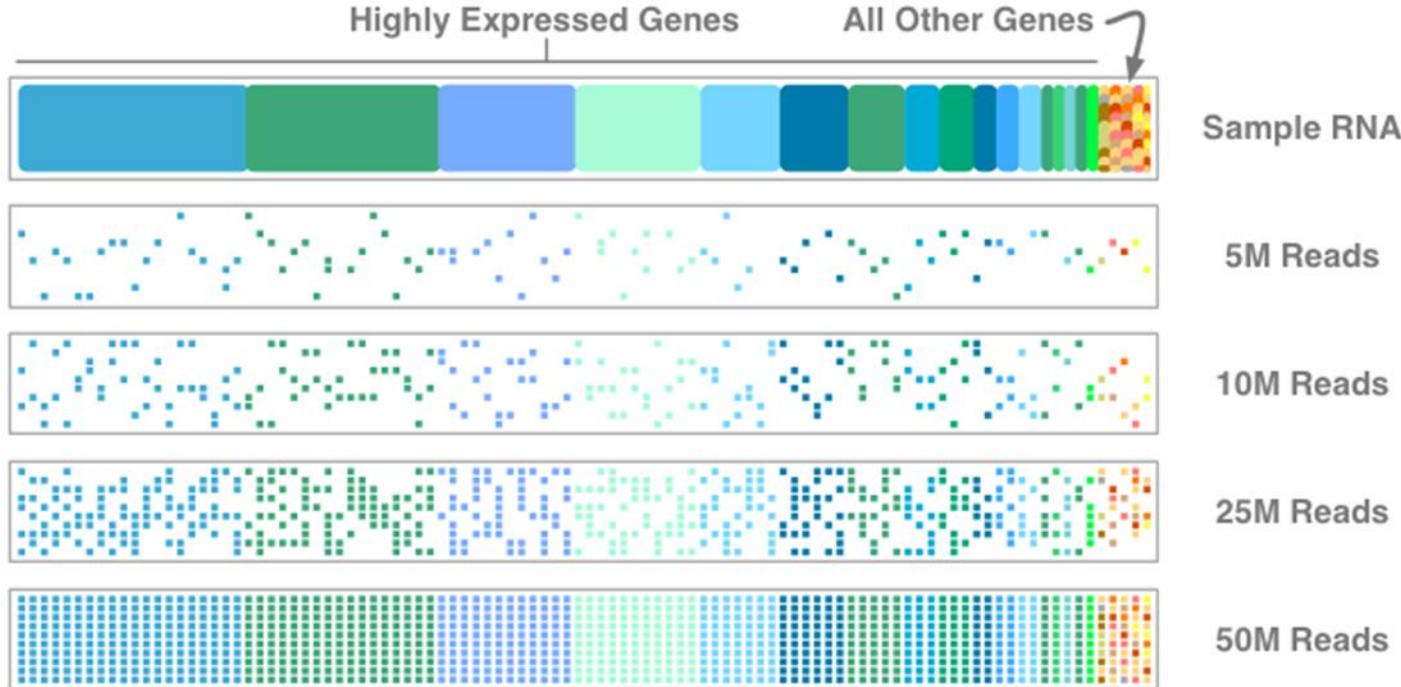


1. *Biological replicates* : Include multiple sampling within the population
2. *Technical replicates* : Include multiple preparation and re-sequencing of the same sample

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

NCBI Resources ▾ How To ▾ Sign in to NCBI

SRA SRA Advanced Search Help

SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started

- [How to Submit](#)
- [How to search and download](#)
- [How to use SRA in the cloud](#)
- [Submit to SRA](#)

Tools and Software

- [Download SRA Toolkit](#)
- [SRA Toolkit Documentation](#)
- [SRA-BLAST](#)
- [SRA Run Browser](#)
- [SRA Run Selector](#)

Related Resources

- [Submission Portal](#)
- [Trace Archive](#)
- [dbGaP Home](#)
- [BioProject](#)
- [BioSample](#)

Sources for RNA-seq datasets

SRA

GTEx Portal

Home Datasets Expression QTLs & Browsers Sample Data Documentation

Search Gene or SNP ID... Sign In

2020-11-20
NHGRI AnVIL Cloud Platform Now Supports Free Export of GTEx Data
One of the most widely-used resources for studying the relationship between genetic variation and gene expression is the Genotype-Tissue Expression (GTEx) project.

Resource Overview

Current Release (V8)

- Tissue & Sample Statistics
- Tissue Sampling Info (Anatomogram)
- Access & Download Data
- Release History
- How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA Seq. Remaining samples are available.

Explore GTEx

Browse

- By gene ID
- By variant or rs ID
- By Tissue
- Histology Image Viewer

Expression

- Multi-Gene Query
- Top 50 Expressed Genes
- Transcript Browser

Sources for RNA-seq datasets

NCBI Resources How To Sign in to NCBI

SRA GTEx Portal

About GTEX Publications Access Biospecimens FAQs Contact

Home EMBL-EBI Services Research Training About us EMBL-EBI Hinxton

ArrayExpress

Search Examples: E-MEXP-31, cancer, p53, Geuvadis advanced search

Contact Us Login

ArrayExpress – functional genomics data

ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

 Browse ArrayExpress

Data Content

Updated today at 02:00

- 74184 experiments
- 2510260 assays
- 60.30 TB of archived data

Latest News

1 October 2020 - **ArrayExpress is moving to BioStudies**

The European Bioinformatics Institute (EMBL-EBI) is building and maintaining the [BioStudies Database](#), a resource for encapsulating all the data associated with a biological study. One of the goals of BioStudies is to accept and archive data generated in experiments that can be characterized as "multi-omics". To streamline the

Sources for RNA-seq datasets

NCBI Resources How To Sign in to NCBI

SRA GTEx Portal

About GTEX Publications Access Biospecimens FAQs Contact

Home EMBL-EBI Services Research Training About us EMBL-EBI Hinxton

ENCODE Data Encyclopedia Materials & Methods Help New Search... Sign in / Create account

ENCODE: Encyclopedia of DNA Elements

3D Chromatin Structure Chromatin Accessibility Chromatin Interactions Methylome Chromatin Modification Transcriptome RNA Binding

Hypersensitive Sites CH₃ CH₃ CO RNA polymerase RNA Binding

Genes

About ENCODE Project Getting Started Experiments

Search ENCODE portal ?

ENCODE Q Functional Characterization Experiments

About ENCODE Encyclopedia candidate Cis-Regulatory Elements

Search for candidate Cis-Regulatory Elements ? Hosted by SCREEN

Sources for RNA-seq datasets

The screenshot shows the homepage of the Genomic Data Commons Data Portal (GDC Data Portal). The header includes the NIH logo, the GDC Data Portal name, and links for Home, Projects, Exploration, Analysis, and Repository. A search bar and a 'Sign in / Create account' button are also present. The main content area features a large image of a human body with internal organs highlighted in various colors. To the right is a bar chart titled 'Cases by Major Primary Site' showing the number of cases for different cancer types. Below the chart is a summary section with counts for Projects, Primary Sites, Cases, Files, Genes, and Mutations.

NIH NATIONAL CANCER INSTITUTE
GDC Data Portal

Home Projects Exploration Analysis Repository

Search... Sign in / Create account

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

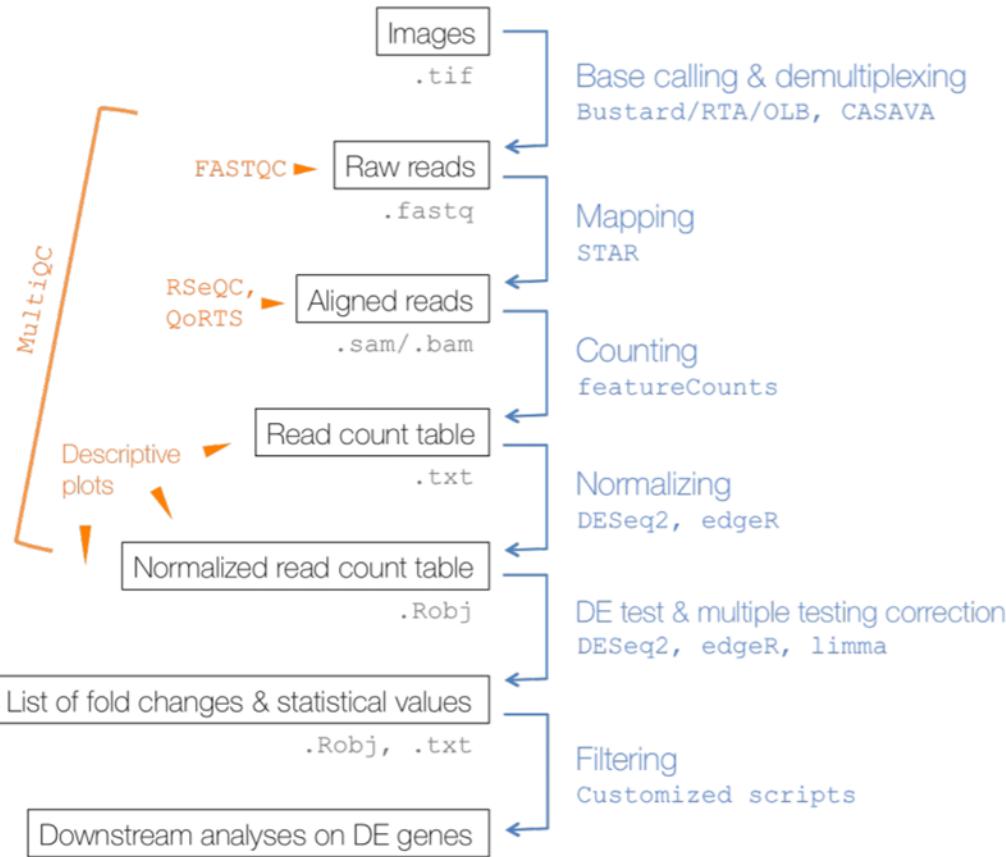
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary Data Release 29.0 - March 31, 2021

Major Primary Site	Cases
Adrenal Gland	1
Bile Duct	1
Bladder	1
Bone	1
Bone Marrow	1
Brain	1
Breast	1
Cervix	1
Colorectal	1
Esophagus	1
Eye	1
Head and Neck	1
Kidney	1
Liver	1
Lung	1
Lymph Nodes	1
Nervous System	1
Ovary	1
Pancreas	1
Pleura	1
Prostate	1
Skin	1
Soft Tissue	1
Stomach	1
Testis	1
Thymus	1
Thyroid	1
Uterus	1

PROJECTS: 68 FILES: 84,609 PRIMARY SITES: 67 GENES: 84,609 CASES: 84,609 MUTATIONS: 84,609

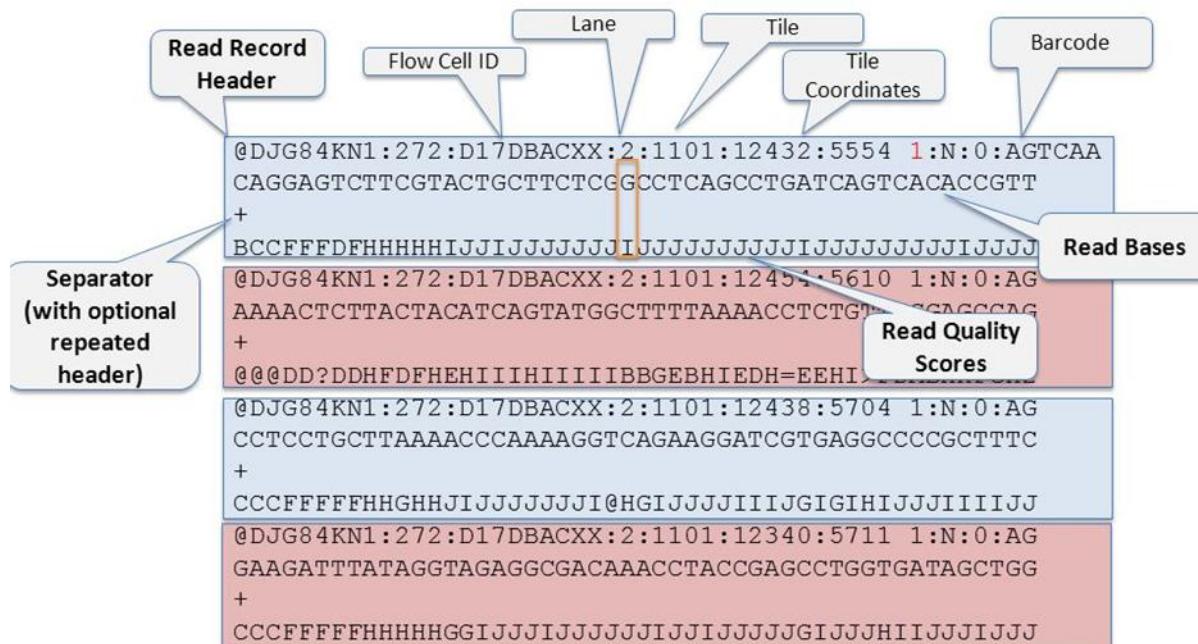
Workflow of differential gene expression analysis



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

