

# Biomarkers for disease identification/outcome

---



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

**Dr. Jaspreet Kaur Dhanjal**

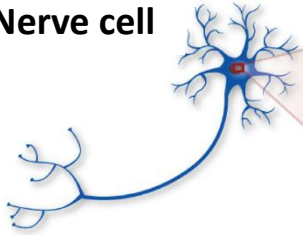
**Assistant Professor, Department of Computational Biology**

**Email ID: [jaspreet@iiitd.ac.in](mailto:jaspreet@iiitd.ac.in)**

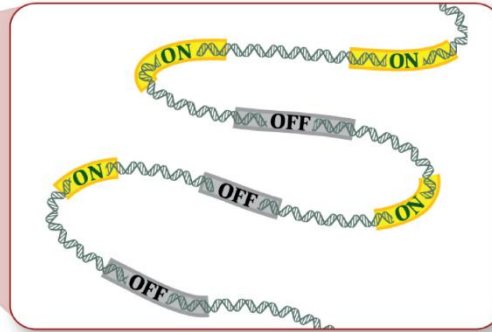
*October 14, 2025*

# Why Transcriptome?

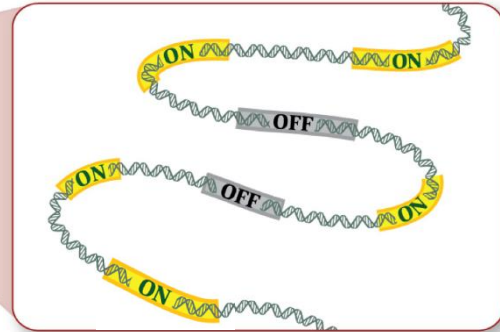
Nerve cell



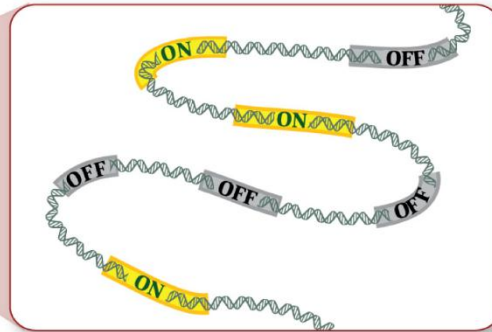
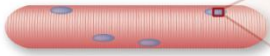
Normal



Diseased



Muscle cell



# Differential gene expression analysis

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

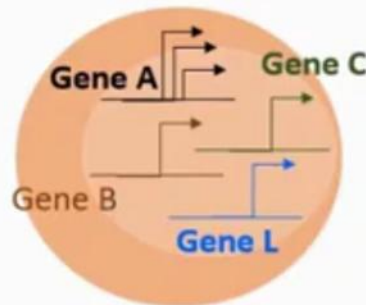
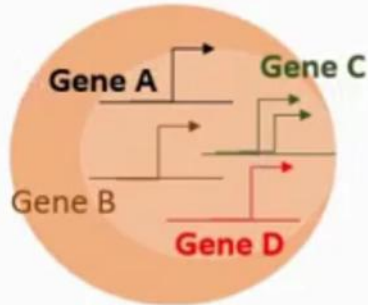


Normal cell

VS



Tumor 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*

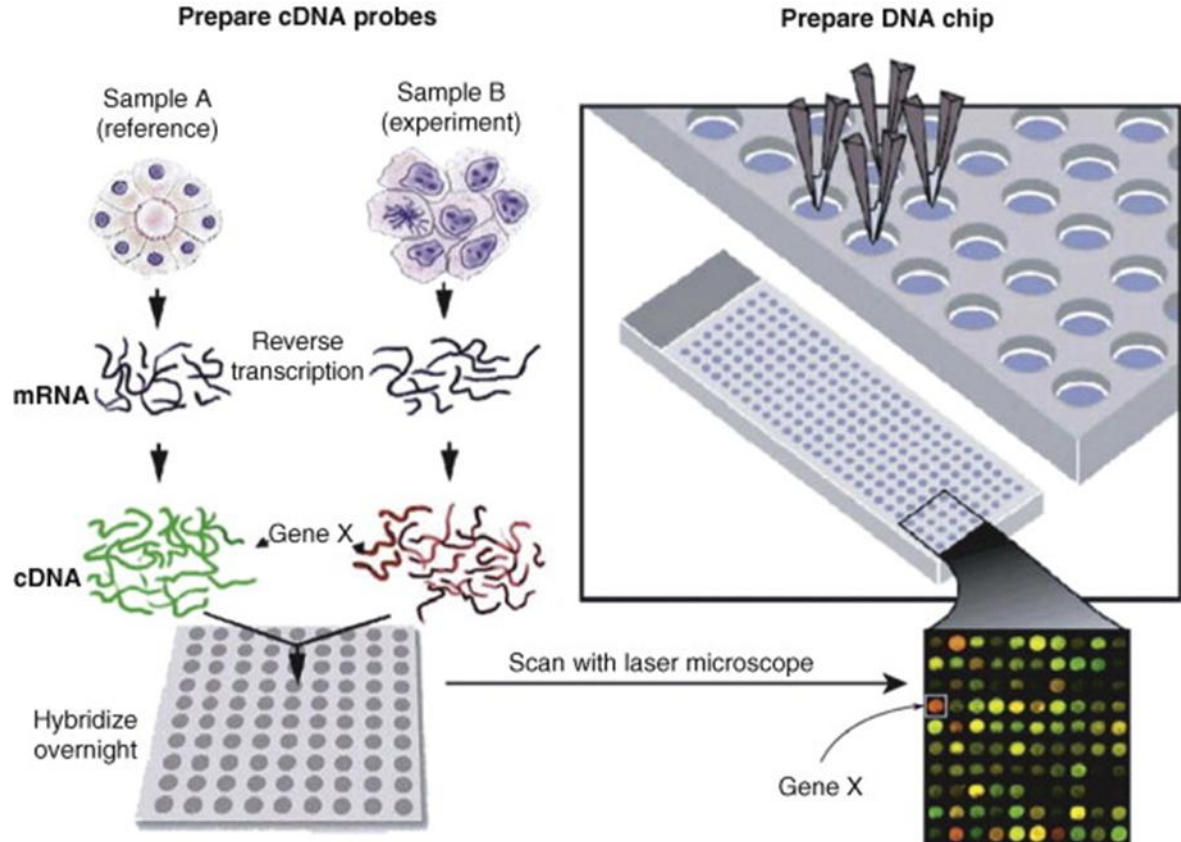
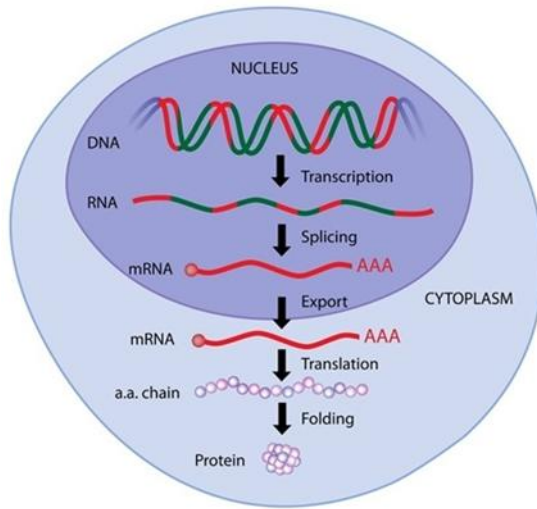
## 3. *Some molecular features can only be observed at the RNA level*

Alternative isoforms, fusion transcripts, RNA editing

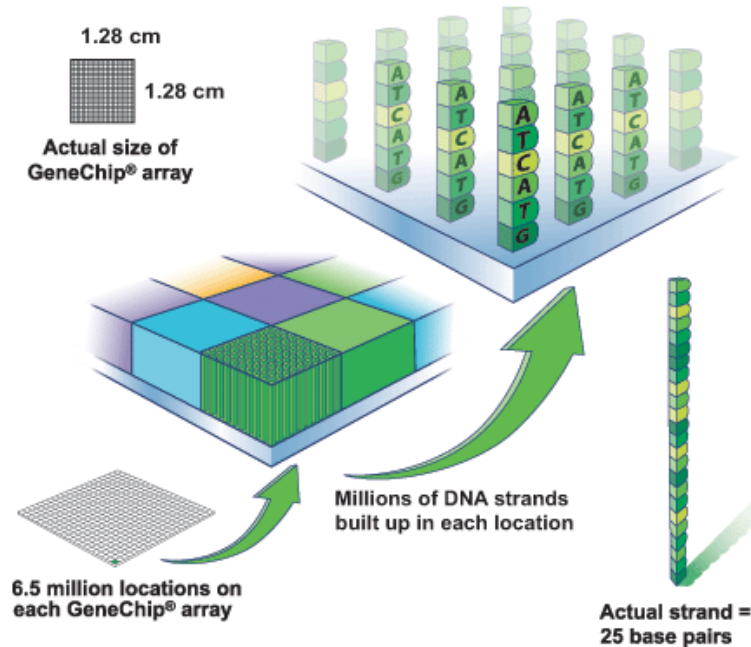
## 4. *Understand allele specific expression*

# Gene expression-based biomarker identification

## DNA Microarray

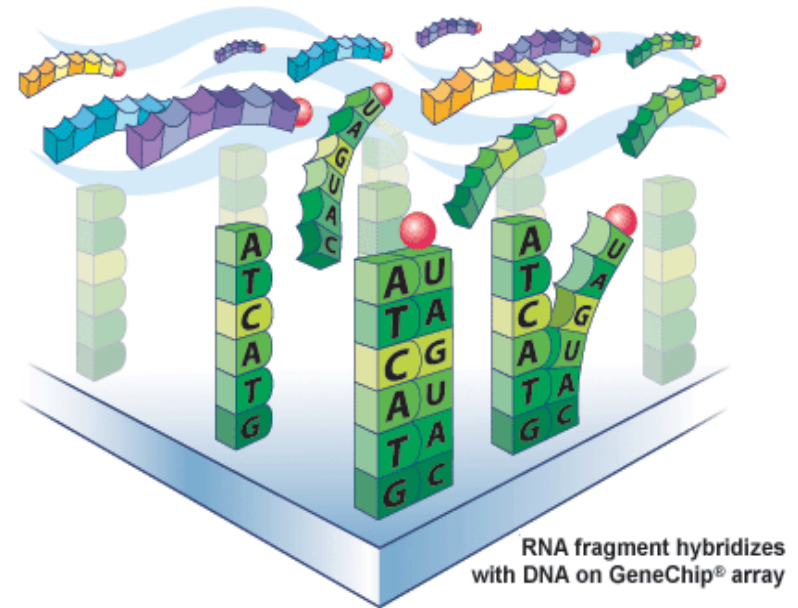


# Microarray analysis



## 1. Micro Array Features

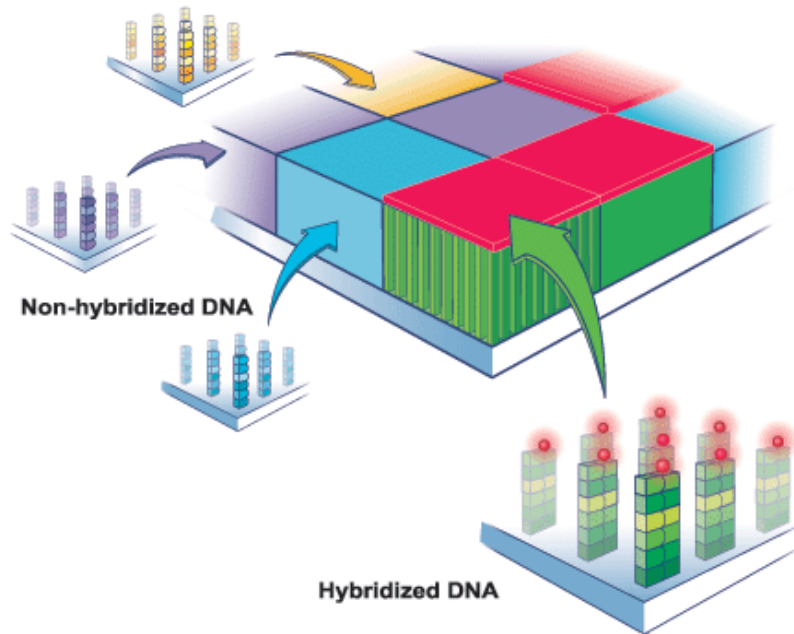
RNA fragments with fluorescent tags from sample to be tested



## 2. Hybridization (Pairing)

# Microarray analysis

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



### 3. Detection

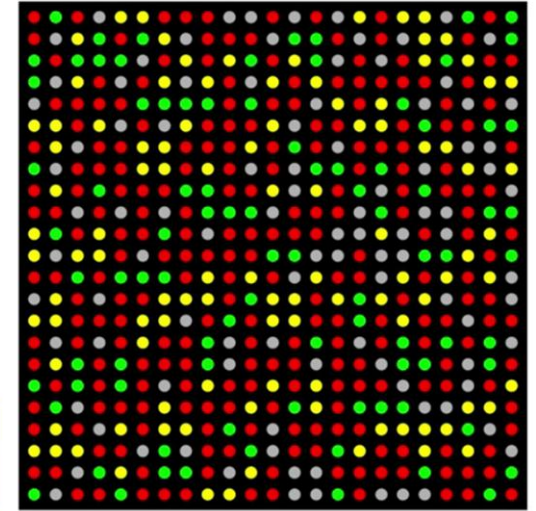
cDNAs from tissue 1  
were labeled red

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

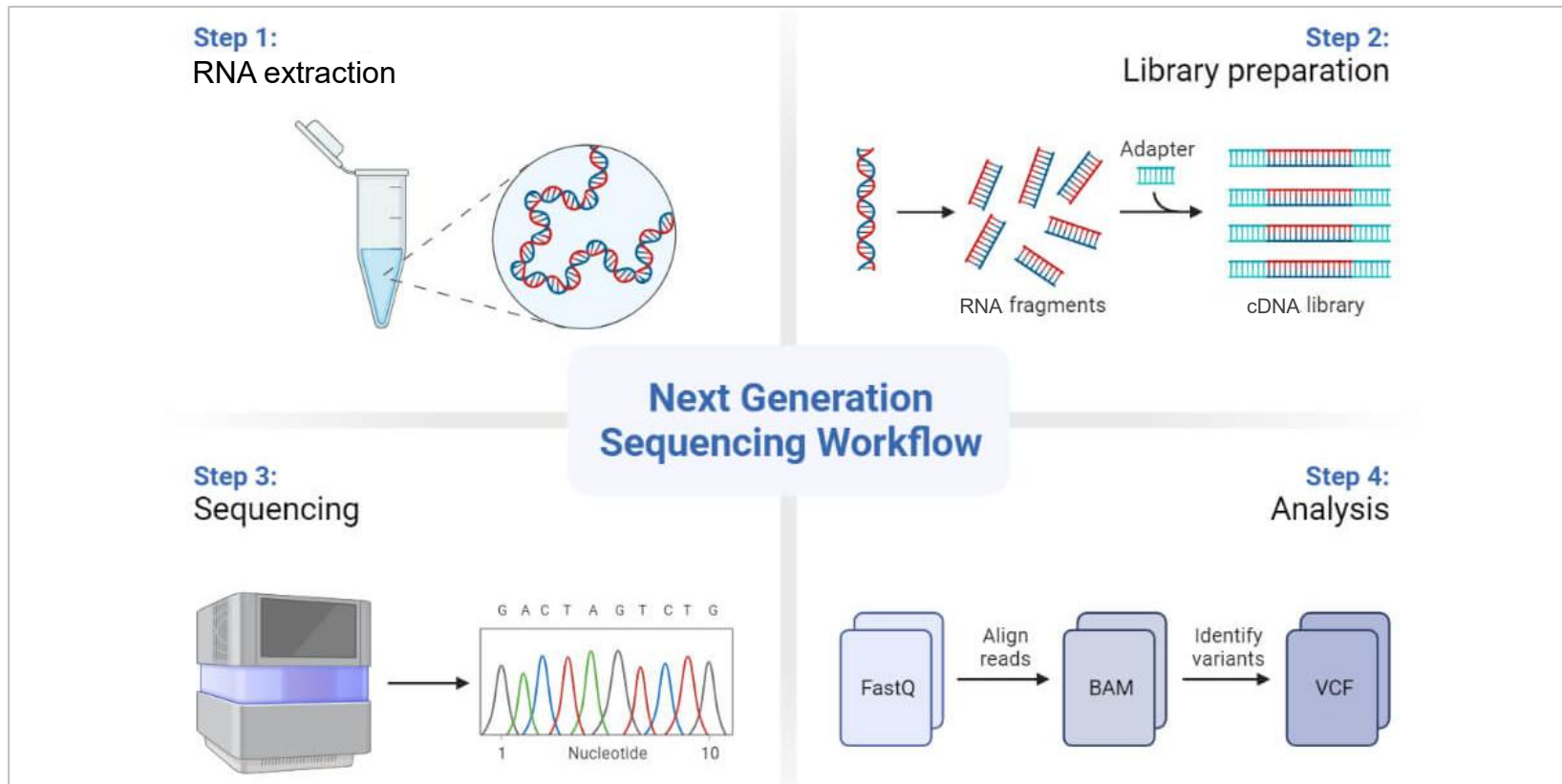


### Limitations:

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.



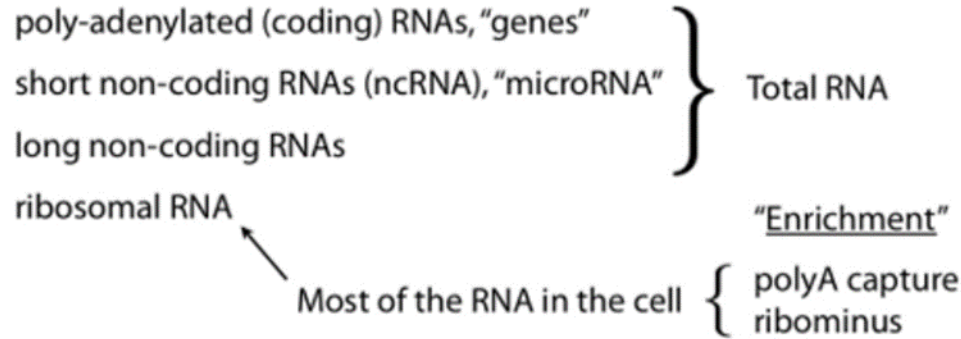
# Next generation RNA-sequencing



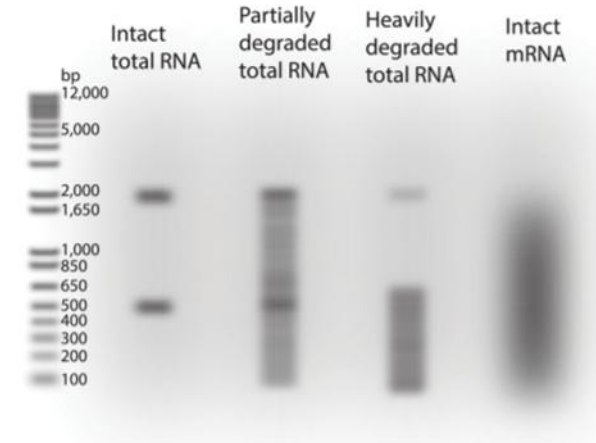


# RNA-seq experiment workflow

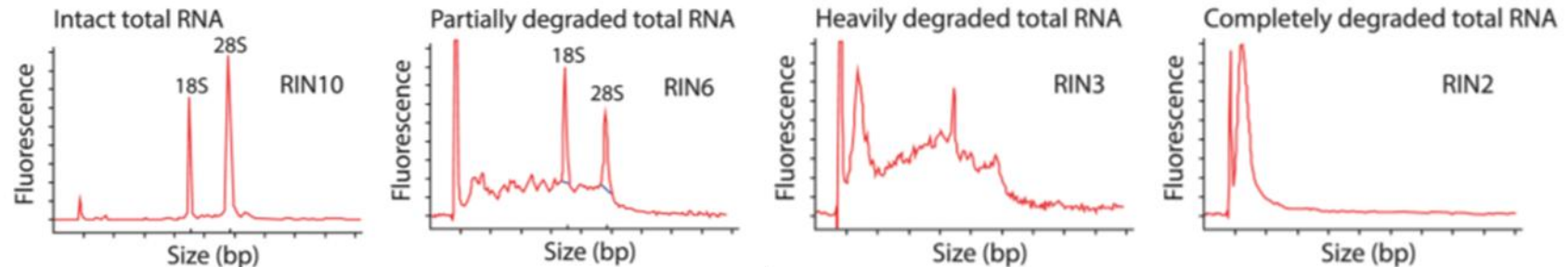
## Types of RNA



(a) Gel electropherogram

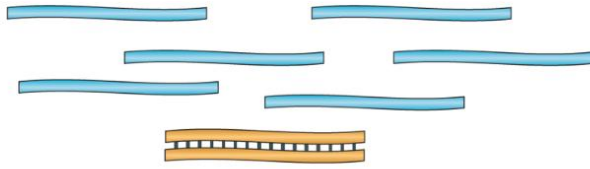


(b) Capillary electropherogram

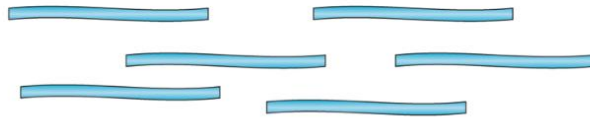


# Library Preparation

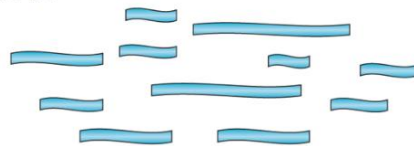
① mRNA or total RNA



② Remove contaminant DNA

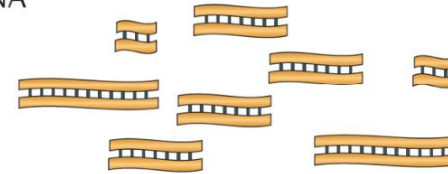


③ Fragment RNA

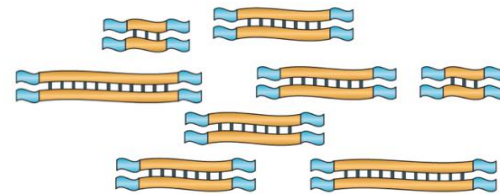


Remove rRNA?  
Select mRNA?

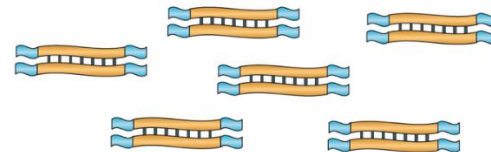
④ Reverse transcribe  
into cDNA



⑤ Ligate sequence adaptors



⑥ Select a range of sizes



PCR amplification?

## Library preparation

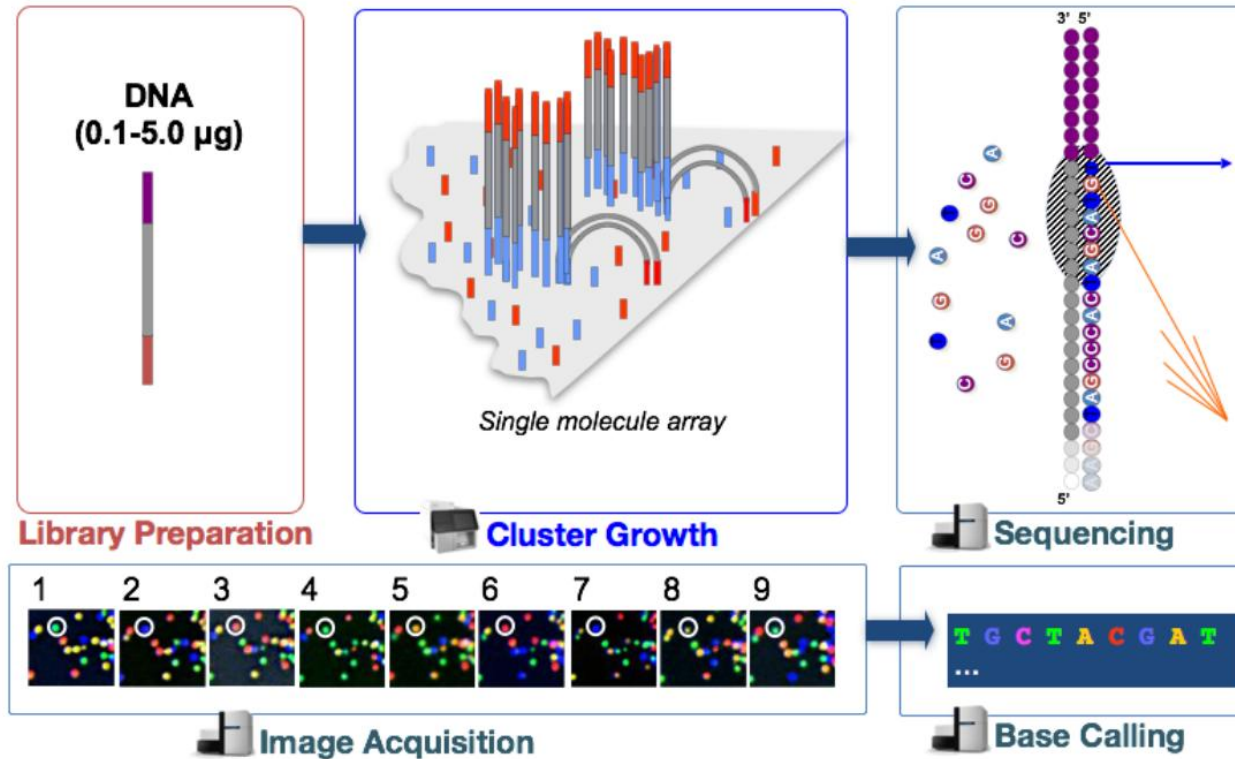


# Illumina sequencing platforms

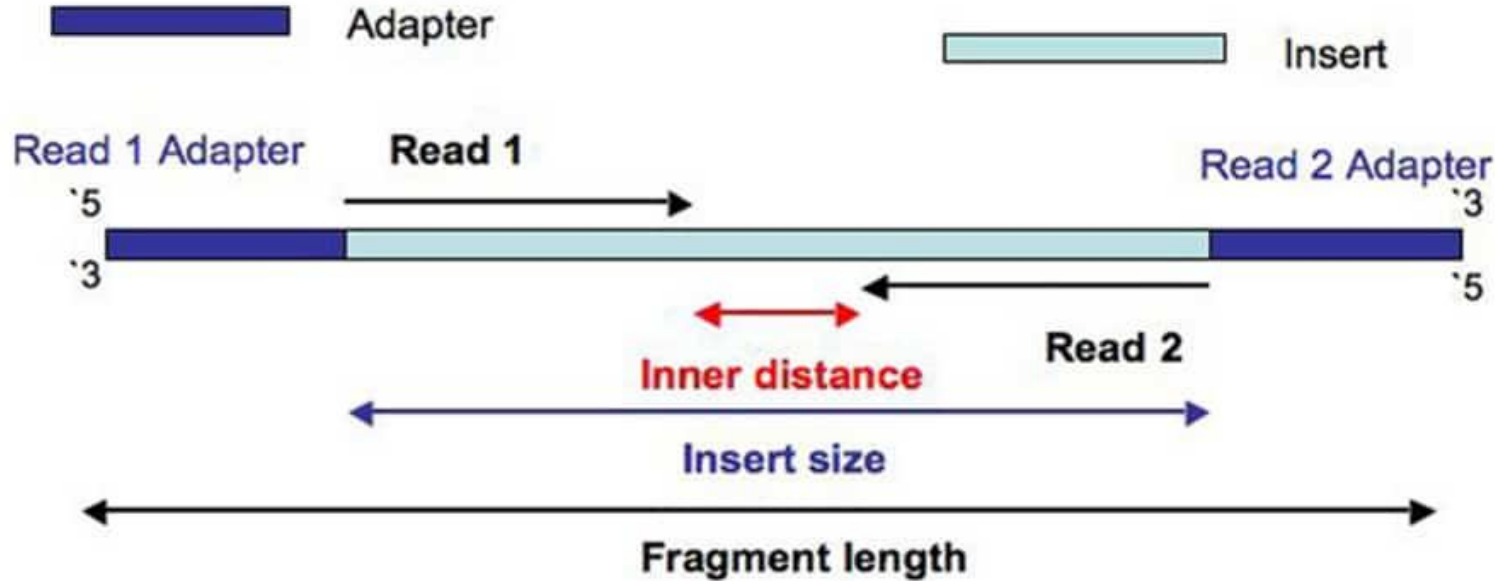


**Other sequencing platforms:** Pacific Bioscience, Oxford Nanopore, 10X Genomics

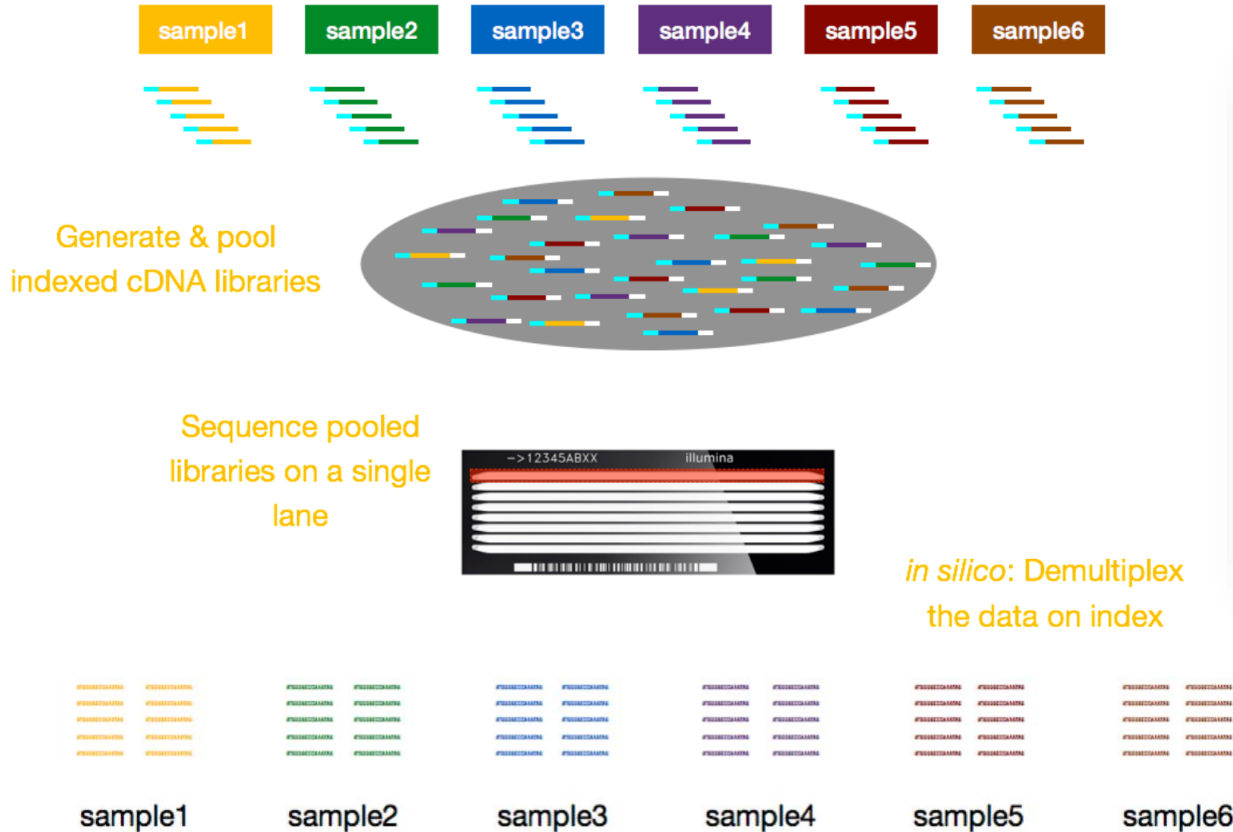
# Sequencing by synthesis



# Single- and Paired-end sequencing



# Multiplexing



## FASTQ sequence files

```
@HWI-ST338:304:HB4SHADXX:1:1181:1162:2855
NAGAACTGGCGGGAATGGCTGACCGCTTCTCGTCTTACGGATCGCCGCTCCGATTGCGAGCGATCGCTTCTAT
+
#1=DOFFFHHHGHJ3333333GEGAFGBHHEHGFBBFDEDECBDA==CB@B000007;B==CB000>BBB00085-qDDC
@HWI-ST338:304:HB4SHADXX:2:2111:20110:84312
GTCGAGGTGCCCTAAGCACTAAATCGGAACCTTAAGGGAGCCCCGATTAGAGCTTGACGGGGAAGCCGCGAAGCTGG
+
@0=FFFFDFH-DEGFEI3GJ3D9;CFCG;B;9700CDBAHGF@84AD87CD>3qCAACBBB00@7790))5855(22
@HWI-ST338:304:HB4SHADXX:1:1214:1417:35291
CTCCAGACTCCGATCGTACAGCTTGAATTCACATCTGAGGGCAGCAGAGACCCACGGGAGGCCACAGGAAAAGCATGG
-bash-4.2$ head -n 100 Mov10_ne_1_subset.fq
@HWI-ST338:304:HB4SHADXX:1:1181:1162:2855
NAGAACTGGCGGGAATGGCTGACCGCTTCTCGTCTTACGGATCGCCGCTCCGATTGCGAGCGATCGCTTCTAT
+
#1=DOFFFHHHGHJ3333333GEGAFGBHHEHGFBBFDEDECBDA==CB@B000007;B==CB000>BBB00085-qDDC
@HWI-ST338:304:HB4SHADXX:2:2111:20110:84312
GTCGAGGTGCCCTAAGCACTAAATCGGAACCTTAAGGGAGCCCCGATTAGAGCTTGACGGGGAAGCCGCGAAGCTGG
+
@0=FFFFDFH-DEGFEI3GJ3D9;CFCG;B;9700CDBAHGF@84AD87CD>3qCAACBBB00@7790))5855(22
@HWI-ST338:304:HB4SHADXX:1:1214:1417:35291
CTCCAGACTCCGATCGTACAGCTTGAATTCACATCTGAGGGCAGCAGAGACCCACGGGAGGCCACAGGAAAAGCATGG
+
BCCDFDHHHHHGHJ3333333GA;9CDFBGGHGGHGH167;C;CH7B";@CA7B>=B/;;77AB79A<77<AB99ACD
@HWI-ST338:304:HB4SHADXX:2:2122:4967:77898
AAGCATGGGCGATAGCAGCCAGCGCCGCTCAACTAGGGCCGGTGGACCCGAGGGGCTGGCGGTTGGAGGGAAG
+
CCCFDFDHHHHGHJ3333333J3333333FHGHGHGH1338;=OFFECEED0000000000000009-64(86)<@B
@HWI-ST338:304:HB4SHADXX:2:2111:20388:84387
CTCATTTGCTCGATCAGGTAGCCAGATTGATATAACATTTAGTGTTTAGGGCAATTGCTGCTCATATTTTATACAG
+
-BBDOFFFHHGHFCEEGGHH1BFE78877FC7F17F-FEBHG1086.778CBFGG=D==CDEA+C;71;@CEE72>@C
```

*in silico*: Demultiplex  
the data on index



# Microarray vs. RNA-seq

---

---

## Microarray

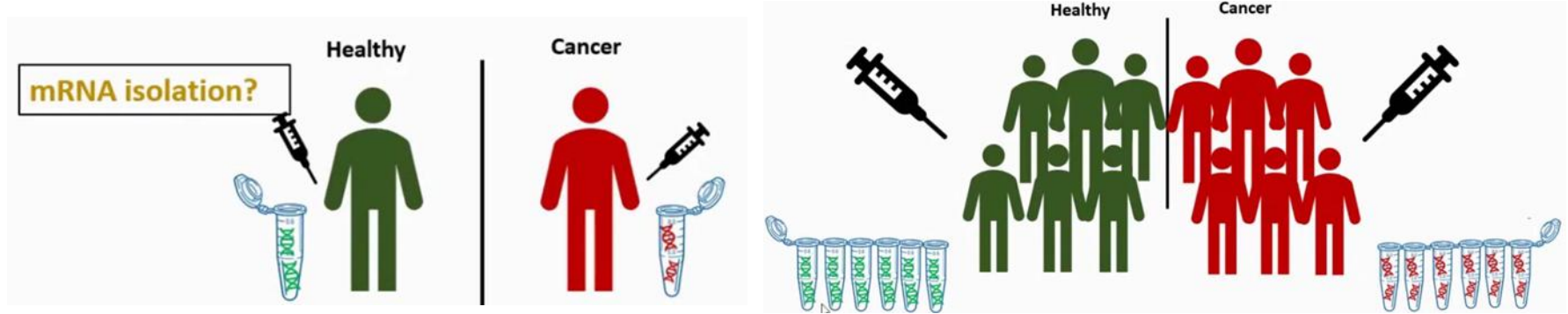
- Limited probe-set based on prior knowledge of the transcriptome
- Higher throughput
- Analysis is more user-friendly than RNA-seq currently

## RNA-seq

- Comprehensive overview of the transcriptome
- Best dynamic range
- Gene fusion, isoform, SNPs detection

# RNA-seq experiment workflow

## Sample preparation

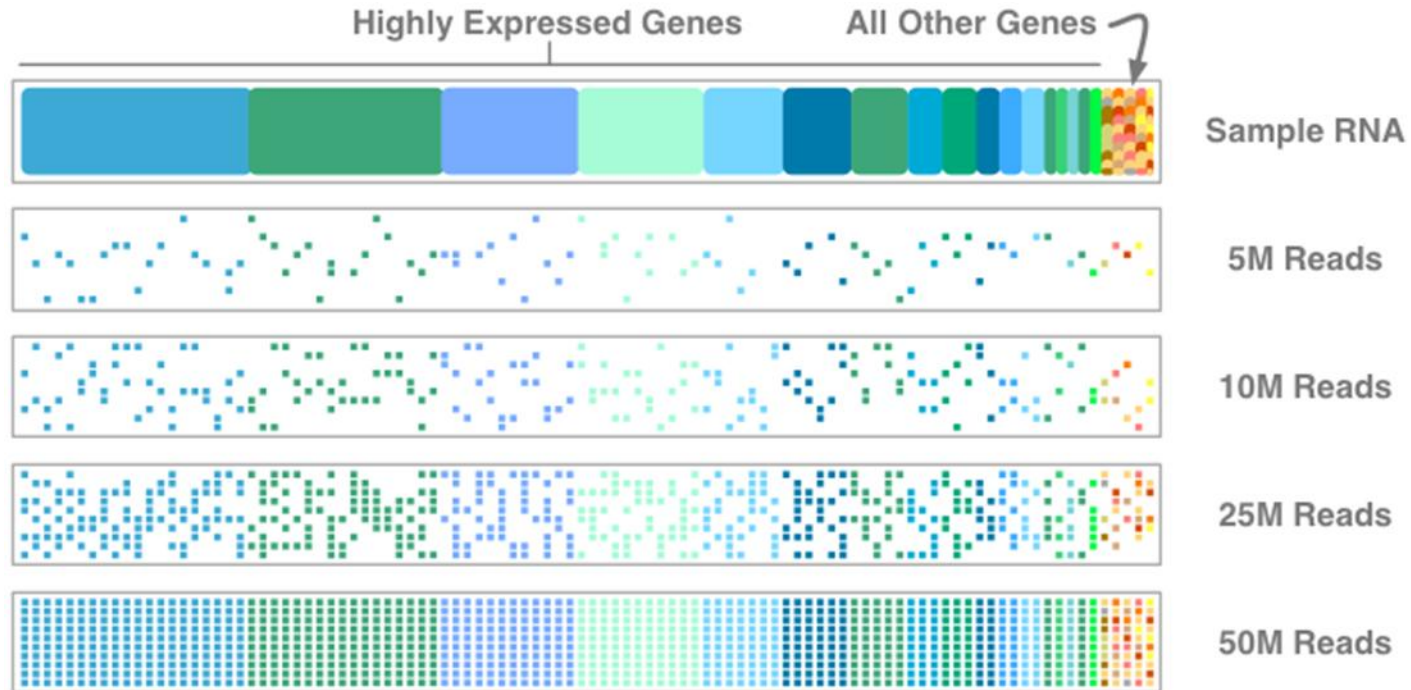


1. *Biological replicates* : Include multiple sampling within the population
2. *Technical replicates* : Include multiple preparation and re-sequencing of the same sample




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   [Advanced](#) [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

The screenshot displays the GTEx Portal website. At the top, there is a blue navigation bar with the NCBI logo, links for Resources and How To, and a Sign in to NCBI button. Below this is a dark grey header with the SRA and GTExPortal logos, and links for About GTEx, Publications, Access Biospecimens, FAQs, and Contact. A secondary navigation bar contains links for Home, Datasets, Expression, QTLs & Browser, Sample Data, and Documentation, along with a search bar and a Sign In button. A banner image shows a stylized human figure with a DNA helix, and a text box on the right mentions the NHGRI AnVIL Cloud Platform. The main content area is divided into two columns: 'Resource Overview' on the left and 'Explore GTEx' on the right. The 'Resource Overview' column includes links for Current Release (V8), Tissue & Sample Statistics, Tissue Sampling Info (Anatomogram), Access & Download Data, Release History, and How to cite GTEx?, followed by a paragraph about the GTEx project. The 'Explore GTEx' column features a 'Browse' section with options to search by gene ID, variant or rs ID, tissue, and histology images, and an 'Expression' section with links for Multi-Gene Query, Top 50 Expressed Genes, and Transcript Browser.

NCBI Resources How To Sign in to NCBI

SRA GTExPortal

About GTEx Publications Access Biospecimens FAQs Contact

Home Datasets Expression QTLs & Browser 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.

Gett...

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

Multi-Gene Query

Top 50 Expressed Genes

Transcript Browser

Expression

Browse and search all data by gene

Browse and search all data by variant

Browse and search all data by tissue

Browse and search GTEx histology images

Browse and search expression by gene and tissue

Visualize the top 50 expressed genes in each tissue

Visualize transcript expression and isoform structures

# Sources for RNA-seq datasets

The screenshot displays the ArrayExpress website interface. At the top, there are navigation links for NCBI, Resources, and How To. Below this, the GTEX Portal is visible with links for About GTEx, Publications, Access Biospecimens, FAQs, and Contact. The main navigation bar includes EMBL-EBI, Services, Research, Training, and About us. The ArrayExpress logo is prominently displayed on the left, with a search bar on the right. The search bar contains the text "Search" and "Examples: E-MEXP-31, cancer, p53, Geuvadis". Below the search bar, there are links for "advanced search", "Contact Us", and "Login". The main content area is divided into two columns. The left column features the title "ArrayExpress – functional genomics data" and a description: "ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community." Below this is a button labeled "Browse ArrayExpress". The right column is titled "Data Content" and includes the text "Updated today at 02:00" followed by a list of statistics: "74184 experiments", "2510260 assays", and "60.30 TB of archived data". At the bottom, there is a section titled "Latest News" with a date "1 October 2020" and the headline "ArrayExpress is moving to BioStudies". The text below the headline states: "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

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

Home Browse Submit Help About ArrayExpress 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 GTExPortal About GTEx Publications Access Biospecimens FAQs Contact

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

Diagram illustrating the ENCODE project's focus on understanding DNA elements. The diagram shows a DNA double helix with various features labeled: Hypersensitive Sites,  $\text{CH}_3$ ,  $\text{CH}_3\text{CO}$ , and RNA polymerase. Below the DNA, a row of boxes represents different data types: 3D Chromatin Structure, Chromatin Accessibility, Chromatin Interactions, Methylome, Chromatin Modification, Transcriptome, and RNA Binding. Arrows indicate how these data types relate to the DNA elements and genes. At the bottom, a gene structure is shown with exons in yellow and introns in red.

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

NCBI Resources How To Sign in to NCBI

SRA GTEX Portal About GTEx Publications Access Biospecimens FAQs Contact

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

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

NIH NATIONAL CANCER INSTITUTE GDC Data Portal Home Projects Exploration Analysis Repository

Harmonized Cancer Datasets  
Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

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

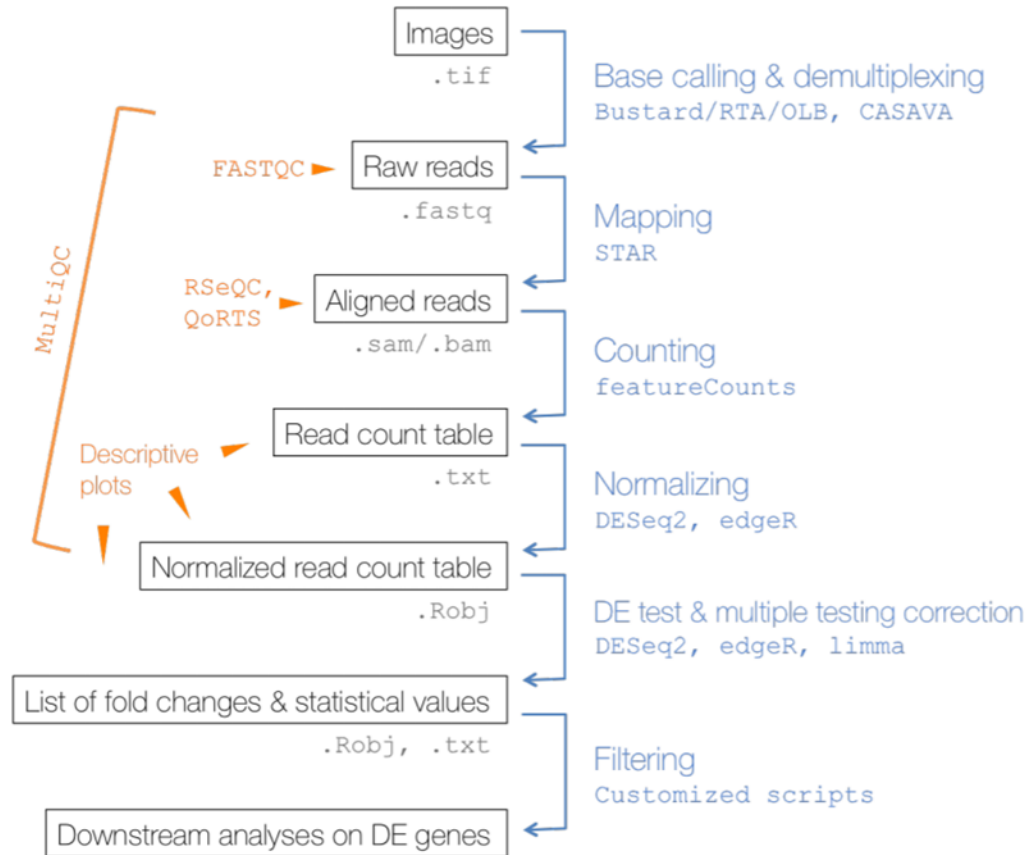
Data Portal Summary Data Release 29.0 - March 31, 2021

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

Cases by Major Primary Site

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

# Workflow of differential gene expression analysis





## Problems in sequencing

- [illegible]

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%