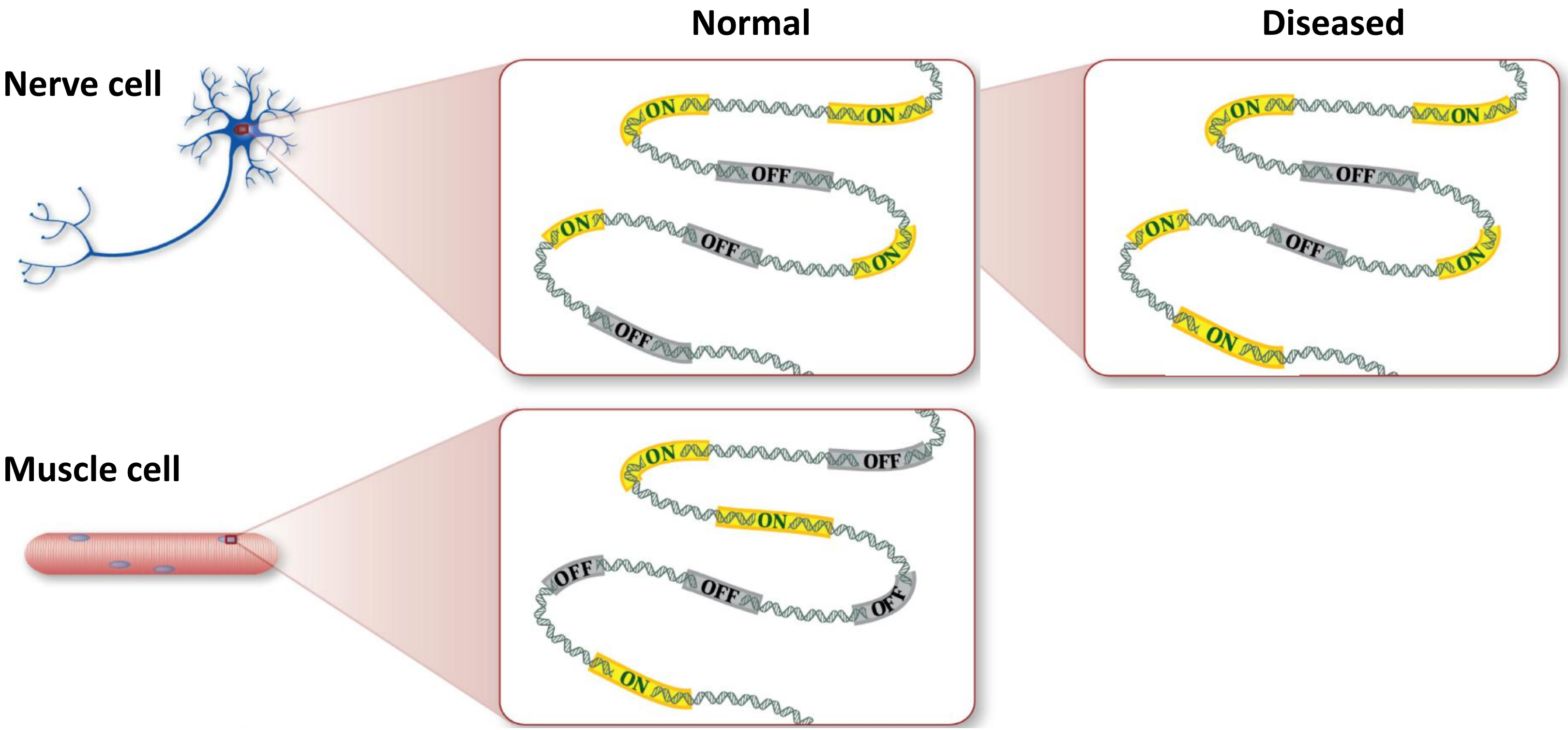
Biomarkers for disease identification/outcome

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?



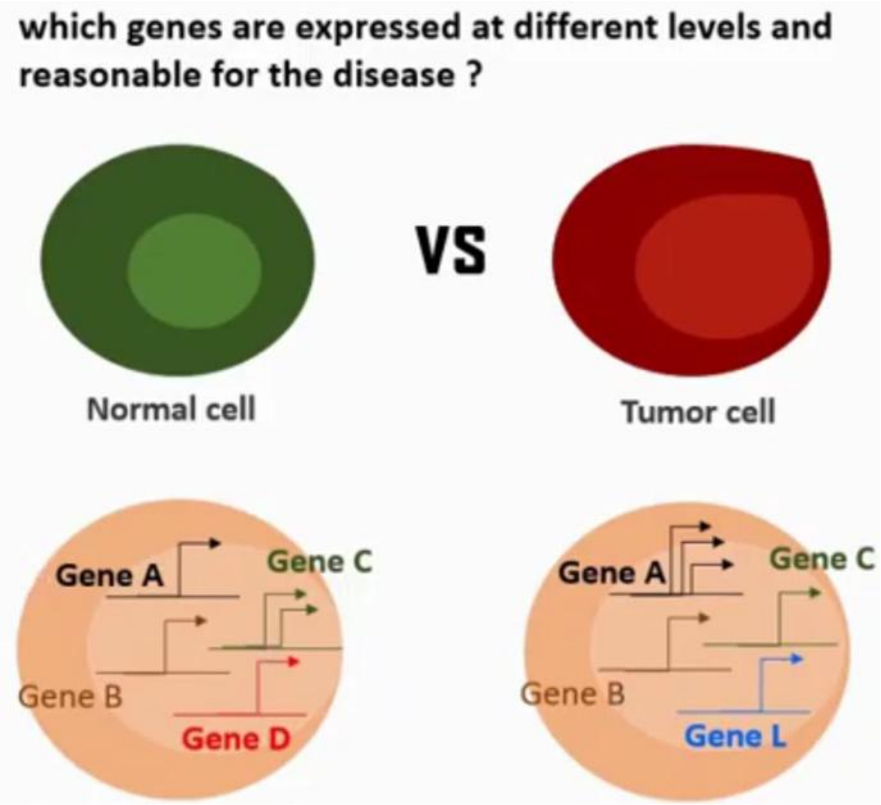
Normal

Diseased

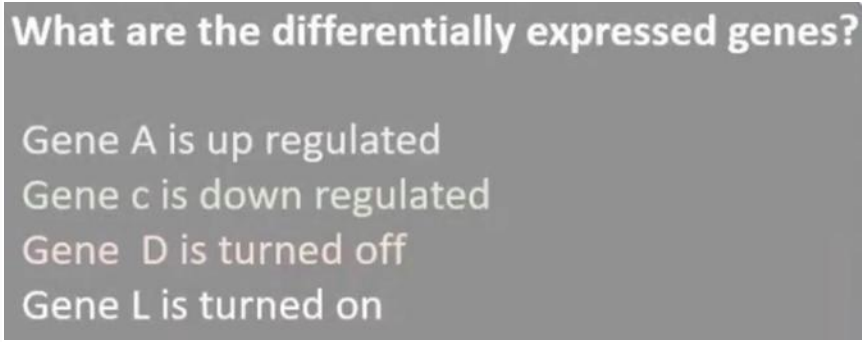
Nerve cell

Muscle cell

Differential gene expression analysis



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



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

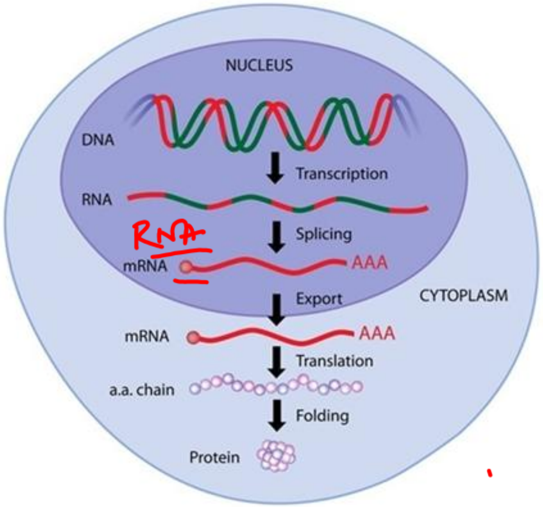
3. Some molecular features can only be observed at the RNA level

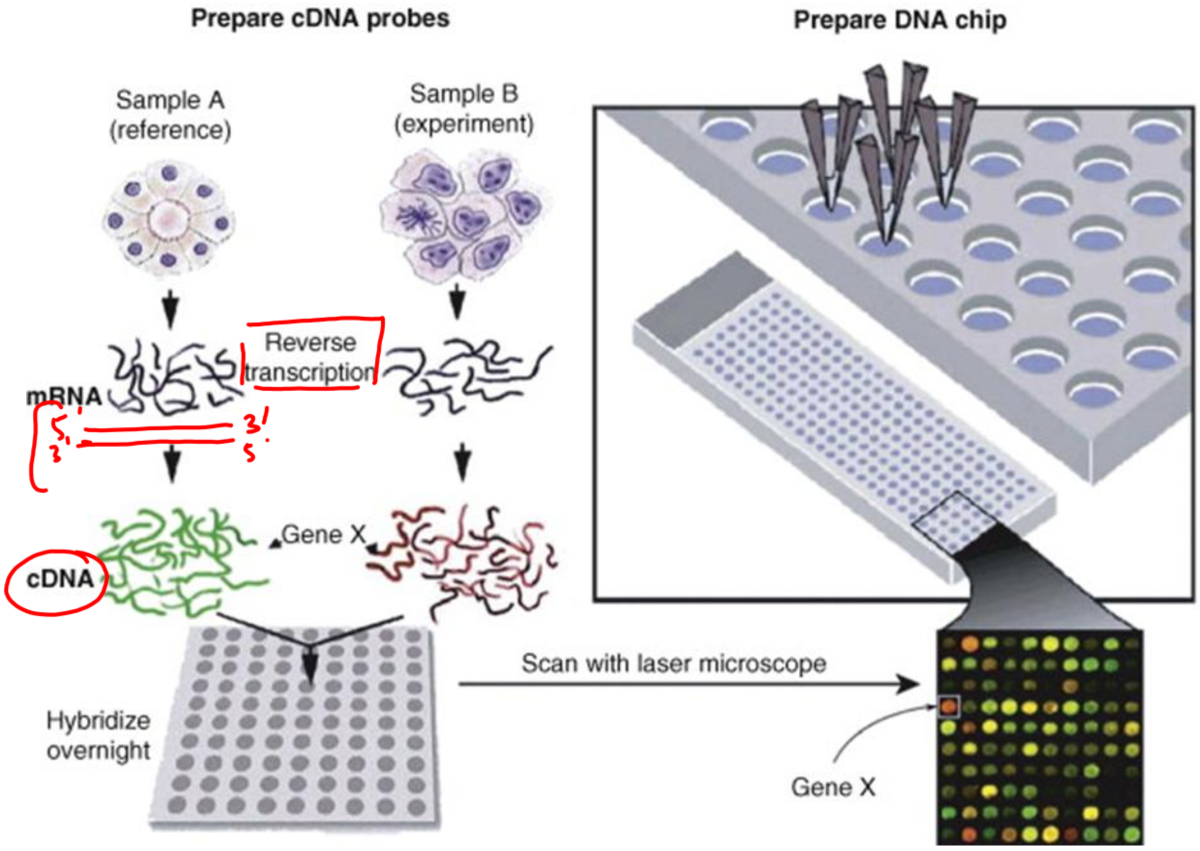
Alternative isoforms, fusion transcripts, RNA editing

4. Understand allele specific expression

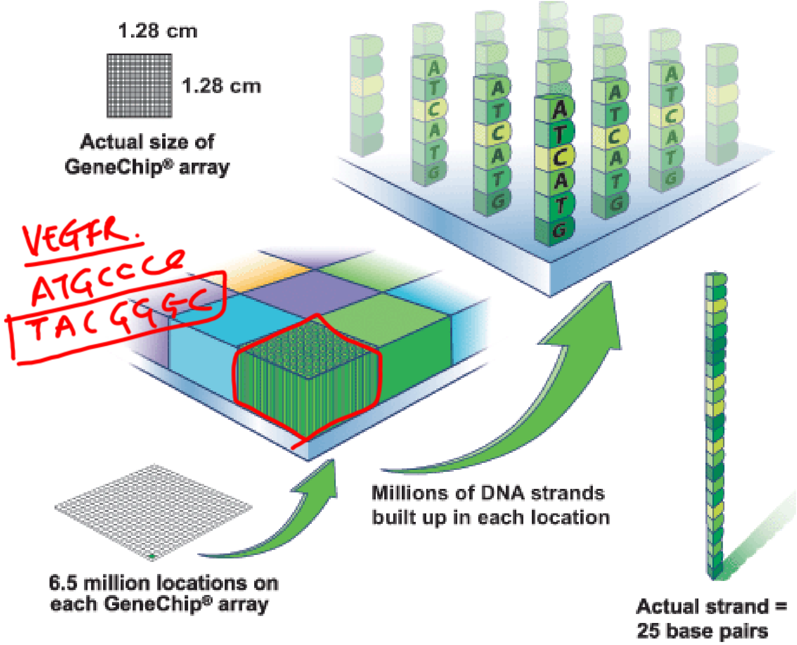
Gene expression-based biomarker identification

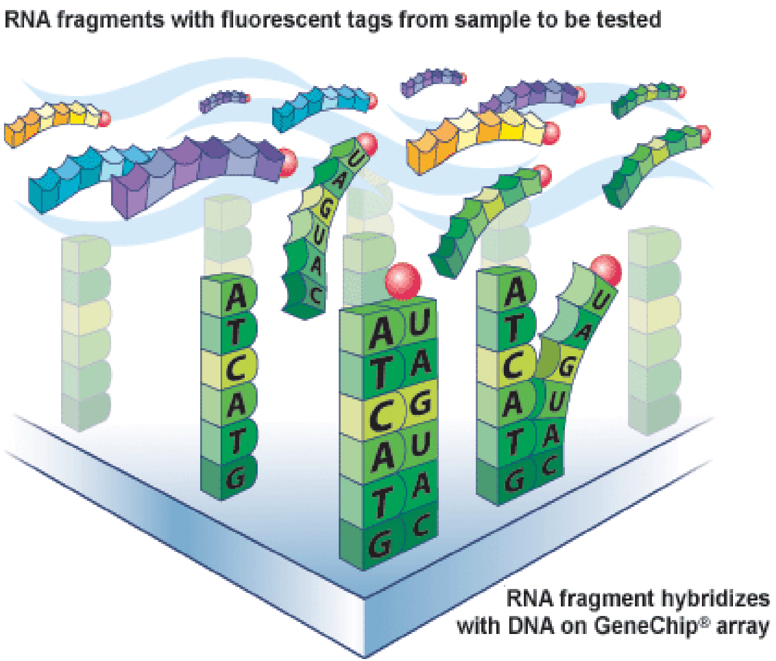
DNA Microarray





Microarray analysis

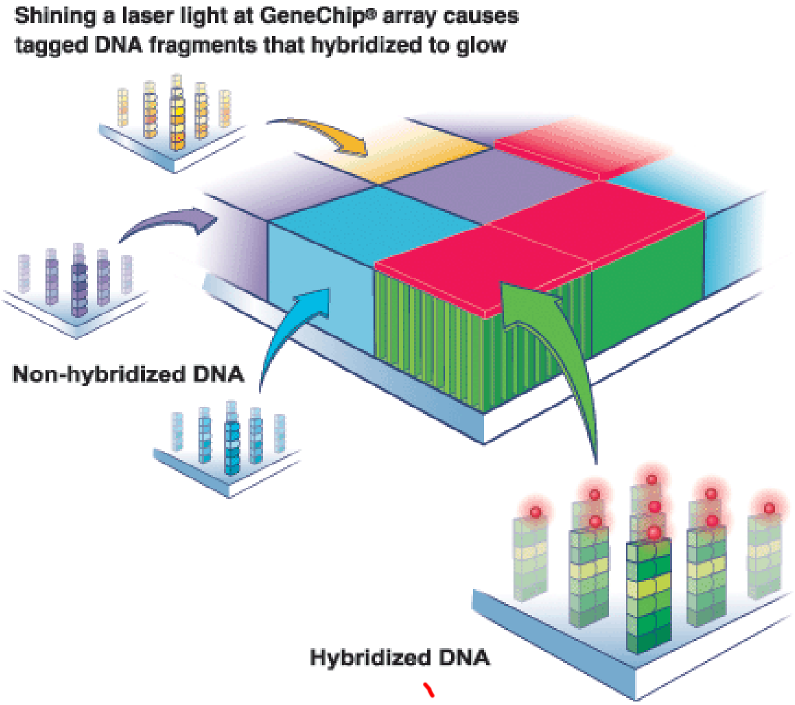

1.28 cm
1.28 cm
Actual size of GeneChip® array
VEGFR.
ATGCCCG
TACGGGC
Millions of DNA strands built up in each location
6.5 million locations on each GeneChip® array
Actual strand = 25 base pairs
A T C A T G


RNA fragments with fluorescent tags from sample to be tested
A T C A T G
U A G U A C
RNA fragment hybridizes with DNA on GeneChip® array

1. Micro Array Features

2. Hybridization (Pairing)

Microarray analysis


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

3. Detection

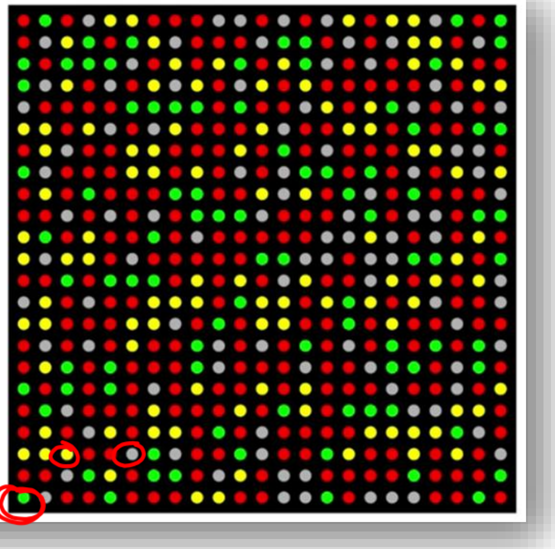
cDNAs from tissue 1 were labeled red

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

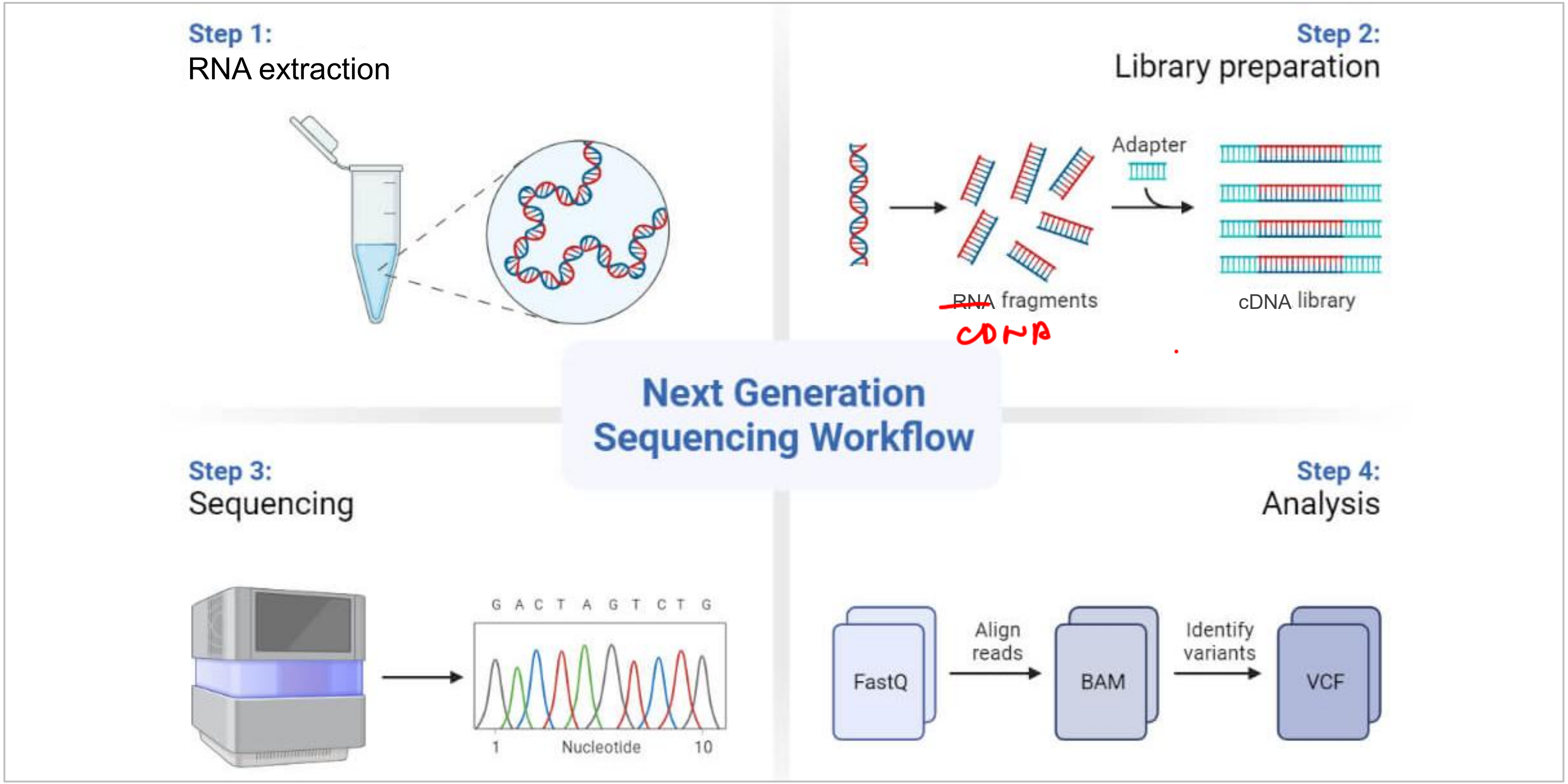
yellow spot means both cDNAs bind and gene is expressed in both tissues



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

poly-adenylated (coding) RNAs, “genes”

short non-coding RNAs (ncRNA), “microRNA”

long non-coding RNAs

} Total RNA

ribosomal RNA

rRNA

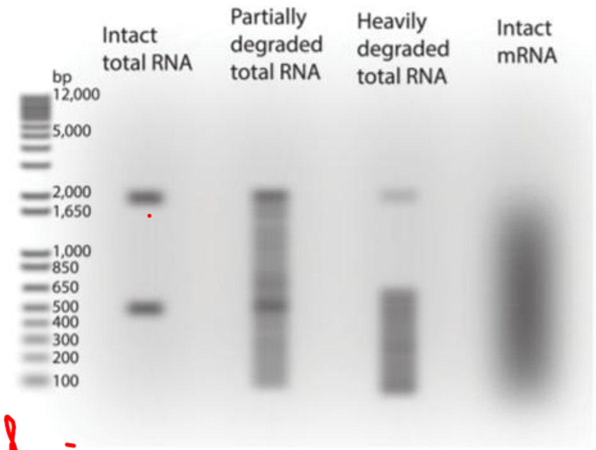
Most of the RNA in the cell

“Enrichment”

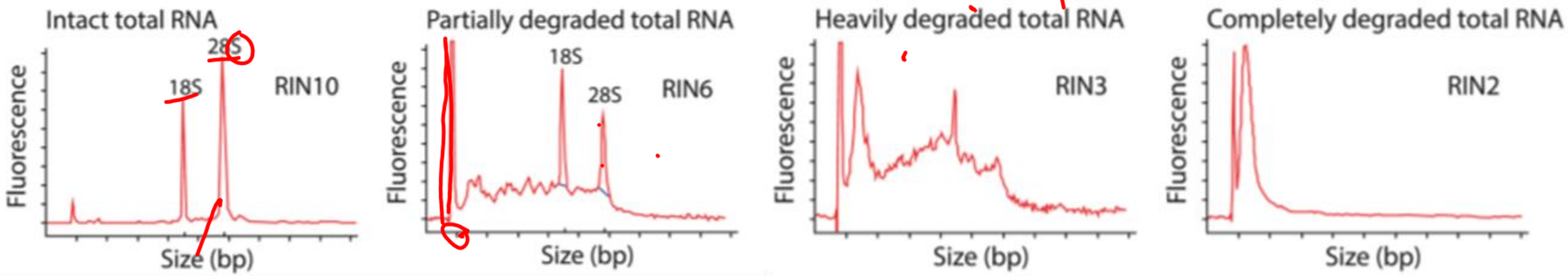
{ polyA capture
ribominus

RNAse

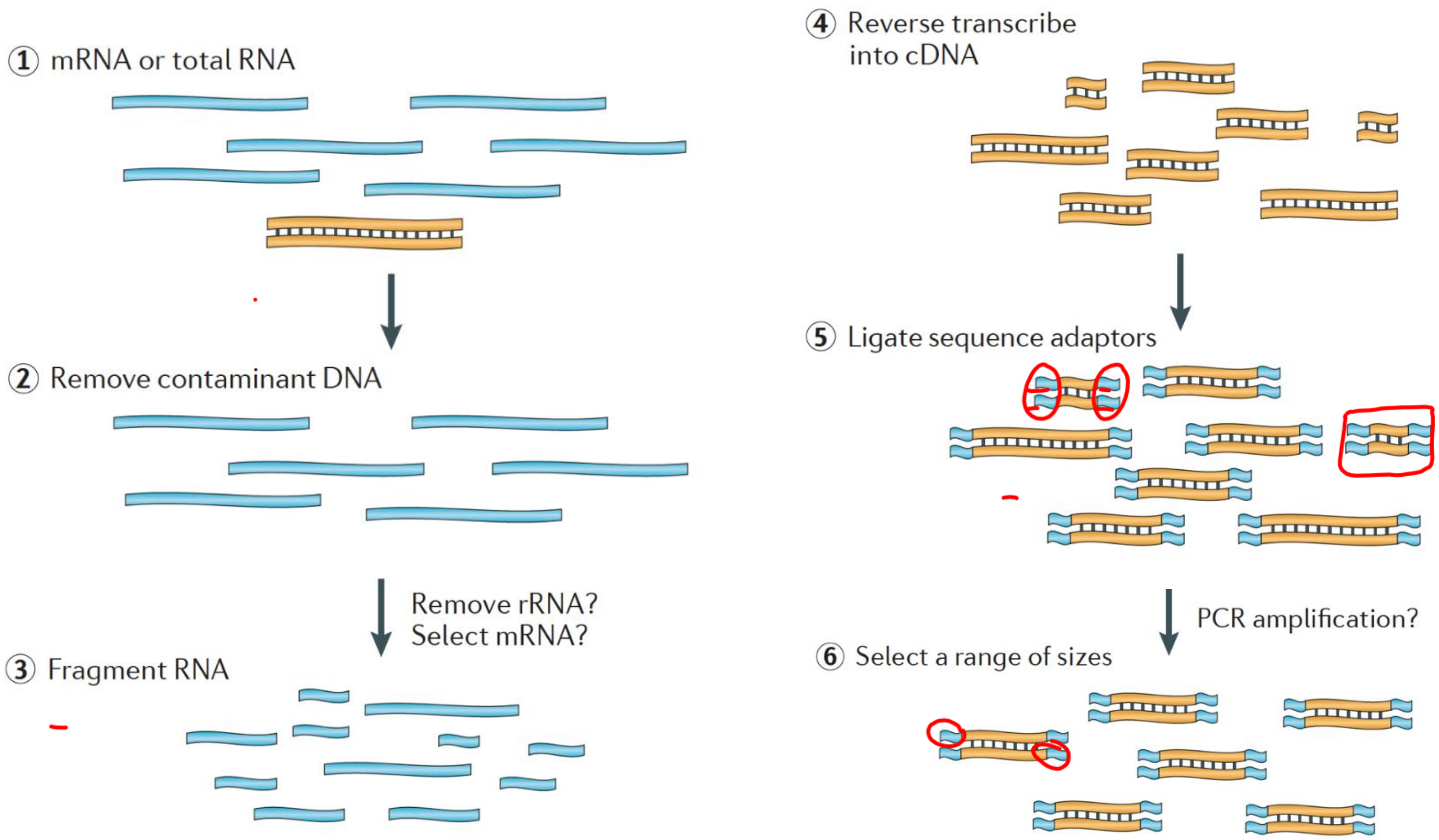
(a) Gel electropherogram



(b) Capillary electropherogram

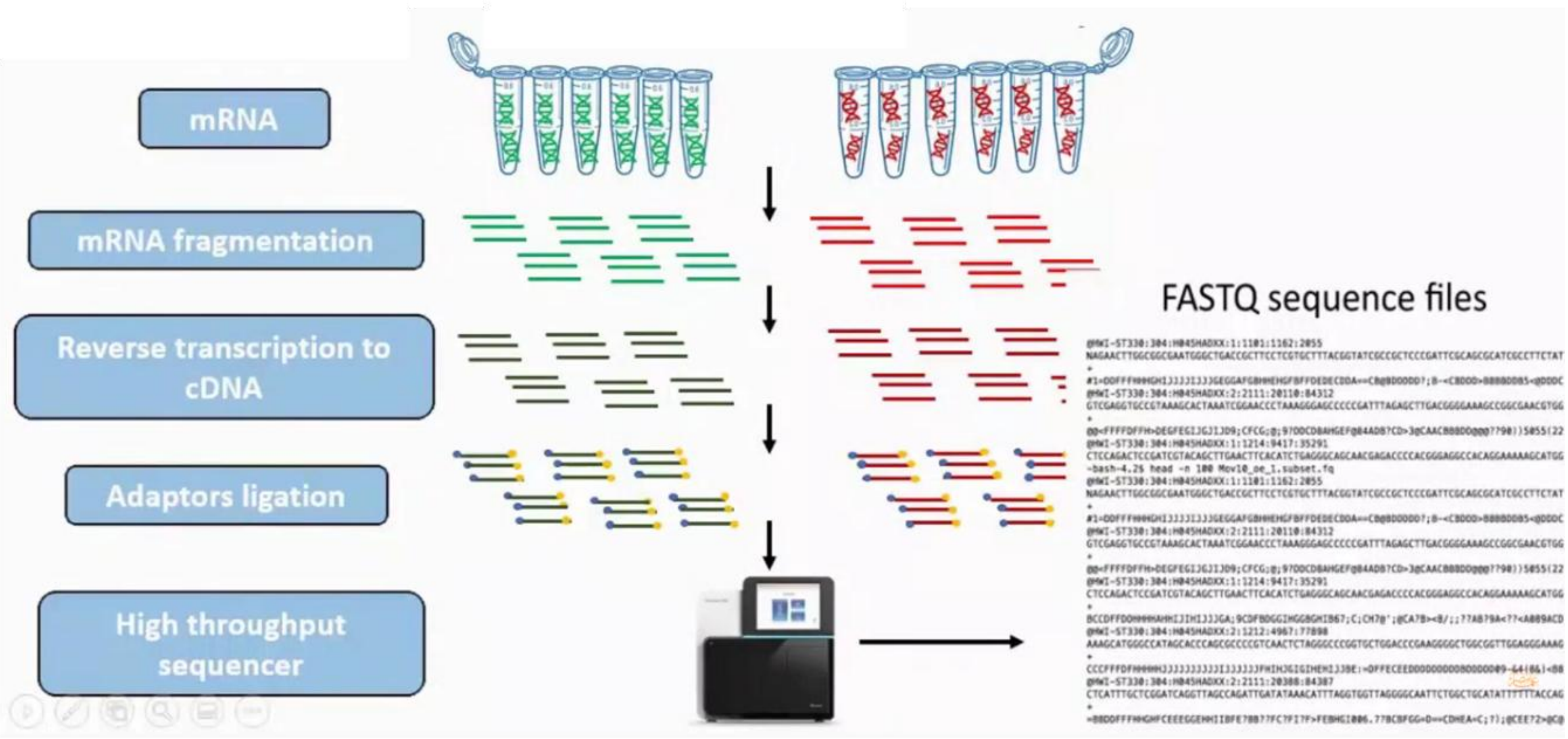


Library Preparation



RNA-seq experiment workflow

Library preparation



mRNA

mRNA fragmentation

Reverse transcription to cDNA

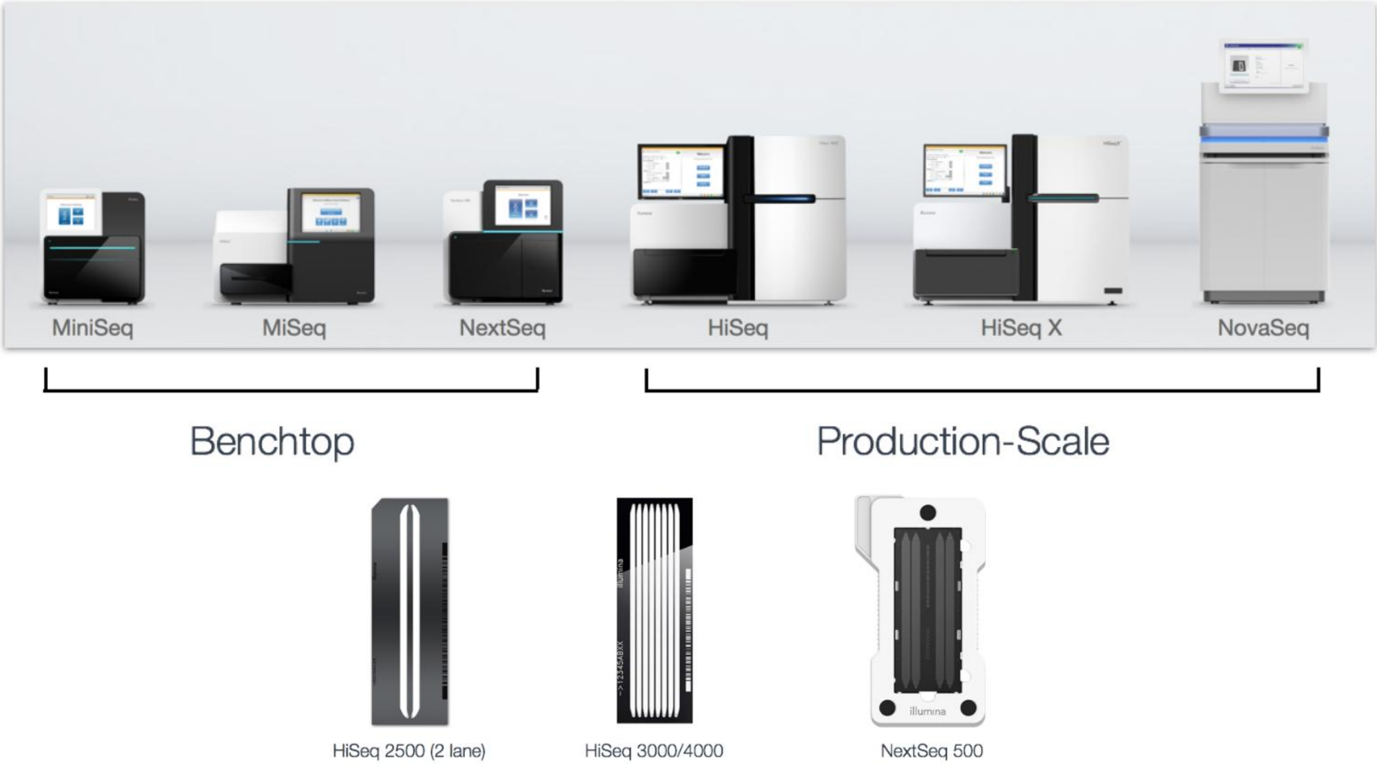
Adaptors ligation

High throughput sequencer

FASTQ sequence files

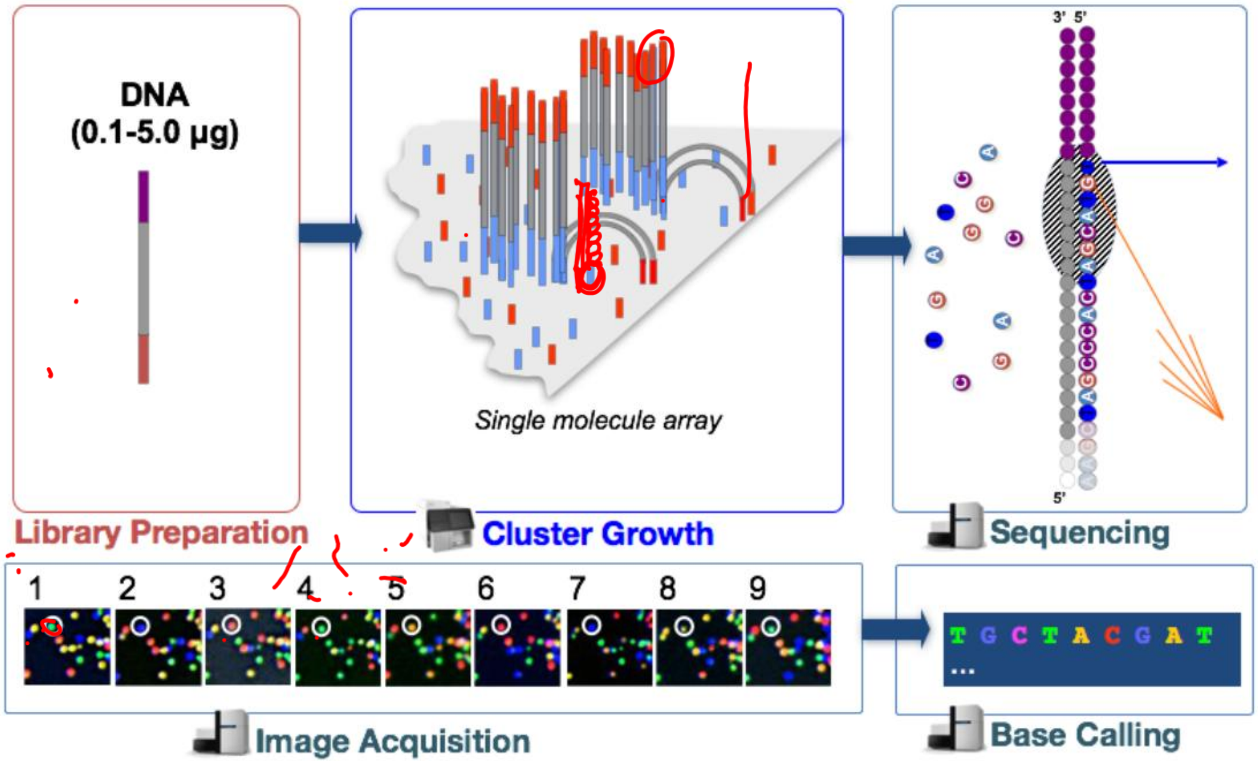
```
@HWI-ST330:304:H045HADXX:1:1101:1162:2055
NAGAACTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTAT
+
#1=DDFFFHHHGHIJJJJIJJJGEGGAFGBHHEHGFBFFDEDECDDA==CB@BDDDDD?;B-<CBDDD>BBBDDB5<@DDDC
@HWI-ST330:304:H045HADXX:2:2111:20110:84312
GTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGG
+
@@<FFFFDFFH>DEGFEGIJGJIJD9;CFCG;@;97DDCDBAHGEF@B4ADB?CD>3@CAACBBBDD@@@??9B))5B55(22
@HWI-ST330:304:H045HADXX:1:1214:9417:35291
CTCCAGACTCCGATCGTACAGCTTGAACTTCACATCTGAGGGCAGCAACGAGACCCCACGGGAGGCCACAGGAAAAAGCATGG
-bash-4.2$ head -n 100 Mov10_oe_1.subset.fq
@HWI-ST330:304:H045HADXX:1:1101:1162:2055
NAGAACTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTAT
+
#1=DDFFFHHHGHIJJJJIJJJGEGGAFGBHHEHGFBFFDEDECDDA==CB@BDDDDD?;B-<CBDDD>BBBDDB5<@DDDC
@HWI-ST330:304:H045HADXX:2:2111:20110:84312
GTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGG
+
@@<FFFFDFFH>DEGFEGIJGJIJD9;CFCG;@;97DDCDBAHGEF@B4ADB?CD>3@CAACBBBDD@@@??9B))5B55(22
@HWI-ST330:304:H045HADXX:1:1214:9417:35291
CTCCAGACTCCGATCGTACAGCTTGAACTTCACATCTGAGGGCAGCAACGAGACCCCACGGGAGGCCACAGGAAAAAGCATGG
+
BCCDFFDDHHHHAHHIJIHIJJJGA;9CDFBDGGIHGGBGHIB67;C;CH7@';@CA?B><B/;;??AB?9A<??<AB89ACD
@HWI-ST330:304:H045HADXX:2:1212:4967:77898
AAAGCATGGGCCATAGCACCCAGCGCCCCGTCAACTCTAGGGCCCGGTGCTGGACCCGAAGGGGCTGGCGGTTGGAGGGAAAG
+
CCCFFFDFHHHHHJJJJJJJJJJIJJJJJJFHIHJGIGIHEHIJJBE:=DFFECEEDDDDDDDDDBDDDDD@9-44(B&)<BB
@HWI-ST330:304:H045HADXX:2:2111:20388:84387
CTCATTTGCTCGGATCAGGTTAGCCAGATTGATATAAACATTTAGGTGGTTAGGGGCAATTCTGGCTGCATATTTTTTACCAG
+
=BBDDFFFHHGHFCEEEGGEHHIIBFE?BB??FC?FI?F>FEBHGI@@6.??BCBFGG=D==CDHEA<C;?);@CEE?2>@C@
```

Illumina sequencing platforms



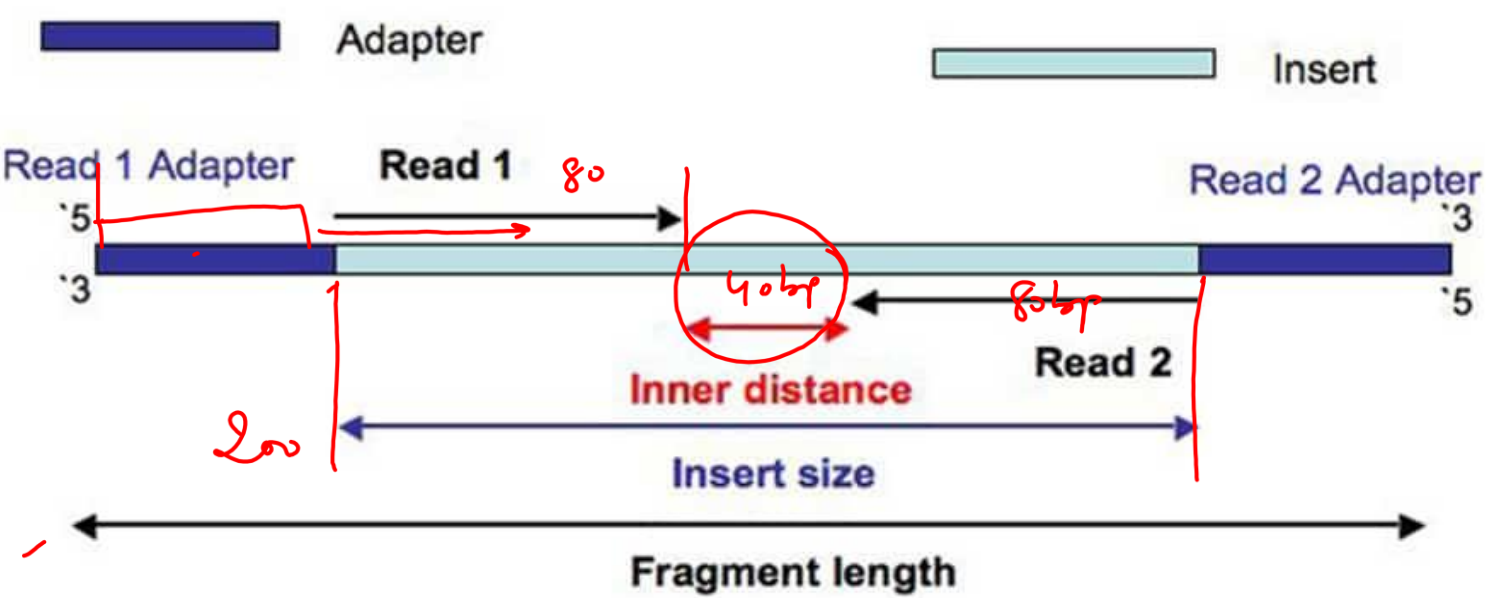
<u>Other sequencing platforms</u>: Pacific Bioscience, Oxford Nanopore, 10X Genomics

Sequencing by synthesis

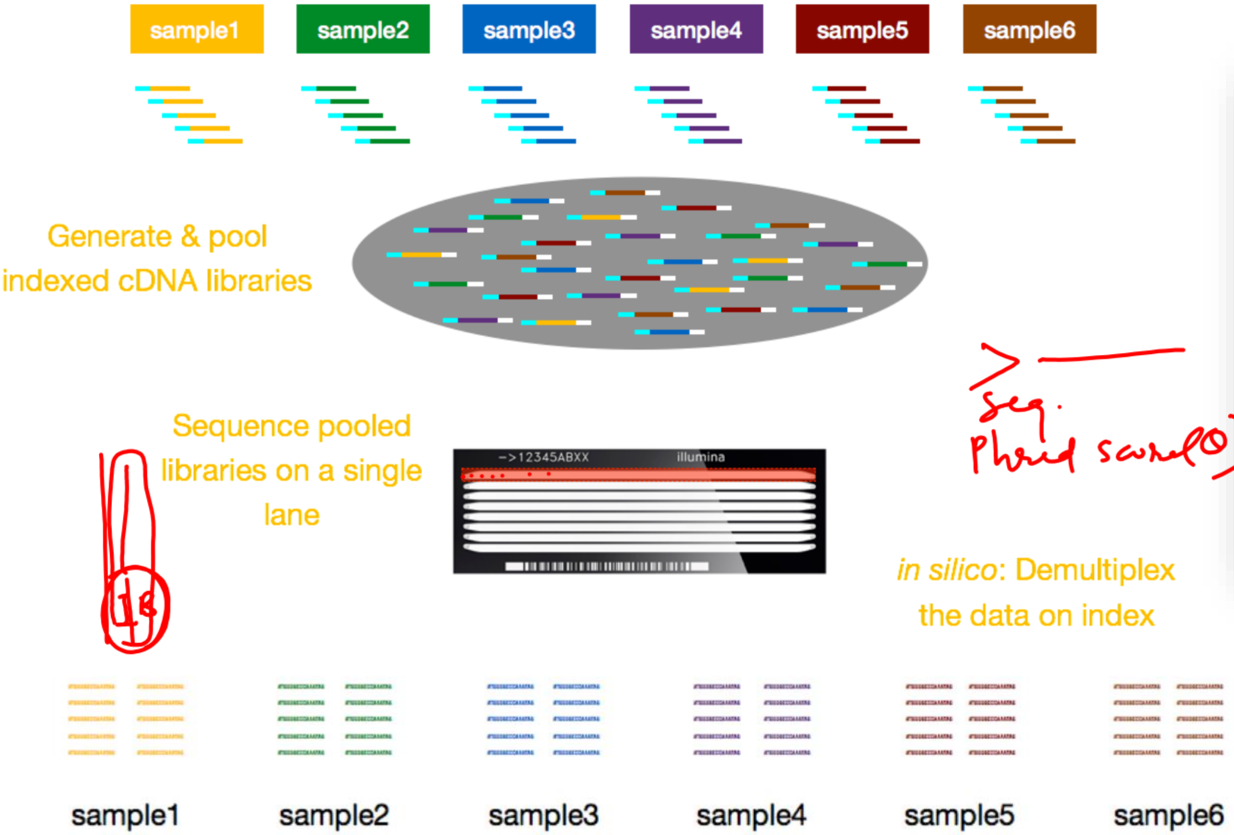


Illumina sequencing video URL: https://www.youtube.com/watch?v=womKfikWlxM

Single- and Paired-end sequencing



Multiplexing



Generate & pool indexed cDNA libraries

Sequence pooled libraries on a single lane

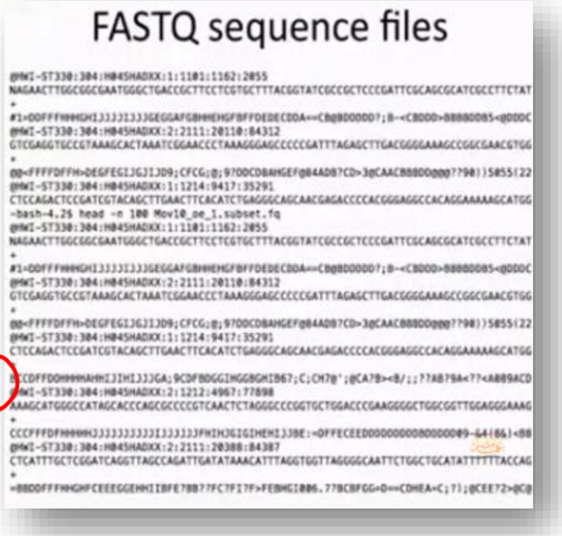
18

in silico: Demultiplex the data on index

> ——

Seq.

Phred score(Q)



FASTQ sequence files

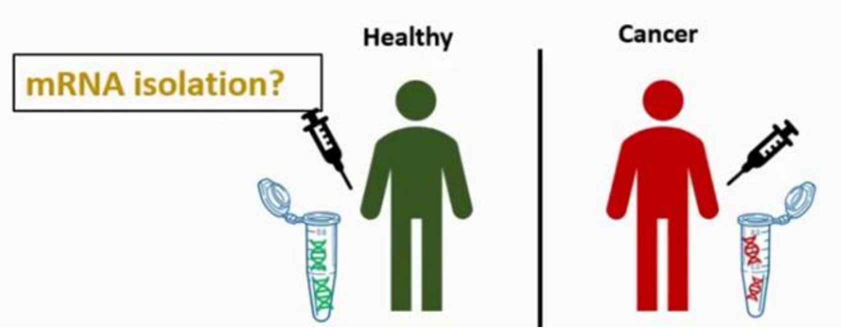
```
@HWI-ST330:304:H045HADXX:1:1101:1162:2055
NAGAACTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTAT
+
#1=DDFFFHHHGHIJJJJIJJJGEGGAFGBHHEHGFBFFDEDECDDA==CB@BDDDDD?;B-<CBDDD>BBBBDDB5<@DDDC
@HWI-ST330:304:H045HADXX:2:2111:20110:84312
GTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGG
+
@@<FFFFDFFH>DEGFEGIJGJIJD9;CFCG;@;97DDCDBAHGEF@B4ADB?CD>3@CAACBBBDD@@@??9B))5B55(22
@HWI-ST330:304:H045HADXX:1:1214:9417:35291
CTCCAGACTCCGATCGTACAGCTTGAACTTCACATCTGAGGGCAGCAACGAGACCCCACGGGAGGCCACAGGAAAAGCATGG
-bash-4.2$ head -n 100 Mov10_oe_1.subset.fq
@HWI-ST330:304:H045HADXX:1:1101:1162:2055
NAGAACTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTAT
+
#1=DDFFFHHHGHIJJJJIJJJGEGGAFGBHHEHGFBFFDEDECDDA==CB@BDDDDD?;B-<CBDDD>BBBBDDB5<@DDDC
@HWI-ST330:304:H045HADXX:2:2111:20110:84312
GTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGG
+
@@<FFFFDFFH>DEGFEGIJGJIJD9;CFCG;@;97DDCDBAHGEF@B4ADB?CD>3@CAACBBBDD@@@??9B))5B55(22
@HWI-ST330:304:H045HADXX:1:1214:9417:35291
CTCCAGACTCCGATCGTACAGCTTGAACTTCACATCTGAGGGCAGCAACGAGACCCCACGGGAGGCCACAGGAAAAGCATGG
+
BCCDFFDDHHHHAHHIJIHIJJJGA;9CDFBDGGIHGGBGHIB67;C;CH7@';@CA?B><B/;;??AB?9A<??<ABB9ACD
@HWI-ST330:304:H045HADXX:2:1212:4967:77898
AAAGCATGGGCCATAGCACCCAGCGCCCCGTCAACTCTAGGGCCCGGTGCTGGACCCGAAGGGGCTGGCGGTTGGAGGGAAAG
+
CCCFFFDFHHHHHJJJJJJJJJJIJJJJJJFHIHJGIGIHEHIJJBE:=DFFECEEDDDDDDDDDBDDDDD@9-&4(B&)<BB
@HWI-ST330:304:H045HADXX:2:2111:20388:84387
CTCATTTGCTCGGATCAGGTTAGCCAGATTGATATAAACATTTAGGTGGTTAGGGGCAATTCTGGCTGCATATTTTTTACCAG
+
=BBDDFFFHHGHFCEEEGGEHHIIBFE?BB??FC?FI?F>FEBHGIB06.??BCBFGG=D==CDHEA=C;?);@CEE?2>@C@
```

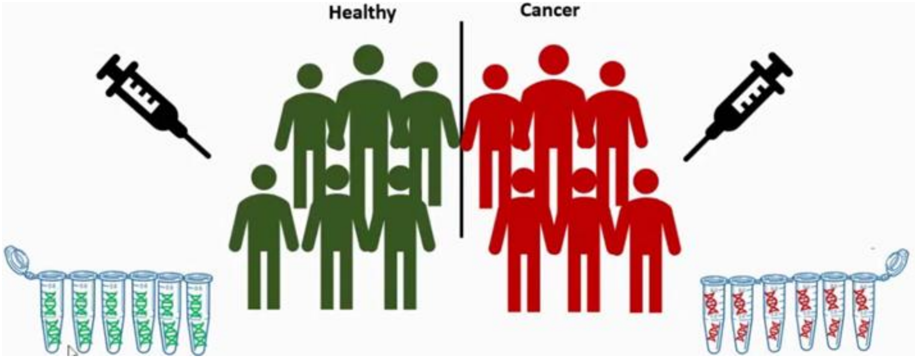
Microarray *vs.* RNA-seq

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



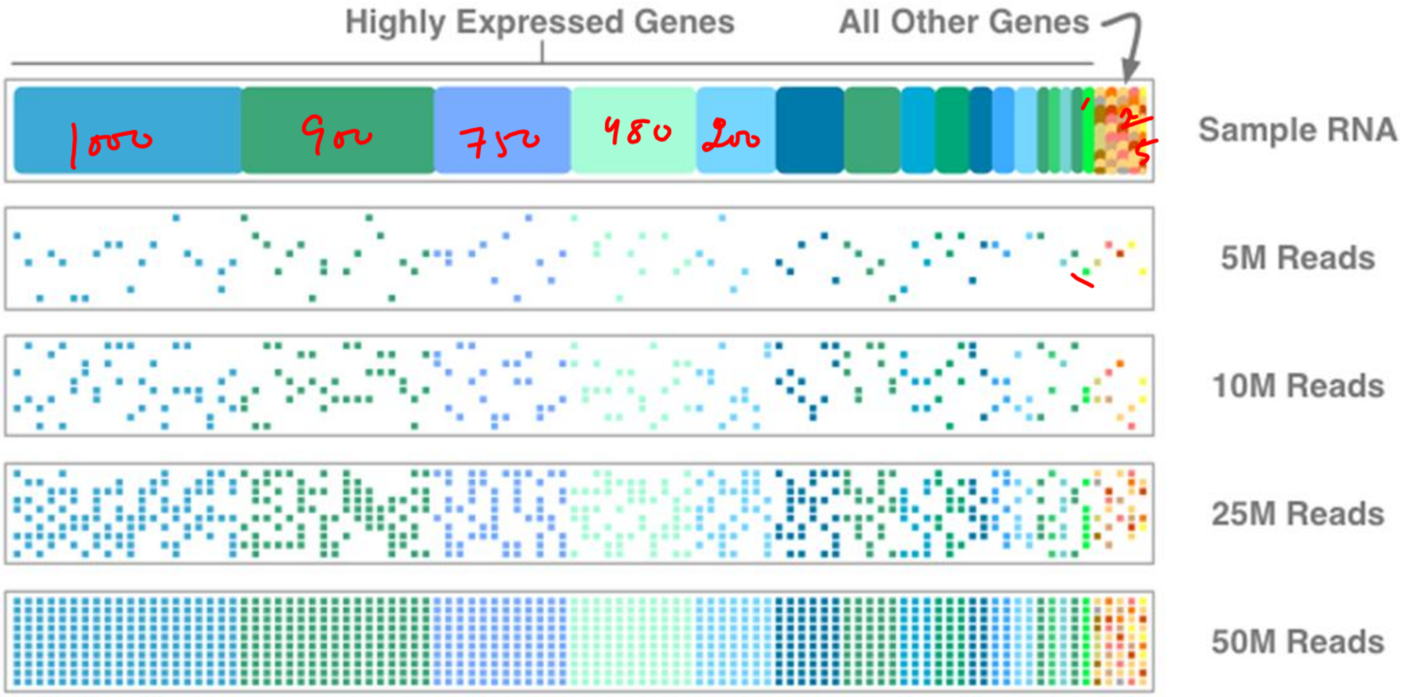


1. *Biological replicates* : Include multiple sampling within the population
2. *Technical replicates* : Include multiple preparation and re-sequencing of the same sample

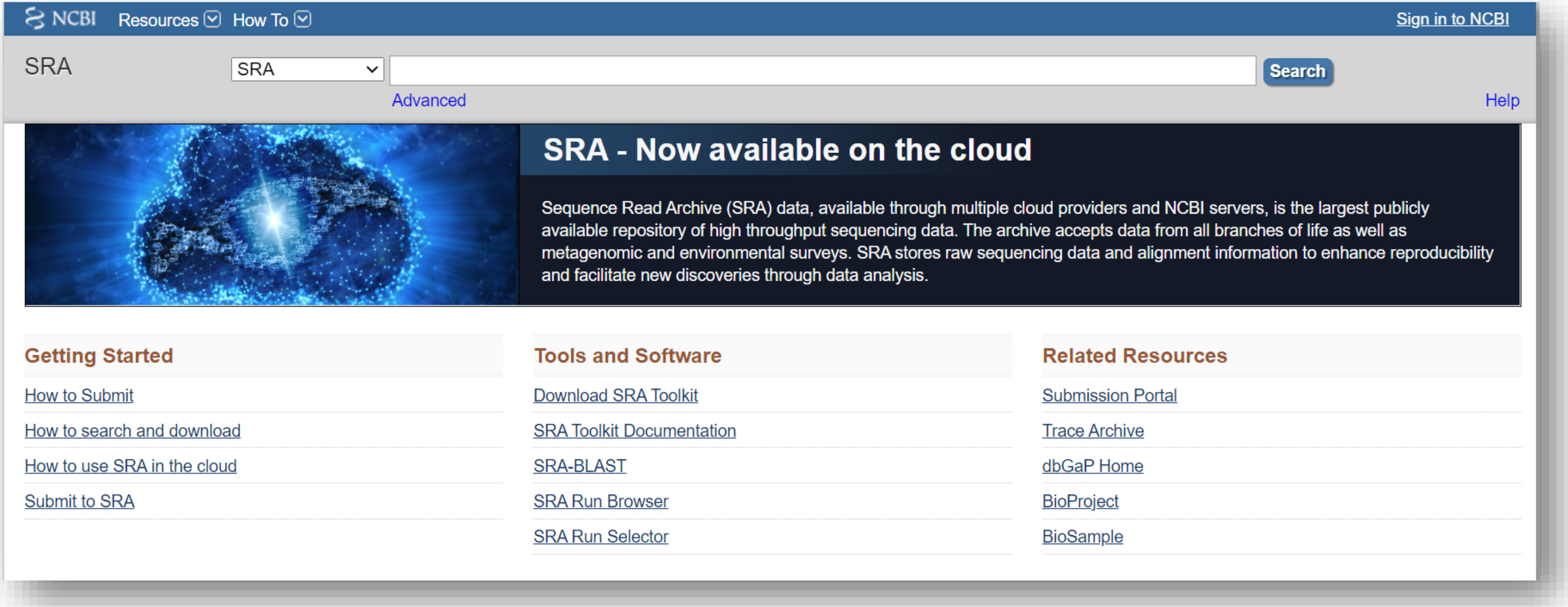
Biological replicates generally increase statistical power more than technical replicates

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

Sequencing depth



Sources for RNA-seq datasets



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

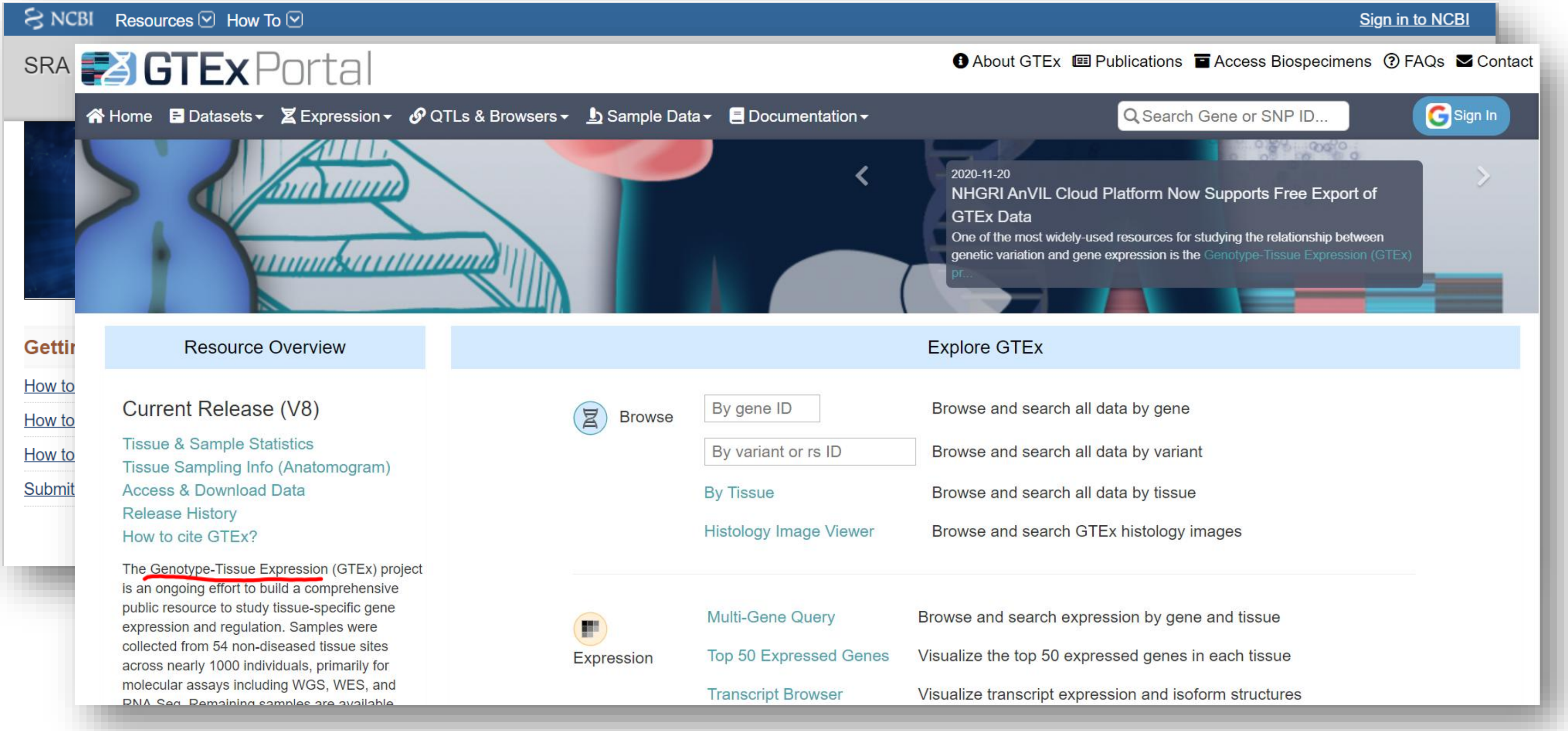
Tools and Software

- 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



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) pr...

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

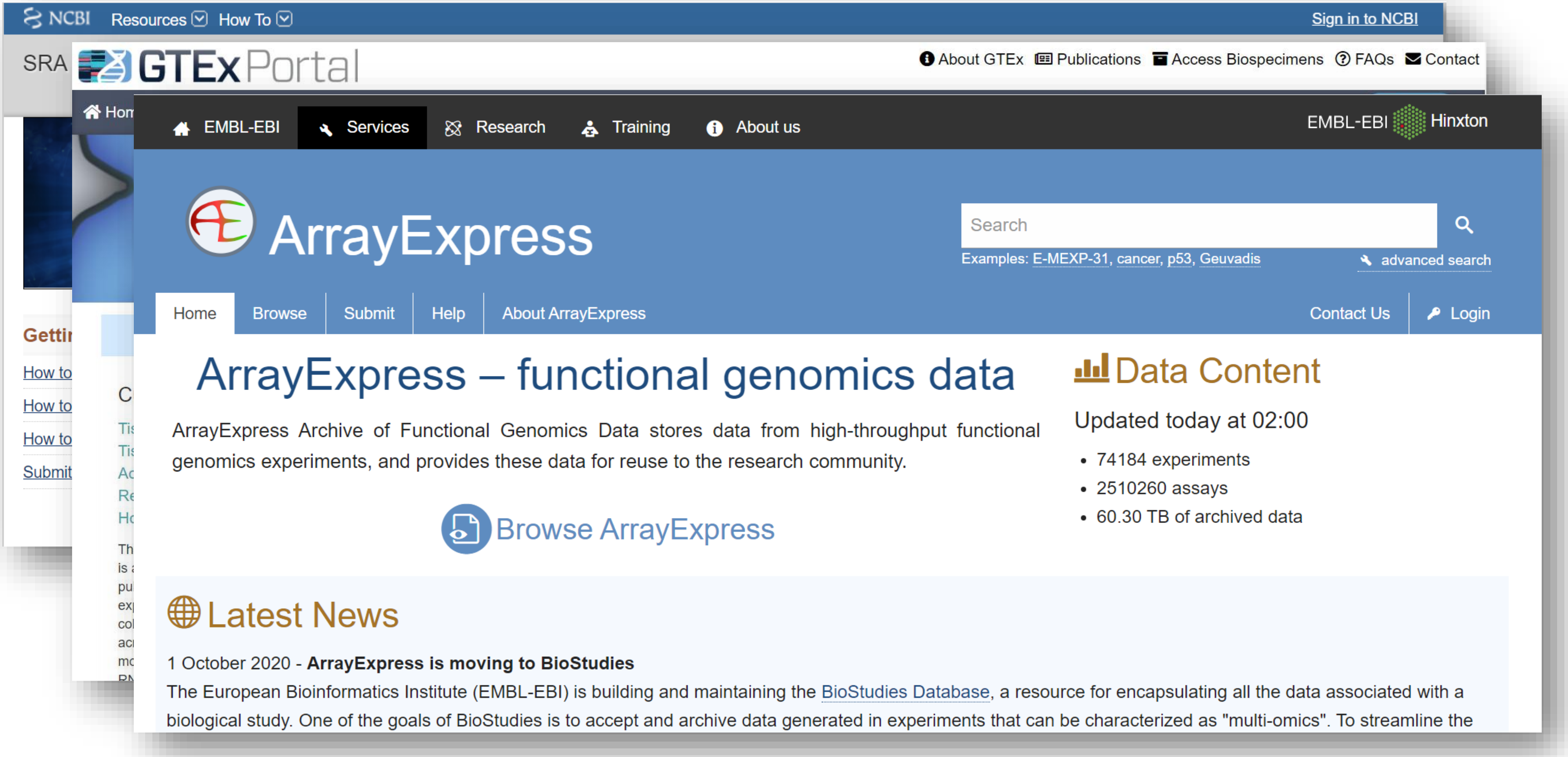
Browse

By gene ID	Browse and search all data by gene
By variant or rs ID	Browse and search all data by variant
By Tissue	Browse and search all data by tissue
Histology Image Viewer	Browse and search GTEx histology images

Expression

Multi-Gene Query	Browse and search expression by gene and tissue
Top 50 Expressed Genes	Visualize the top 50 expressed genes in each tissue
Transcript Browser	Visualize transcript expression and isoform structures

Sources for RNA-seq datasets



Examples: E-MEXP-31, cancer, p53, Geuvadis

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

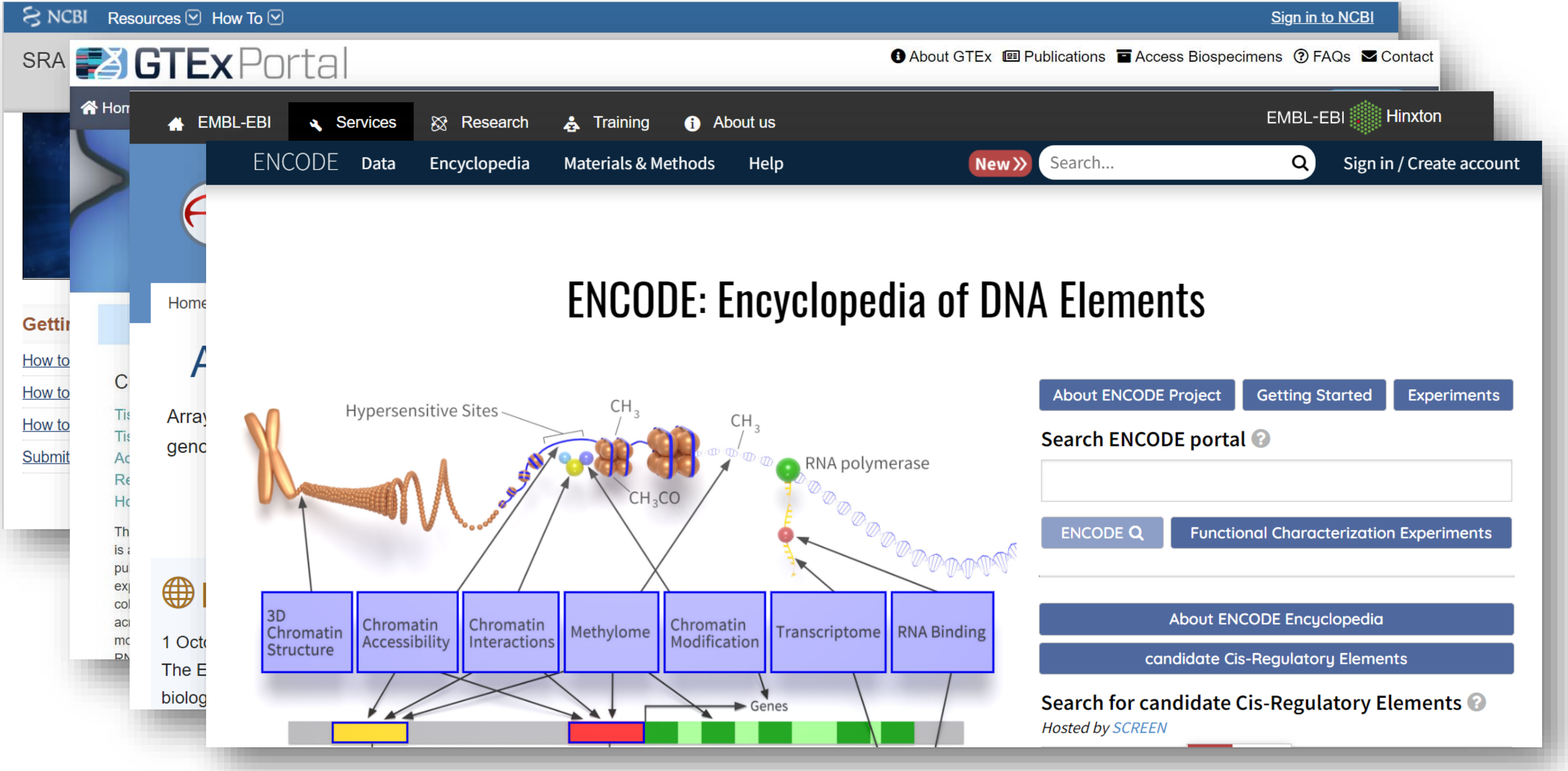
- 74184 experiments
- 2510260 assays
- 60.30 TB of archived data

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

Gettir

How to

How to

How to

Submit

GTEx Portal

About GTEx Publications Access Biospecimens FAQs Contact

Hon

C

Tis
Tis
Ac
Re
Ho

Th
is a
pu
exp
col
acr
mo
RN

EMBL-EBI Services Research Training About us

EMBL-EBI Hinxton

Home

A

Array
genc

1 Octo
The E
biolog

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

ENCODE: Encyclopedia of DNA Elements

About ENCODE Project | Getting Started | Experiments

Search ENCODE portal ?

ENCODE | Functional Characterization Experiments

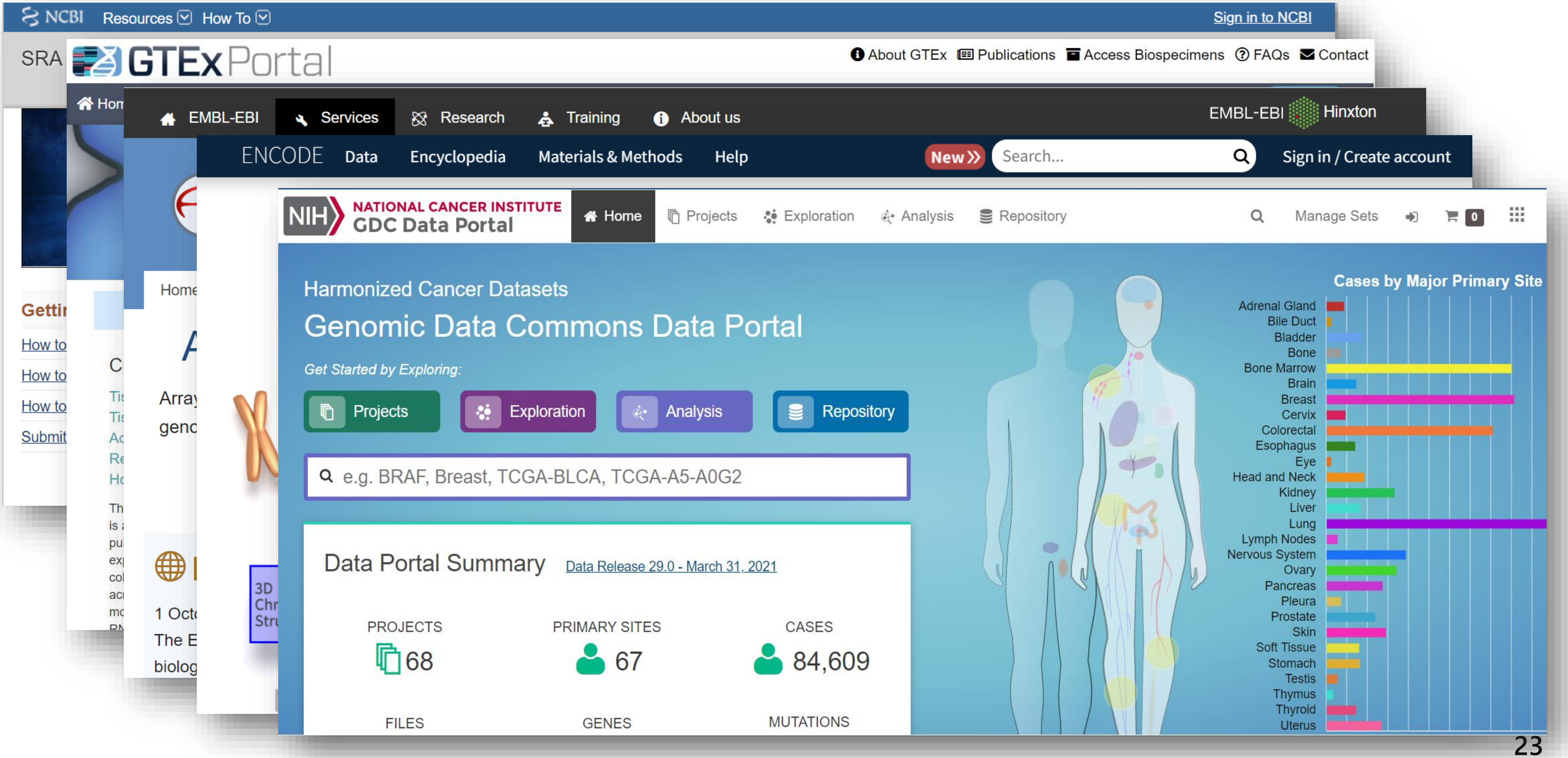
About ENCODE Encyclopedia

candidate Cis-Regulatory Elements

Search for candidate Cis-Regulatory Elements ?

Hosted by SCREEN

Sources for RNA-seq datasets



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]

PROJECTS 68

PRIMARY SITES 67

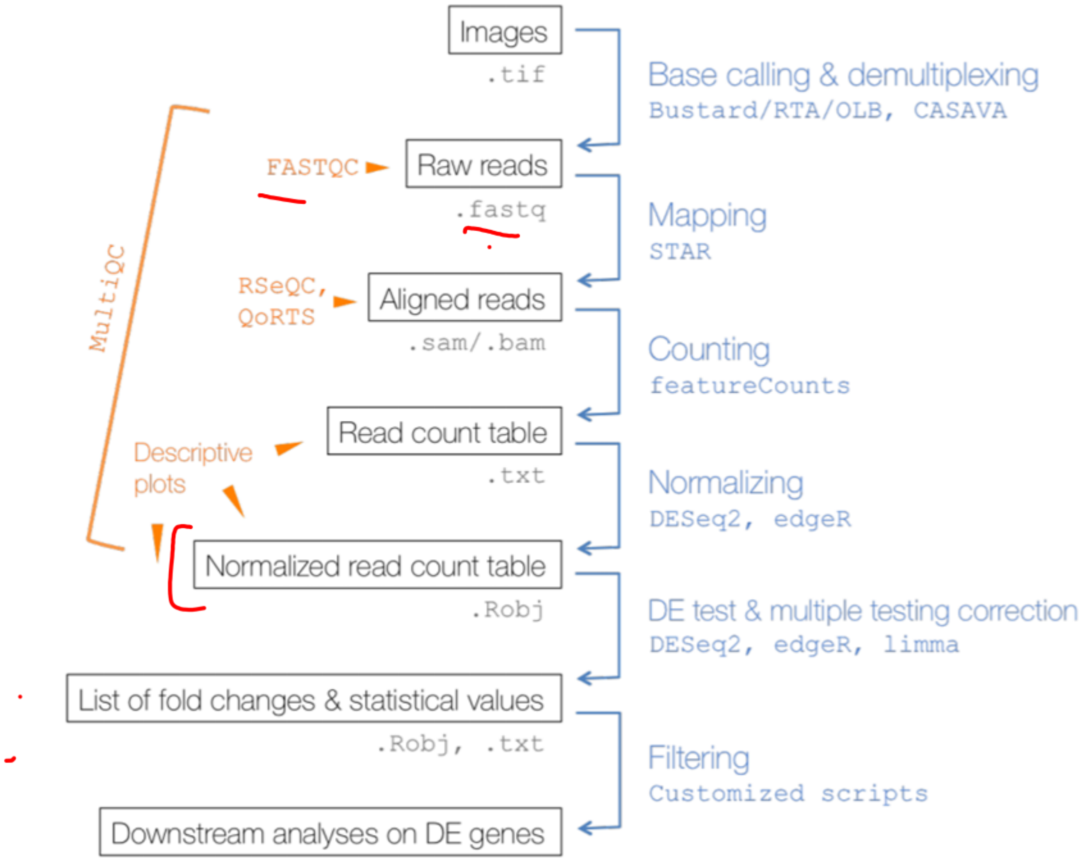
CASES 84,609

FILES

GENES

MUTATIONS

Workflow of differential gene expression analysis



Images
.tif

Base calling & demultiplexing
Bustard/RTA/OLB, CASAVA

MultiQC

FASTQC

Raw reads
.fastq

Mapping
STAR

RSeQC,
QoRTS

Aligned reads
.sam/.bam

Counting
featureCounts

Read count table
.txt

Descriptive
plots

Normalizing
DESeq2, edgeR

Normalized read count table
.Robj

DE test & multiple testing correction
DESeq2, edgeR, limma

List of fold changes & statistical values
.Robj, .txt

Filtering
Customized scripts

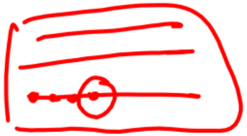
Downstream analyses on DE genes

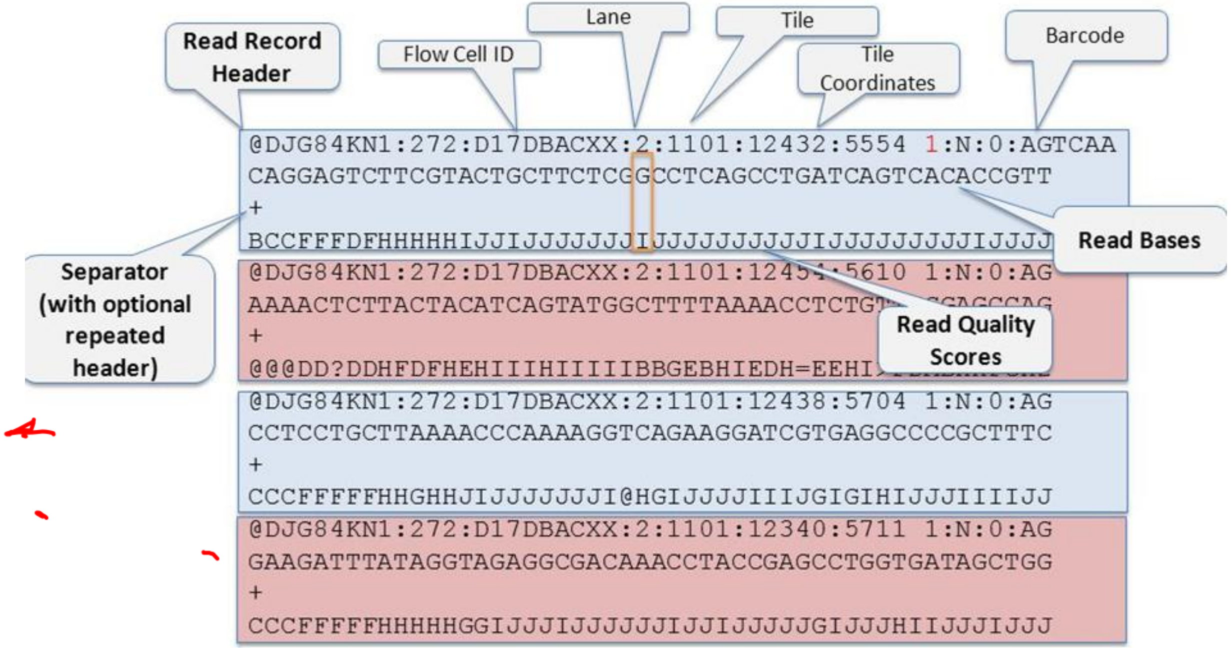
Gene expression-based biomarker identification

Problems in sequencing

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

10 — 60





```
@DJG84KN1:272:D17DBACXX:2:1101:12432:5554 1:N:0:AGTCAA
CAGGAGTCTTCGTACTGCTTCTCGGCCTCAGCCTGATCAGTCACACCGTT
+
BCCFFFDFHHHHHIJJIJJJJJJJIJJJJJJJJJIJJJJJJJJIJJJJ
@DJG84KN1:272:D17DBACXX:2:1101:12454:5610 1:N:0:AG
AAAACTCTTACTACATCAGTATGGCTTTTAAAACCTCTGT  CGAGCCAG
+
@@@DD?DDHFDFHEHIIIHIIIIIBBGEBHIEDH=EEHI>
@DJG84KN1:272:D17DBACXX:2:1101:12438:5704 1:N:0:AG
CCTCCTGCTTAAAACCCAAAAGGTCAGAAGGATCGTGAGGCCCCGCTTTC
+
CCCFFFFFHHGHHJIJJJJJJJI@HGIJJJJIIIJGIGIHIJJJIIIIJJ
@DJG84KN1:272:D17DBACXX:2:1101:12340:5711 1:N:0:AG
GAAGATTTATAGGTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTGG
+
CCCFFFFFHHHHHGGIJJJIJJJJJJIJJIJJJJGIJJJHIIJJJIJJJ
```

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

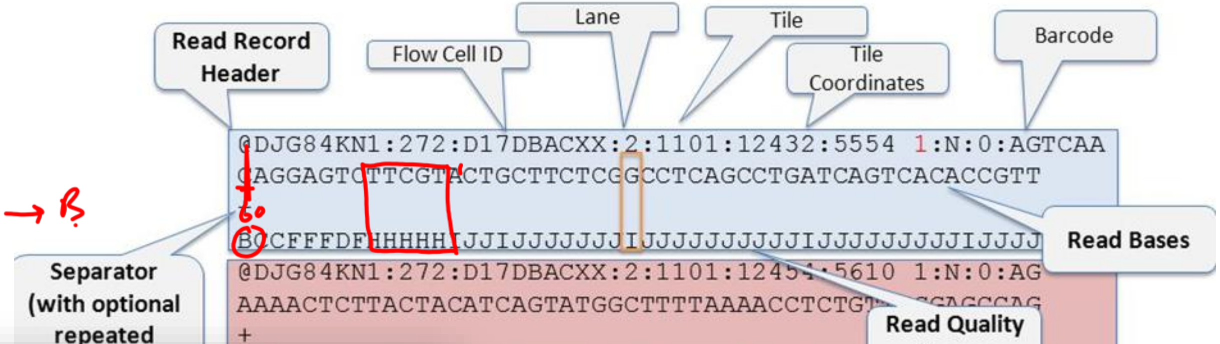
Gene expression-based biomarker identification

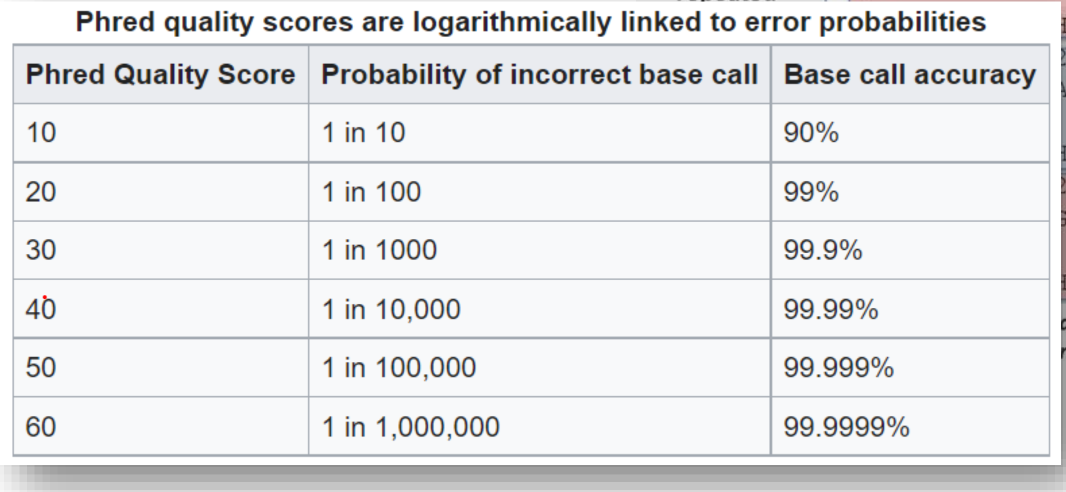
Problems in sequencing

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

ASCII → B

B -33


Read Record Header
Flow Cell ID
Lane
Tile
Tile Coordinates
Barcode
Read Bases
Separator (with optional repeated
Read Quality Scores
@DJG84KN1:272:D17DBACXX:2:1101:12432:5554 1:N:0:AGTCAA
CAGGAGTCTTCGTACTGCTTCTCGGCCTCAGCCTGATCAGTCACACCGTT
+
60
BCCFFFDFHHHHHIJJIJJJJJJJIJJJJJJJJJJJIJJJJJJJJJJIJJJJ
@DJG84KN1:272:D17DBACXX:2:1101:12454:5610 1:N:0:AG
AAAACTCTTACTACATCAGTATGGCTTTTAAAACCTCTGT
+
EHIIIHIIIIIBBGEBHIEDH=EEHI
:D17DBACXX:2:1101:12438:5704 1:N:0:AG
ACCCAAAAGGTCAGAAGGATCGTGAGGCCCCGCTTTC
HJIJJJJJJJI@HGIJJJJIIIJGIGIHIJJJIIIIJJ
:D17DBACXX:2:1101:12340:5711 1:N:0:AG
GTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTGG
HGGIJJJIJJJJJJIJJIJJJJJGIJJJHIIJJJIJJJ
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%