

# Biomarkers for disease identification/outcome

---



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

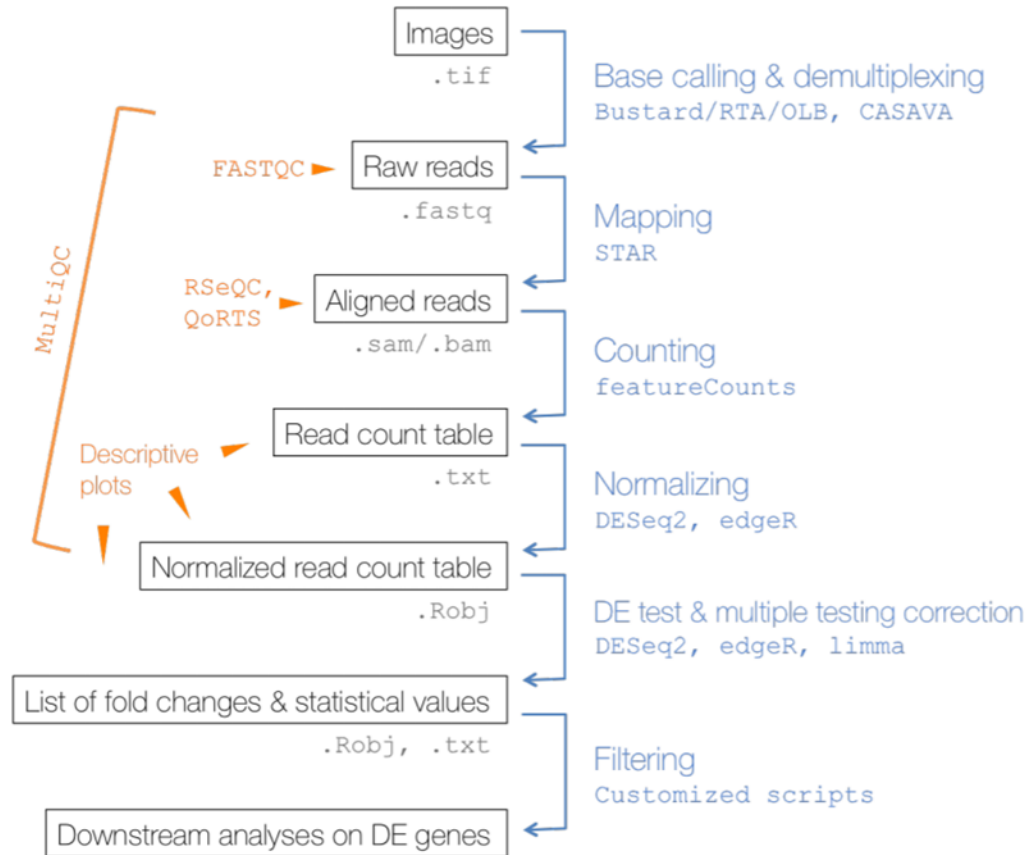
**Dr. Jaspreet Kaur Dhanjal**

**Assistant Professor, Department of Computational Biology**

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

*October 28, 2025*

# Workflow of differential gene expression analysis



# 3

## Problems in sequencing

- 
- The diagram illustrates the structure of a FASTQ file. It shows a sequence of reads, each consisting of a header line, a separator, the read bases, and quality scores. Callouts identify the following components:
- Read Record Header:** The first line of a read record, containing sample and sequencing information.
  - Flow Cell ID:** A unique identifier for the sequencing flow cell.
  - Lane:** The specific lane within the flow cell where the read was generated.
  - Tile:** The specific tile on the flow cell where the read was generated.
  - Tile Coordinates:** The specific coordinates on the tile where the read was generated.
  - Barcode:** A unique identifier for the read, used for multiplexing.
  - Separator:** A line starting with a plus sign (+) that separates the header from the read bases.
  - Read Bases:** The sequence of nucleotide bases (A, C, G, T) for the read.
  - Read Quality Scores:** A string of characters representing the quality of each base call.
- The diagram also shows a table with two columns: "Base call" and "Base call accuracy". The table shows the relationship between the quality score and the base call accuracy.
- | Base call | Base call accuracy |
|-----------|--------------------|
| 90%       |                    |
| 99%       |                    |
| 99.9%     |                    |
| 99.99%    |                    |

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%

# Quality check

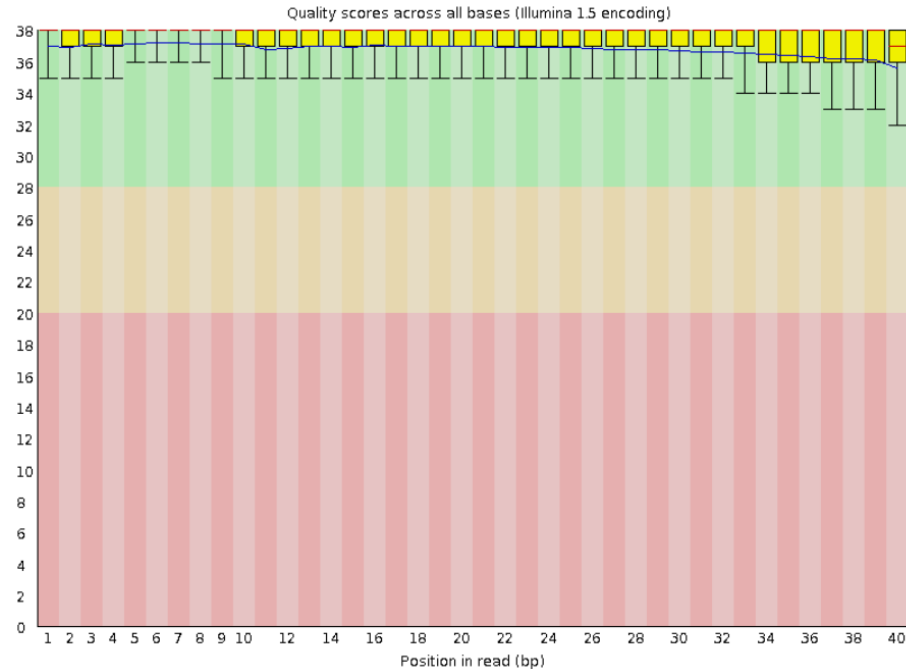
## FastQC Report

Good Data

### Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✓ [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

### ✓ Per base sequence quality



# Quality check

## FastQC Report

### Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✓ [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

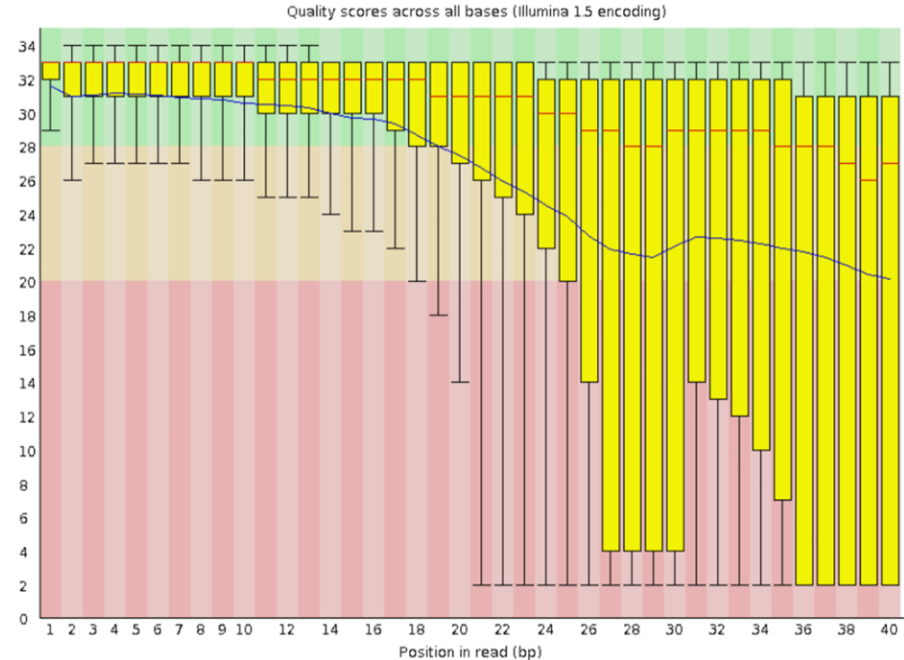
## FastQC Report

### Summary

- ✓ [Basic Statistics](#)
- ✗ [Per base sequence quality](#)
- ✗ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ⚠ [Per base sequence content](#)
- ⚠ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ⚠ [Sequence Duplication Levels](#)
- ⚠ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

Bad Data

### ✗ Per base sequence quality



# Quality check

## Adapter dimer contaminated run

### FastQC Report

#### Summary

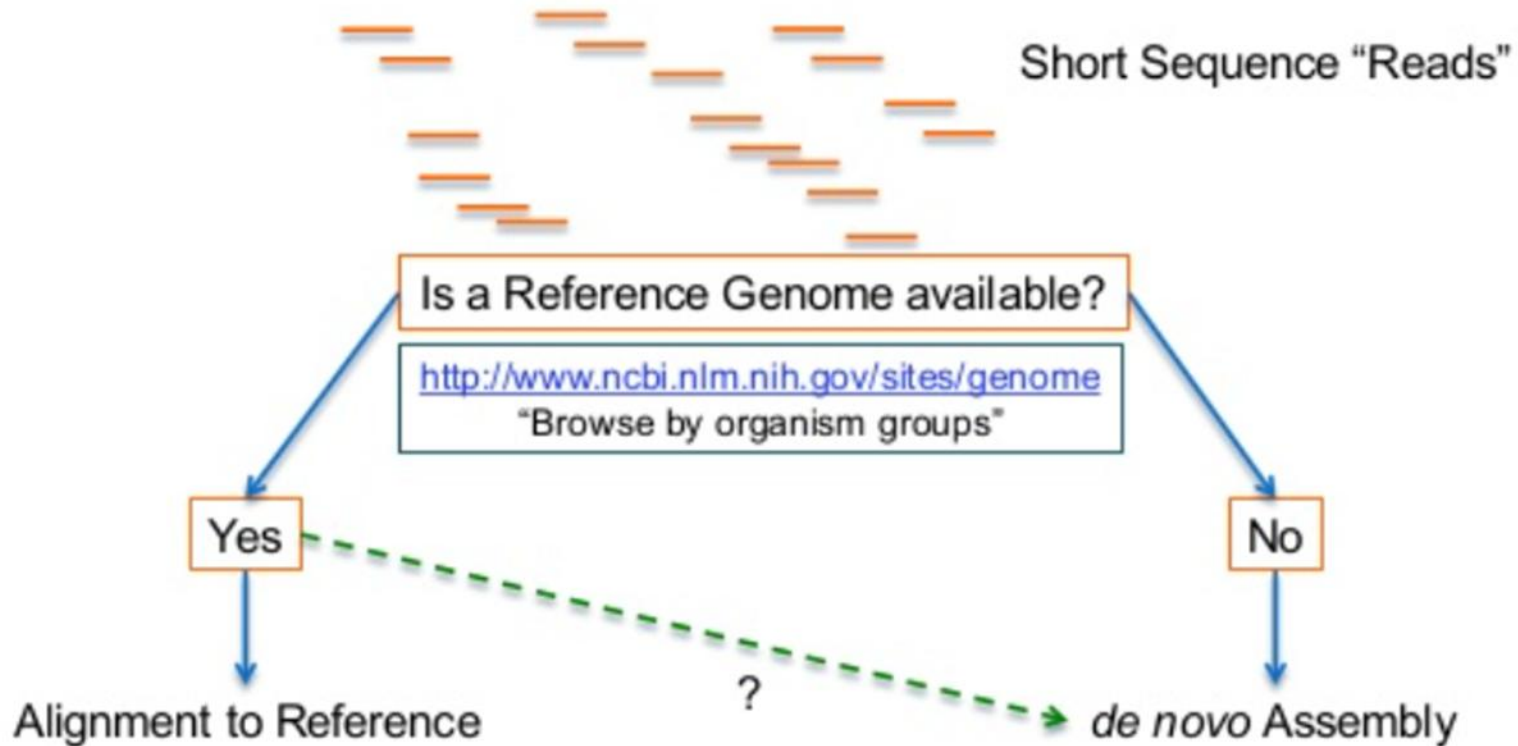
- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✗ [Per base sequence content](#)
- ✗ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✗ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

#### ✗ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCAACGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCAACGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.2350000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTACTTCTACCACTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

Link: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

# Mapping of reads





# Alignment to reference genome

reference genome sequences  
(FASTA files)  
annotations (GTF file)  
FASTQ sequence reads

Reads mapping to the genome  
(STAR, GSNAP, TopHat)

Read

ATGTTGTACCCTAGTTTACCCCCAGAGTG x Billions

Reference

GCGATATAGTATGATATGATAACCCACACCCACACGTGCACATG  
CTCATGTTGTACCCTAGTTTACCCCCAGAGTGCGCGCGTTATA  
CTCTCTTTTTCATCAAACTACTACACACATTTTCATCACTACCA  
CTGGTGAGCATGGGTACGACTGGACTGACTGACTGGACTGA  
CTGCATGCATCGACCTACTGACTCGACTACGCATGCATGCATG  
CTAGCTGACTGACCATGACTGACTGACTAGCTAGCCAATGCAT  
ACTGACTGACTAGCTGACTAGCCCTAGCATGCATGCCTGCACT  
CTGACTAGCTGGCTGACTGACCTGACCTGACTGCCGCCACGT  
CGACTGACTGCCATGCCCGCGCCGTTTATATATTATGCGCACT  
CGTATGCATTGCGATCGACTGAAACTGACTGACGCCTGCCTT  
CTTCGACCGACTAGCGCGCGGCAGTCTACGGCATGCATGGTT  
CGTGAGGATCGACGACCTGACACTGACCTGATAGCCATGCAT  
CGTACGACTGCATGCATCGCAGCACTAGCCATGCATCGATCGC  
TGCTAGCATGCCCTGCTGACCATATATTCTGCATGCATGCACA

x Millions



Reference genome

ATCGGACGTAGCACCTAGTTCAGGTATGCCCTTGG

Reference genome + Annotation file

Gene 1 Gene 2 Gene 3 Gene 4

Sample 1



Sample 2



CTCTGCACGCGTGGGTTTGAATCCACCTTCGTCTGA

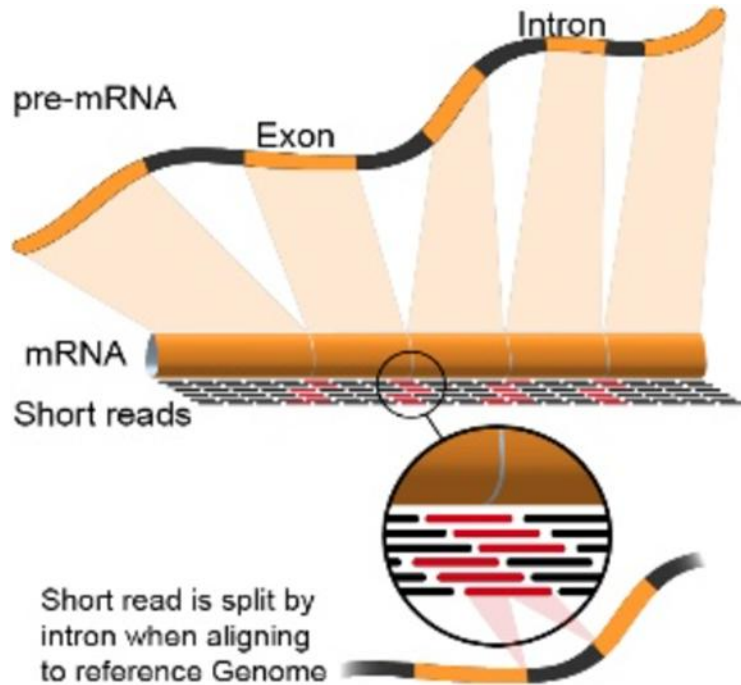
Coordinate:

chr6 27,373,801

chr6



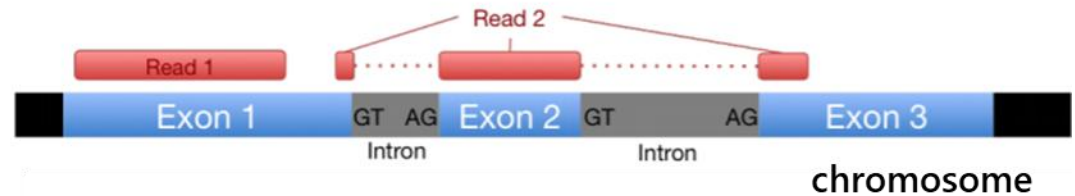
# Alignment to reference



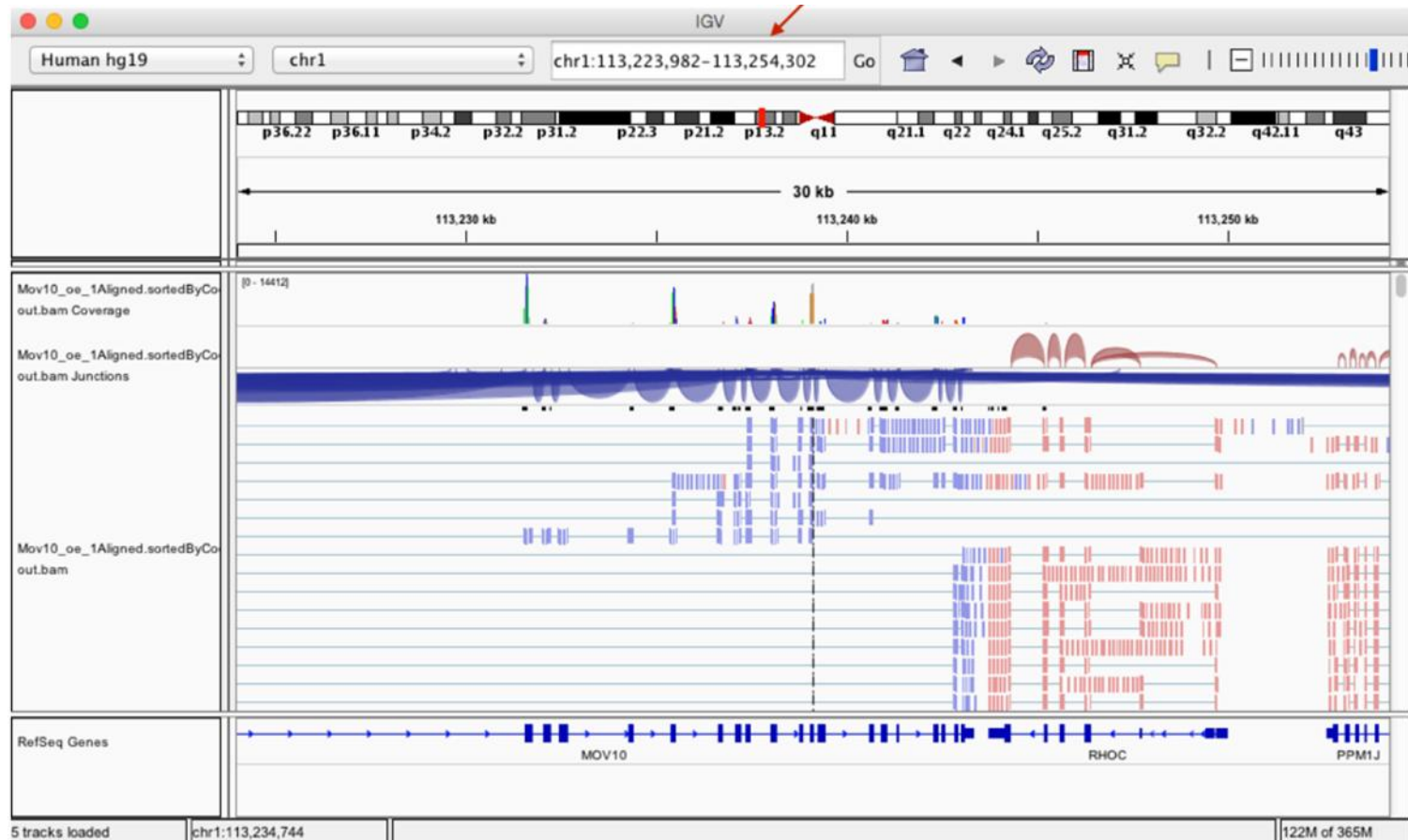
(a) Aligning to the transcriptome



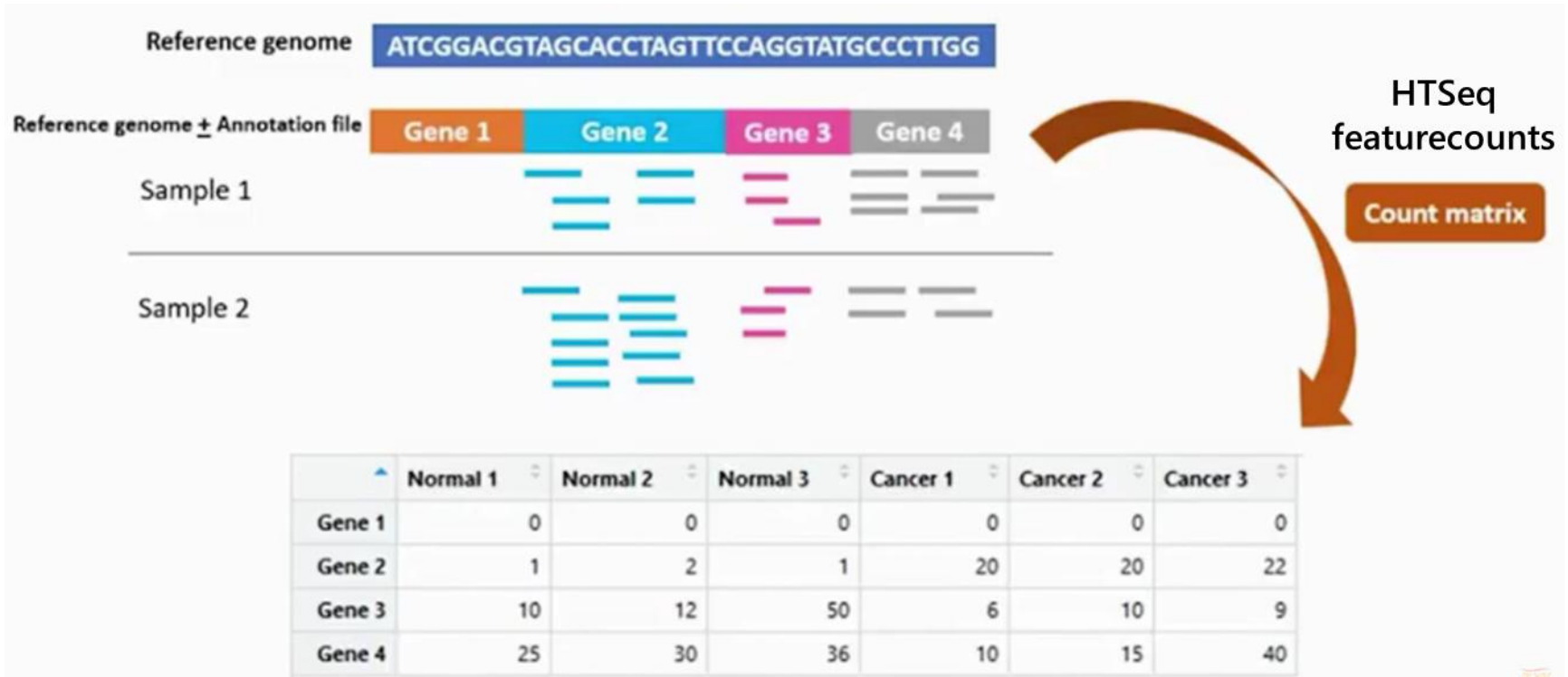
(b) Aligning to the genome



# Alignment to reference genome

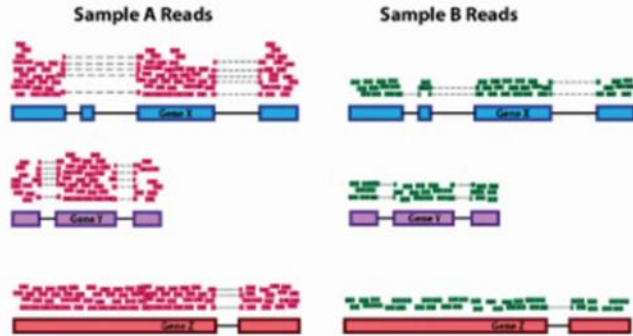


# Read counts

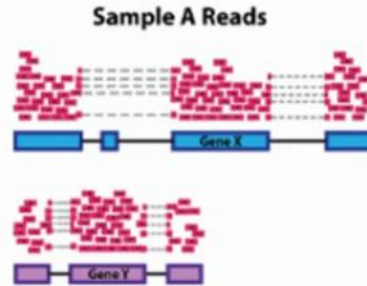


# Normalization

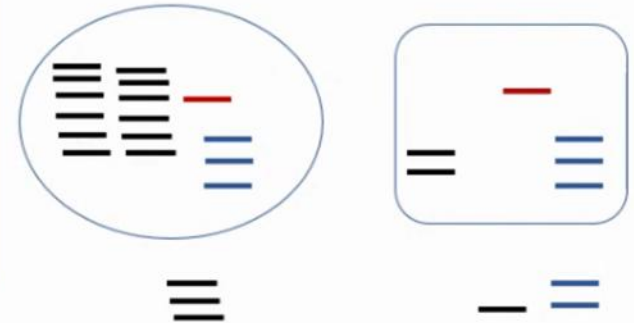
## A. Sequencing depth (library size)



## B. Gene length



## C. RNA composition



Read 1: CGGATTACGTGGACCATG (read length of 18)  
Read 2: ATTACGTGGACCATGAATTGCTGACA  
Read 3: ACCATGAATTGCTGACATTCGTCA  
Read 4: TGAATTGCTGACATTCGTCA

Depth: 11122222222333344333333333332222221

# Normalization

RPKM - Reads Per Kilobase of transcript per Million reads mapped

FPKM - Fragments Per Kilobase of transcript per Million mapped reads

CPM - Counts Per Million

TPM - Transcripts Per Million

Gene B is twice as long as gene A, and that might explain why it always gets twice as many reads, regardless of replicate.

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Genes

Replicates (Samples)

Rep3 has way more reads than the other replicates, regardless of the gene.

# Normalization

RPKM – step 1: normalize for read depth.

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Total reads: 35                      45                      106

Tens of reads: 3.5                      4.5                      10.6

These are our “per million” scaling factors

For the purpose of this 4 gene example, we’re scaling the total read counts by 10 instead of 1,000,000.

RPM - scaled  
using the “per  
million” factors.

Gene Name	Rep1 RPM	Rep2 RPM	Rep3 RPM
A (2kb)	2.86	2.67	2.83
B (4kb)	5.71	5.56	5.66
C (1kb)	1.43	1.78	1.43
D (10kb)	0	0	0.09



# Normalization

RPKM – step 2: normalize for gene length.

Gene Name	Rep1 RPM	Rep2 RPM	Rep3 RPM
A (2kb)	2.86	2.67	2.83
B (4kb)	5.71	5.56	5.66
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.09



Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009

Reads are scaled  
for depth (M)  
and gene length  
(K).



# Normalization

TPM – step 1: normalize for gene length

Original data:

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

RPK – scaled by  
gene length:

Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

TPM – step 2: normalize for sequencing depth

Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

Total RPK: 15 20.25 45.1  
Tens of RPK: 1.5 2.025 4.51

TPM – scaled by  
gene length and  
sequencing  
depth (M):

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

# Normalization

## RPKM vs TPM

Both TPM  
RPKM (and FPKM)  
correct for biases in gene  
length and sequencing  
depth. But....

RPKM

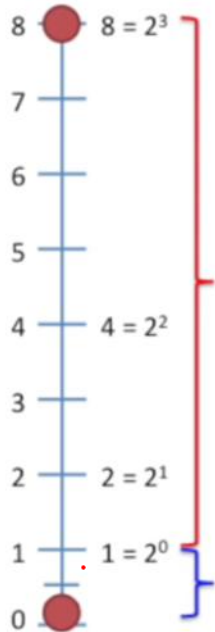
... the sums of each  
column are very  
different.

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

TPM

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10

# Log fold change

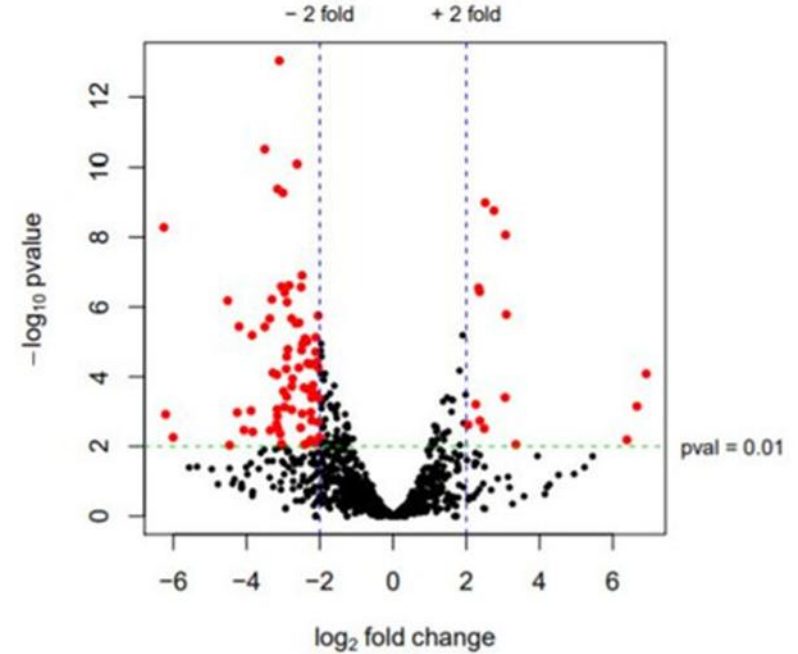
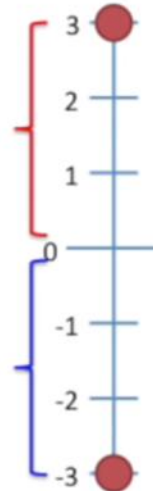


Take home message so far...

- 1) "logs" isolate exponents.

$$\log_2(8) = \log_2(2^3) = 3$$

- 2) Use a log scale/axis when talking about fold change. This puts positive and negative fold changes on a symmetric scale.



# Differential gene expression analysis

---

## Statistical hypothesis

You need to make some assumptions to analyze this data, this assumption is called a hypothesis.

e.g. Is there any difference between the cases (e.g. Control/disease)?

So, if there is no difference, we call this null hypothesis  $H_0$ :

$$H_0: d = 0$$

If there is a difference, we call this the alternative hypothesis  $H_1$ :

$$H_1: d \neq 0$$

# Differential gene expression analysis

---

## Testing of hypothesis

### Parametric test

If your data, follow the normal distribution

Examples: T test ANOVA test

### Nonparametric test

If your data doesn't follow the normal distribution

Examples: Wilcoxon test Kruskal-Wallis Test

The output of both tests is a statistical value (p-value).

A p-value less than 0.05 is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than 5% probability that the null hypothesis is correct or the error probability in the alternative hypothesis is less than 5%.

# Differential gene expression analysis

---

## Multiple hypothesis testing



In RNA-seq differential expression, we usually run hundreds or thousands of comparison tests in a single study, as each gene is tested separately for being differentially expressed that will increase the chance of false positives.

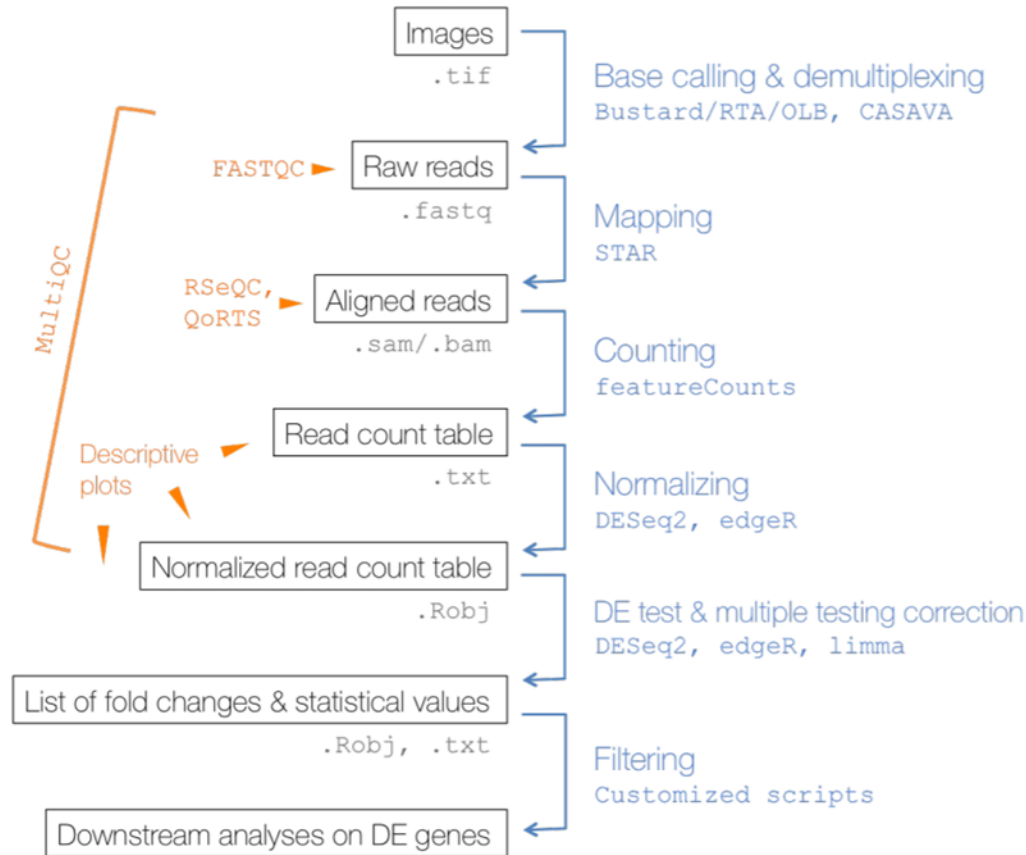


E.g. suppose you have 20,000 genes on RNA-seq experiment. If p value is 0.05 You would expect  $20,000 \times 0.05 = 1,000$  of them to have a p-value < 0.05 by chance.



Individual p-values of e.g 0.05 no longer correspond to significant findings, it's need to be adjusted for multiple testing using methods like False discovery rate (FDR)

# Workflow of differential gene expression analysis



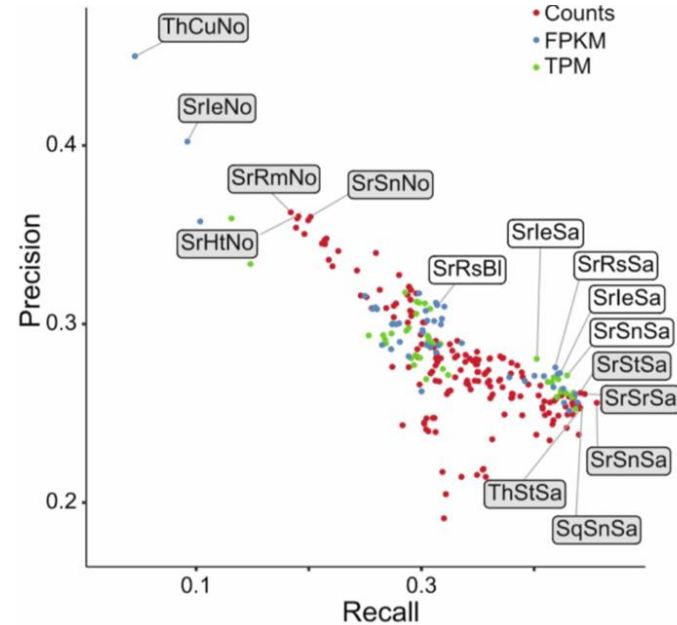
# Choice of tools matters

## Precision vs. recall tradeoff when using RNA-seq analysis pipelines

**Table 1** Analysis tools used in this study

Read aligner	RA code	Expression modeler	EM code	Differential expression	DE code
Bowtie2	Bw	BitSeq	Bs	Ballgown	Bl
HISAT2	Hs	cufflinks	Cu	BitSeq	Bs
Kallisto	Ka	htseq	Ht	baySeq	By
Salmon-FMD	Sf	IsoEM	Ie	cuffdiff	Cd
Sailfish	Sl	kallisto	Ka	DESeq2	De
SeqMap	Sm	RSEM	Rm	EBseq	Eb
Salmon-Quasi	Sq	rSeq	Rs	edgeR	Er
STAR	St	Sailfish	Sl	limma + voom	Lo
TopHat2	Th	Salmon	Sn	limma + vst	Lv
		STAR	Sr	NBPseq	Nb
		Stringtie	St	NOISeqBio	No
		eXpress	Xs	SAMseq	Sa
				Sleuth	Su

Williams, et.al. BMC Bioinformatics 2017  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5240434/>



**Precision** ( $tp/(tp+fp)$ ) for a class is the *number of true positives divided by the total number of elements labeled as belonging to the positive class.*

**Recall** ( $tp/(tp+fn)$ ) is defined as the *number of true positives divided by the total number of elements that actually belong to the positive class*



# Biomarkers identified using RNA-seq

From: [RNA sequencing: new technologies and applications in cancer research](#)

Cancer type	Biomarker name	Biomarker type	Up/Down	Value	References
Liver cancer	tRNA-ValTAC-3/tRNA-GlyTCC-5/tRNA-ValAAC-5/tRNA-GluCTC-5	tsRNA	Up	Diagnostic	[92]
Lung cancer	ACVR2B-AS1	LncRNA	Up	Prognostic/therapeutic target	[93]
	LINC01537	LncRNA	Down	Prognostic/therapeutic target	[94]
	circFARSA	CircRNA	Up	Noninvasive biomarker	[95]
	LINC01123	LncRNA	Up	Prognostic/therapeutic target	[96]
Gastric cancer	CTD2510F5.4	LncRNA	Up	Diagnostic/prognostic	[97]
	MEF2C-AS1/FENDRR	LncRNA	Down	Diagnostic/prognostic	[98]
Prostate cancer	PSLNR	LncRNA	Down	Diagnostic/therapeutic target	[99]
Colorectal cancer	RAMS11	LncRNA	Up	Therapeutic target	[100]
	CRCAL-1/CRCAL-2 /CRCAL-3/ CRCAL-4	LncRNA	Up	Therapeutic target	[101]
Colon cancer	AFAP1-AS1	LncRNA	Up	Prognostic/ therapeutic target	[102]
Head and neck squamous cell carcinoma	LINC00460	LncRNA	Up	Prognostic	[103]
	HCG22	LncRNA	Down	Prognostic	[104]
	HOXA11-AS/LINC00964/MALAT1	LncRNA	Up	Diagnostic	[105]
Clear cell renal cell carcinomas	SLINKY	LncRNA	Up	Prognostic	[106]
Leukemia	LUCAT1	LncRNA	Up	Therapeutic target	[107]

Link: <https://jhoonline.biomedcentral.com/articles/10.1186/s13045-020-01005-x>