

Biomarkers for disease identification/outcome



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

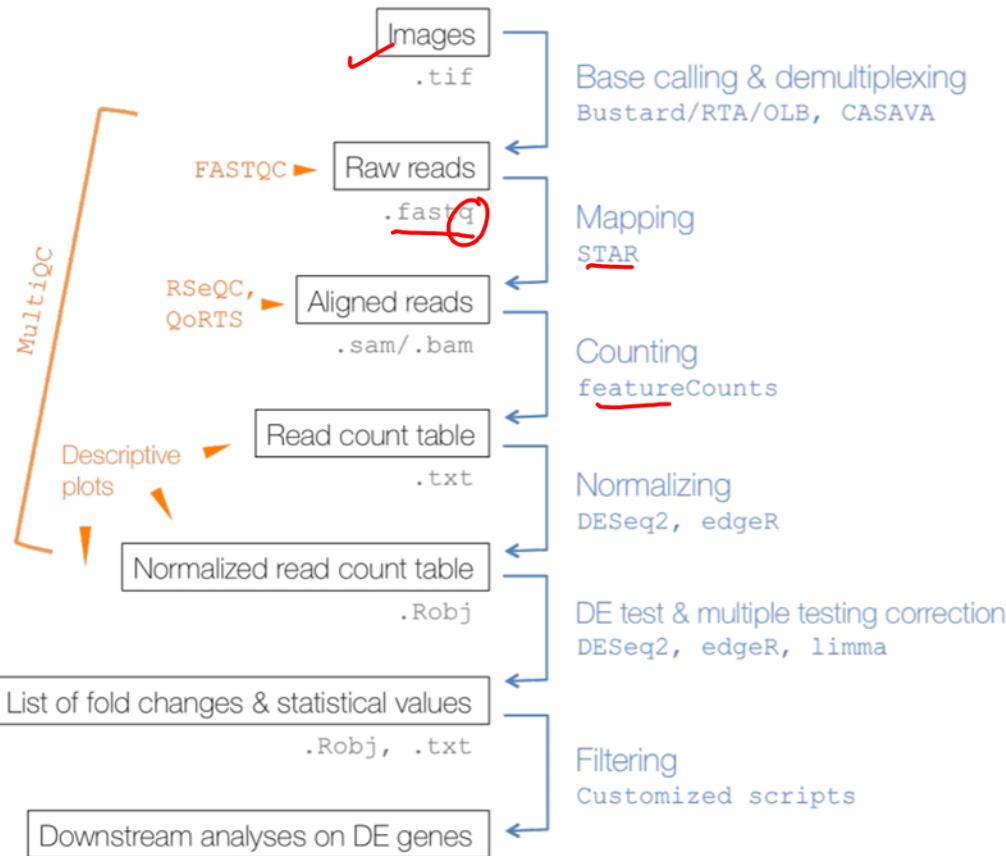
Dr. Jaspreet Kaur Dhanjal

Assistant Professor, Department of Computational Biology

Email ID: jaspreet@iiitd.ac.in

October 28, 2025

Workflow of differential gene expression analysis



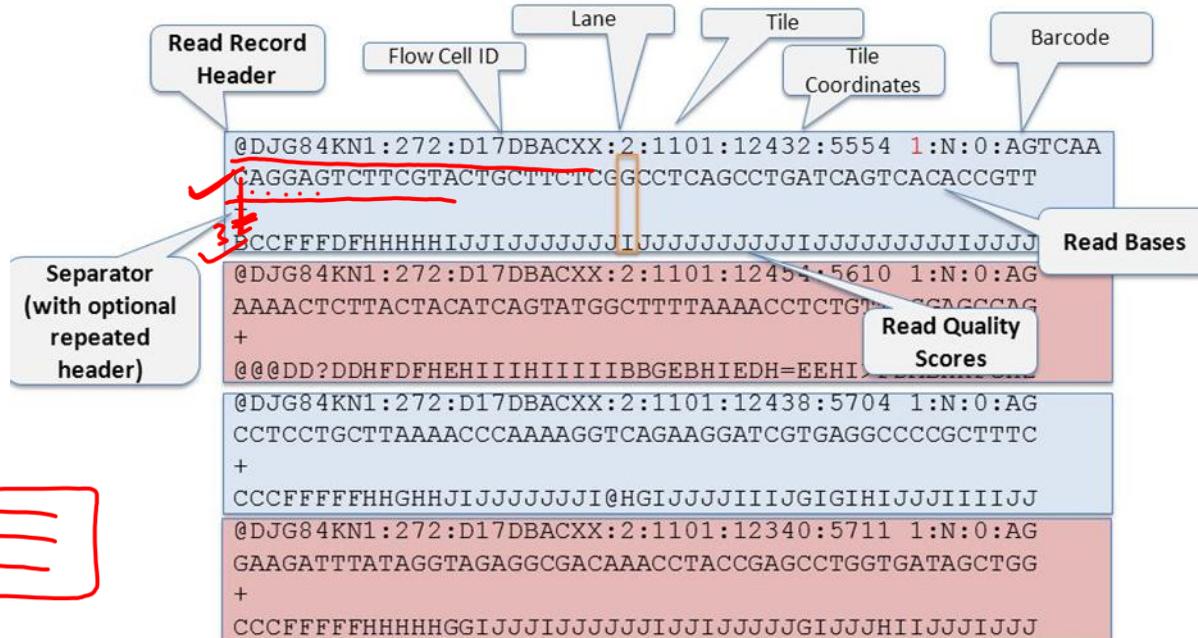
Gene expression-based biomarker identification

Problems in sequencing

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

$$65 - 33$$

Phred
10-60

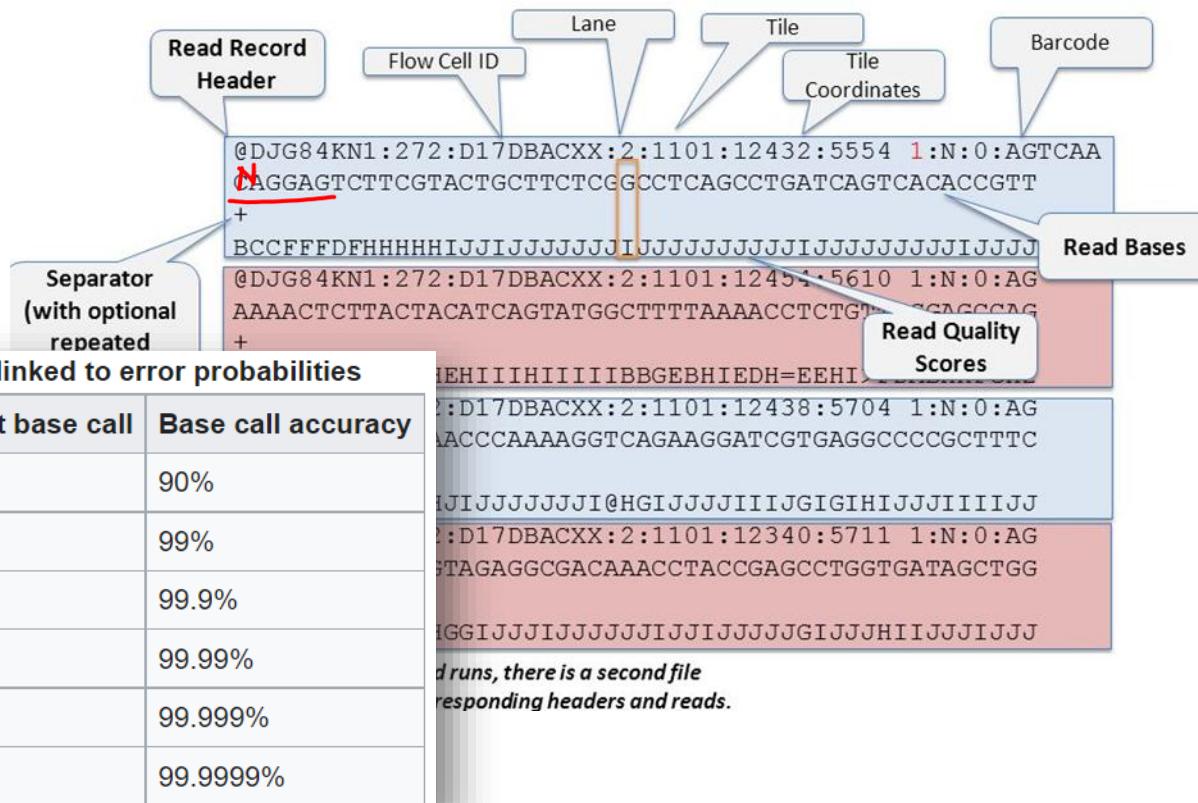


NOTE: for paired-end runs, there is a second file with one-to-one corresponding headers and reads.

Gene expression-based biomarker identification

Problems in sequencing

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



Quality check

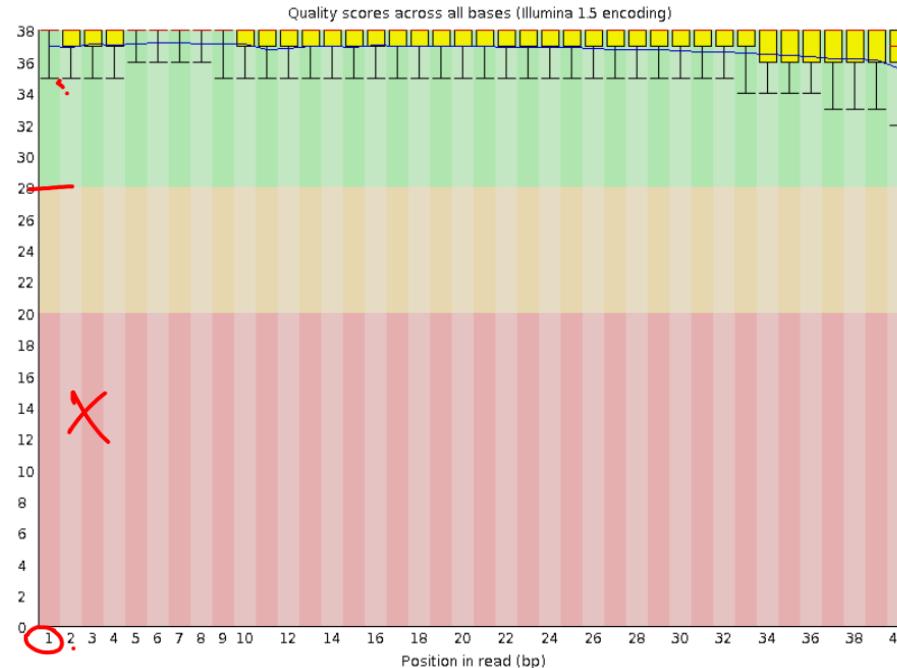
FastQC Report

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

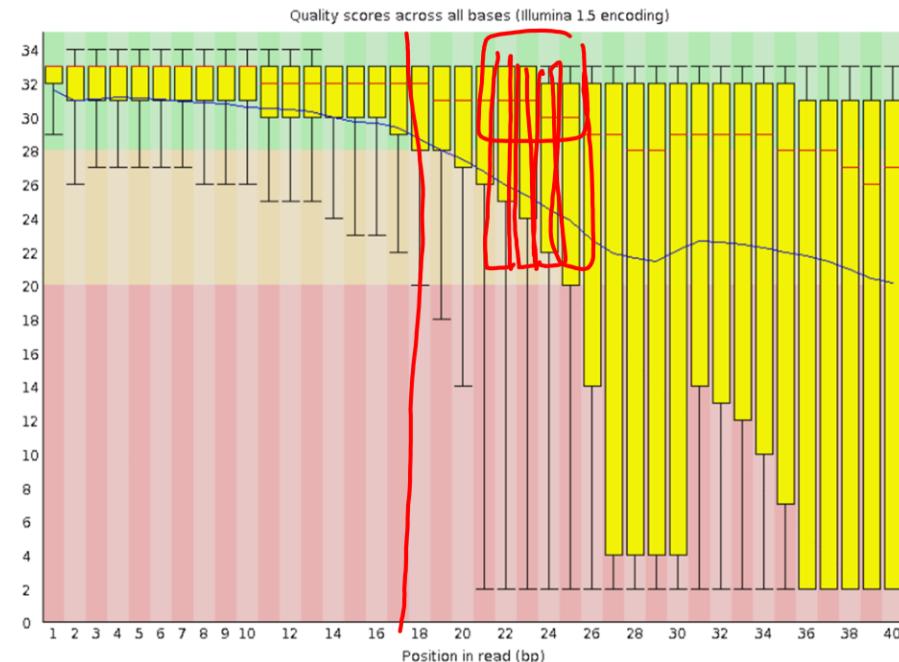
FastQC Report

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

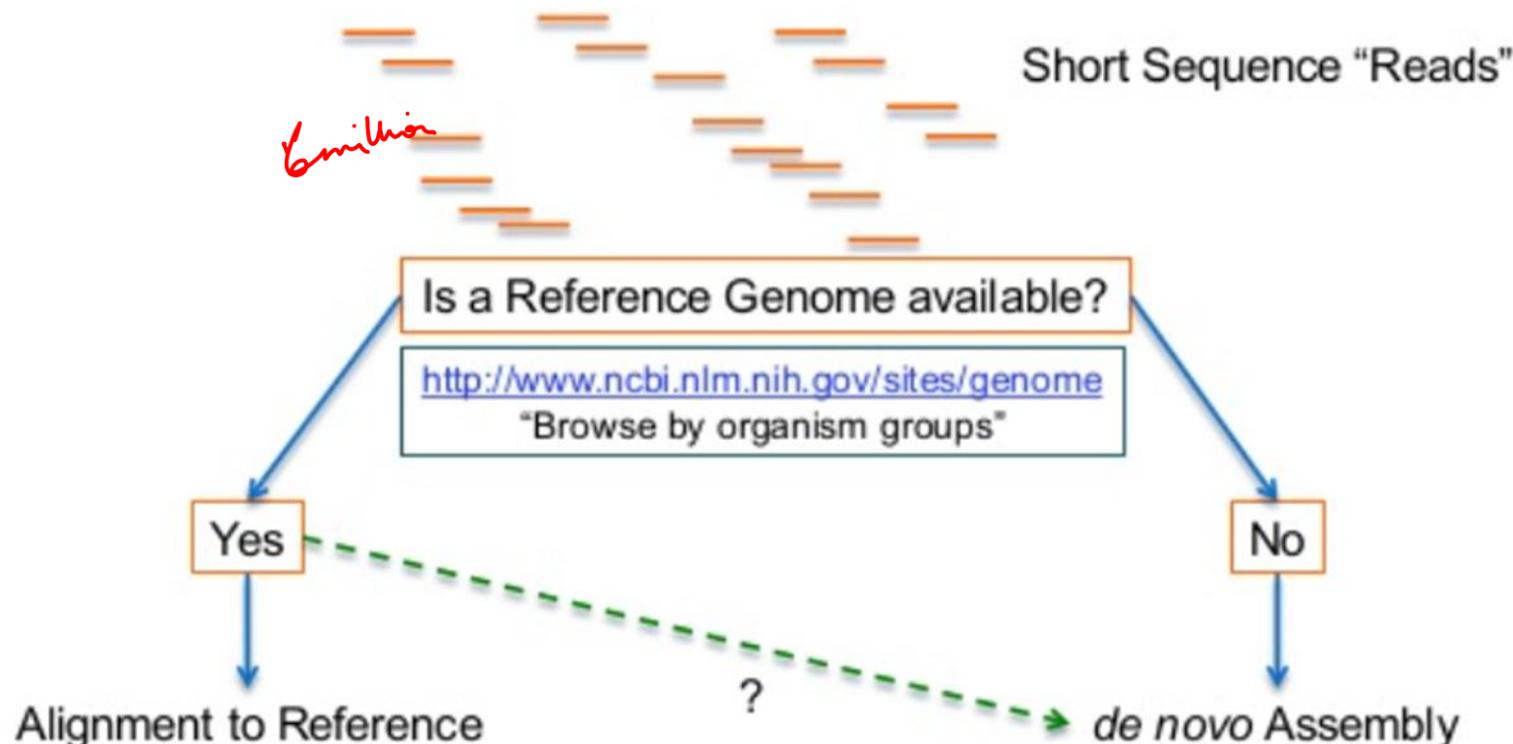
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented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

 **Overrepresented sequences**

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATA CGCGCACCACCGAGATCTACACTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCACCACCGAGATCTACACTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGATCGGAAG	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGGATCGGA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGGTGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGACCGGA ACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTACCA CCTATATCTACACTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCA CGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTCA CGCAGGAATGCCGAGATCGGAAGAGCGGTTCA CGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

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

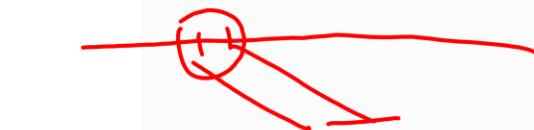
Mapping of reads



Alignment to reference genome

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

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

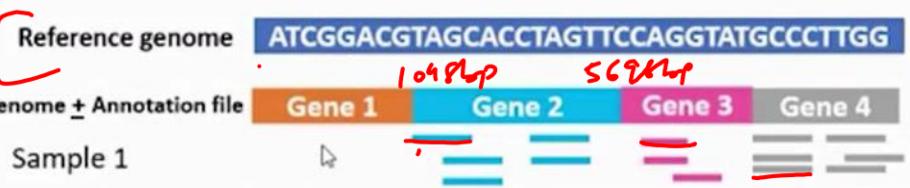


Read

ATGTTGTACCTAGTTACCCCCAGAGTG x Billions

Reference hg38 DNAphy

hg38 DNAphy
GCGATATAGTATGATA[GATAACCACACCCACACG]GCACATG x Millions
CTCATGGTGTACCTAGTTACCCCCAGAGTGC[GCGCG]TATA
CTCTCTCATCAAAC[TACTA]ACACATTTCATCACTACCA
CTGGTGAGCATGGTAGCTGGACTGACTGACTGACTGGACTGA
CTGCATGCATGACCTACTGACTGACTACGCGATGCATGCATG
CTAGCTGACTGACCATGACTGACTGACTAGCTAGCCAATGCAT
ACTGACTGACTGACTGACTAGCCCTAGCATGCATGCCCTGCACT
CTGACTAGCTGGCTGACTGACCTGACCTGACTGCCGCCACGT
CGACTGACTGCCATGCCCGCGCCGTTATATATTATGCGCACT
CGTATGCATTGCGATCGACTGAAACTGACTGACGCCCTGCTT
CTTCGACCGACTAGCGCGCGGAGCTACGGCATGCATGGTT
CGTGAGGATCGACGACCTGACACTGACCTGATAGCCATGCA
CGTACGACTGCATGCATGCACGACTAGCCATGCATGATCGC
TGCTAGCATGCCCTGCTGACCATATATTCTGCATGCATGCACA



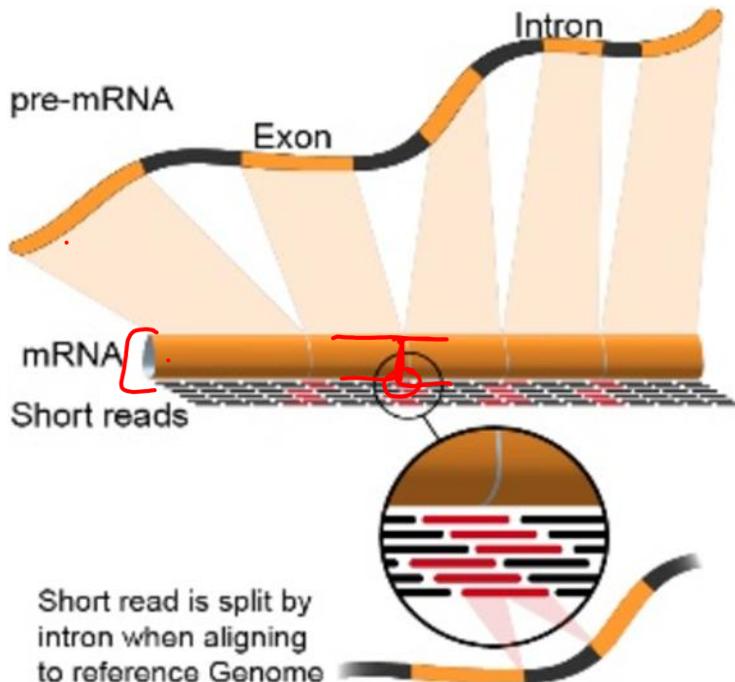
CTCTGCACGCGTGGTTCTGAATCCCACCTTCGTCGA

Coordinate:

chr6 27,373,801



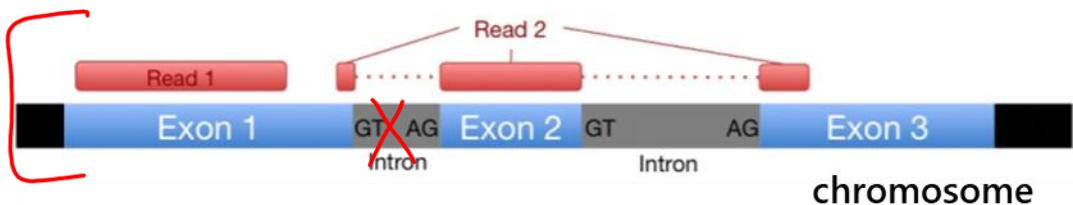
Alignment to reference



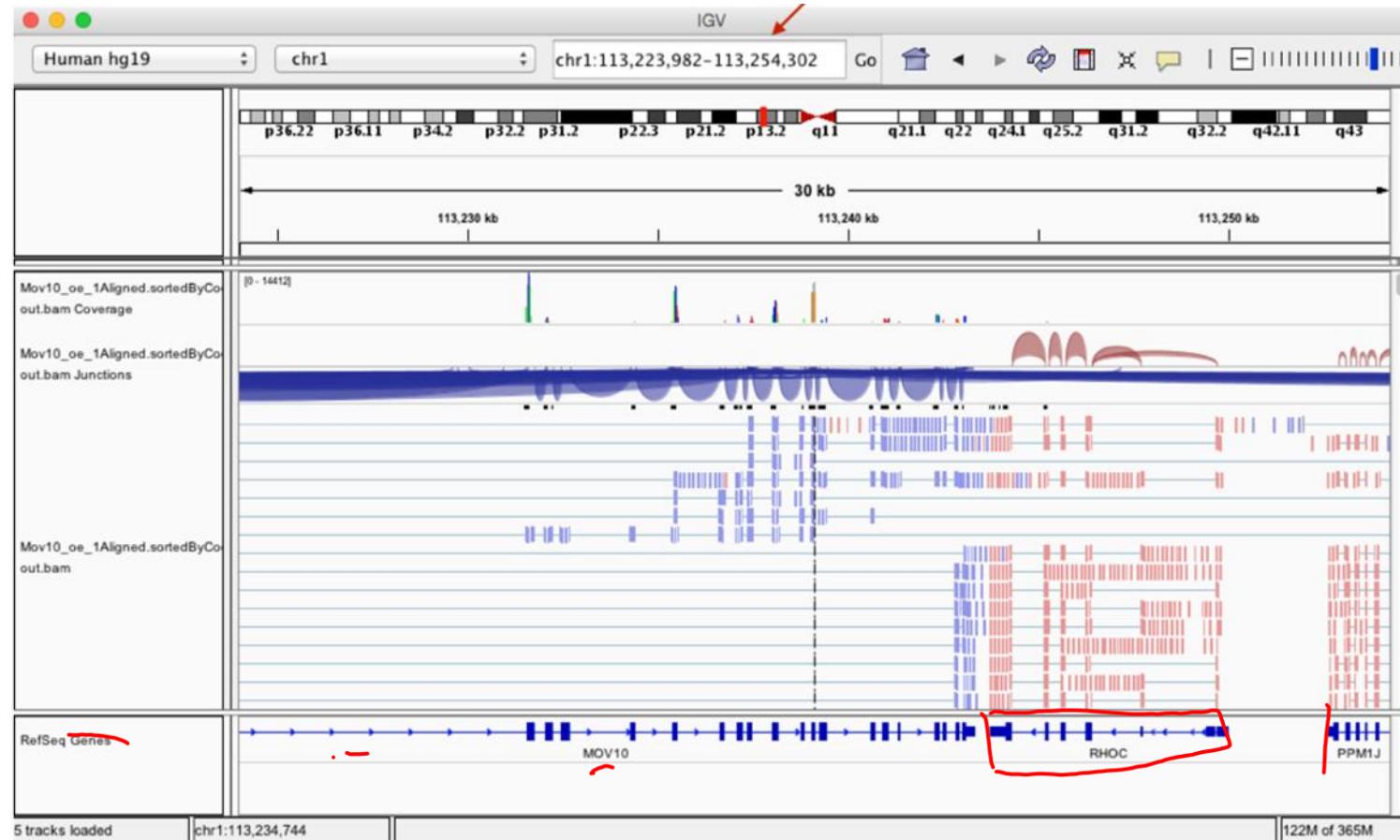
(a) Aligning to the transcriptome



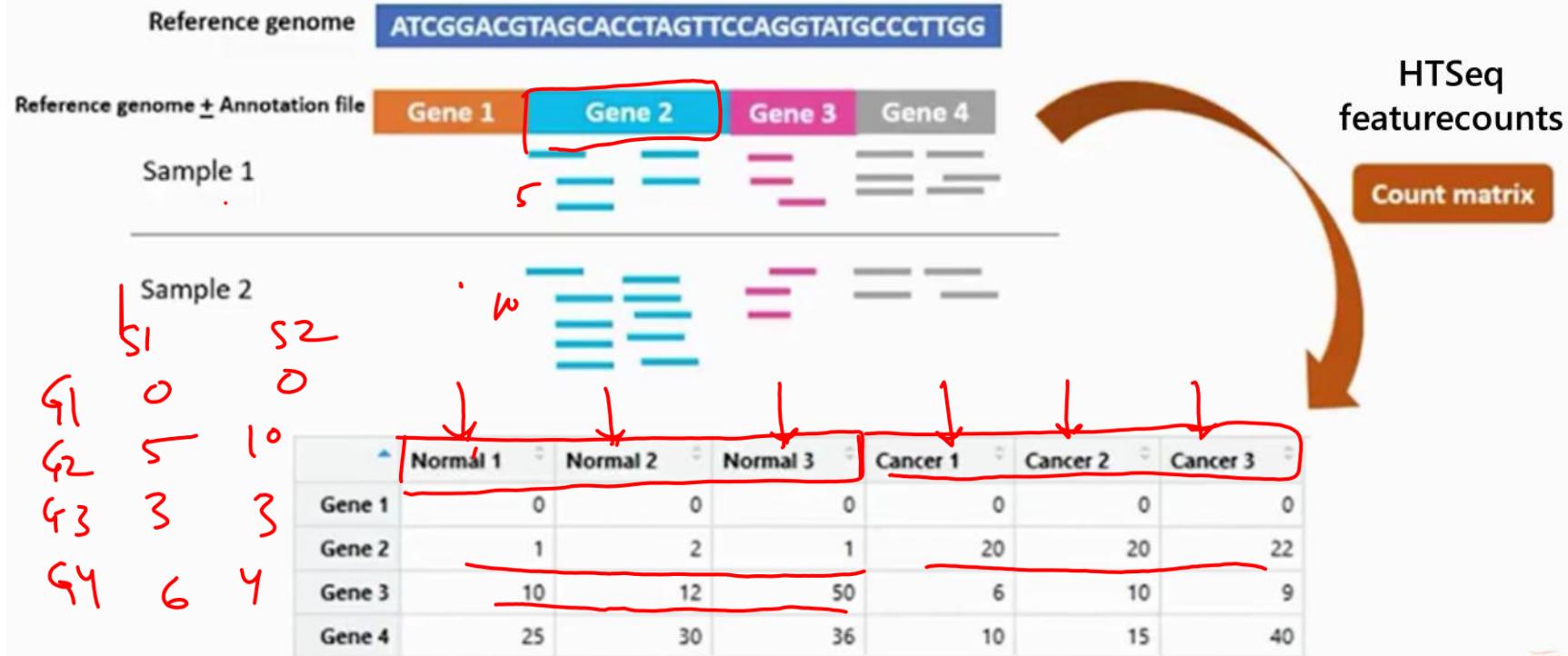
(b) Aligning to the genome



Alignment to reference genome

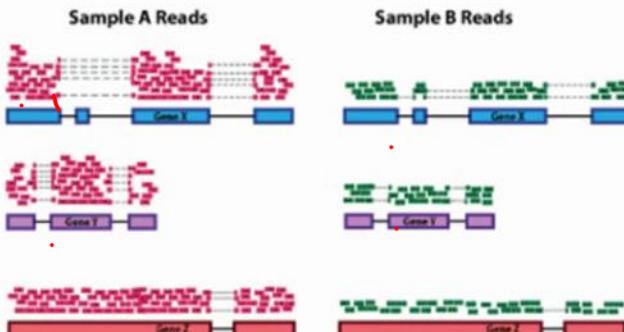


Read counts

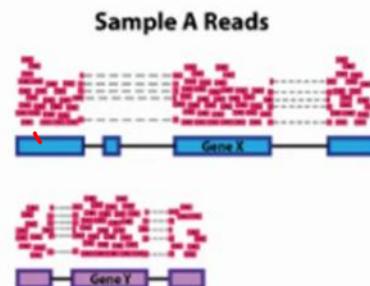


Normalization

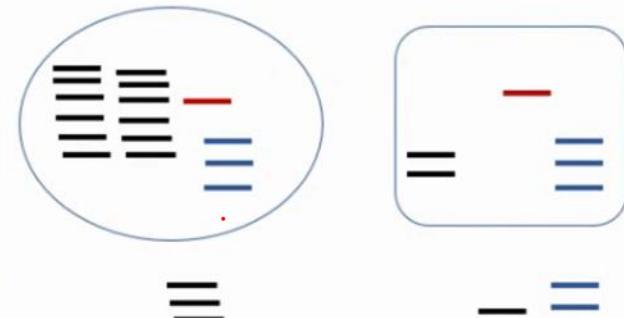
A. Sequencing depth (library size)



B. Gene length



C. RNA composition



Read 1: CGGATTACGTGGACCATG (read length of 18)
Read 2: ATTACGTGGACCATGAATTGCTGACA
Read 3: ACCATGAATTGCTGACATTCTGTCA
Read 4: TGAATTGCTGACATTCTGTCA

Depth: 1112222222233334433333333332222221

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.

Genes	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

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



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

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

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

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

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 – 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

$$GA - \frac{\text{AvgD}}{\text{AvgN}} = \frac{8}{8} - 8$$

the sums of each column are very different.

$$GC - \frac{\text{AvgD}-1}{\text{AvgN}=8} = \frac{1}{8} \text{ TPM}$$

$$= 0.125$$

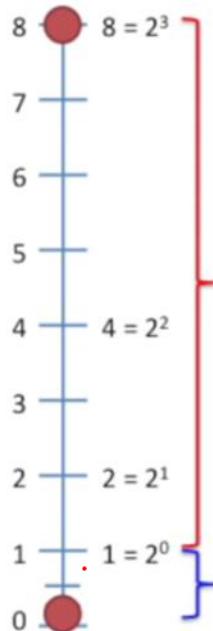
Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

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

Arg D.

D1	D2	D3
3.3	3.1	3.0
15.0	15.2	14.8
0.2	1.6	2.6

Log fold change

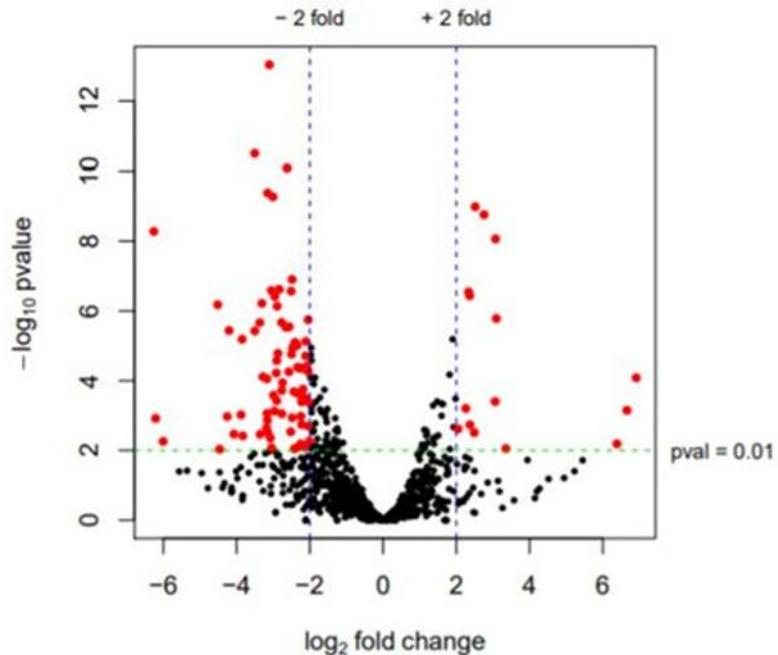
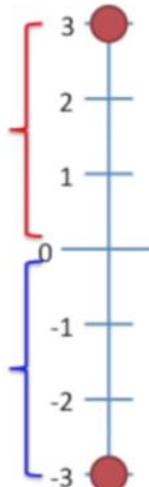


Take home message so far...

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 H0:

$H_0: d = 0$

If there is a difference, we call this the alternative hypothesis H1:

$H_1: d \neq 0$

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.

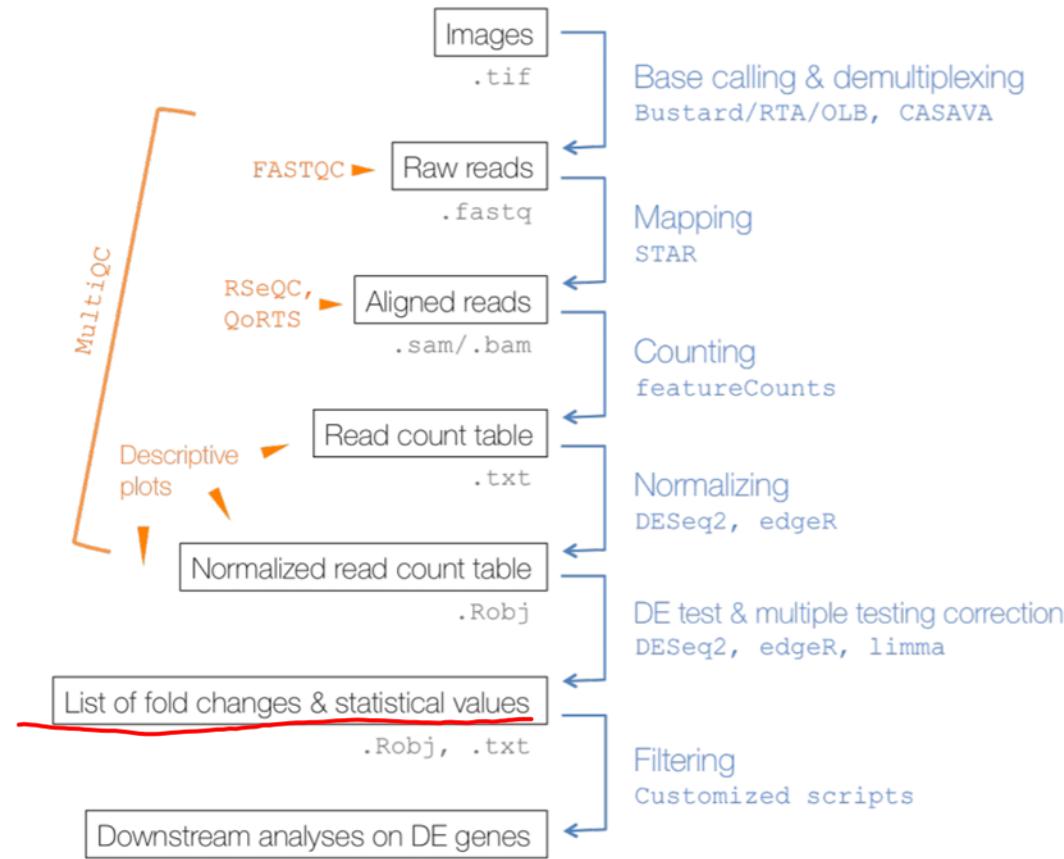


E.g. suppose you have 20,000 genes on RNA-seq experiment. If p value is 0.05 You would expect $20,000 \times 0.05 = 1,000$ of them to have a p-value < 0.05 by chance.



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

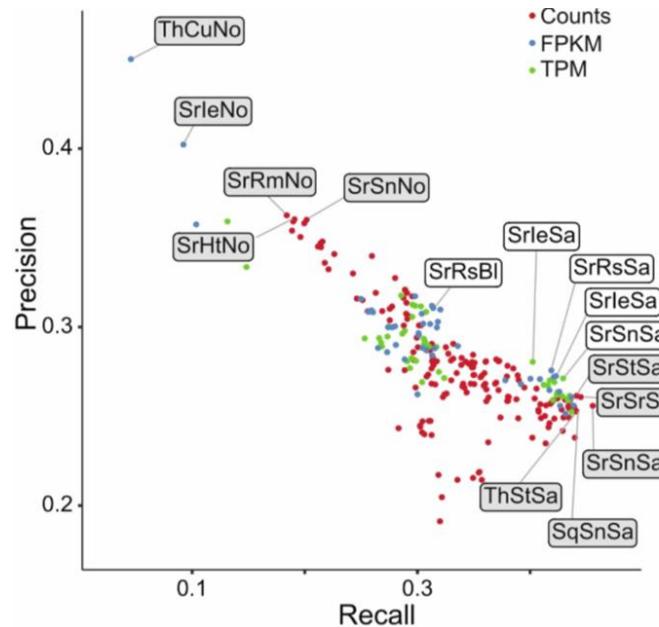


Choice of tools matters

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	Sr	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



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]
	ACVR2B-AS1	LncRNA	Up	Prognostic/therapeutic target	[93]
Lung cancer	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>