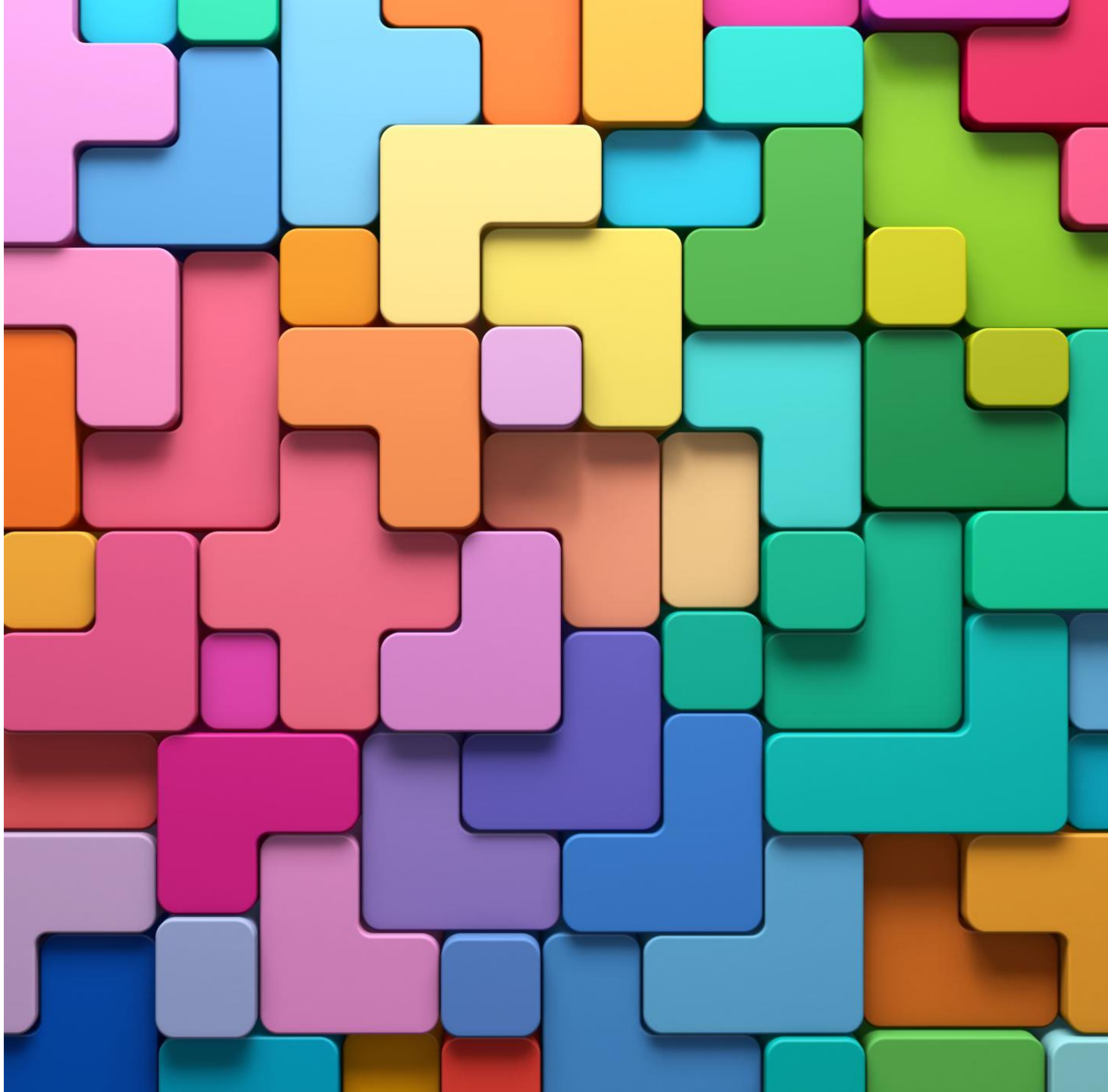
Bulk RNA sequencing analysis

Dr. Aatish Thennavan MDS, PhD

THE UNIVERSITY OF TEXAS
MD Anderson ~~Cancer~~ Center
Making Cancer History®

NAVINLAB



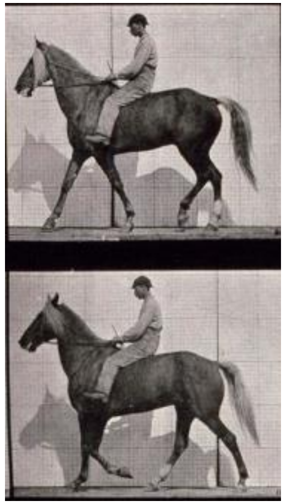
Lecture Overview

- Introduction – Why RNA sequencing & flowchart of steps
- Bulk RNA sequencing applications
- Steps involved in bulk RNAseq computational analysis
- Terminology and file formats
- Mapping and alignment pipelines for bulk RNAseq
- Data Normalization
  - Log normalization
  - Quantile Normalization
  - DESeq2 Normalization – VST
- Dimensionality reduction
  - Principal component analysis (PCA)
  - Hierarchical Clustering
- Basic principles of DESeq2 Differential expressed gene (DEG) analysis
- Basic principles of pathway analysis – GSEA and GSVA

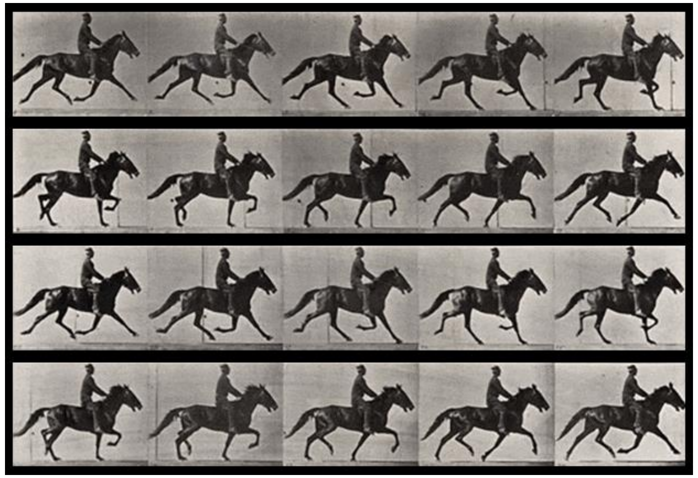
Introduction – Why RNA sequencing

- The cells in our bodies become structurally and functionally diverse by activating different combinations of genes.
- By studying the RNA that is transcribed from these genes, we can find out which genes are active in a particular cell type.
- Measuring **DNA gives us the static painting, but RNA measurements gives us the dynamic motion picture**

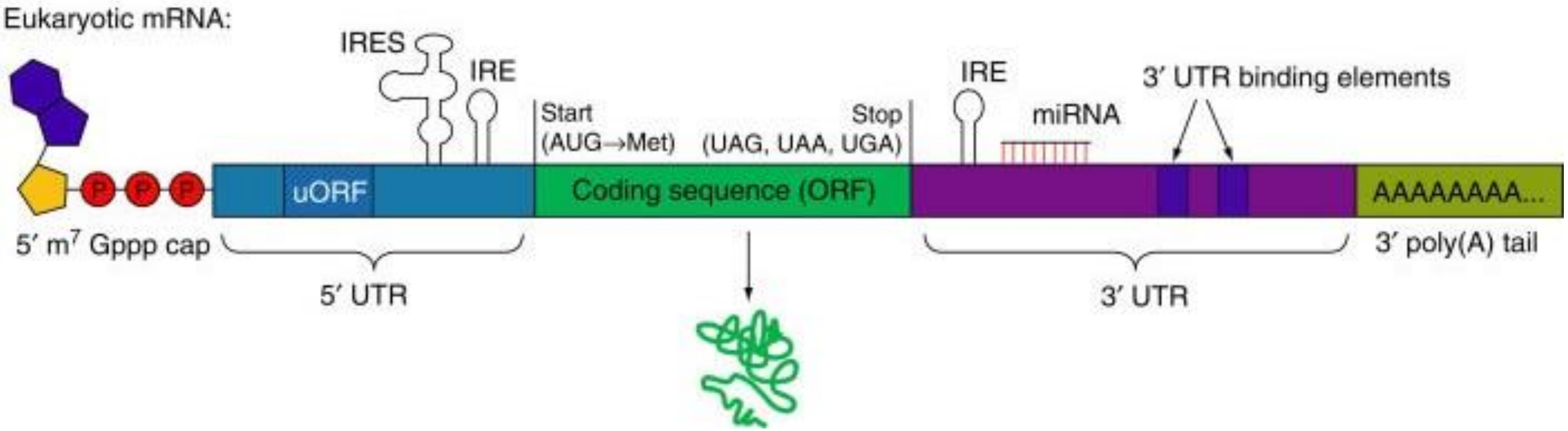
DNA



RNA



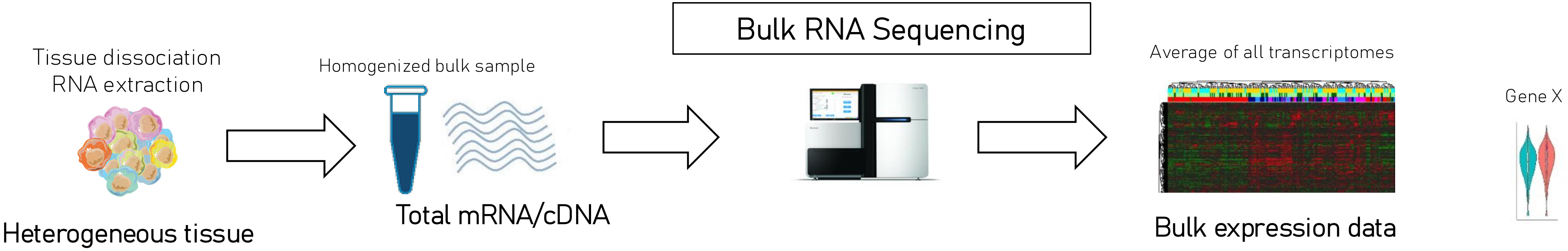
Introduction – What are we sequencing?



- Poly A tail – regulation of mRNA translation, stability, and export
- Three prime untranslated region (3′ UTR) – contains the regulatory regions that post-transcriptionally influence gene expression. Promote proteins and microRNA association with mRNA
- Coding region – codes for proteins
- 5′ untranslated region (5′ UTR) & 5′ Cap

https://microbenotes.com/types-of-rna/

Bulk RNA sequencing – Applications



Tissue dissociation
RNA extraction

Heterogeneous tissue

Homogenized bulk sample

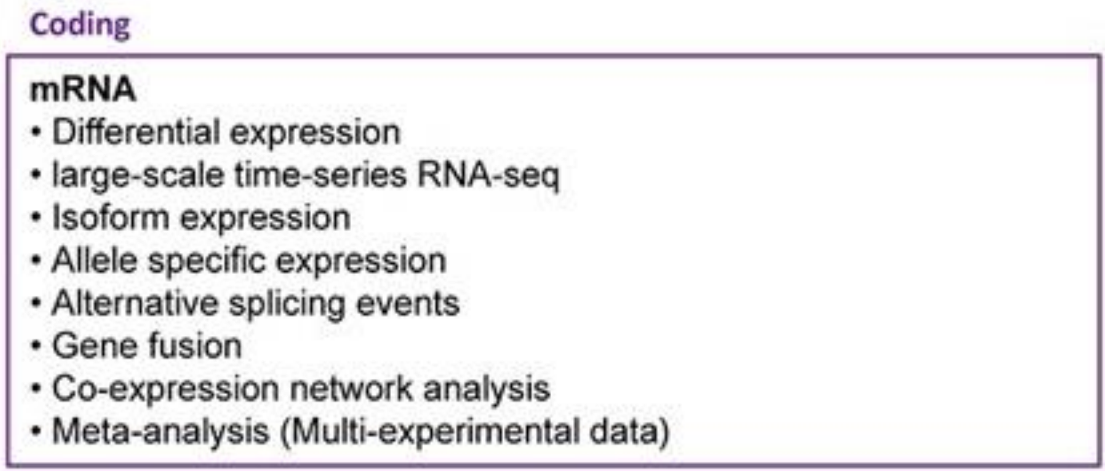
Total mRNA/cDNA

Bulk RNA Sequencing

Average of all transcriptomes

Bulk expression data

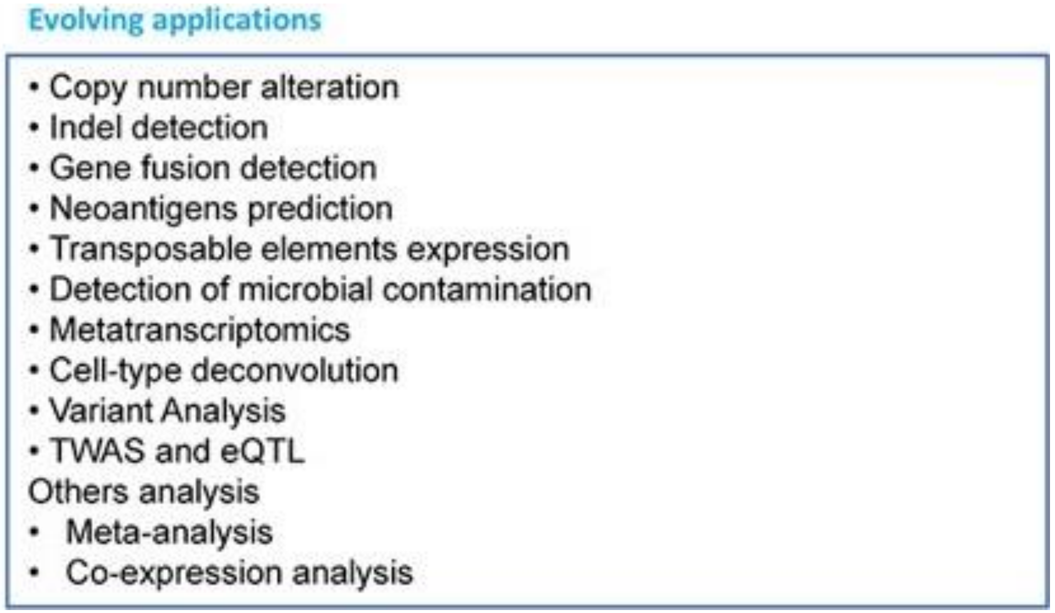
Gene X



Coding

mRNA

- Differential expression
- large-scale time-series RNA-seq
- Isoform expression
- Allele specific expression
- Alternative splicing events
- Gene fusion
- Co-expression network analysis
- Meta-analysis (Multi-experimental data)



Evolving applications

- Copy number alteration
- Indel detection
- Gene fusion detection
- Neoantigens prediction
- Transposable elements expression
- Detection of microbial contamination
- Metatranscriptomics
- Cell-type deconvolution
- Variant Analysis
- TWAS and eQTL

Others analysis

- Meta-analysis
- Co-expression analysis

Briefings in Bioinformatics, Volume 22, Issue 6, November 2021,
bbab259, https://doi.org/10.1093/bib/bbab259

Introduction – Flowchart of every RNA sequencing

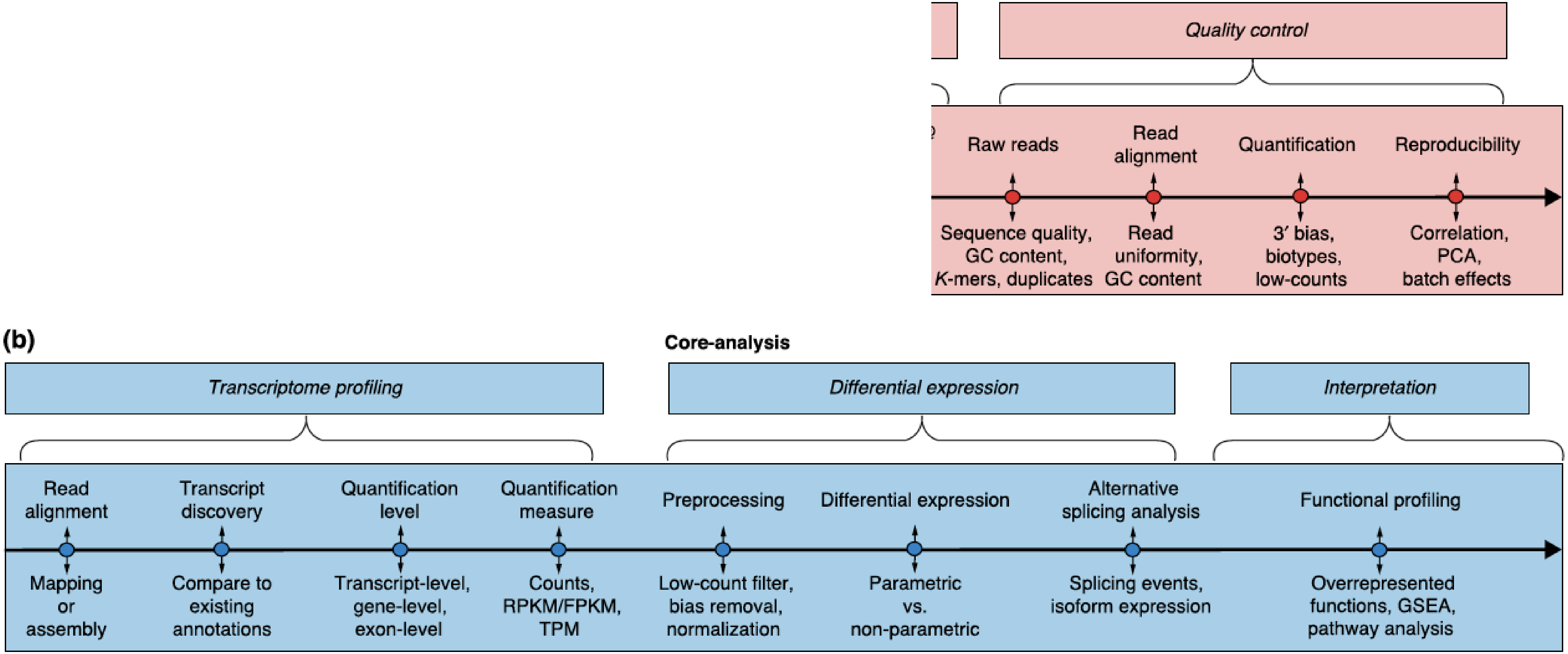
Experimental

1. mRNA isolation/extraction techniques (cell isolation)
2. Quality Check/Quantity
3. Reverse transcription into cDNA
4. Adapted Ligation
5. Amplification
6. Sequencing

Analysis

1. Alignment to a reference genome
2. Quantify transcripts
3. Quality Control
4. Normalization
5. Dimension Reduction
6. Specific analysis –clustering, differential gene expression (DE) analysis etc.

Schematic of bulk RNAseq computational analysis



Quality control

Raw reads — Sequence quality, GC content, *K*-mers, duplicates

Read alignment — Read uniformity, GC content

Quantification — 3′ bias, biotypes, low-counts

Reproducibility — Correlation, PCA, batch effects

(b)

Core-analysis

Transcriptome profiling

Read alignment — Mapping or assembly

Transcript discovery — Compare to existing annotations

Quantification level — Transcript-level, gene-level, exon-level

Quantification measure — Counts, RPKM/FPKM, TPM

Differential expression

Preprocessing — Low-count filter, bias removal, normalization

Differential expression — Parametric vs. non-parametric

Alternative splicing analysis — Splicing events, isoform expression

Interpretation

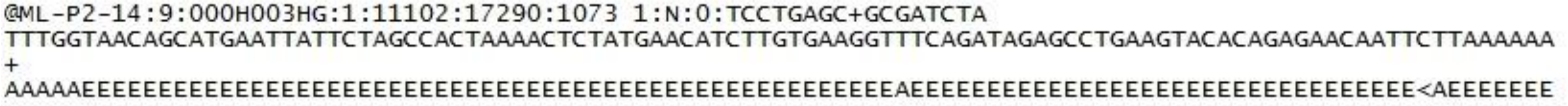
Functional profiling — Overrepresented functions, GSEA, pathway analysis

Conesa, A., Madrigal, P., Tarazona, S. et al. A survey of best practices for RNA-seq data analysis. Genome Biol 17, 13 (2016). https://doi.org/10.1186/s13059-016-0881-8

Frequently used file terms and formats

- BCL file – Binary base call files (.bcl). The file produced via Illumina sequencing. This file contains the base information added in each sequencing cycle and the confidence in the call as a quality score to base call. It is the true raw data output from a sequencing run.

- FASTQ file – (.fastq.gz or .fq.gz) It is generated from the .bcl file. It is text file that consists of 4 lines - A sequence identifier with information about the sequencing run and the cluster, the sequence (the base calls; A, C, T, G and N), A separator, which is simply a plus (+) sign, the base call quality scores. These are Phred +33 encoded, using ASCII characters to represent the numerical quality scores.



```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEEE
```

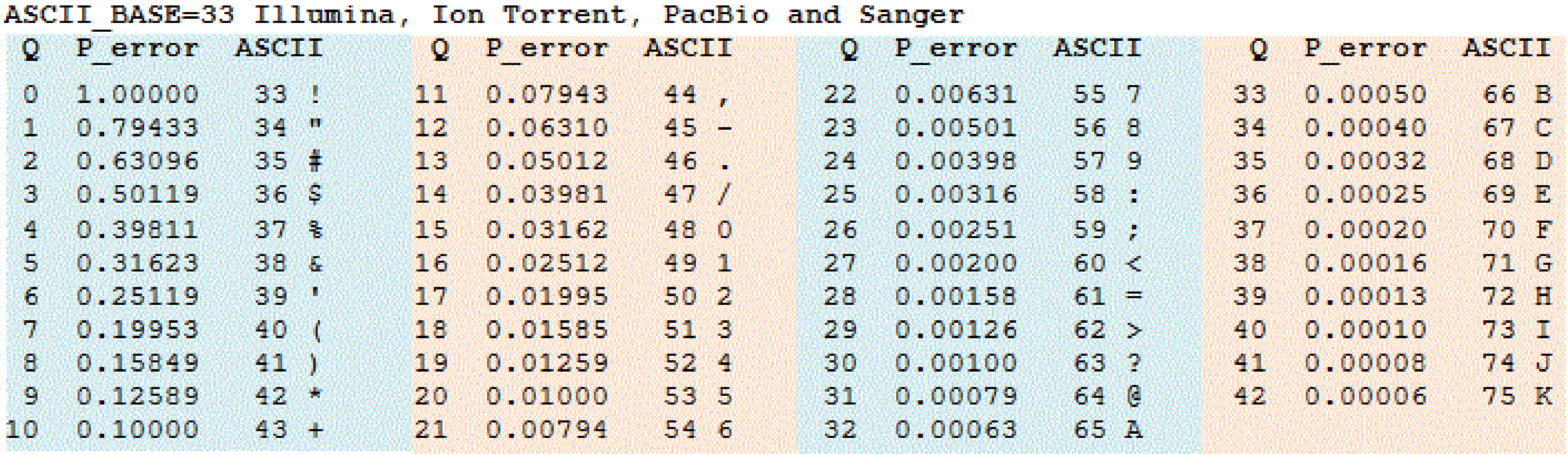
Typically, you will have 2 fastq files per sample in most bulk RNAseq data – paired-end sequencing

Frequently used file terms and formats

The quality score of a base, also known as a Phred or Q score (represented by American Standard Code for Information Interchange – ASCII scores), is an integer value representing the estimated probability of an error, i.e. that the base is incorrect. If P is the error probability, then:

P = 10-Q/10

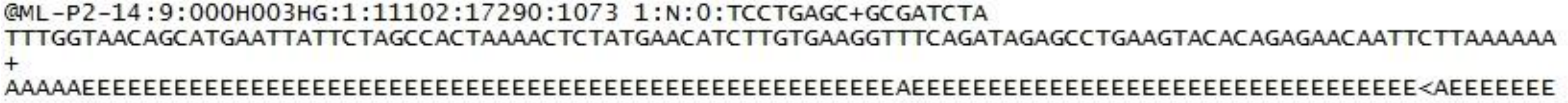
Q = -10 log10(P)



ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

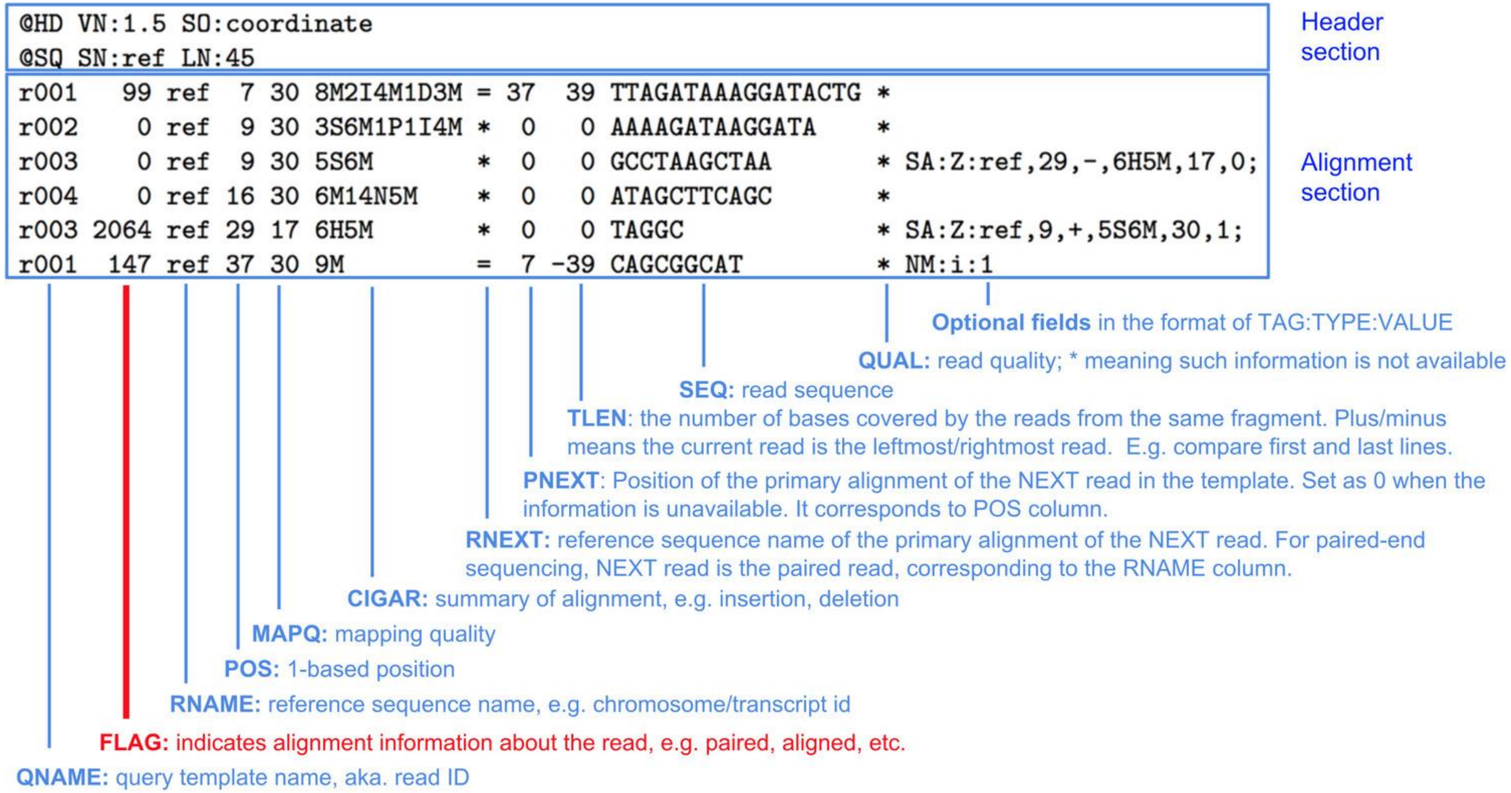
https://drive5.com/usearch/manual/quality_score.html



```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

Frequently used file terms and formats

- **SAM file – Sequence alignment/map files (.sam)**. It is a file format to save alignment information of short reads mapped against reference sequences. Comes from using **samtools** for alignment. It also uses ASCII format, It will have a header section starting with (@). Each alignment line will have 11 mandatory fields: QNAME FLAG RNAME POS MAPQ CIGAR RNEXT PNEXT TLEN SEQ QUAL TAGS



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

https://samtools.github.io/hts-specs/SAMv1.pdf

Frequently used file terms and formats

- BAM file – Binary alignment/map files (.bam). It is a compressed binarized .SAM file. Comes from using samtools for alignment. It will also have a header section starting with (@) and alignment fields. It is much smaller than a .sam file.
- CRAM file - Compressed Reference-oriented Alignment Map (.cram). It is an even more compressed version of the bam file

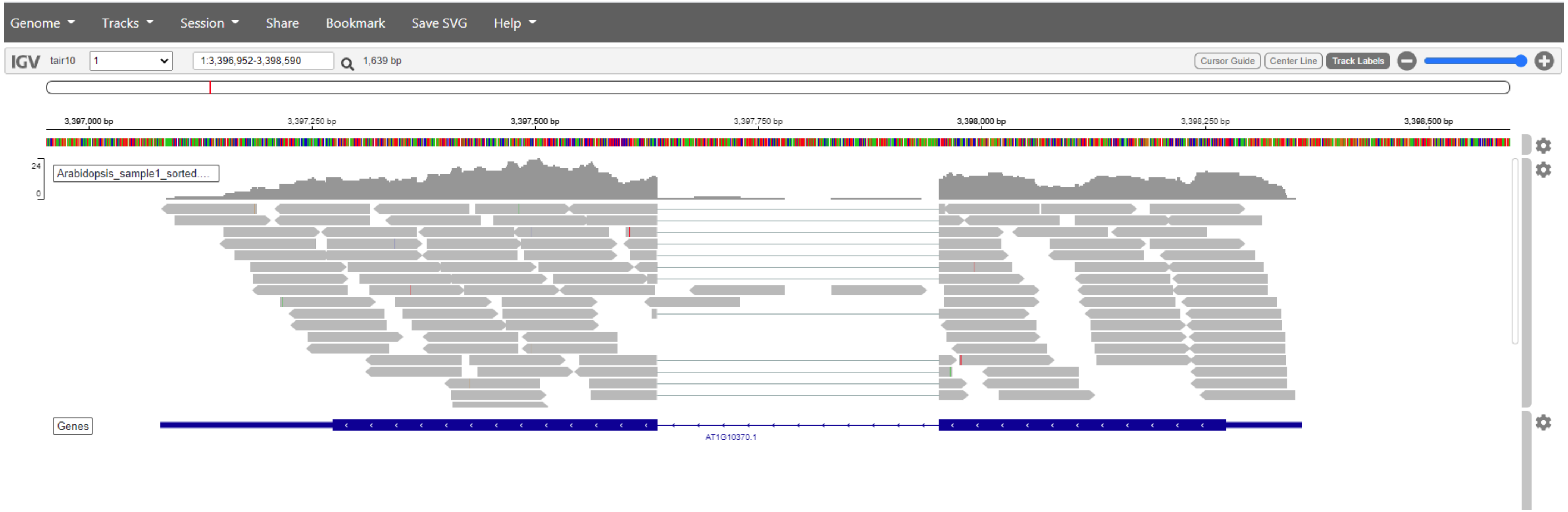
```
## ERR188273.4711308 73 chrX 30 5S70M = 21649 0
CGGGTGATCACGAGGTCAGGAGATCAAGACCATCCTGGCCAACACAGTGAAACCCCATCTCTACTAAAAATACAA
@@@F=DDFFHGHBHIFFHIGGIFGEGHFHIGIGIFIIIGIGIGGDHIIGIIC@>DGHCHHHGHHFFFFFDEACC@ AS:i:-
5 ZS:i:-5 XN:i:0 XM:i:0 XO:i:0 XG:i:0 YT:Z:UP NH:i:2 MD:Z:70 NM:i:0
```

https://scienceparkstudygroup.github.io/rna-seq-lesson/04-bioinformatic-workflow/index.html

https://samtools.github.io/hts-specs/SAMv1.pdf

Frequently used file terms and formats

- **BAM file – Binary alignment/map files (.bam)**. It is a compressed binarized .SAM file. Comes from using samtools for alignment. It will also have a header section starting with (@) and alignment fields. It is much smaller than a .sam file.
- Once indexed (.bam.bai) it can be viewed using an interactive genome viewer (IGV).



https://samtools.github.io/hts-specs/SAMv1.pdf

Frequently used file terms and formats

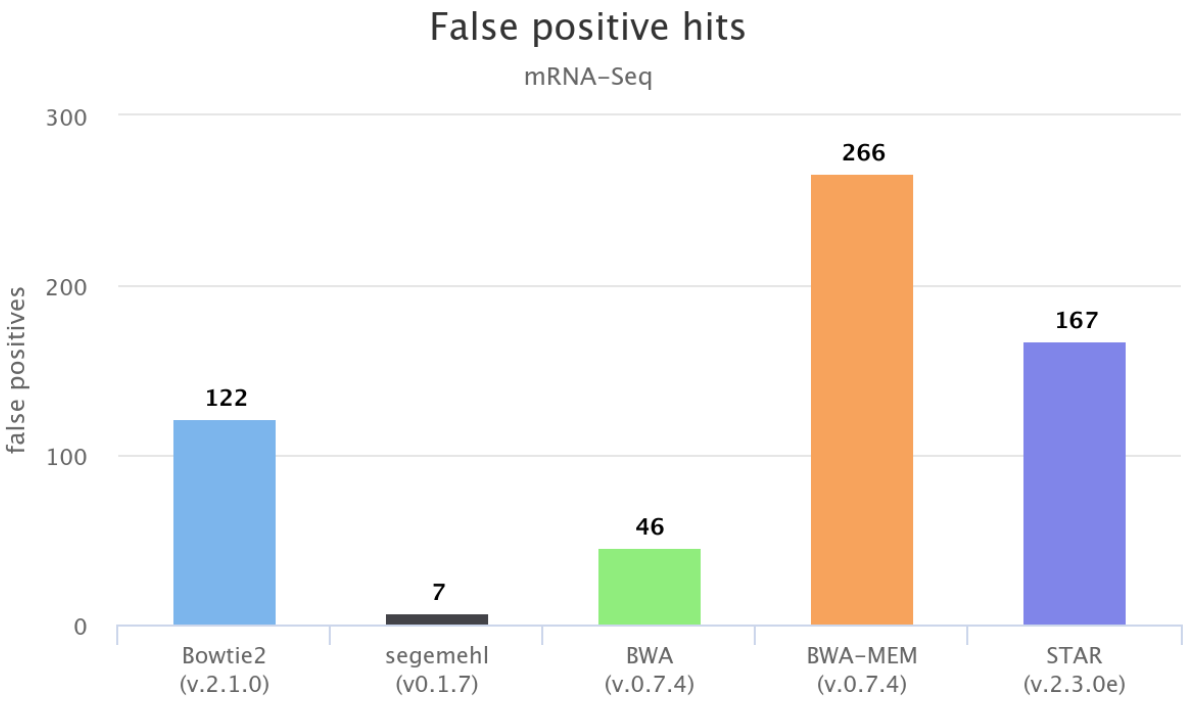
- **Gene/transcript/count file – (.txt) File containing** 'abundance estimates' which predict the relative abundance of different genes/isoforms in the form of three possible metrics (FPKM, RPKM and TPM).
  - RPKM (Reads Per Kilobase Million) – Used for single end reads. Count the total reads in a sample and divide that number by 1,000,000 (scaling factor). Divide the RPM values by the length of the gene, in kilobases.
  - FPKM (Fragments Per Kilobase Million) – Is for paired end reads. It takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice).
  - TPM (Transcripts Per Kilobase Million) – Divide the read counts by the length of each gene in kilobases (RPK). Count all the RPK values in a sample and divide this number by 1,000,000.

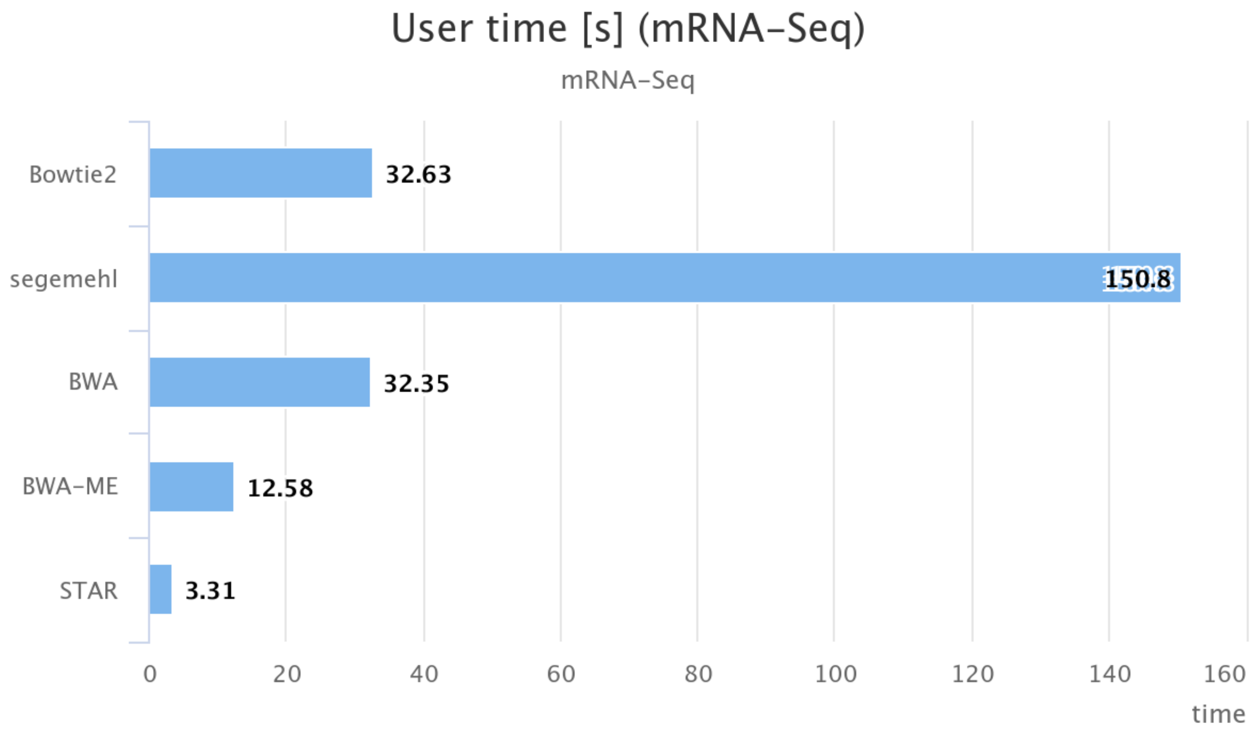
```
Name                  Length  EffectiveLength  TPM    NumReads
ENST00000632684.1     12       3.00168          0      0
ENST00000434970.2     9       2.81792           0      0
ENST00000448914.1     13       3.04008          0      0
ENST00000415118.1     8       2.72193           0             0
```

https://www.youtube.com/watch?v=TTUrtCY2k-w&t=236s

Softwares and Pipeline

- Bowtie2, Tophat2, BWA, BWA-mem, STAR. Mapping algorithms can largely be grouped into two categories based on properties of their indices: algorithms based on hash tables, and algorithms based on the Burrows-Wheeler transform

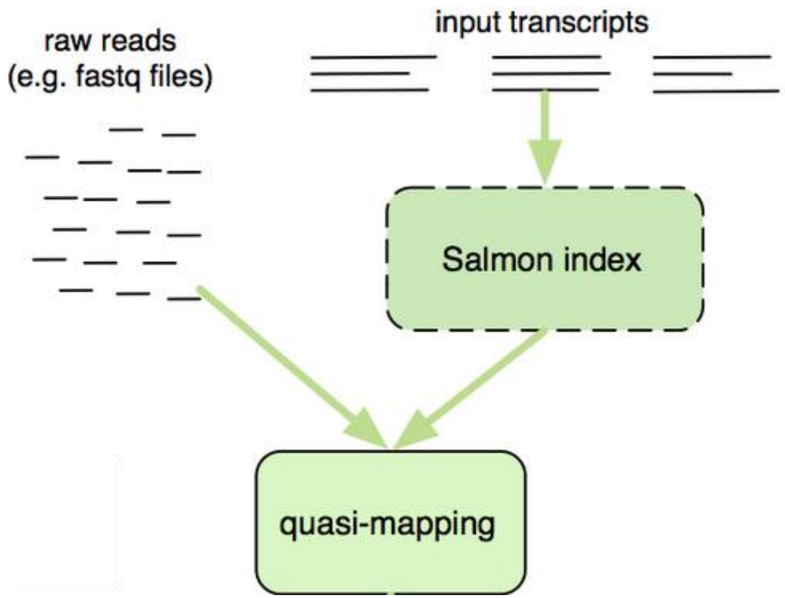




https://www.ecseq.com/support/benchmark

Softwares and Pipeline – Mapping & quantification

- **Aligners** typically align against the entire **genome.** Used for detecting novel genes/transcripts. Cannot be used to detect non-coding RNAs or splicing variants (unless the reference used is splicing aware). Typically need quantification algorithms like **Cufflinks or HTSeq.**
- **PseudoAligners** assign reads to the most appropriate **transcript.** Can't find novel genes/transcripts. Extremely fast! E.g. **Salmon and Kalisto**



Best for RNAseq – STAR + Salmon

https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/04_quasi_alignment_salmon.html

Softwares and Pipeline – Pseudocode Syntax

```
#!/bin/bash

#BSUB -p short
#BSUB -t 0-12:00
#BSUB --mem 8G
#BSUB –J salmon_in_serial
#BSUB -o %j.out
#BSUB -e %j.err
#BSUB --reservation=HBC

cd ~/rnaseq/results/salmon

module load salmon
```

```
for fq in ~/rnaseq/raw_data/*.fq

do

# create a prefix
base=`basename $fq .fq`

# run salmon
salmon quant -i /n/groups/hbctraining/rna-
seq_2019_02/reference_data/salmon.ensembl38.idx.09-06-2019 \
-l A \
-r $fq \
-p 6 \
-o $base.salmon \
--seqBias \
--useVBOpt \
--numBootstraps 30 \
--validateMappings

done
```

https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/04_quasi_alignment_salmon.html

Softwares and Pipeline – Pseudocode Syntax

```
# Step 1 Use bcl-convert (recommended by Illumina)
bcl-convert --bcl-input-directory <path_to_run_folder> \
          --output-directory <path_to_output_fastqs> \
          --sample-sheet <path_to_samplesheet.csv>
```

```
# Step 2a: Use bwa aligner Map the FASTQ files to a
reference genome and save as SAM
bwa mem <path_to_reference.fasta> \
<path_to_output_fastqs>/<SampleID>_R1.fastq.gz \
<path_to_output_fastqs>/<SampleID>_R2.fastq.gz >
<SampleID>.sam
```

```
# Step 2b: Use samtools to convert the SAM file to a sorted
and indexed BAM file #
samtools view -bS <SampleID>.sam | samtools sort -o
<SampleID>.sorted.bam samtools index
<SampleID>.sorted.bam
```

Softwares and Pipeline – Quality control – Picard & MultiQC

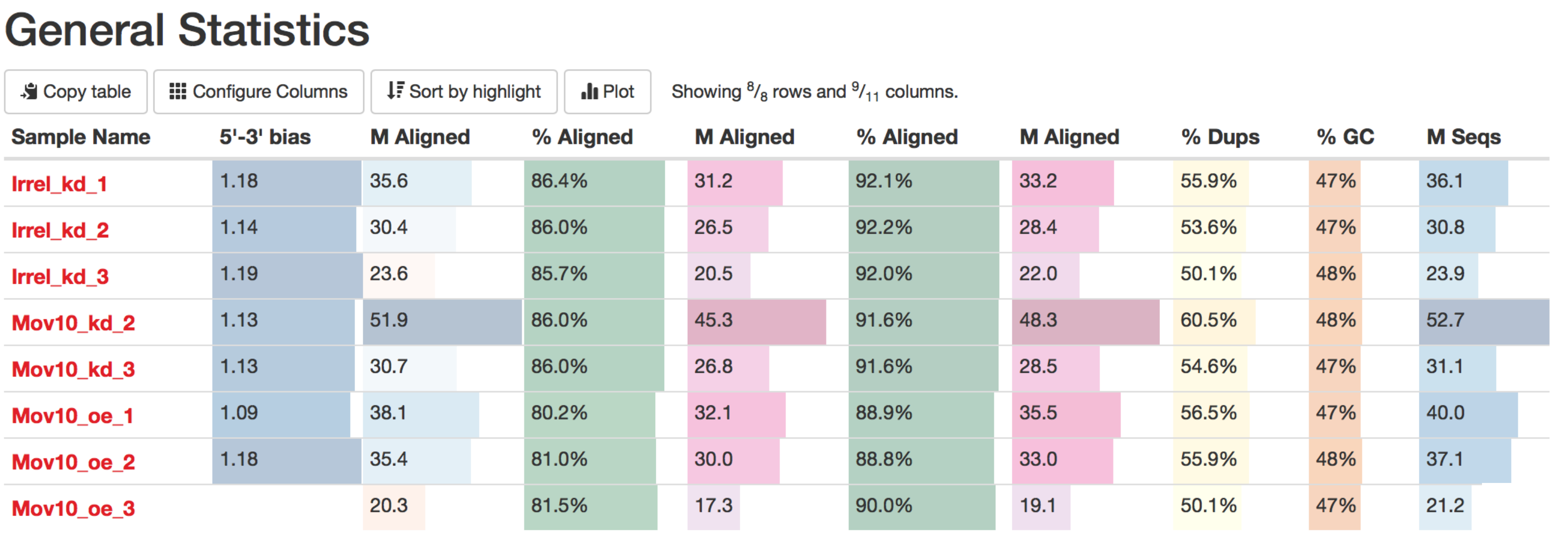
- Once we have aligned our reads and quantified gene expression with Salmon, it is then possible to run some additional quality checks on our data.

- Picard Tools is a suite of tools for analysing and manipulating sequencing data. It is maintained by the Broad Institute and comprises 88 different tools for doing jobs such as generating QC metrics, modifying bam files in various ways, or converting files between different formats. E.g. MarkDuplicates – Finds duplicate reads marked by **1024 as a sam flag**

- MultiQC is a tool for collating multiple QC results files into a single report.

https://bioinformatics-core-shared-training.github.io/Bulk_RNASeq_Course_March23/Bulk_RNAseq_Course_Base/Markdowns/04_Quality_Control_practical.html
https://broadinstitute.github.io/picard/

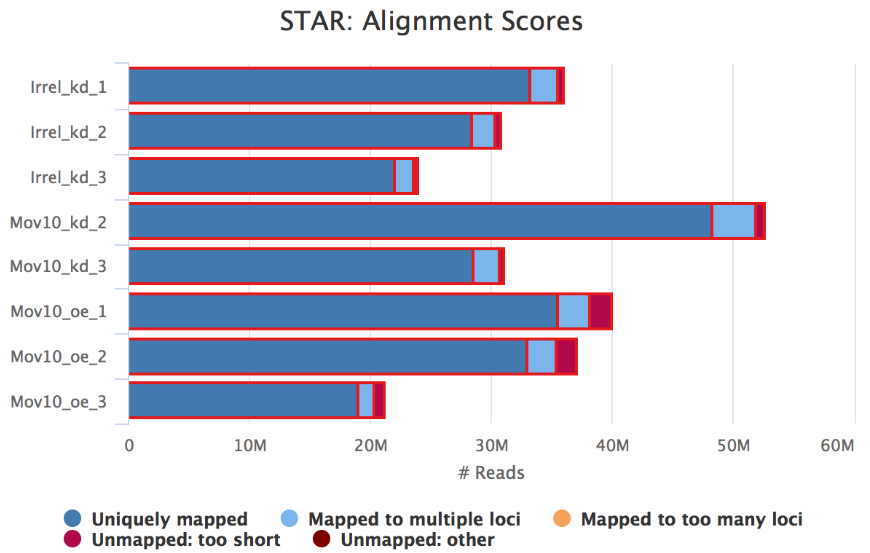
Softwares and Pipeline – Quality control – MultiQC example report



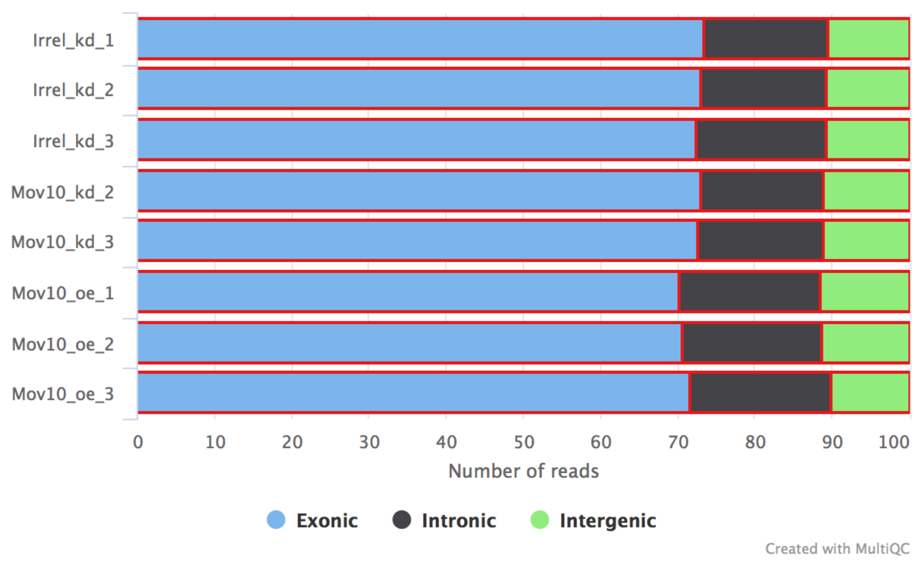
General Statistics

Showing $^8/_8$ rows and $^9/_{11}$ columns.

Sample Name	5'-3' bias	M Aligned	% Aligned	M Aligned	% Aligned	M Aligned	% Dups	% GC	M Seqs
Irrel_kd_1	1.18	35.6	86.4%	31.2	92.1%	33.2	55.9%	47%	36.1
Irrel_kd_2	1.14	30.4	86.0%	26.5	92.2%	28.4	53.6%	47%	30.8
Irrel_kd_3	1.19	23.6	85.7%	20.5	92.0%	22.0	50.1%	48%	23.9
Mov10_kd_2	1.13	51.9	86.0%	45.3	91.6%	48.3	60.5%	48%	52.7
Mov10_kd_3	1.13	30.7	86.0%	26.8	91.6%	28.5	54.6%	47%	31.1
Mov10_oe_1	1.09	38.1	80.2%	32.1	88.9%	35.5	56.5%	47%	40.0
Mov10_oe_2	1.18	35.4	81.0%	30.0	88.8%	33.0	55.9%	48%	37.1
Mov10_oe_3		20.3	81.5%	17.3	90.0%	19.1	50.1%	47%	21.2



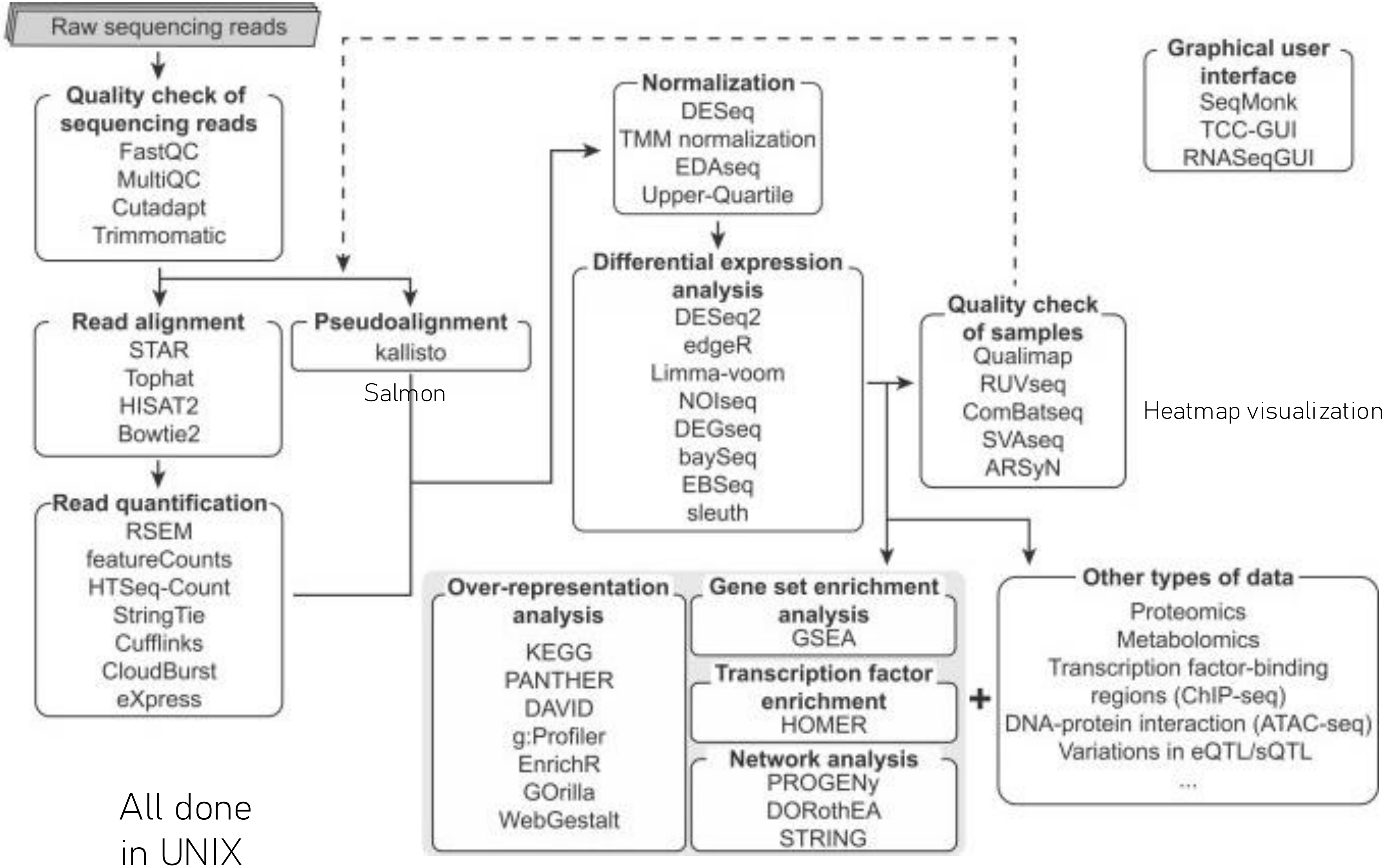
STAR: Alignment Scores



Qualimap RNAseq: Genomic Origin

https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/05_multiQC.html

Analysis after getting the counts matrix



Salmon

Heatmap visualizations

All done
in UNIX

All done in R/Python or other online platforms e.g. GenePattern

Lee GY, Ham S, Lee SV. Brief guide to RNA sequencing analysis for nonexperts in bioinformatics. Mol Cells. 2024;47(5):100060. doi:10.1016/j.mocell.2024.100060

Analysis after getting the counts matrix

Name	Length	EffectiveLength	TPM	NumReads
ENSG00000121410.12_4	509.732	325.991	3.22494	322.674
ENSG00000268895.6_6	1823.71	1633.86	0.9255	464.119
ENSG00000148584.15_4	5354.1	5164.27	0	0
ENSG00000175899.14_4	4544.77	4354.95	0.039651	53
A2M-AS1	2592.39	2402.54	0.008136	5.999
A2ML1	1749	1561.55	0	0
SLC7A2	452	269.66	0	0
ENSG00000001461.12_NIPAL3	386	208.766	0	0
ENSG00000001497.12_LAS1	1715	1526.05	0	0
ENSG00000001617.7_SEMA3F	1023	833.15	0	0
ENSG00000003096.9_KLHL13	1457.48	1269.51	3.23046	1258.74

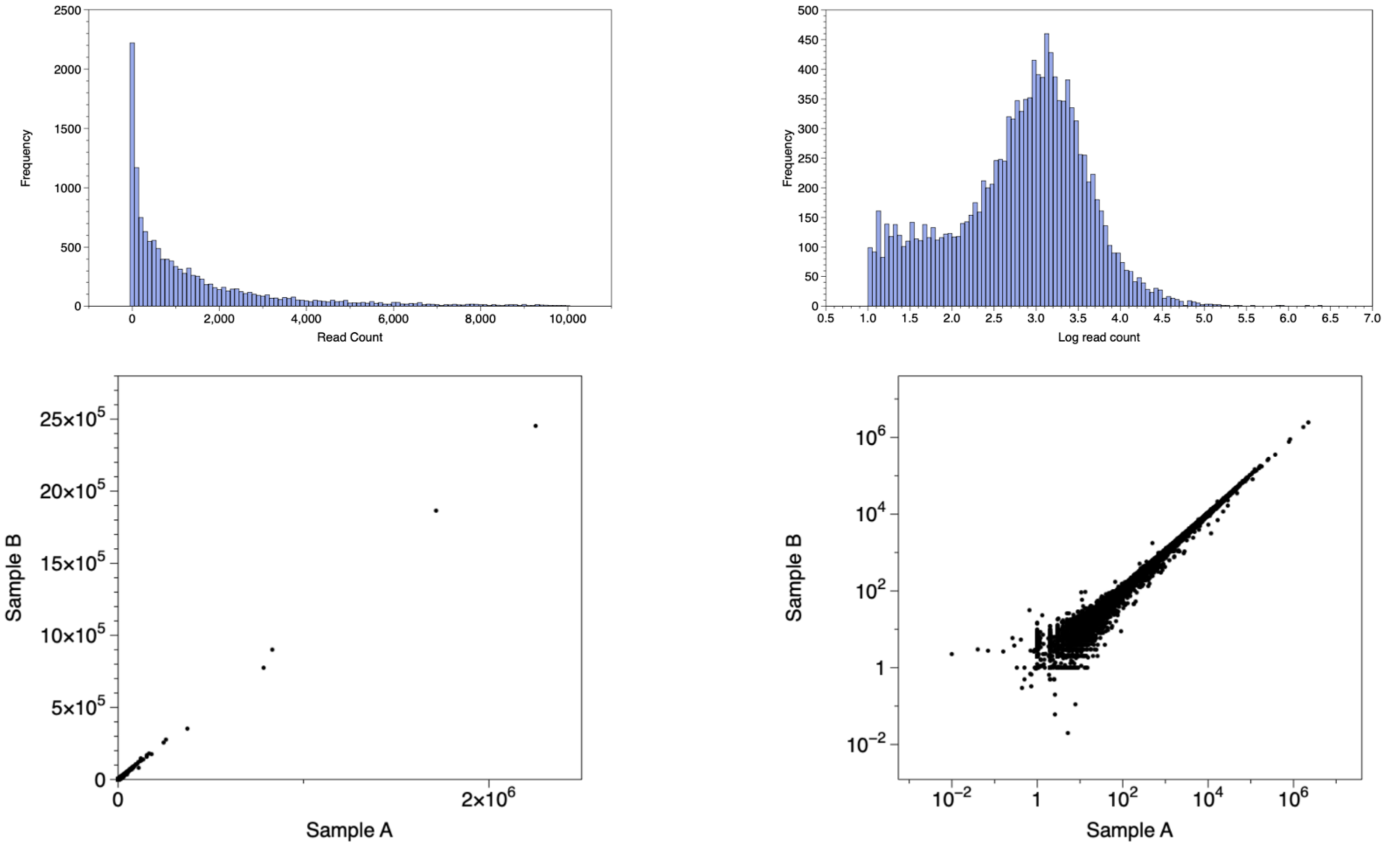
Different ways of annotating the genes

Not always integers -
may not be acceptable to some programs

https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/04_quasi_alignment_salmon.html

Filtering and Log transformation

- Because of its vast dynamic ranges RNASEQ data is typically filtered for low transcripts and log transformed in order to: provide better visualizations and to present analysis software with a more "normal distribution".



https://bioinformatics.ccr.cancer.gov/docs/b4b/RNASeq_Overview/06.Quantitation/

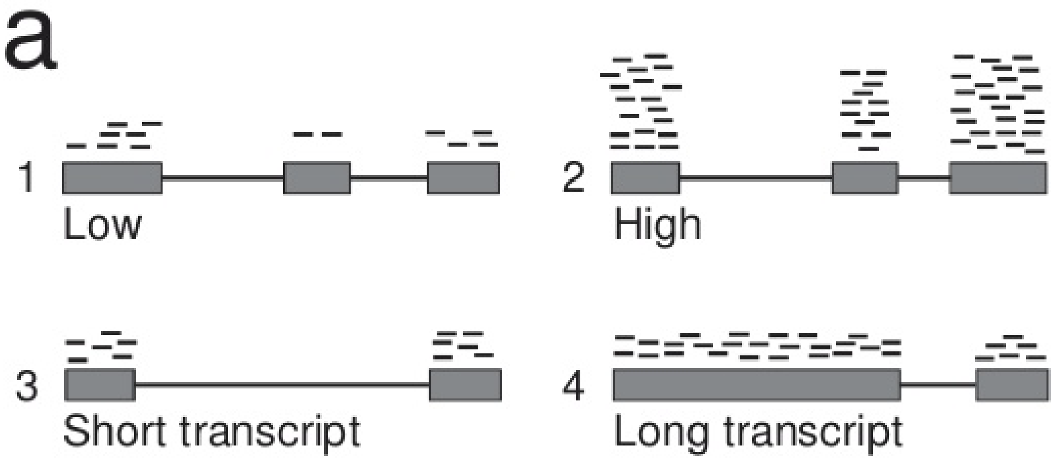
Normalization

- RNASeq transcriptomic expression data may vary based on a variety of often-uncontrollable experimental conditions

- Consequently, RNASeq raw data needs to be adjusted so that comparisons are based on biological truth. This mathematical adjustment is known as **normalization.**

- Number of reads aligned to a gene gives a measure of its level of expression

- Normalization of the count data
  - Sequencing depth
  - Length bias



https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/04_quasi_alignment_salmon.html

Normalization

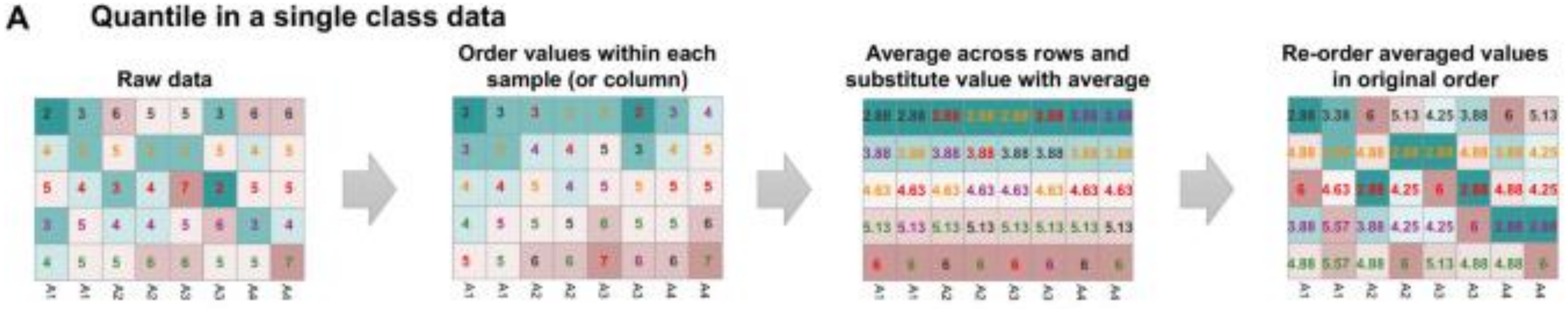
- A general strategy, common to many normalization techniques, is to re-distribute signal intensities across all samples such that they now all have the same distribution (e.g. same mean and/or standard deviation).

- Basic assumption of any differential gene expression (DGE) analysis is that there is no difference between the two populations. Therefore, normalization is a necessity!

- Common examples of normalization techniques include
  - linear scaling (also known as min–max scaling),
  - Z-normalization, and
  - rank-scaling (also known as linear interpolation).
  - Specialized approaches for removing batch effects (a form of technical variation) such as ComBat

Normalization – Quantile normalization

- The quantile normalization (QN) procedure is simple. It involves first ranking the gene of each sample by magnitude, calculating the average value for genes occupying the same rank, and then substituting the values of all genes occupying that particular rank with this average value.
- Finally, we reorder the genes of each sample in their original order.



Raw data

A1	A1	A2	A2	A3	A3	A4	A4
2	3	6	5	5	3	6	6
4	[illegible]	5	[illegible]	[illegible]	5	4	5
5	4	3	4	7	2	5	5
3	5	4	4	5	6	3	4
4	5	5	6	6	5	5	7

Order values within each sample (or column)

A1	A1	A2	A2	A3	A3	A4	A4
2	3	3	[illegible]	[illegible]	2	3	4
3	[illegible]	4	4	5	3	4	5
4	4	5	4	5	5	5	5
4	5	5	5	6	5	5	6
5	5	6	6	7	6	6	7

Average across rows and substitute value with average

A1	A1	A2	A2	A3	A3	A4	A4
2.88	2.88	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3.88	[illegible]	3.88	3.88	3.88	3.88	[illegible]	[illegible]
4.63	4.63	4.63	4.63	4.63	4.63	4.63	4.63
5.13	5.13	5.13	5.13	5.13	5.13	5.13	5.13
6	6	6	6	6	6	6	6

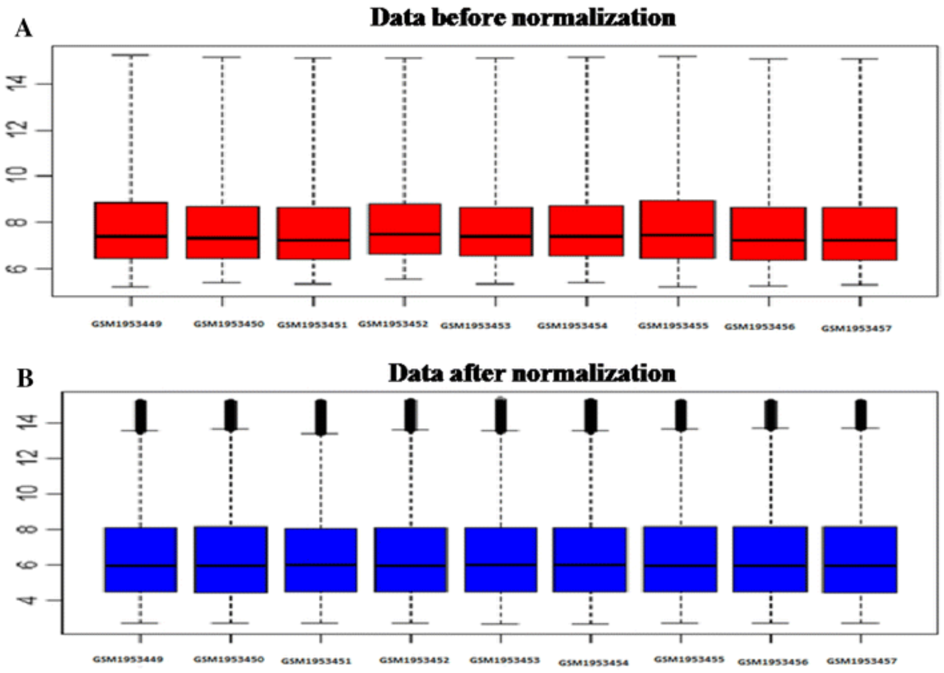
Re-order averaged values in original order

A1	A1	A2	A2	A3	A3	A4	A4
2.88	3.38	6	5.13	4.25	3.88	6	5.13
4.88	[illegible]	4.88	[illegible]	[illegible]	4.88	[illegible]	4.25
6	4.63	[illegible]	4.25	6	[illegible]	4.88	4.25
3.88	5.57	3.88	4.25	4.25	6	[illegible]	[illegible]
4.88	5.57	4.88	6	5.13	4.88	4.88	6

Zhao, Y., Wong, L. & Goh, W.W.B. How to do quantile normalization correctly for gene expression data analyses. Sci Rep 10, 15534 (2020). https://doi.org/10.1038/s41598-020-72664-6

Normalization – Upper Quantile normalization

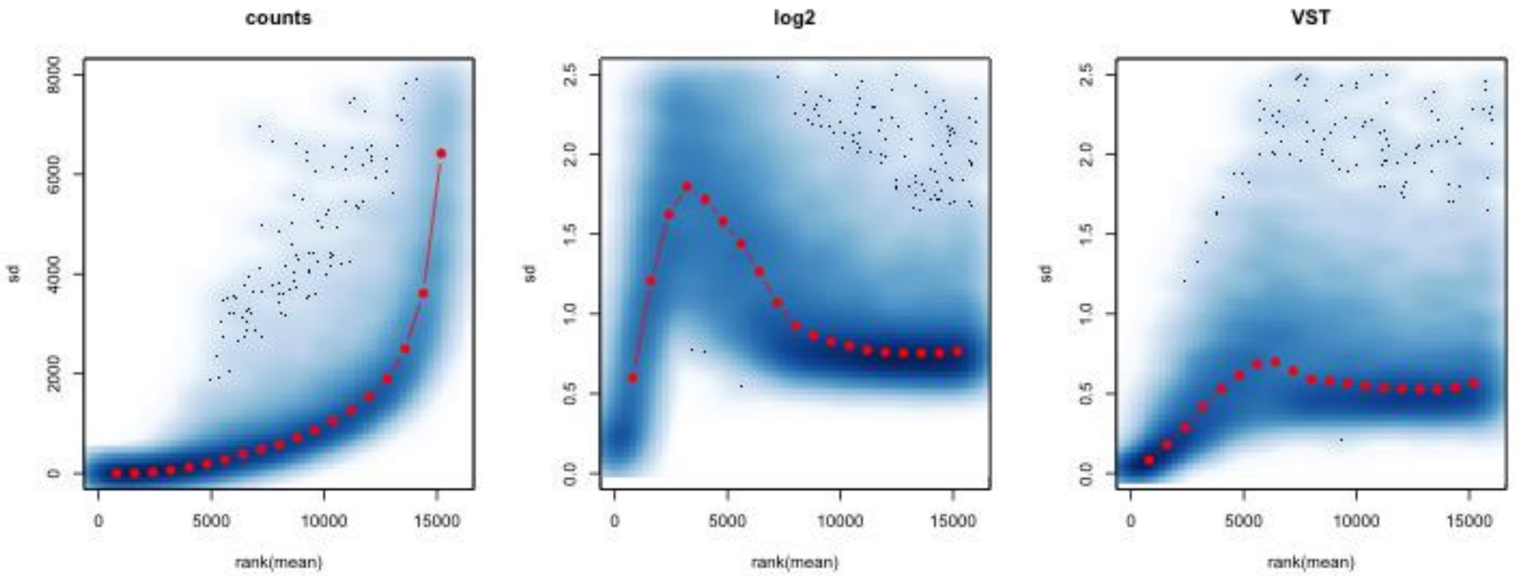
- The total number of counts for a sample can is very dependent on a small number of highly abundant genes.
- To solve this problem, Divide by the 75th percentile of the total number of counts. This provides a sequencing depth scaling factor that does not depend on the small proportion (less than 5%) of highly expressed genes.
- Advantages:
  - Robustness to outliers
  - Preserves the distribution of the data
  - Easy to implement
  - Consistent results



https://fastercapital.com/content/Quartile-Normalization--Adjusting-Data-for-Statistical-Analysis.html
https://www.youtube.com/watch?v=ecjN6Xpv6SE

Normalization – DESeq2

- Most used R library for differential gene expression (DGE) analysis
- Uses a unique normalization – Variance stabilizing transformation (VST); aims at generating a matrix of values for which variance is constant across the range of mean values (homoscedastic), especially for low mean.
- The transformation also normalizes with respect to library size.
- The input to DESeq2 is absolute raw count matrix



https://seqqc.wordpress.com/2015/02/16/should-you-transform-rna-seq-data-log-vst-voom/comment-page-1/

Normalization strategies – DESeq2

- Step 1 - Create a pseudo-reference sample (row-wise geometric mean)

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**
…	…	…	…

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Normalization strategies – DESeq2

- Step 2 - **Calculate ratio of each sample to the reference**

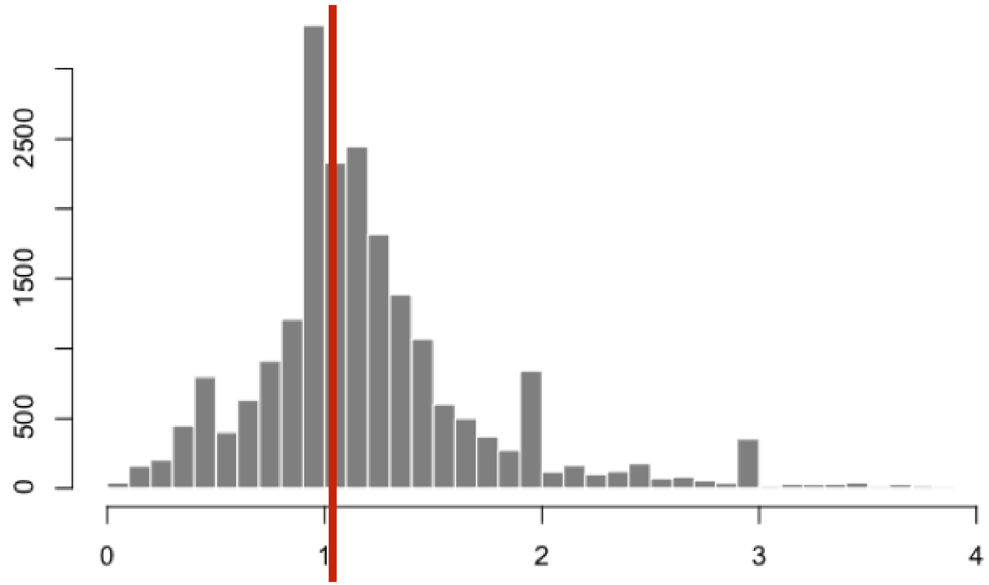
gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Normalization strategies – DESeq2

- Step 3 – Calculate the normalization factor for each sample (size factor)

```
normalization_factor_sampleA <- median(c(1.28, 1.3, 1.39, 1.35, 0.59))
normalization_factor_sampleB <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```



sample 1 / pseudo-reference sample

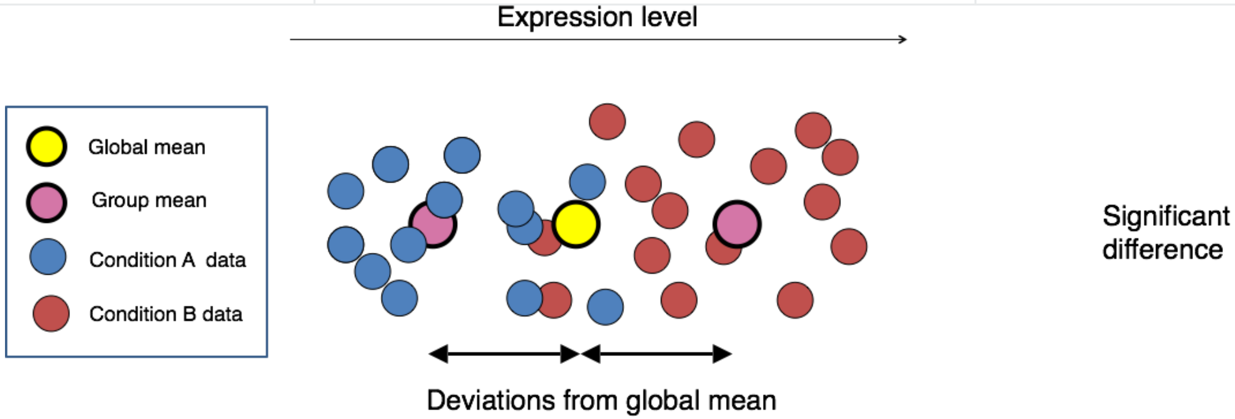
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Normalization strategies – DESeq2

- Step 4 – **Calculate the normalized count values using the normalization factor**

SampleA median ratio = 1.3
SampleB median ratio = 0.77

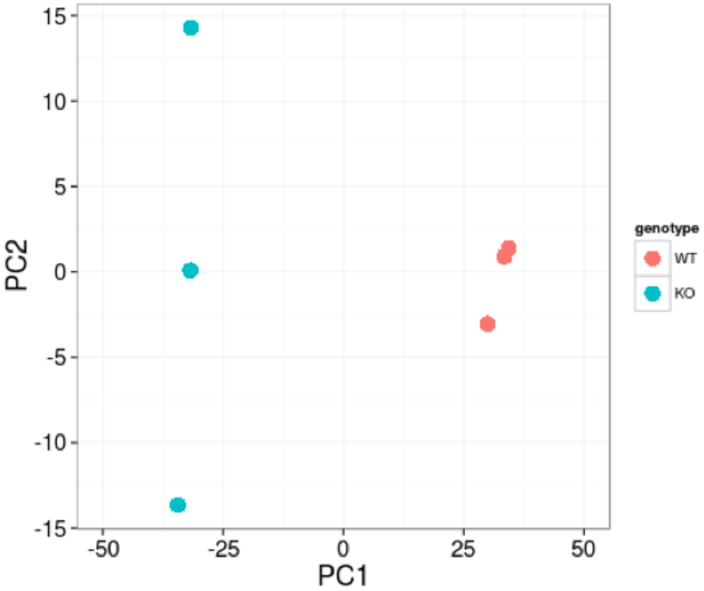
gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
...	...	...



https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Dimension reductionality – Principal Component analysis (PCA)

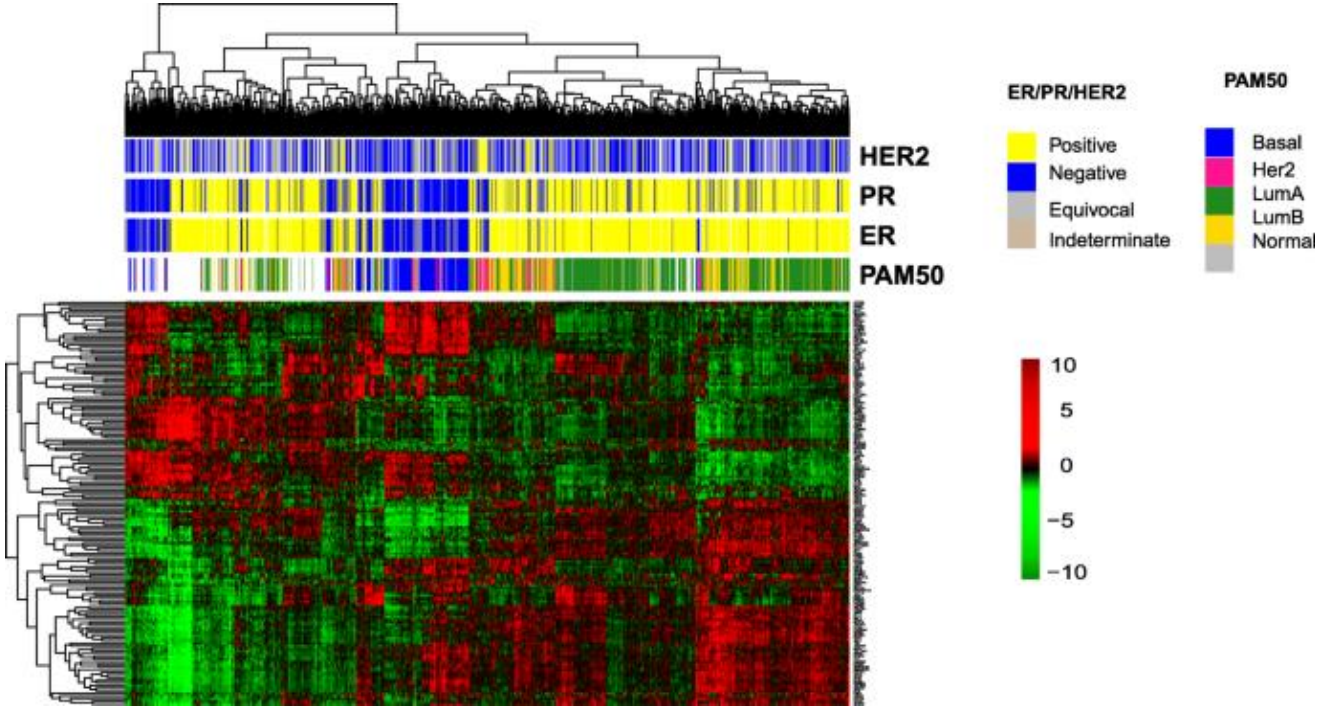
- Principal component analysis (PCA) is a multivariate statistical method that combines information from several variables observed on the same subjects into fewer variables, called principal components (PCs).
- Information is measured by the total variance of the original variables (genes), and the PCs optimally account for the major part of that variance.
- For bulk RNAseq, most variance is captured in 1:4 PCs (Typically we use 1:2 PCs for visualization)



Greenacre, M., Groenen, P.J.F., Hastie, T. et al. Principal component analysis. Nat Rev Methods Primers 2, 100 (2022). https://doi.org/10.1038/s43586-022-00184-w

Dimension reductionality – Hierarchical Clustering (HC)

- This is another way to capture the pattern of informative genes in the data
- Take the top 2000-5000 most variable genes and cluster all the RNAseq samples (unsupervised HC)
- The heatmap displays **the correlation of gene expression for all pairwise combinations of samples** in the dataset. Strong "clusters" of genes will have a high correlation value.



Dong, C., Liu, J., Chen, S.X. et al. Highly robust model of transcription regulator activity predicts breast cancer overall survival. BMC Med Genomics 13 (Suppl 5), 49 (2020). https://doi.org/10.1186/s12920-020-0688-z

DESeq2 – pairwise comparisons

- The final step in the DESeq2 workflow is fitting the Negative Binomial model for each gene and performing differential expression testing. Once the model is fit, coefficients are estimated for each sample group

raw count for gene i, sample j

The mean is taken as "normalized counts" scaled by a normalization factor

one dispersion per gene

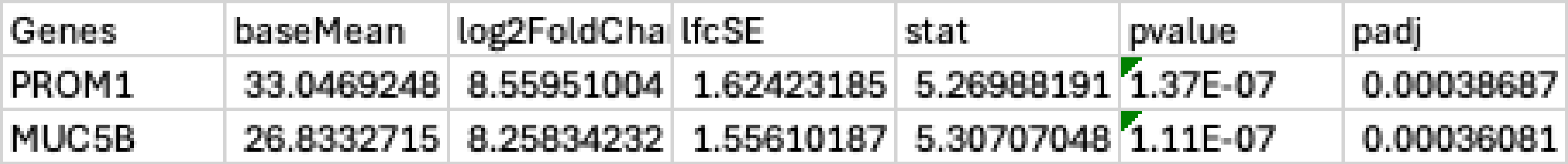
$$K_{ij} \sim \mathrm{NB}(s_{ij} q_{ij}, \alpha_i)$$

normalized counts for gene i, sample j

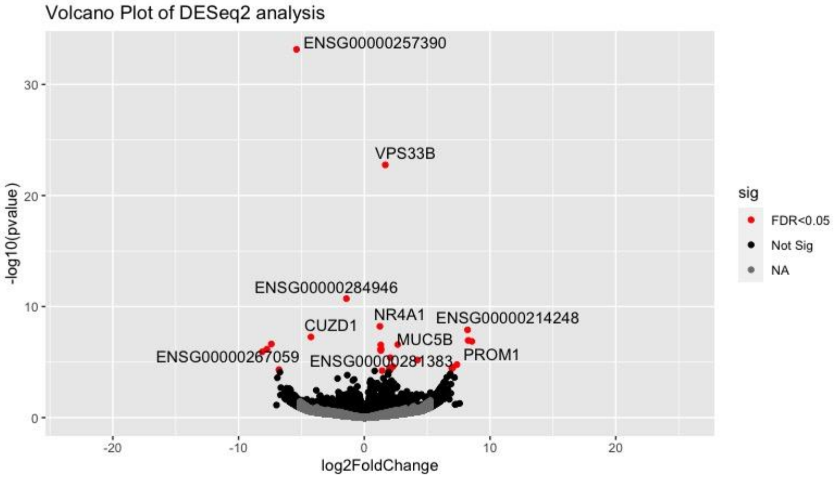
log2 fold change between conditions

$$\log2 q_{ij} = \Sigma_r x_{jr} \beta_{ir}$$

- With DESeq2, the Wald test is commonly used for hypothesis testing when comparing two groups.



Genes	baseMean	log2FoldCha	lfcSE	stat	pvalue	padj
PROM1	33.0469248	8.55951004	1.62423185	5.26988191	1.37E-07	0.00038687
MUC5B	26.8332715	8.25834232	1.55610187	5.30707048	1.11E-07	0.00036081



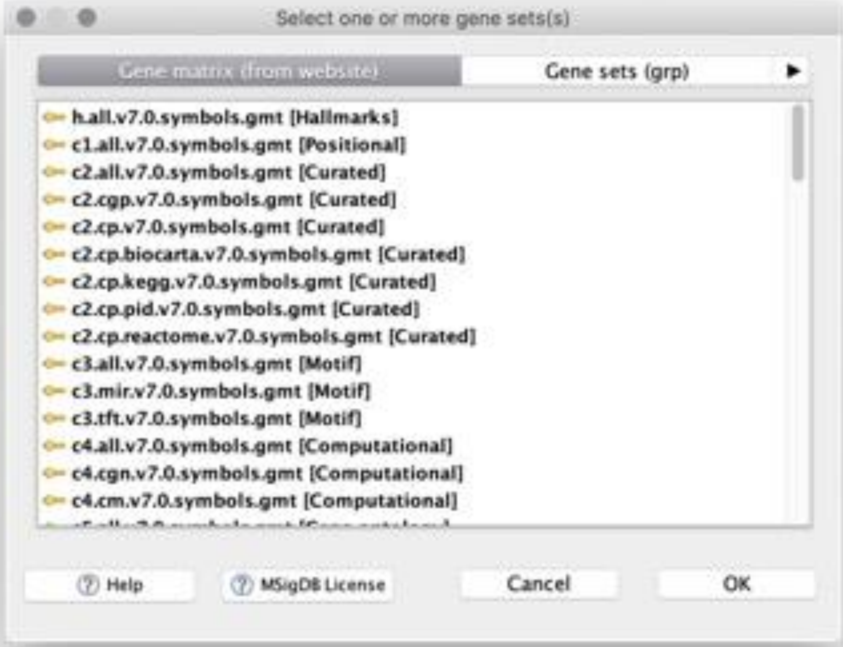
https://hbctraining.github.io/DGE_workshop/lessons/05_DGE_DESeq2_analysis2.html

Gene set enrichment analysis (GSEA)

- GSEA is a computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes). Uses a 2-way Kolmogrov-Smirnov (K-S) Test to test the cumulative density of genes. Must be performed on normalized data

- Genesets are housed in MSigDB

- https://www.gsea-msigdb.org/gsea/msigdb

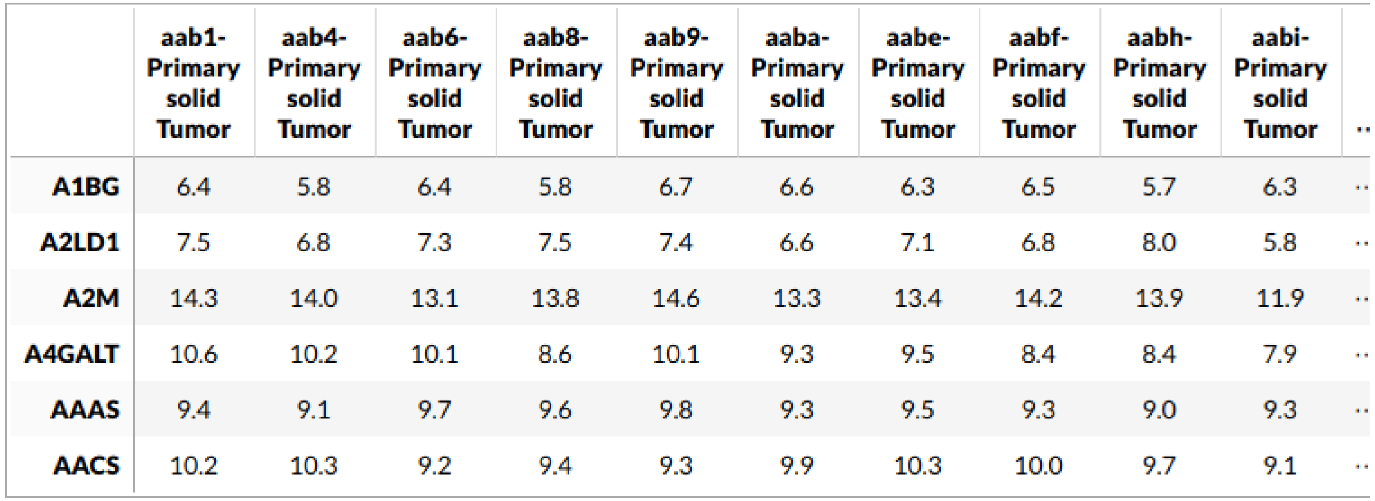


h.all.v7.0.symbols.gmt [Hallmarks]
c1.all.v7.0.symbols.gmt [Positional]
c2.all.v7.0.symbols.gmt [Curated]
c2.cgp.v7.0.symbols.gmt [Curated]
c2.cp.v7.0.symbols.gmt [Curated]
c2.cp.biocarta.v7.0.symbols.gmt [Curated]
c2.cp.kegg.v7.0.symbols.gmt [Curated]
c2.cp.pid.v7.0.symbols.gmt [Curated]
c2.cp.reactome.v7.0.symbols.gmt [Curated]
c3.all.v7.0.symbols.gmt [Motif]
c3.mir.v7.0.symbols.gmt [Motif]
c3.tft.v7.0.symbols.gmt [Motif]
c4.all.v7.0.symbols.gmt [Computational]
c4.cgn.v7.0.symbols.gmt [Computational]
c4.cm.v7.0.symbols.gmt [Computational]

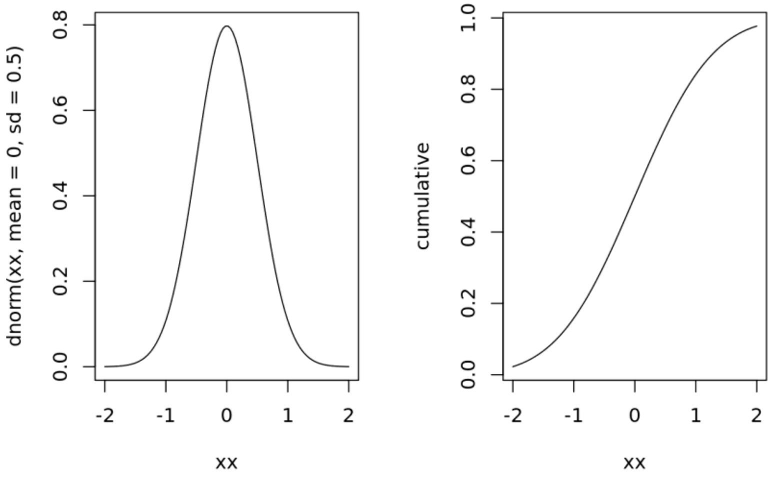
https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html

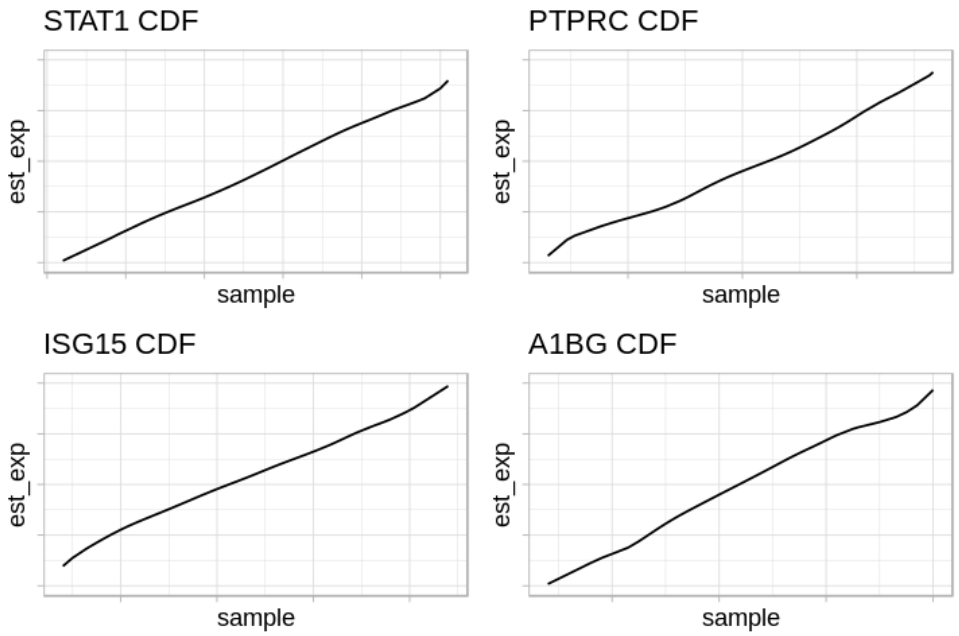
Gene set enrichment analysis (GSEA)

- Cumulative density frequency (CDF) is calculated for each gene



	aab1-Primary solid Tumor	aab4-Primary solid Tumor	aab6-Primary solid Tumor	aab8-Primary solid Tumor	aab9-Primary solid Tumor	aaba-Primary solid Tumor	aabe-Primary solid Tumor	aabf-Primary solid Tumor	aabh-Primary solid Tumor	aabi-Primary solid Tumor	··
A1BG	6.4	5.8	6.4	5.8	6.7	6.6	6.3	6.5	5.7	6.3	··
A2LD1	7.5	6.8	7.3	7.5	7.4	6.6	7.1	6.8	8.0	5.8	··
A2M	14.3	14.0	13.1	13.8	14.6	13.3	13.4	14.2	13.9	11.9	··
A4GALT	10.6	10.2	10.1	8.6	10.1	9.3	9.5	8.4	8.4	7.9	··
AAAS	9.4	9.1	9.7	9.6	9.8	9.3	9.5	9.3	9.0	9.3	··
AACS	10.2	10.3	9.2	9.4	9.3	9.9	10.3	10.0	9.7	9.1	··





STAT1 CDF

PTPRC CDF

ISG15 CDF

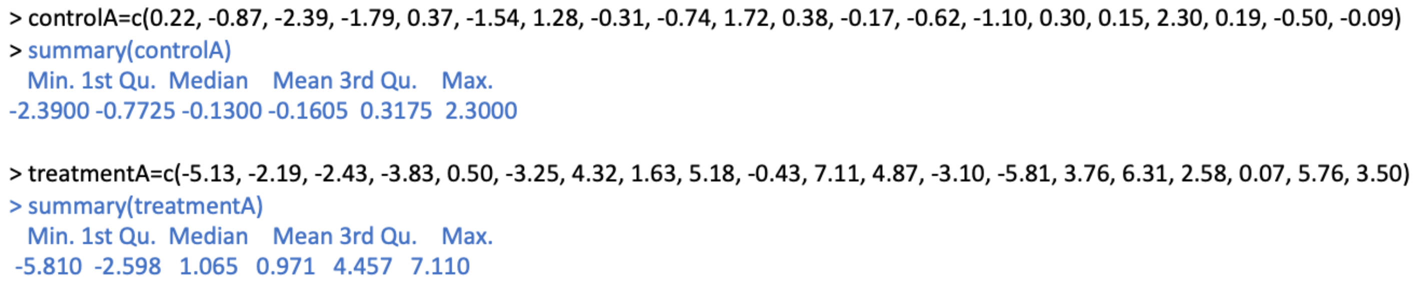
A1BG CDF

- Rank each gene in each sample based on the CDFs

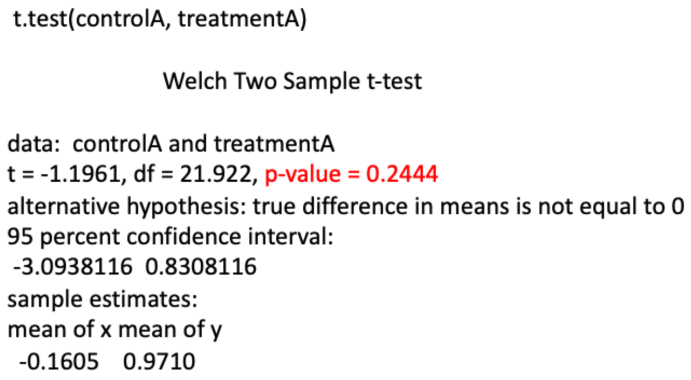
https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html

Gene set enrichment analysis (GSEA)

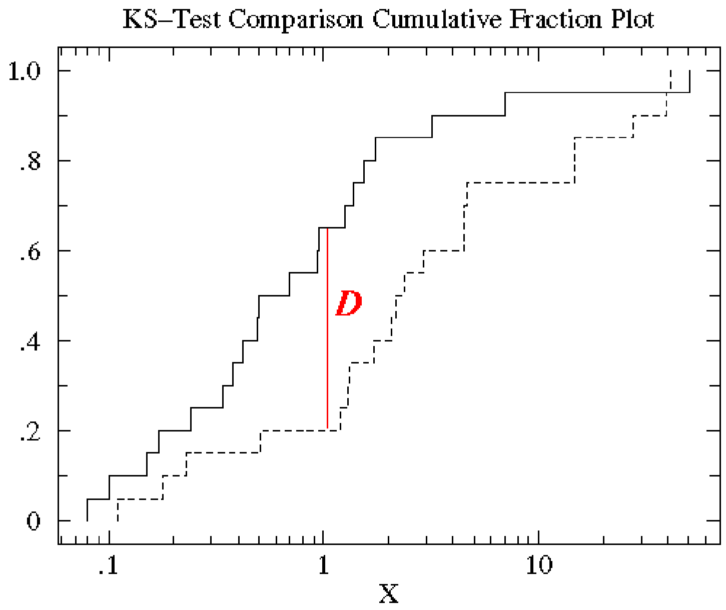
- Why K-S test instead of t-test or Fisher's exact test?

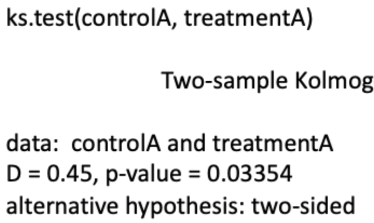


```
> controlA=c(0.22, -0.87, -2.39, -1.79, 0.37, -1.54, 1.28, -0.31, -0.74, 1.72, 0.38, -0.17, -0.62, -1.10, 0.30, 0.15, 2.30, 0.19, -0.50, -0.09)
> summary(controlA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-2.3900 -0.7725 -0.1300 -0.1605  0.3175  2.3000 

> treatmentA=c(-5.13, -2.19, -2.43, -3.83, 0.50, -3.25, 4.32, 1.63, 5.18, -0.43, 7.11, 4.87, -3.10, -5.81, 3.76, 6.31, 2.58, 0.07, 5.76, 3.50)
> summary(treatmentA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -5.810  -2.598   1.065   0.971   4.457   7.110 
```



```
 t.test(controlA, treatmentA)

	Welch Two Sample t-test

data:  controlA and treatmentA
t = -1.1961, df = 21.922, p-value = 0.2444
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0938116  0.8308116
sample estimates:
mean of x mean of y 
  -0.1605    0.9710 
```



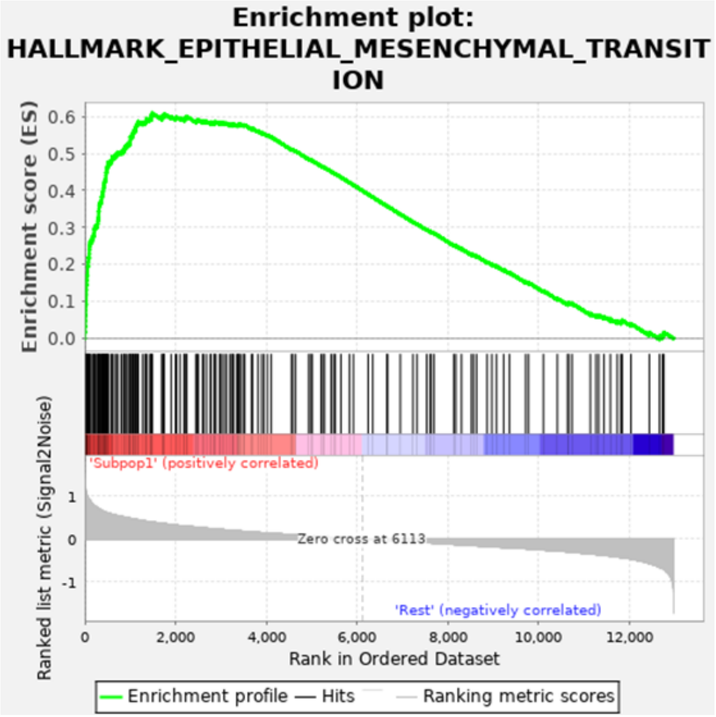


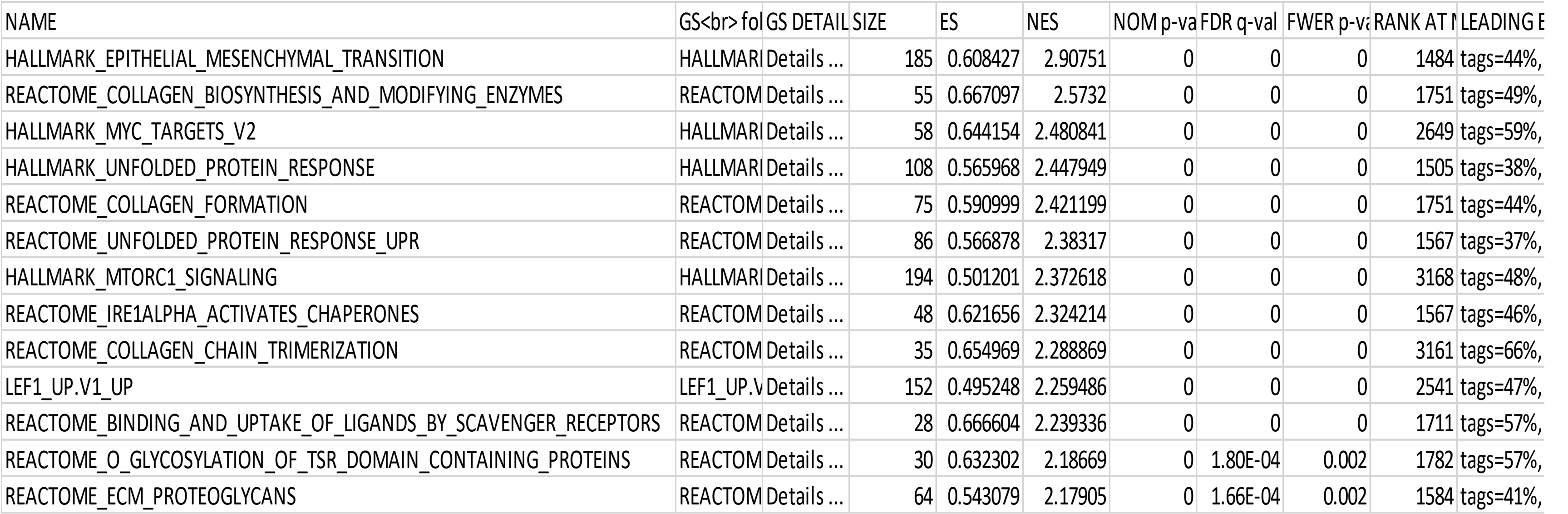
```
ks.test(controlA, treatmentA)

	Two-sample Kolmogorov-Smirnov test

data:  controlA and treatmentA
D = 0.45, p-value = 0.03354
alternative hypothesis: two-sided
```

- The Kolmogorov-Smirnov (KS) test of goodness-of-fit tests whether the observed data is consistent with a given cumulative density function (CDF).

https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html

Gene set enrichment analysis (GSEA)



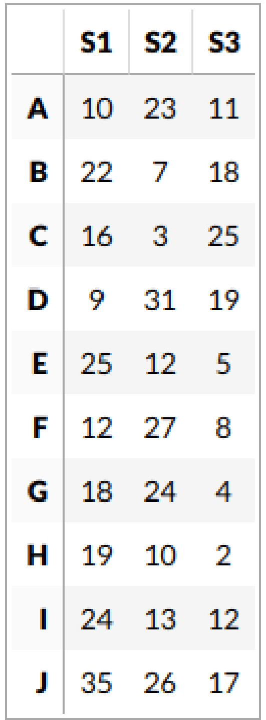


NAME	GS fol	GS DETAIL	SIZE	ES	NES	NOM p-va	FDR q-val	FWER p-va	RANK AT	LEADING E
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	HALLMARI	Details ...	185	0.608427	2.90751	0	0	0	1484	tags=44%,
REACTOME_COLLAGEN_BIOSYNTHESIS_AND_MODIFYING_ENZYMES	REACTOM	Details ...	55	0.667097	2.5732	0	0	0	1751	tags=49%,
HALLMARK_MYC_TARGETS_V2	HALLMARI	Details ...	58	0.644154	2.480841	0	0	0	2649	tags=59%,
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	HALLMARI	Details ...	108	0.565968	2.447949	0	0	0	1505	tags=38%,
REACTOME_COLLAGEN_FORMATION	REACTOM	Details ...	75	0.590999	2.421199	0	0	0	1751	tags=44%,
REACTOME_UNFOLDED_PROTEIN_RESPONSE_UPR	REACTOM	Details ...	86	0.566878	2.38317	0	0	0	1567	tags=37%,
HALLMARK_MTORC1_SIGNALING	HALLMARI	Details ...	194	0.501201	2.372618	0	0	0	3168	tags=48%,
REACTOME_IRE1ALPHA_ACTIVATES_CHAPERONES	REACTOM	Details ...	48	0.621656	2.324214	0	0	0	1567	tags=46%,
REACTOME_COLLAGEN_CHAIN_TRIMERIZATION	REACTOM	Details ...	35	0.654969	2.288869	0	0	0	3161	tags=66%,
LEF1_UP.V1_UP	LEF1_UP.V	Details ...	152	0.495248	2.259486	0	0	0	2541	tags=47%,
REACTOME_BINDING_AND_UPTAKE_OF_LIGANDS_BY_SCAVENGER_RECEPTORS	REACTOM	Details ...	28	0.666604	2.239336	0	0	0	1711	tags=57%,
REACTOME_O_GLYCOSYLATION_OF_TSR_DOMAIN_CONTAINING_PROTEINS	REACTOM	Details ...	30	0.632302	2.18669	0	1.80E-04	0.002	1782	tags=57%,
REACTOME_ECM_PROTEOGLYCANS	REACTOM	Details ...	64	0.543079	2.17905	0	1.66E-04	0.002	1584	tags=41%,

https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html

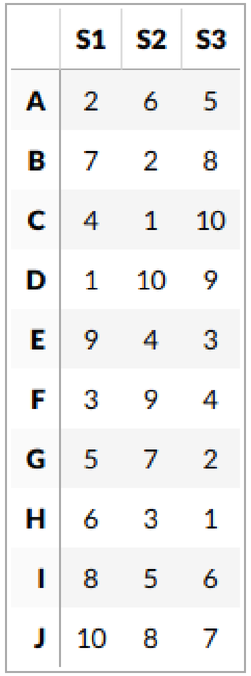
Gene set variation analysis (GSVA)

- GSEA relies on phenotypic data and samples are looked at in a way that two groups of samples whose phenotype are already known have to be compared.
- GSVA is done on single sample level and on normalized data
- *How much do the genes in the gene set of interest vary relative to the genes not in the gene set in the data*



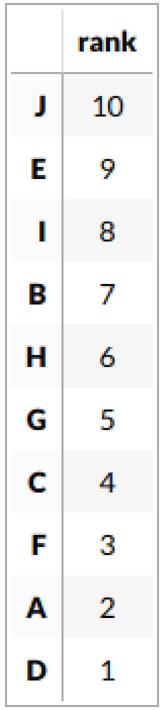
	S1	S2	S3
A	10	23	11
B	22	7	18
C	16	3	25
D	9	31	19
E	25	12	5
F	12	27	8
G	18	24	4
H	19	10	2
I	24	13	12
J	35	26	17

CDF ranked

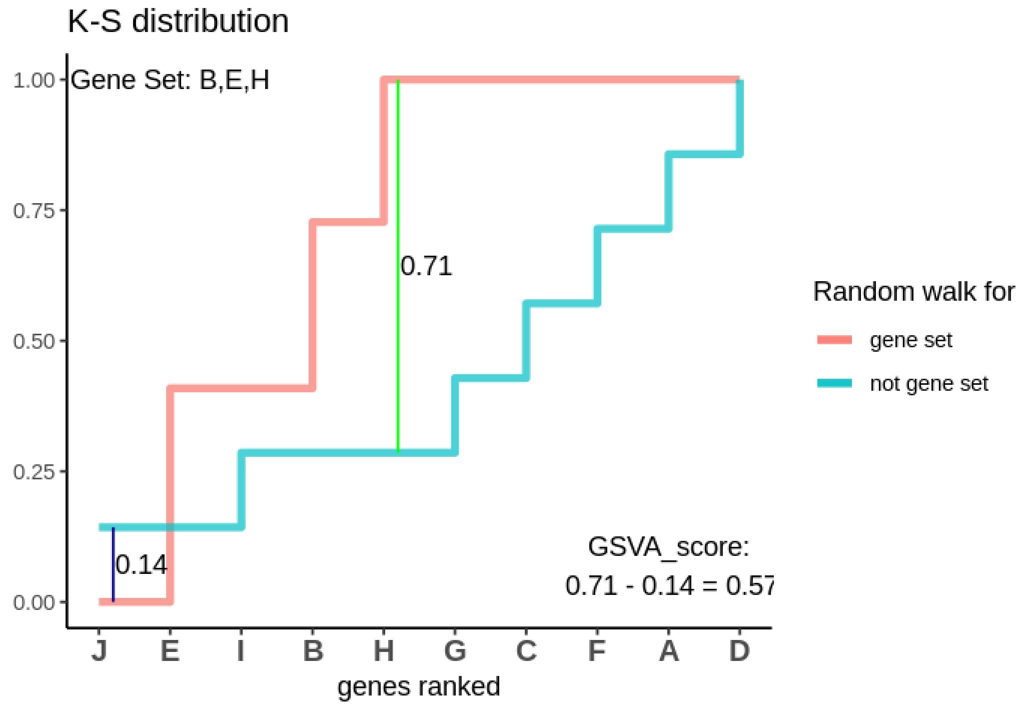


	S1	S2	S3
A	2	6	5
B	7	2	8
C	4	1	10
D	1	10	9
E	9	4	3
F	3	9	4
G	5	7	2
H	6	3	1
I	8	5	6
J	10	8	7

Order Per sample



	rank
J	10
E	9
I	8
B	7
H	6
G	5
C	4
F	3
A	2
D	1



https://medium.com/data-science/decoding-gene-set-variation-analysis-8193a0cfda3

Suggested Readings and Learning links

https://www.acgt.me/blog/2014/12/16/understanding-mapq-scores-in-sam-files-does-37-42

https://www.zymoresearch.com/blogs/blog/what-are-sam-and-bam-files?srsltid=AfmBOoodZOqnlsfZBPYoSoIkBx-IQoGdNBH5nxWcapeK0KscPlZ-rNkz

https://hutchdatascience.org/Choosing_Genomics_Tools/bulk-rna-seq-1.html

https://timd.one/blog/genomics/cigar.php

https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

https://onlinelibrary.wiley.com/doi/full/10.1002/qub2.78

https://www.pnas.org/doi/10.1073/pnas.0506580102

https://davetang.github.io/muse/gsva.html

https://pluto.bio/resources/Learning%20Series/gsea-vs-ora-two-key-pathway-analysis-approaches-for-next-gen-sequencing-data