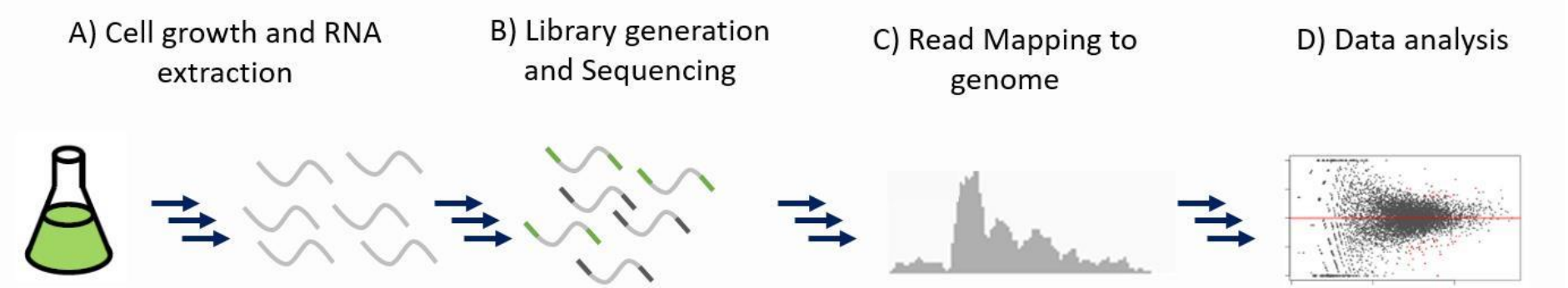
Bulk RNA sequencing Downstream analysis

Presenter: Nguyen Le Duc Minh

Protocol for RNA-seq Expression Analysis in Yeast



Background

- 1 µg down to 10 ng input RNA is sufficient for downstream amplification and library generation.
- RNA-seq can be applied to any population of extracted RNA, independent of the source organism.

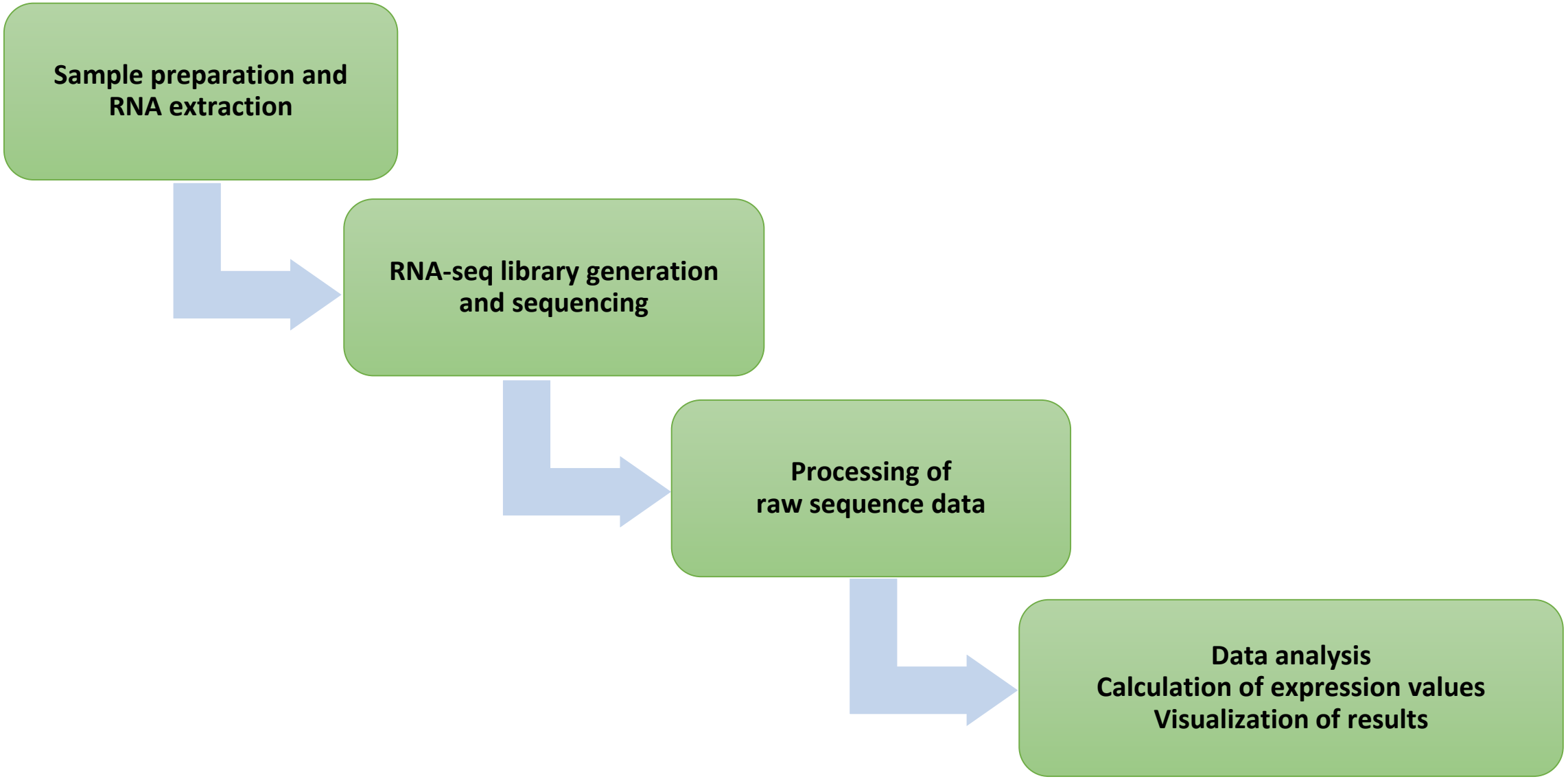
Background

- RNA-seq sequencing is performed by the detection of fluorescently labeled nucleic acids bound to the surface of flowcells.
- The RNA fragments are converted into a cDNA library and amplified, and flowcell adapters are introduced.
- During each sequencing cycle, DNA polymerases attach fluorescently labeled nucleotides to the flowcell-bound library molecules, which are then detected by the sequencer, typically generating read lengths of 150–300 bp to several Kbp.

Background

- RNA-seq expression analysis is a powerful and commonly used tool to identify genes that are up- or downregulated in a stressed sample.
  - *E.g.*, in the presence of genomic mutations, UV light, drugs, chemical or nutrient stress) as compared with a relaxed sample (*e.g.*, wild-type cell population
- A gene is “upregulated” or “downregulated,” respectively, when more or less of its RNA is measured (*i.e.*, expressed in the cell) under the stressed conditions as compared with the wild type.

Procedure for RNA-seq Expression Analysis in Yeast



Processing of raw sequence data

Preparation of data processing

- R
- Bowtie2
- Samtools
- Biocmanager + DESeq2-package

Processing of raw sequence read files (1)

1. Download the *S. cerevisiae* genome assembly and gene annotation
   - http://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/bigZips/
   - sacCer3.fa
   - sacCer3.ensGene.gtf

Processing of raw sequence read files (2)

2. Create the index files based on the genome assembly (“sacCer3.fa”)

   - bwa index sacCer3.fa

   (Caution: bwa is not a splice-aware alignment tool. If splice events need to be considered during analysis, aligners like “tophat” need to be used for index creation as well as alignment)

Processing of raw sequence read files (3)

3. Remove the random primer sequence, adapter contamination, and low-quality tails.
   - bbmap/bbduk.sh in=sample1.fastq out=sample1_trimmed.fastq ref=bbmap/resources/polyA.fa.gz,bbmap/resources/truseq.fa.gz k=13 ktrim=r forcetrimleft=11 useshortkmers=t mink=5 qtrim=t trimq=10 minlength=20; done
   - polyA-tail sequence
   - Illumina-specific adapter sequence information

Processing of raw sequence read files (4)

4. Create alignments of the pre-processed sequence reads using an alignment tool.
   - bwa mem sacCer3.fa sample1_trimmed.fastq > sample1_trimmed_aligned.sam

Processing of raw sequence read files (5)

5. Filter the data based on their quality by MAPQ filtering using samtools. All reads with an average base read quality score less than 50 will be removed.
   - samtools view -bq 50 sample1_trimmed_aligned.sam > sample1_trimmed_aligned_mapq50.bam

Processing of raw sequence read files (6)

6. Sort the filtered, aligned reads and create the index files using samtools.
   - samtools sort sample1_trimmed_aligned_mapq50.bam -o sample1_trimmed_aligned_mapq50_sorted.bam
   - samtools index sample1_trimmed_aligned_mapq50_sorted.bam

Data analysis, calculation of expression values, and visualization of results

- Differential gene expression (DGE) analysis aims to determine which, if any, genes show a higher or lower amount of aligned reads across the tested conditions.
- Reads belonging to a feature (i.e., a gene) are summed for each replicate -> differential expression values are calculated across conditions considering the variance within a condition among replicates.
- It is critical for DGE that several replicates of the same condition are considered.
- Gene expression values are usually reported as ***log2-fold changes***, in conjunction with adjusted ***P-values*** describing the significance of the change. (*P-value < 0.05*)

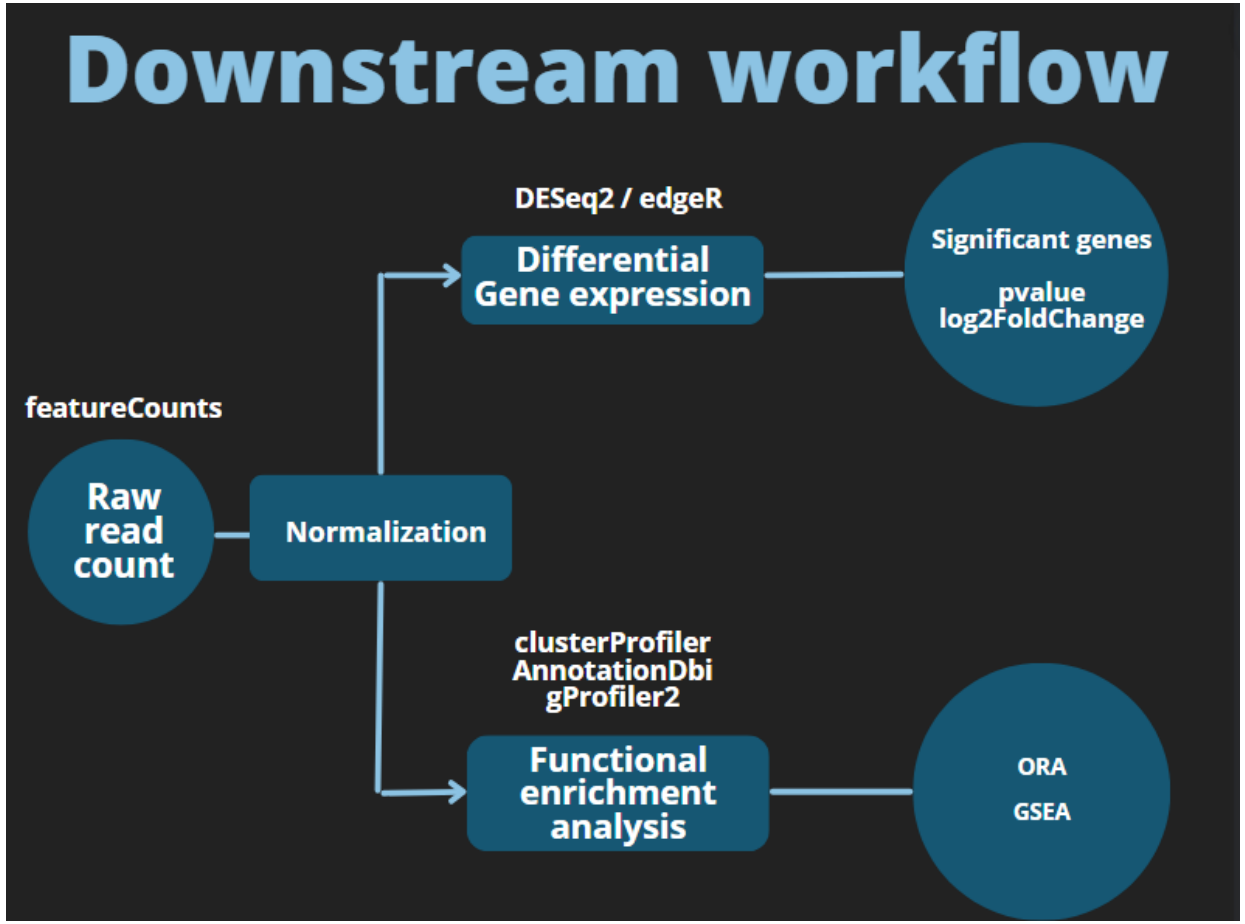
Normalization methods for DGE analysis

- The number of aligned reads can differ strongly between replicates due to technical reasons -> reads must be normalized across replicates and conditions.
- Normalization methods for the calculation of DGE:
  - Reads Per Kilobase of transcript per Million mapped reads (RPKM)
  - Fragments Per Kilobase of transcript per Million mapped reads (FPKM)
  - Transcript per Million reads (TPM)
  - Counts per feature (*i.e.,* genes)

DESeq2 normalization

- This RNA-seq analysis protocol in yeast was normalized using the count-based normalization by DESeq2, based on the assumption that most genes are not differentially expressed across conditions.
- The counts per features are extracted, combined and indexed.
- Finally the counts are normalized, and DGE values are calculated using DESeq2.

Workflows summary for bulk RNA-seq downstream analysis



Generate feature counts matrix for all samples from bam files

1. Extract the counts for each sample using desired tools from the aligned, filtered and sorted bam files combined with the gene annotation file (sacCer3.ensGene.gtf). This will generate a *.txt file containing the number of reads assigned to each gene annotated in the gtf-file.

2. Count-based expression values are calculated using R and DESeq2. This step requires the count data to be assembled in a text document(“count.txt”) as well as an index file (“table.txt”).

region_name	WT_1	WT_2	WT_3	Mut_1	Mut_2	Mut3
YAL069W 1	0	2	0	5	2	4
YAL068W-A	3	0	2	4	1	0
YAL068C 0	27	14	43	32	12	17
…						
…						
…						
YPR202W0	2	0	4	6	11	14
YPR203W0	0	0	0	0	0	0
YPR204W0	0	0	0	0	0	0
YPR204C-A	0	0	0	0	0	0

Importantly, the read counts are not yet normalized to the total number of read counts in each sample, and respective variations are expected.

Labeling each sample with its relevant condition/group

3. Generate a “table.txt”- file for each sample, indexing each column of data in tab-delimited format.

sample_name	condition
WT_1	WT
WT_2	WT
WT_3	WT
Mut_1	Mutant
Mut_2	Mutant
Mut_3	Mutant

3. In R, load the DESeq2 library, the combined counts-file from step 2, and the table-file from step 3.

```
  library(DESeq2)
  count_table                                                    <-
read.delim('counts.txt',sep='\t',header=TRUE,row.names='region_name')
  sample_table                                                   <-
read.delim('table.txt',sep='\t',header=TRUE,row.names='sample_name')
```

Using DESeq2 library to perform DGE analysis

5. Write the RNA-seq expression and p-values to file using DESeq2. The generated .txt file (wt_mutant_p-values.txt) contains the log2-fold change and p-values for the respective condition in tab-delimited format and can be used for further downstream analysis or visualization.

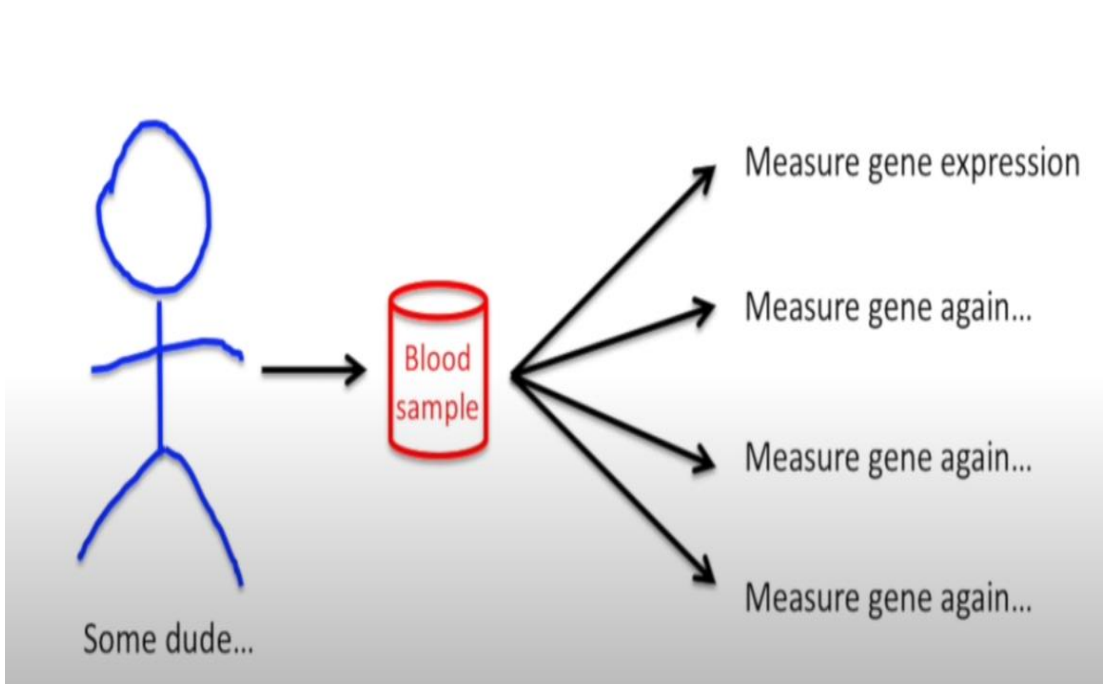
```
dds <- DESeqDataSetFromMatrix(countData = count_table,colData =
sample_table,design = ~ condition)
dds <- DESeq(dds)
res <- results(dds)
resOrdered <- res[order(res$padj),]
plot <- plotMA(res, main = 'mutant', ylim = c(-2,2), xlab = 'mean
count')
write.table(as.data.frame(resOrdered),sep='\t',quote=FALSE,file='wt_m
utant_p-values.txt')
```

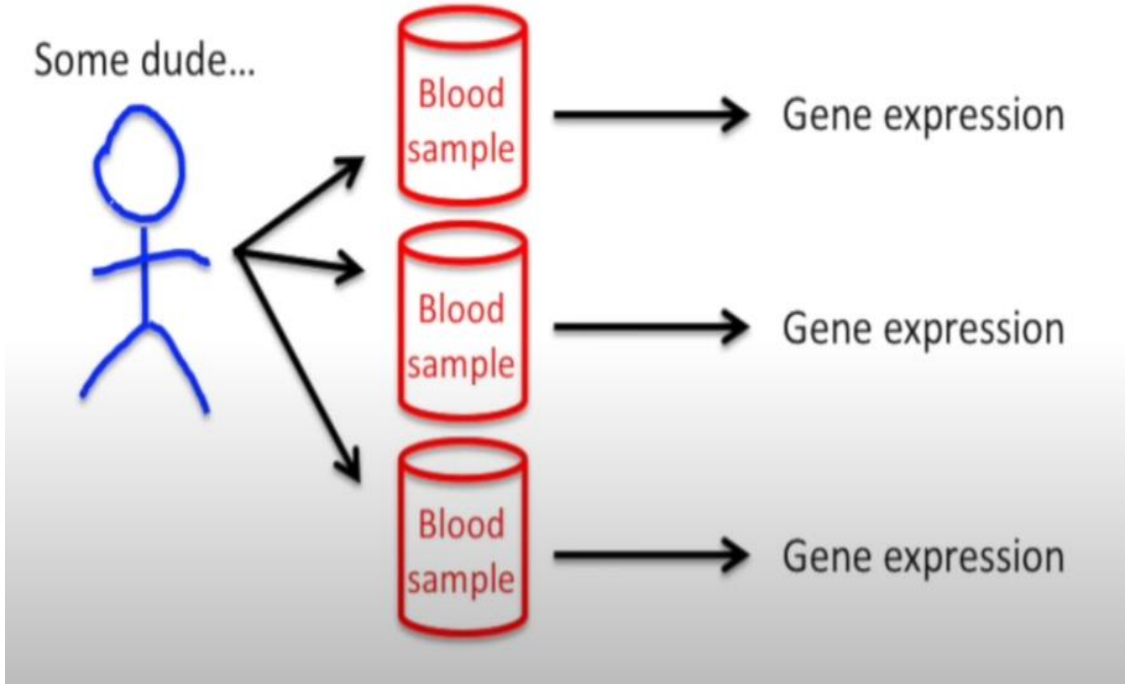
RNA-seq Normalization Methods

RPKM/FPKM and TPM

Technical Replicates

- Samples in which the starting biological sample is the ***same***, but the replicates are processed separately. For example, if a biological sample is divided and two different library preps are processed and sequenced, those two samples would be considered technical replicates.



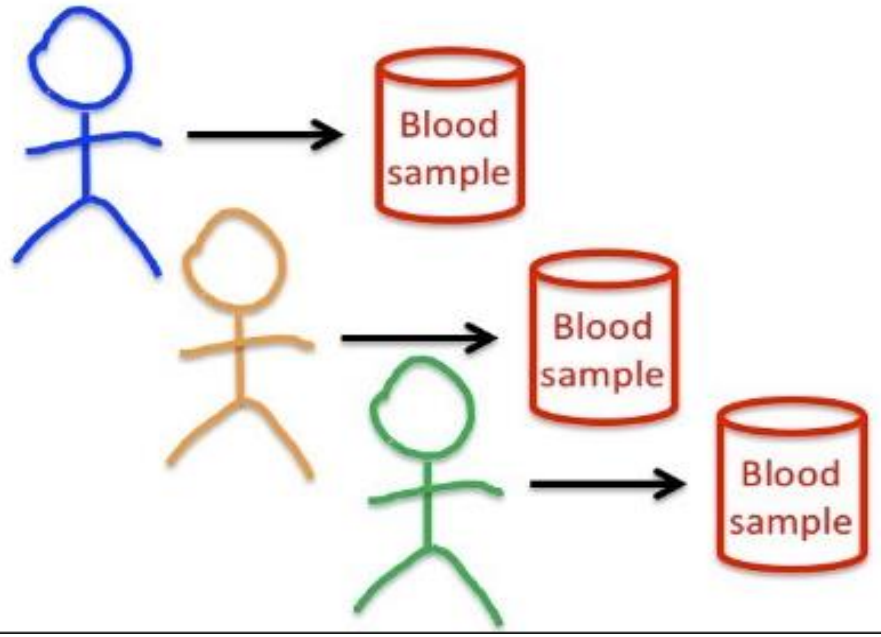


https://www.melbournebioinformatics.org.au/tutorials/tutorials/rna_seq_exp_design/rna_seq_experimental_design/

https://www.youtube.com/watch?v=Exk0OoRG0PQ

Biological Replicates

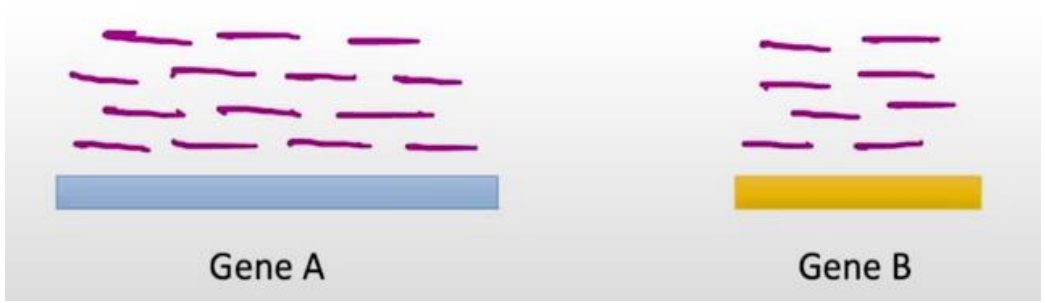
- Samples that have been obtained from ***biologically separate samples***. This can mean different individual organisms (e.g. tissue samples from different mice), different samplings of the same tumour, or different population of cells grown separately from each other but originating from the same cell-line. For example, the samples obtained from three different knock-out mice could be considered biological replicates in a knock-out versus wild-type experiment. A biological replicate combines both technical and biological variability as it is also an independent case of all the technical steps.



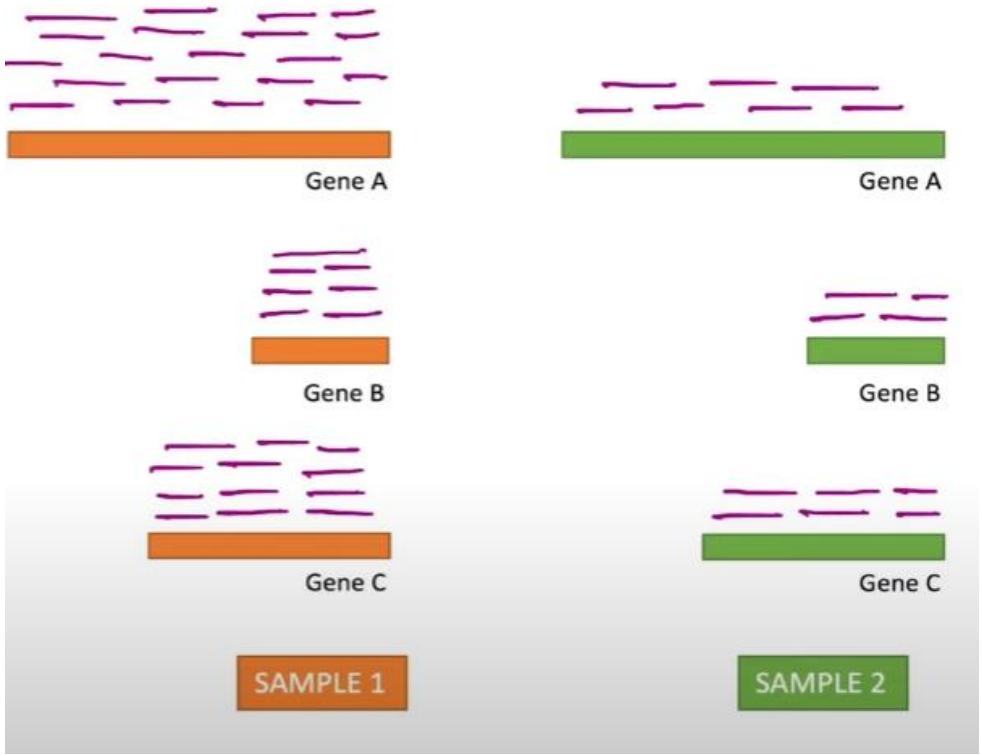
- Use different biological sources of samples (i.e. different people, plants, or cell lines).

https://www.melbournebioinformatics.org.au/tutorials/tutorials/rna_seq_exp_design/rna_seq_experimental_design/

Biases



Gene length



Sequencing depth

RPKM/FPKM and TPM

1. **RPKM** (Reads Per Kilobase of transcript per Million mapped reads)

   - Normalizes for gene length and sequencing depth
   - Higher the RPKM of a gene, higher the expression
   - Used to quantify transcripts from single-end reads
   - **NOT TO BE** used for Differential gene expression analysis (DESeq2/edgeR)

2. **FPKM** (Fragments Per Kilobase of transcript per Million mapped reads)

   - Analogues to RPKM
   - Used for pair-end data, read pair, rather than a single read corresponds to a cDNA fragment
   - Higher the FPKM of a gene, higher the expression
   - **NOT TO BE** used for Differential gene expression analysis (DESeq2/edgeR)

1. **TPM** (Transcript per Million reads)

   - Normalizes for gene length and sequencing depth
   - TPM is better suited to compare expression between samples
   - **NOT TO BE** used for Differential gene expression analysis (DESeq2/edgeR)

Performing RPKM/FPKM normalization

	Library size*	6M	6M	6M
Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5 kb	Gene A	50	25	85
2 kb	Gene B	75	50	90
	…	…	…	…
	Total number of reads mapped = (sequencing depth)	5M	3M	4M

Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5 kb	Gene A	50	25	85
2 kb	Gene B	75	50	90
	Total number of reads mapped = (sequencing depth)	125	75	175

Sequencing depth normalization

- **Step 1**: Normalize for sequencing depth

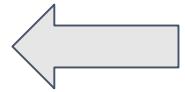
RPM(replicate 1) = (50/125) x scaling_factor (10^6) = 0.4 * 10 = 4

RPM(replicate 2) = (25/75) * 10 = 3.33

…

Gene length normalization

Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5 kb	Gene A	4	3.33	4.85
2 kb	Gene B	6	6.66	5.14



RPM values

- **Step 2**: Normalize for gene length

RPKM = (4 / 1.5) = 2.66

= (6 / 2) = 3

…

RPKM values

Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5 kb	Gene A	2.66	2.22	3.23
2 kb	Gene B	3	3.33	2.57

Rewriting the formula

$$RPKM = \frac{number\ of\ reads\ of\ the\ region}{\frac{total\ reads}{1{,}000{,}000} \times \frac{region\ length}{1{,}000}}$$

Performing TPM normalization

Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5 kb	Gene A	50	25	85
2 kb	Gene B	75	50	90
	Total number of reads mapped = (sequencing depth)	125	75	175

- **Step 1**: Normalize for gene length

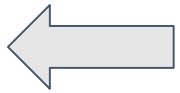
RPK = (50 / 1.5) = 33.33

= (75 / 2) = 37.5

…

Performing TPM normalization

Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5 kb	Gene A	33.33	16.66	56.66
2 kb	Gene B	37.5	25	45



RPK Values

- **Step 2** : Normalize for <u>sequencing depth</u>

(RPK / total RPK in each replicate) * scaling_factor

RPKM = (33.33 / 70.83) * 10 = 4.7

= (25 / 41.66) * 10 = 0.6

...

TPM values

Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5 kb	Gene A	4.71	3.99	5.57
2 kb	Gene B	5.29	6	4.426

Compare and interpret RPKM and TPM normalization values

Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5 kb	Gene A	2.66	2.22	3.23
2 kb	Gene B	3	3.33	2.57
	Sum for each replicate	**5.66**	**5.55**	**5.8**

RPKM Values

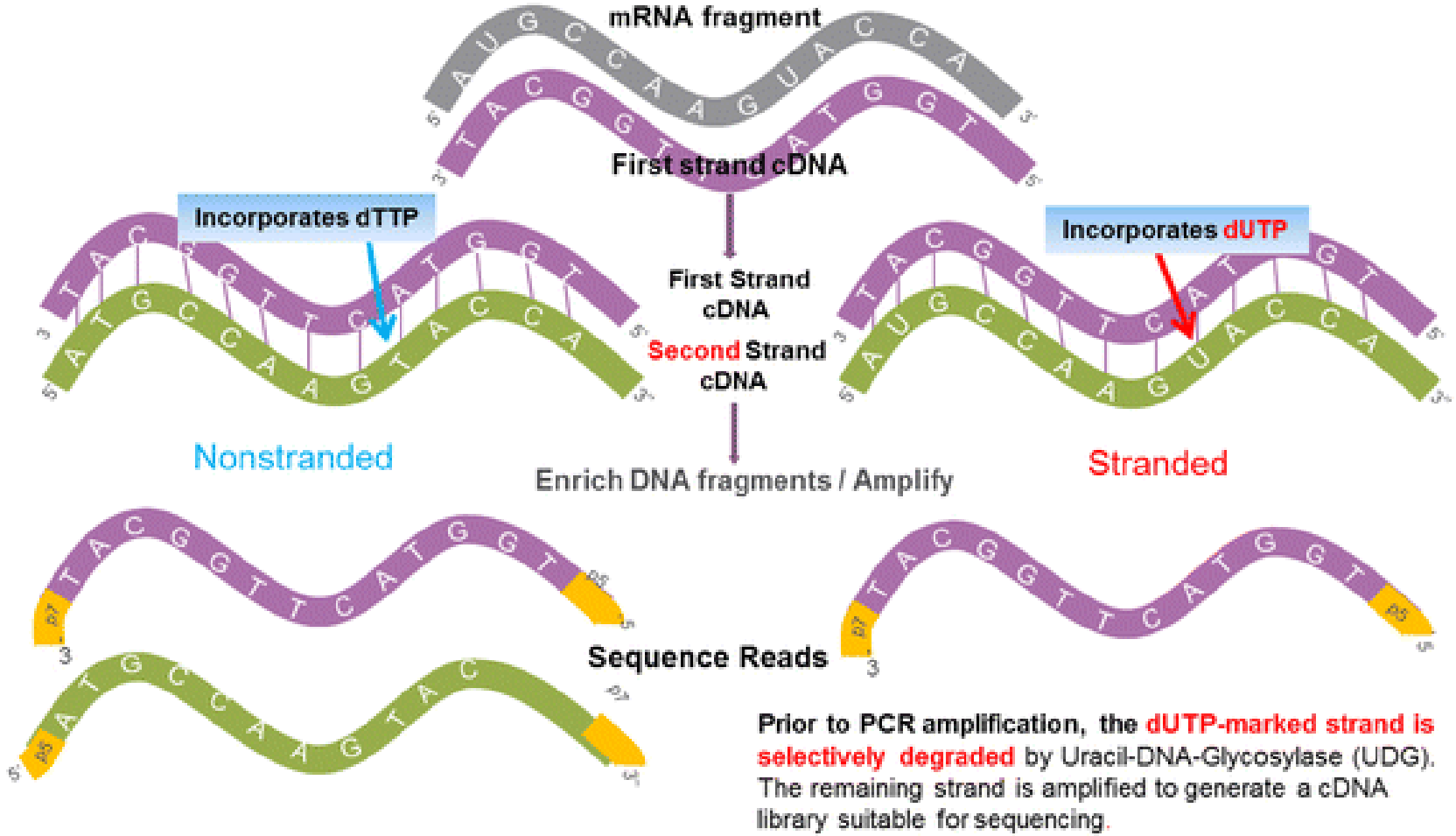
Gene lengths	Genes	Technical replicate 1	Technical replicate 2	Technical replicate 3
1.5 kb	Gene A	4.71	3.99	5.57
2 kb	Gene B	5.29	6	4.426
	Sum for each replicate	**10**	**9.99**	**9.996**

TPM Values

https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

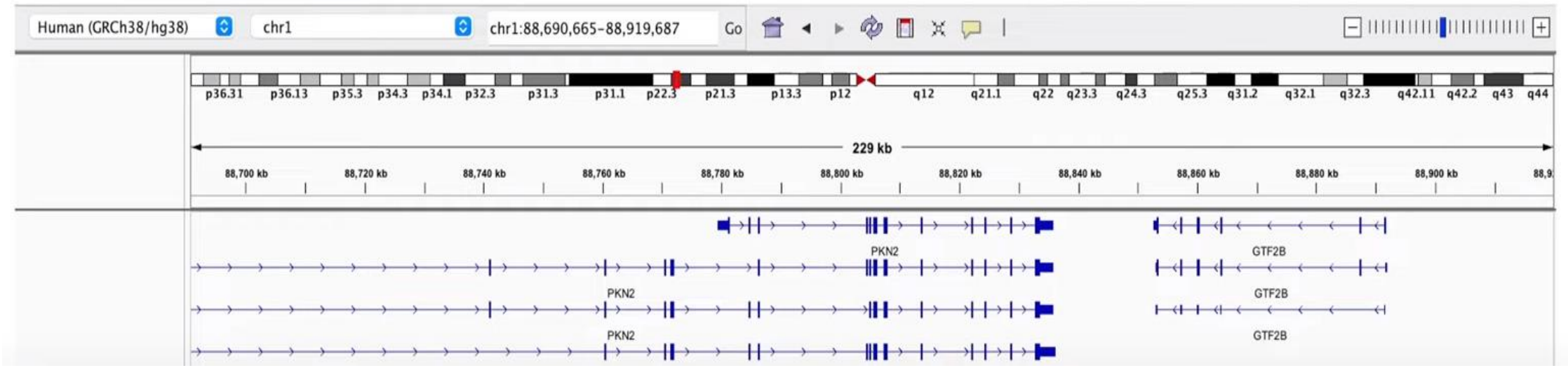
RNA-seq Strandedness

Stranded vs Unstranded



https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1876-7

Why strand-specific information from RNAseq reads is important ?



Strandedness in transcription

5′-**ATGTCGGTCAGCTCGAGCCGCAGCAGATGAGCAGCTAGCTAA**-3′ ← Sense/non-template strand

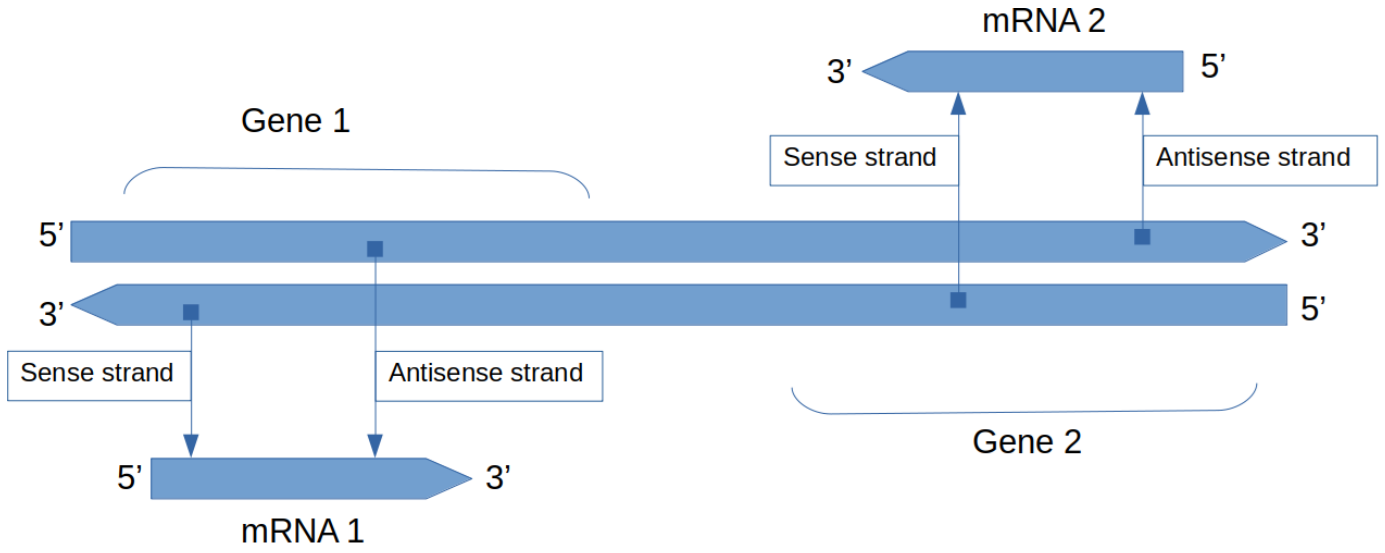
3′-TACAGCCAGTCGAGCTCGGCGTCGTCTACTCGTCGATCGATT-5′ ← Antisense/template strand

Transcription

messenger RNA

5′-**AUGUCGGUCAGCUCGAGCCGCAGCAGAUGAGCAGCUAGCUAA**-3′

- Sense strand is the same as mRNA except T is replaced by U
- Antisense strand acts a template for the mRNA synthesis. Therefore antisense strand is complementary reverse to both, the sense strand and the mRNA.



Library preparation methods

- **Unstranded**

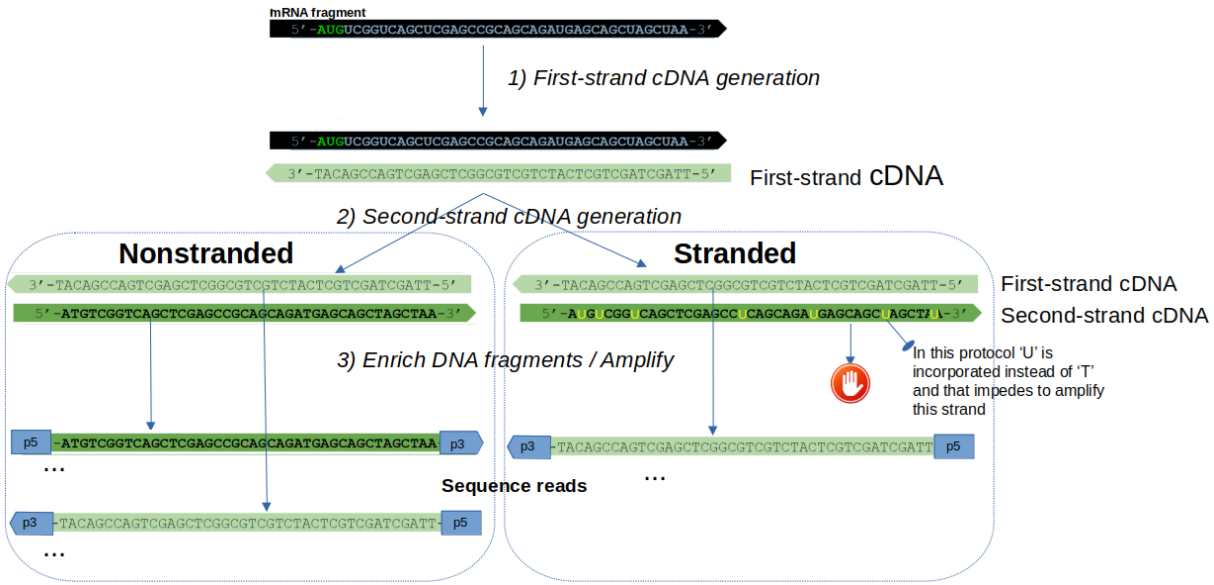
Information regarding the strand is not conserved(it is lost during the amplification of the mRNA fragments).

- **Directional, first strand (reverse)**

The second read (read 2) is from the original RNA strand/template, first read (read 1) is from the opposite strand.

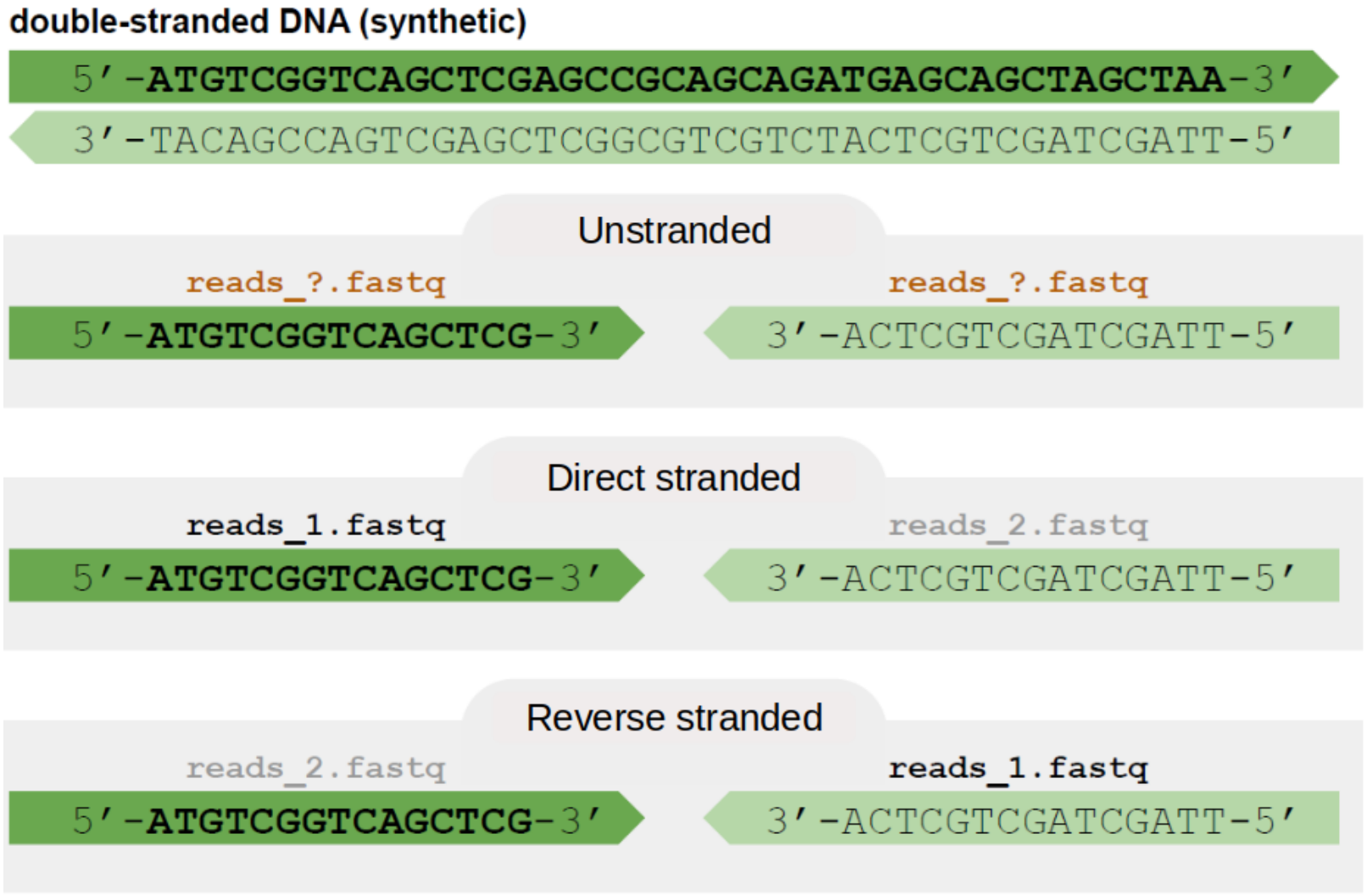
- **Directional, second strand (direct)**

The first read (read 1) is from the original RNA strand/template, second strand is from the opposite strand.



https://bioinfogp.cnb.csic.es/tools/seqnjoy/help/#appendix-2-strandness-and-rna-seq

Reads in .fastq files based on library construction method



Practicalities of the RNAseq experiment

- It is recommended to check the provider's report of the sequencing to know which library prep method was used: *unstranded* (non-stranded), *direct-stranded* or *reverse-stranded*.

Protocol	Synonym	Examples/Kits[1]	Description
Unstranded	non-stranded	Standard Illumina, TruSeq RNA Sample Prep kit, NuGEN OvationV2, SMARTer universal low input (TaKara), GDC normalized TCGA data	Information regarding the strand is not conserved (lost)
Reverse	first_strand	dUTP, NSR, NNSR; TruSeq Stranded (Total RNA/mRNA), NEB Ultra Directional, Agilent SureSelect Strand-Specific	The second read (read 2) is from the original RNA strand/template, first read (read 1) is from the opposite (reverse) strand.
Direct	second_strand	Directional Illumina (Ligation), Standard SOLiD, ScriptSeq v2, SMARTer Stranded Total RNA, NuGEN Encore Complete, NuGEN SoLo, Illumina ScriptSeq	The first read (read 1) is from the original (direct sense) RNA strand/template, second read (read 2) is from the opposite strand.

Terminologies

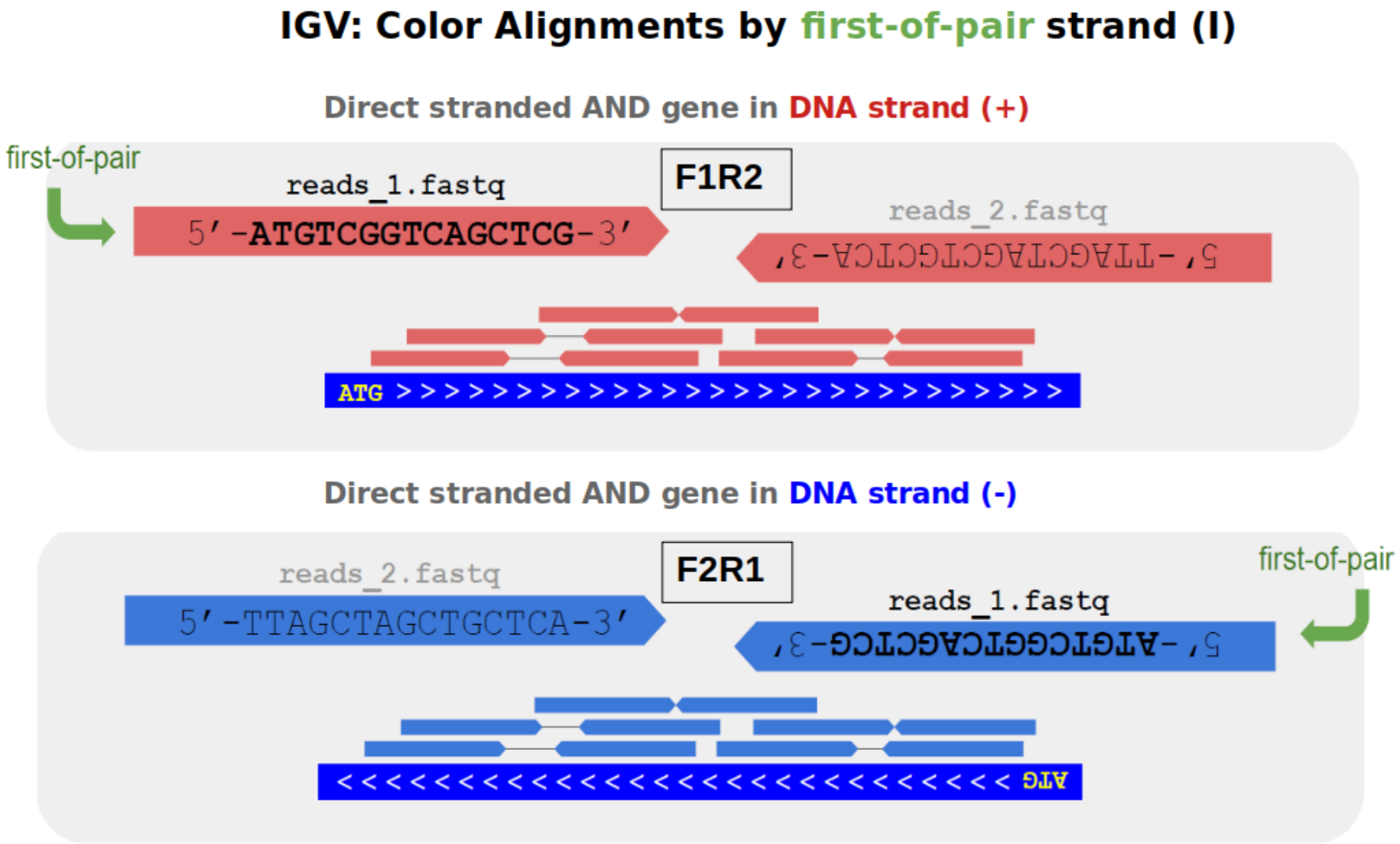
- **Paired-end reads**:
  - RF/fr-firststrand/Reverse Stranded:
    - First read (/1) of fragment pair is sequenced as anti-sense (reverse)
    - Second read (/2) is in the sense strand (forward)
    - Typical of dUTP/UDG sequencing method
  - FR/fr-secondstrand/Direct Stranded:
    - First read (/1) of fragment pair is sequenced as sense (forward)
    - Second read (/2) is in the anti-sense strand (reverse)
    - Typical of Directional Illumina (Ligation) and Standard SOLiD protocols
- **Single-end reads**:
  - F: the single read is in the sense (forward) direction
  - R: the single read is in the anti-sense (reverse) direction

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1876-7

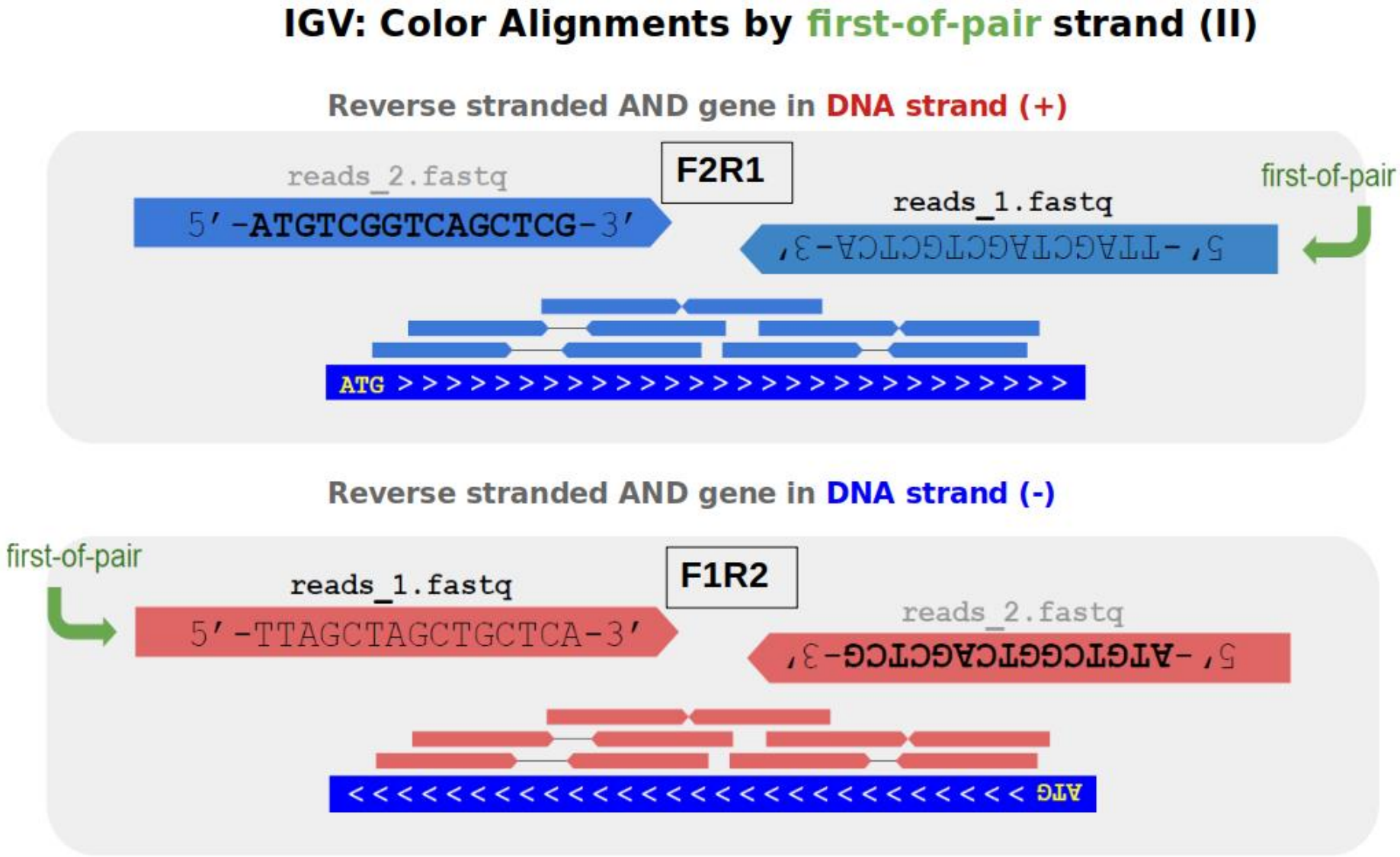
How to determine strandedness from the RNAseq data

- RSeQC: An RNA-seq Quality Control package
- STAR: The --quantMode GeneCounts option will output a file with suffix “ReadsPerGene.out.tab”, which counts the number of reads mapped to each gene.
- Salmon: The --libType A option will allow Salmon to automatically infer the library type.
- Visualization in IGV.

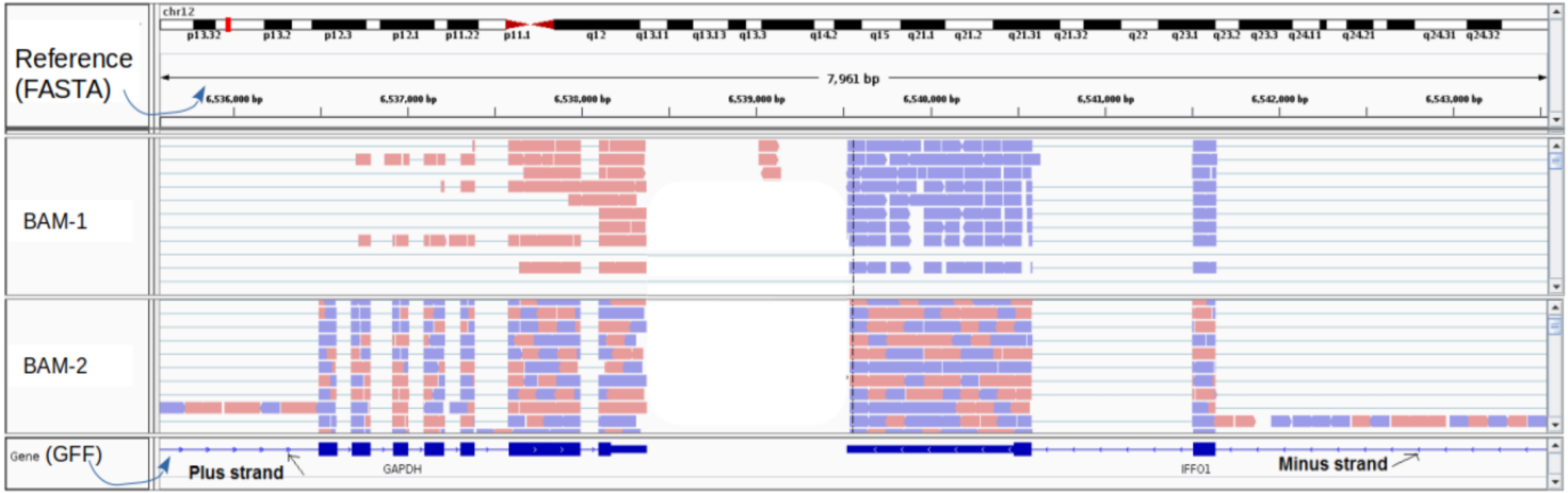
Direct stranded mode



Reverse stranded mode



What are the strandness protocols in BAM-1 and BAM-2 ?

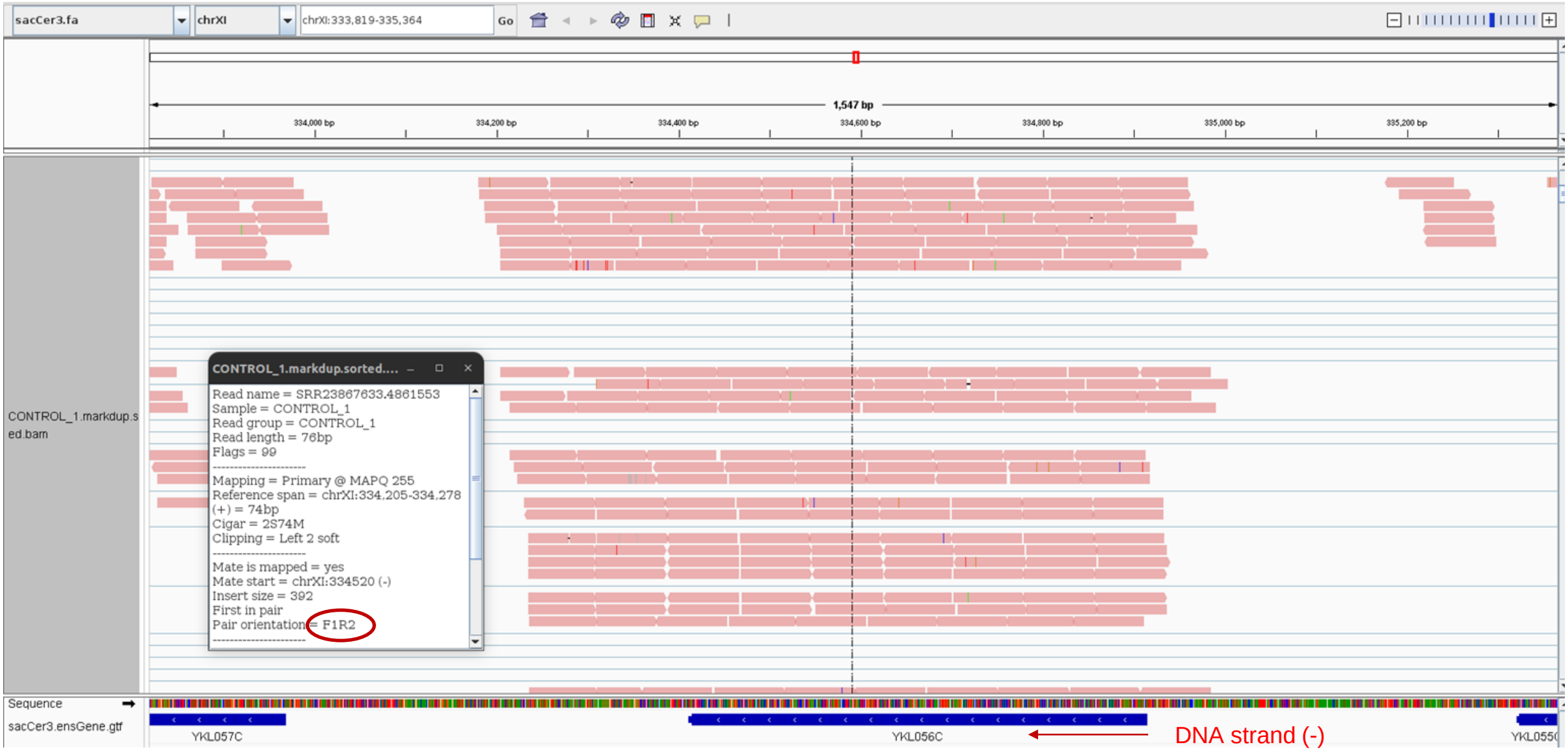


Strand-related settings

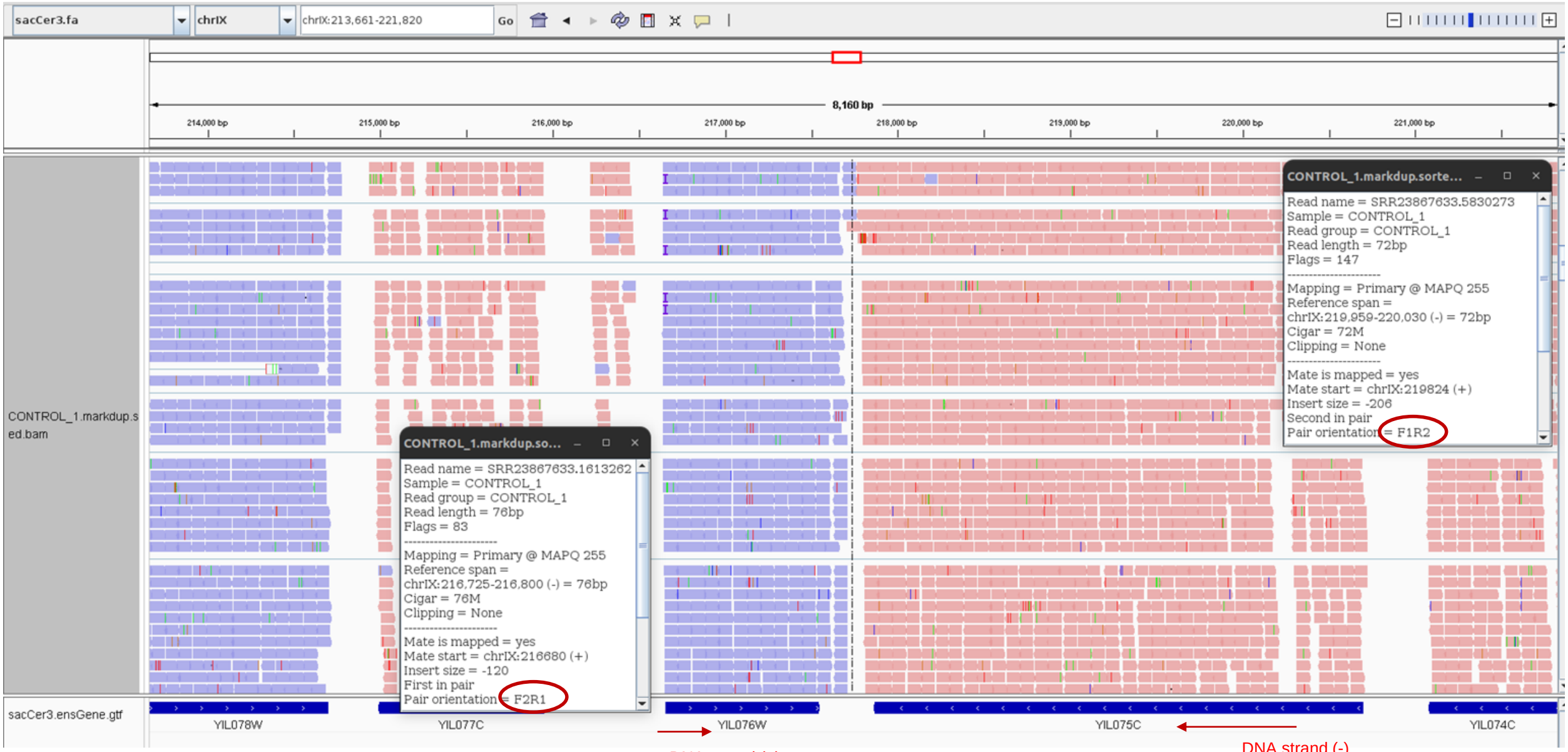
Tool	RF/fr-firststrand stranded (dUTP)	FR/fr-secondstrand stranded (Ligation)	Unstranded
check_strandedness (output)	RF/fr-firststrand	FR/fr-secondstrand	unstranded
IGV (5p to 3p read orientation code)	F2R1	F1R2	F2R1 or F1R2
TopHat (--library-type parameter)	fr-firststrand	fr-secondstrand	fr-unstranded
HISAT2 (--rna-strandness parameter)	R/RF	F/FR	NONE
HTSeq (--stranded/-s parameter)	reverse	yes	no
STAR	n/a (STAR doesn't use library strandedness info for mapping)	NONE	NONE
Picard CollectRnaSeqMetrics (STRAND_SPECIFICITY parameter)	SECOND_READ_TRANSCRIPTION_STRAND	FIRST_READ_TRANSCRIPTION_STRAND	NONE
Kallisto quant (parameter)	--rf-stranded	--fr-stranded	NONE
StringTie (parameter)	--rf	--fr	NONE
FeatureCounts (-s parameter)	2	1	0
RSEM (–forward-prob parameter)	0	1	0.5
Salmon (--libType parameter)	ISR (assuming paired-end with inward read orientation)	ISF (assuming paired-end with inward read orientation)	IU (assuming paired-end with inward read orientation)
Trinity (–SS_lib_type parameter)	RF	FR	NONE
MGI CWL YAML (strand parameter)	first	second	NONE
WASHU WDL YAML (strand parameter)	first	second	unstranded
RegTools (strand parameter)	-s 1	-s 2	-s 0
Example kits	**Example methods/kits:** dUTP, NSR, NNSR, Illumina TruSeq Strand Specific Total RNA, NEBNext Ultra II Directional	**Example methods/kits:** Ligation, Standard SOLiD, NuGEN Encore, 10X 5' scRNA data	**Example kits/data:** Standard Illumina, NuGEN OvationV2, SMARTer universal low input RNA kit (TaKara), GDC normalized TCGA data

https://rnabio.org/module-09-appendix/0009/12/01/StrandSettings/

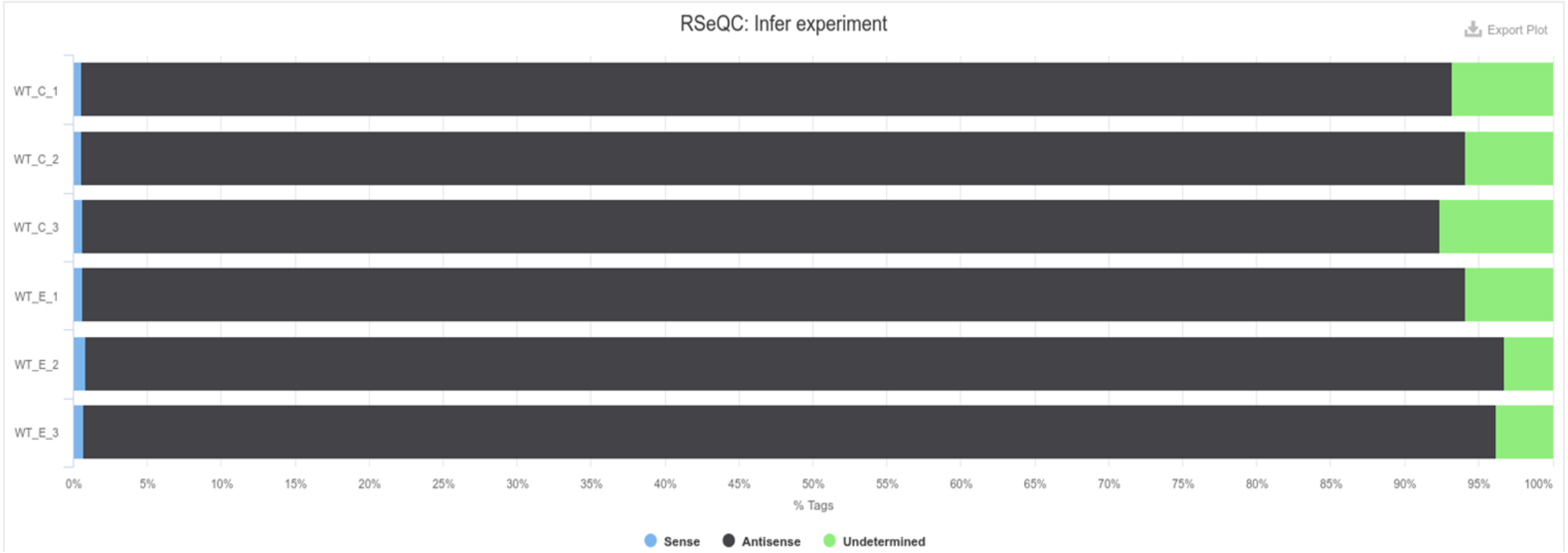
IGV view for strandness



IGV: Based on alignment reads on DNA strand



Infer experiment



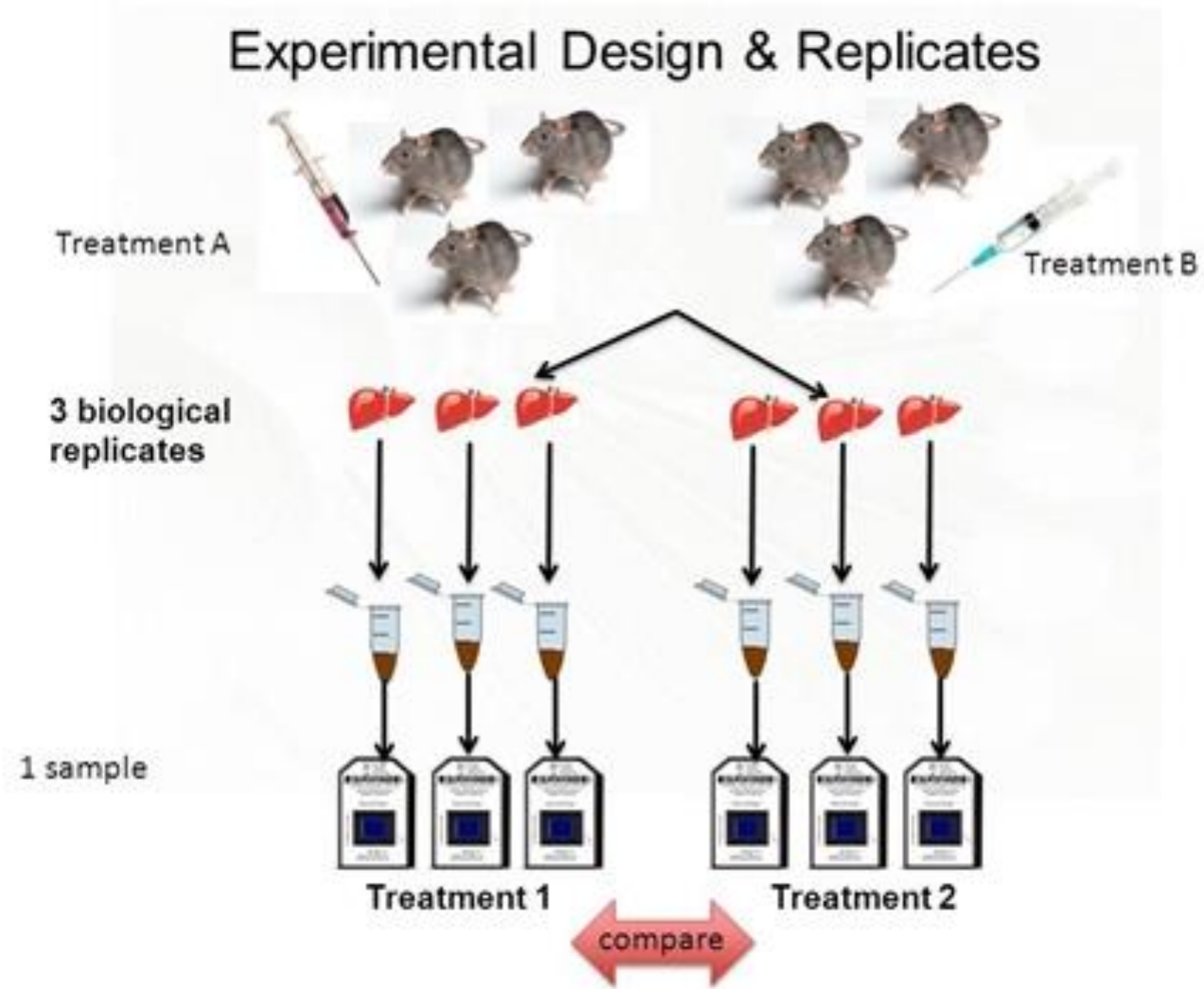
Strandedness identification based on raw reads

```
checking strandedness
Reading reference gene model stranded_test_WT_E_1_R1/genes.bed ... Done
Loading SAM/BAM file ...  Total 200000 usable reads were sampled
This is PairEnd Data
Fraction of reads failed to determine: 0.0296
Fraction of reads explained by "1++,1--,2+-,2-+": 0.0090 (0.9% of explainable reads)
Fraction of reads explained by "1+-,1-+,2++,2--": 0.9614 (99.1% of explainable reads)
Over 90% of reads explained by "1+-,1-+,2++,2--"
Data is likely RF/fr-firststrand
```

Differential Gene Expression Analysis

DESeq2 package

Study design to analyse differential gene expression



Features of RNAseq counts data matrix

countData

gene	ctrl_1	ctrl_2	exp_1	exp_1
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

colData

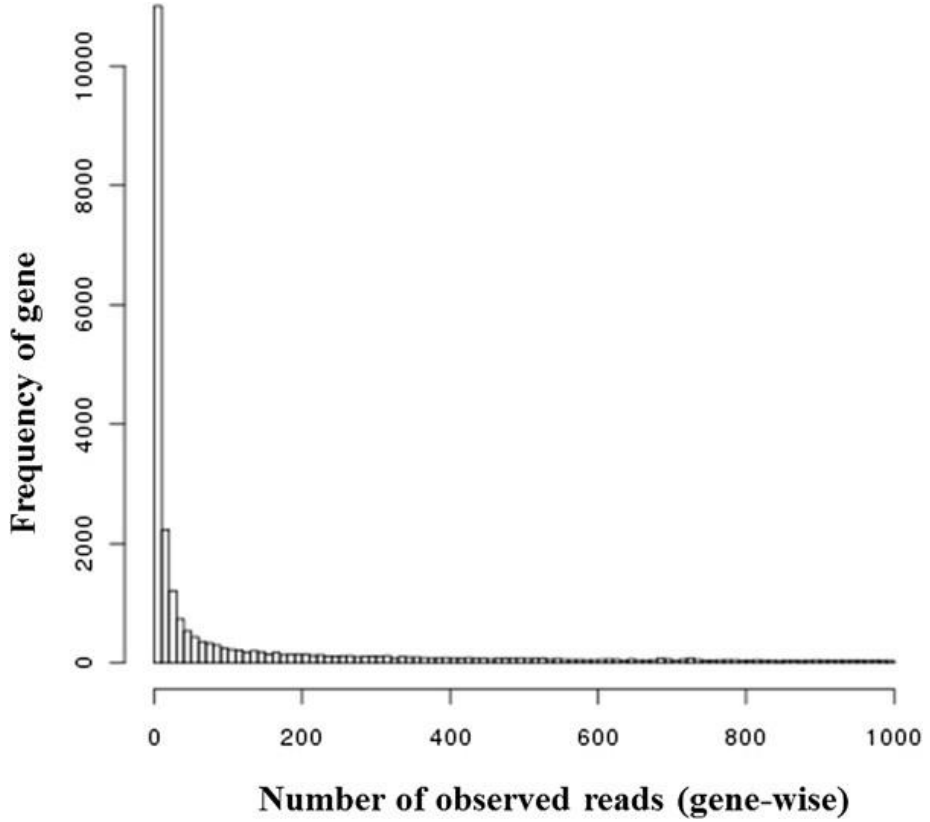
id	treatment	sex
ctrl_1	control	male
ctrl_2	control	female
exp_1	treatment	male
exp_2	treatment	female

Sample names:
ctrl_1, ctrl_2, exp_1, exp_2

countData is the count matrix
(number of reads mapping to each gene for each sample)

colData describes metadata about the *columns* of countData

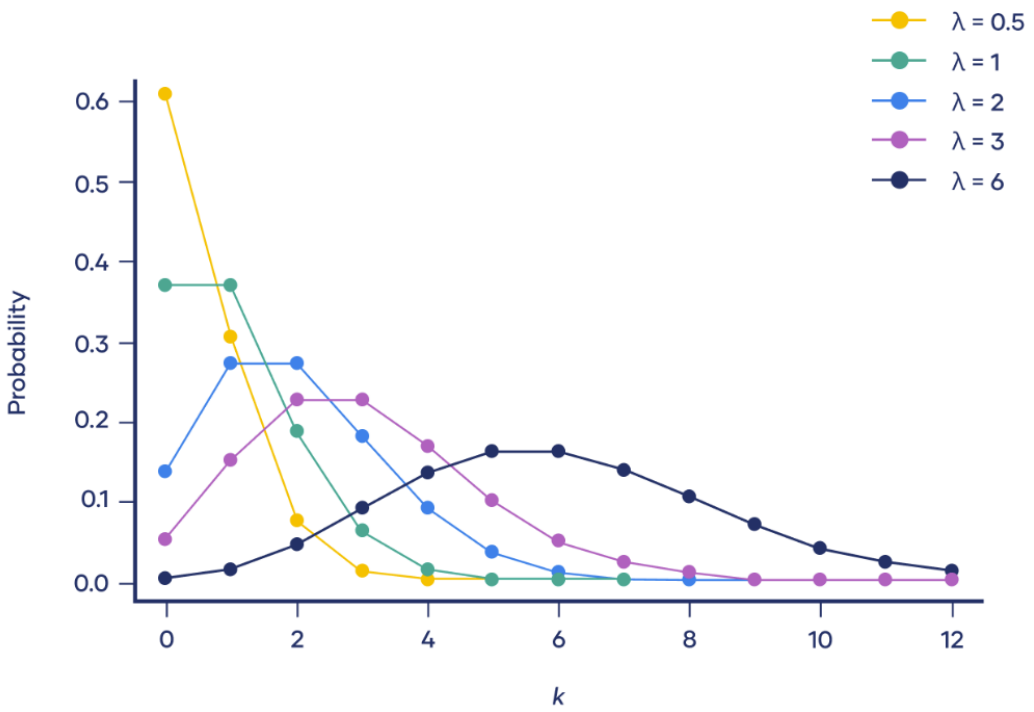
First column of colData must match column names of countData (-1st)



https://www.researchgate.net/figure/Distribution-of-number-of-observed-reads-per-gene-for-genes-with-read-count-less-than_fig5_256188117

https://4va.github.io/biodatasci/r-rnaseq-airway.html

Poisson distribution



https://www.biostars.org/p/84445/

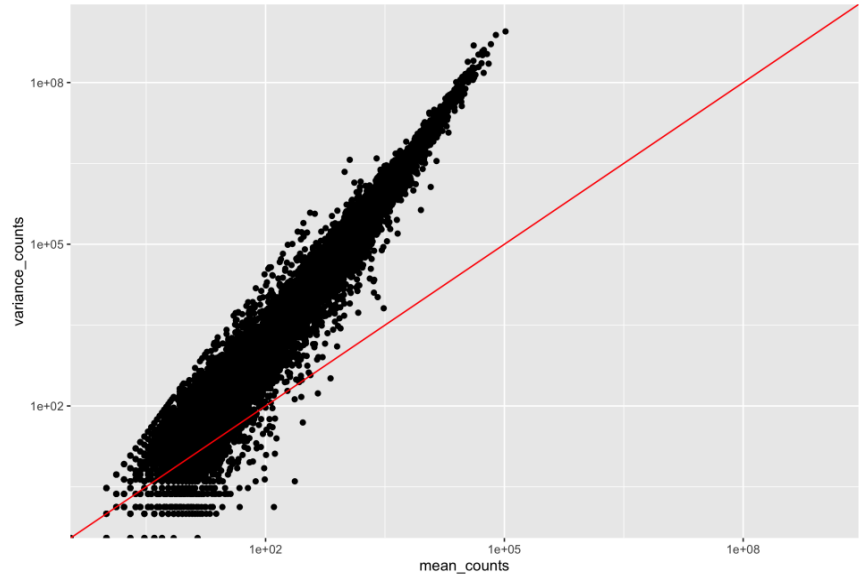
Why Poisson ?

- Number of cases are large, probability of an event happening is low
- In context of RNAseq: Selecting mRNA (event) from a large number of RNA molecules (cases)

Why does Poisson not work ?

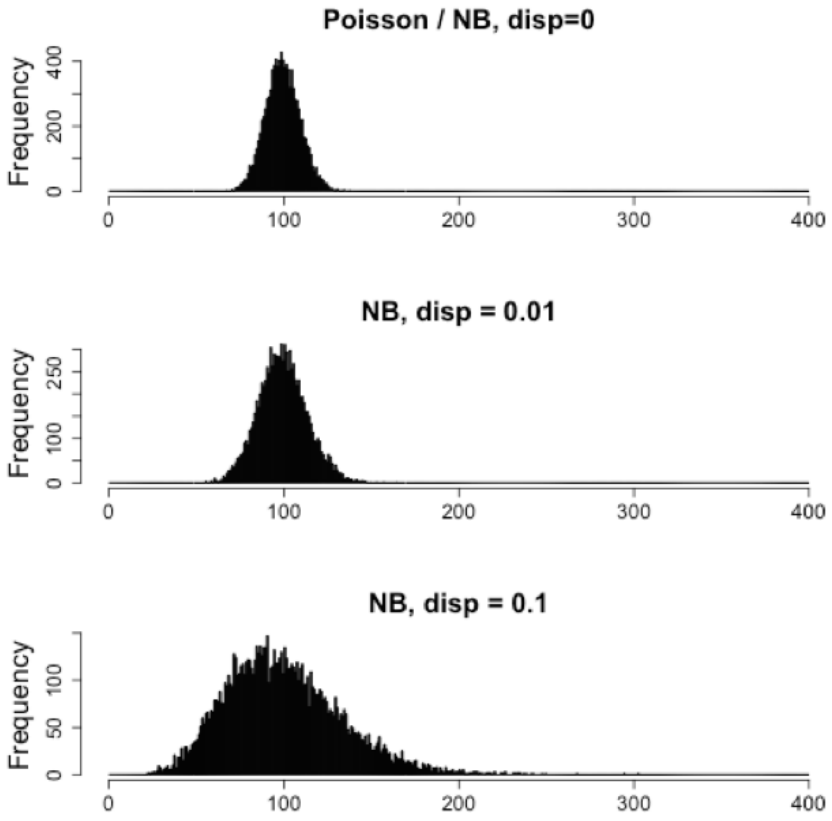
- Poisson distribution has only one parameter indicating its expected mean: λ
- Λ = mean = variance

mean # variance



https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/01c_RNAseq_count_distribution.html

An alternative: The Negative Binomial distribution



K: counts
Gene i
Sample j

Dispersion
“extra variability”

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

Fitted mean

Sample-sp.size factor

$$\mu_{ij} = s_j . q_{ij}$$

~ expected true concentration of fragments for sample j

$$\log_2(q_{ij}) = \sum_r x_j . \beta_{ir}$$

raw count for gene i, sample j

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

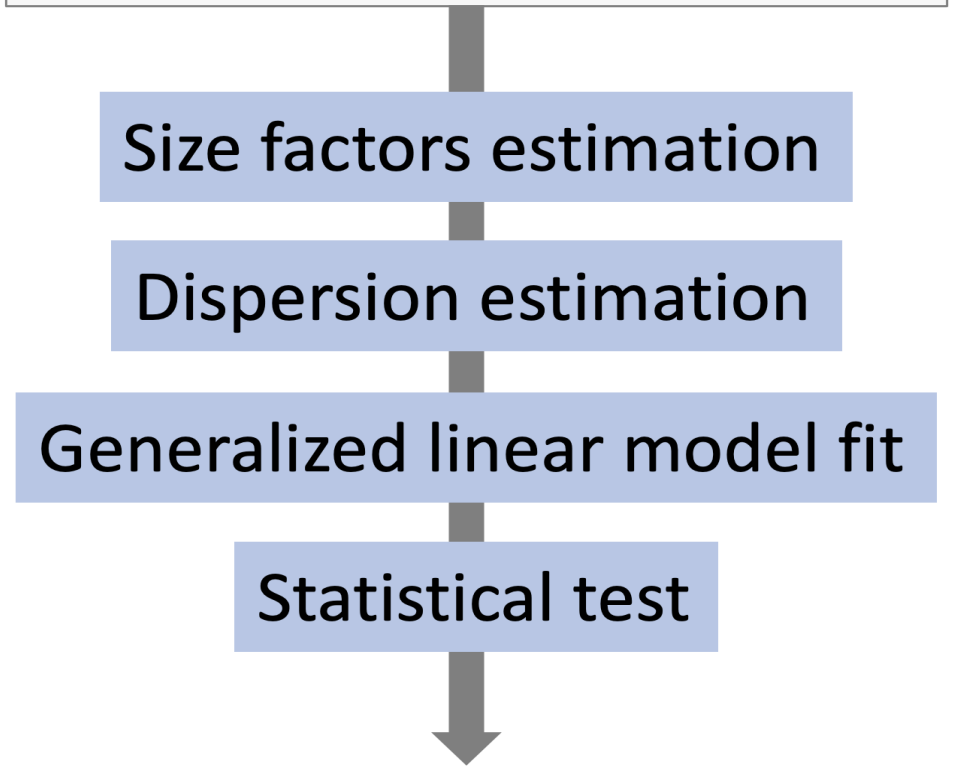
one dispersion per gene

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

DESeq2 processing steps

Count matrix

	sample1	sample2	sample3	sample4	sample5	sample6
ENSG00000223972	0	0	0	0	0	1
ENSG00000227232	14	28	17	40	16	13
ENSG00000278267	8	4	3	1	1	6
ENSG00000243485	0	0	0	0	0	0
ENSG00000284332	0	0	0	0	0	0
ENSG00000237613	0	0	0	0	0	0



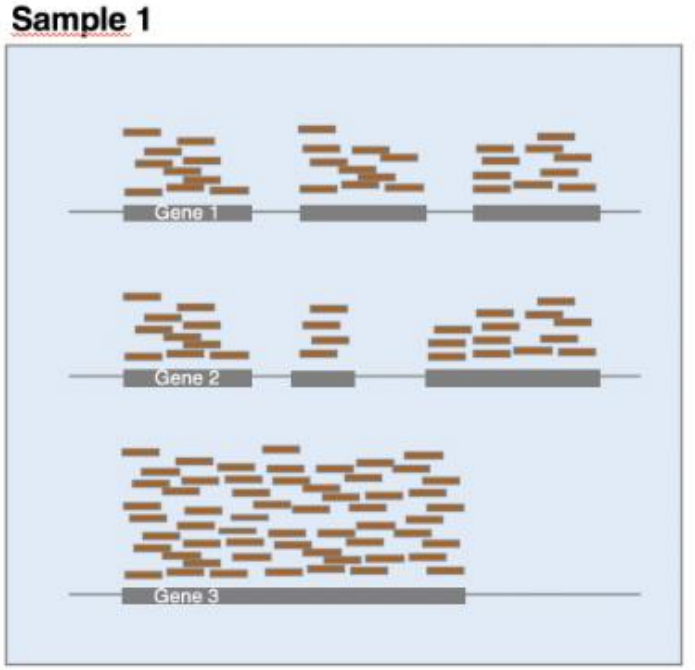
For each gene

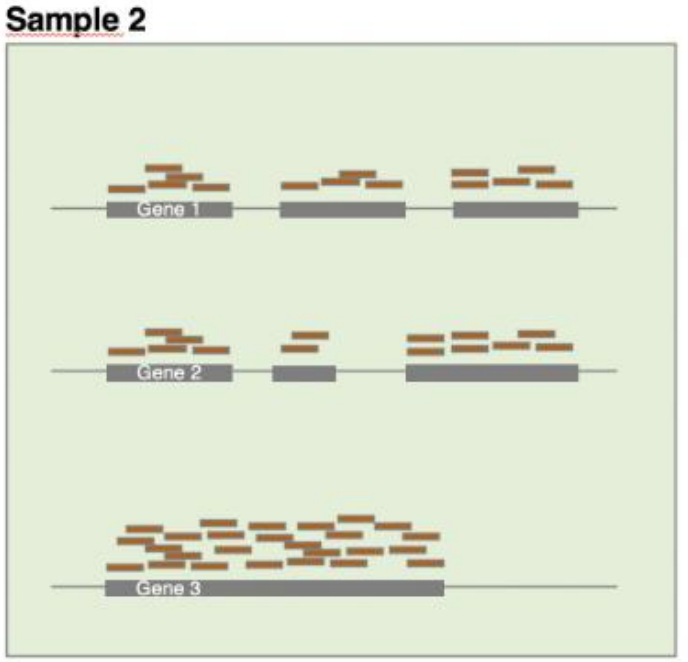
Estimate size factor

Counts

Library size

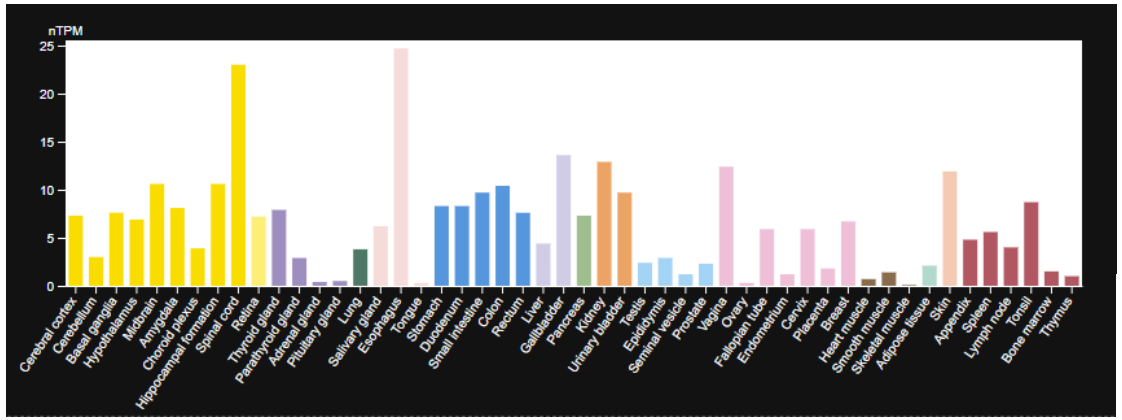
Library Composition

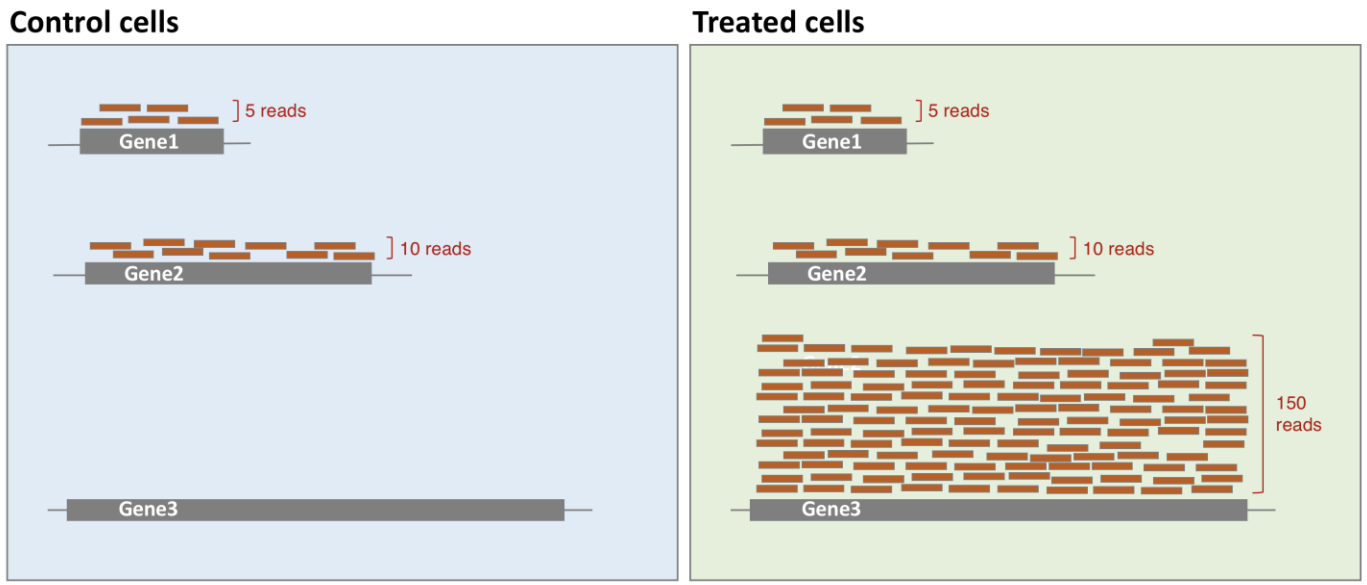




Library composition

- Many genes are organ-specific genes, meaning some genes are more expressed in certain type of tissues.
- These type of genes are needed to be taken into consideration so that we can actually identify differentially expressed genes.





Estimate size factor (median of ratios method)

Gene	Untreated	Treated
A	2	10
B	4	12
C	6	20
D	30	0

Step 1: Geometric mean(GM)

Gene	Untreated	Treated	Pseudo ref
A	2	10	Sqrt(2*10)= 4.47
B	4	12	6.93
C	6	20	10.95
D	30	0	0

Step 2: Counts/GM (ratio)

Gene	Untreated	Treated	Pseudo ref	Untreated/ ref	Treated/ ref
A	2	10	Sqrt(2*10) = 4.47	0.45	2.24
B	4	12	6.93	0.58	1.73
C	6	20	10.95	0.55	1.83
D	30	0	0	0	0

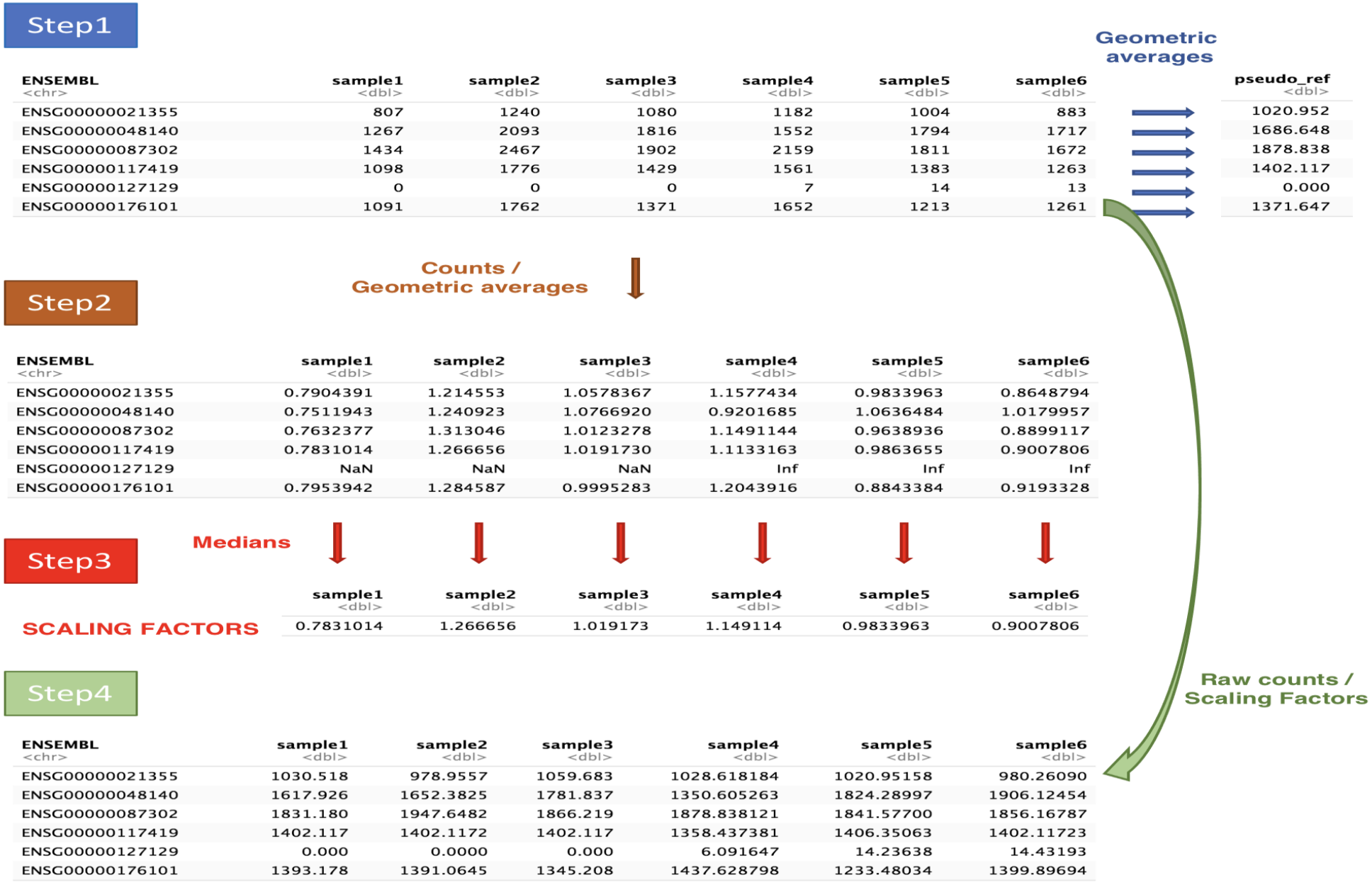
Step 3: Median of ratios (counts/GM)

Normalization factor untreated = median(0.45, 0.58, 0.55) = **0.498**

Normalization factor treated = median(2.24, 1.73, 1.83) = **1.78**

Size factor

DESeq2 normalisation method



ENSEMBL <chr>	sample1 <dbl>	sample2 <dbl>	sample3 <dbl>	sample4 <dbl>	sample5 <dbl>	sample6 <dbl>	pseudo_ref <dbl>
ENSG00000021355	807	1240	1080	1182	1004	883	1020.952
ENSG00000048140	1267	2093	1816	1552	1794	1717	1686.648
ENSG00000087302	1434	2467	1902	2159	1811	1672	1878.838
ENSG00000117419	1098	1776	1429	1561	1383	1263	1402.117
ENSG00000127129	0	0	0	7	14	13	0.000
ENSG00000176101	1091	1762	1371	1652	1213	1261	1371.647

ENSEMBL <chr>	sample1 <dbl>	sample2 <dbl>	sample3 <dbl>	sample4 <dbl>	sample5 <dbl>	sample6 <dbl>
ENSG00000021355	0.7904391	1.214553	1.0578367	1.1577434	0.9833963	0.8648794
ENSG00000048140	0.7511943	1.240923	1.0766920	0.9201685	1.0636484	1.0179957
ENSG00000087302	0.7632377	1.313046	1.0123278	1.1491144	0.9638936	0.8899117
ENSG00000117419	0.7831014	1.266656	1.0191730	1.1133163	0.9863655	0.9007806
ENSG00000127129	NaN	NaN	NaN	Inf	Inf	Inf
ENSG00000176101	0.7953942	1.284587	0.9995283	1.2043916	0.8843384	0.9193328

sample1 <dbl>	sample2 <dbl>	sample3 <dbl>	sample4 <dbl>	sample5 <dbl>	sample6 <dbl>
0.7831014	1.266656	1.019173	1.149114	0.9833963	0.9007806

ENSEMBL <chr>	sample1 <dbl>	sample2 <dbl>	sample3 <dbl>	sample4 <dbl>	sample5 <dbl>	sample6 <dbl>
ENSG00000021355	1030.518	978.9557	1059.683	1028.618184	1020.95158	980.26090
ENSG00000048140	1617.926	1652.3825	1781.837	1350.605263	1824.28997	1906.12454
ENSG00000087302	1831.180	1947.6482	1866.219	1878.838121	1841.57700	1856.16787
ENSG00000117419	1402.117	1402.1172	1402.117	1358.437381	1406.35063	1402.11723
ENSG00000127129	0.000	0.0000	0.000	6.091647	14.23638	14.43193
ENSG00000176101	1393.178	1391.0645	1345.208	1437.628798	1233.48034	1399.89694

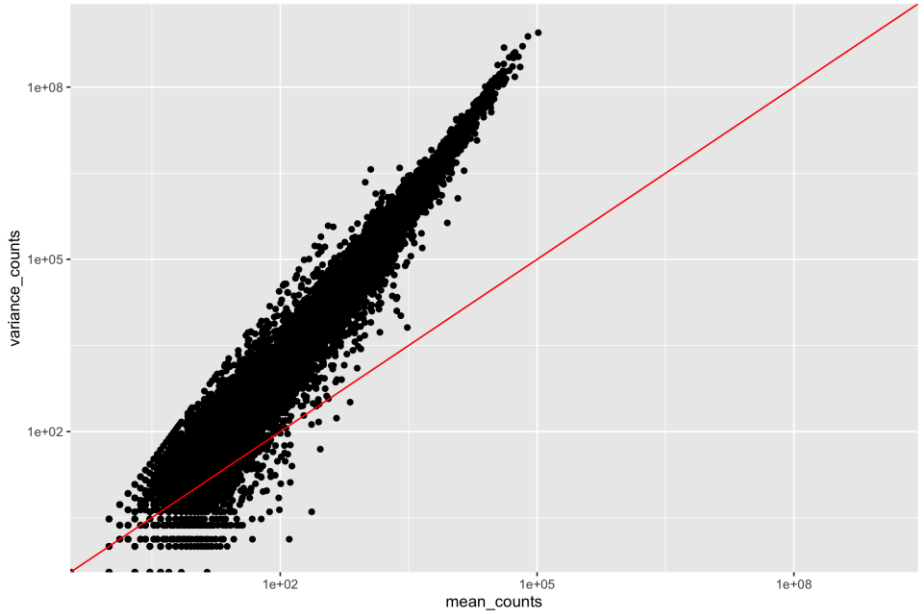
Normalized factor

Size factor	0.498	1.78
Gene	Untreated	Treated
A	2	10
B	4	12
C	6	20
D	30	0

Gene	Norm_untreated	Norm_treated
A	4.016	5.62
B	8.032	6.74
C	12.05	11.24
D	60.24	0

Estimate dispersion

- Refers to the process of estimating variability of gene expression measurements across samples in RNAseq data.
- Disperson estimation is a crucial step as it helps to model the biological and technical variability in the data accurately.
- For low mean counts, the variance estimations have a much larger spread, therefore, the estimate dispersion will differ much more between genes with small means.
- Dispersion ~ Variance
- Dispersion of 0.01 = 10% of variability



Estimate dispersion steps

- **Step 1**: Get dispersion estimates for each gene

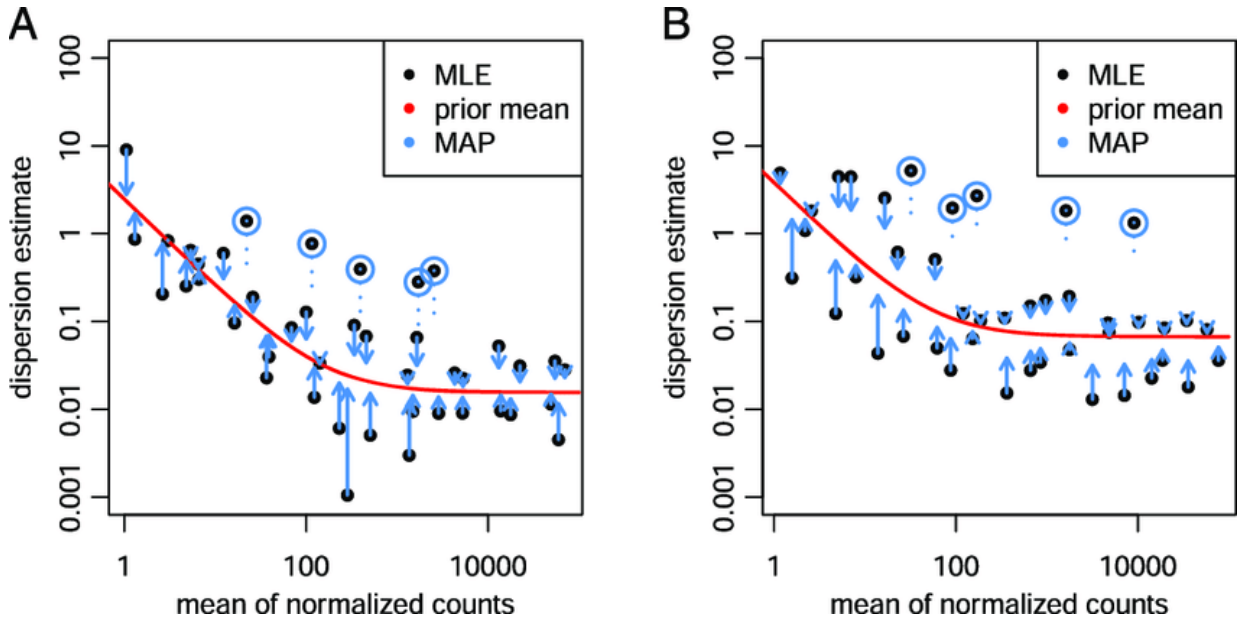
Maximum likelihood estimation (MLE): Calculate the most likely estimate dispersion given the count values of genes across replicates.

- **Step 2**: Fit a curve to gene-wise dispersion estimates

Represent the expected value of dispersion given the expression values of genes.

- **Step 3**: Adjust the dispersion parameter toward the curve(“shrinking”)

Final estimate = maximum a posteriori (MAP)



Estimate dispersion in DESeq2

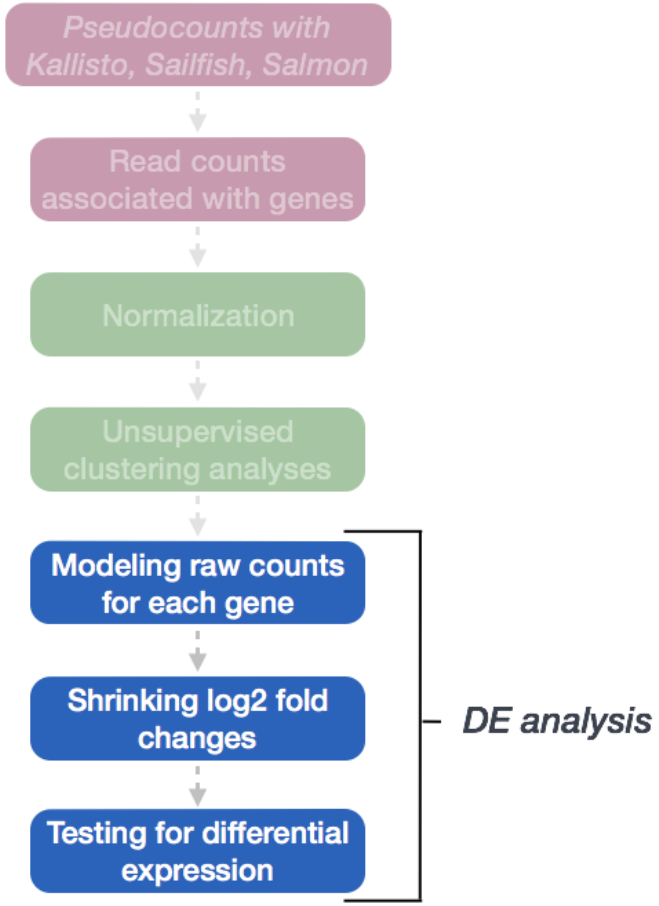
Pooled (“Gene-wise)

- The pooled dispersion estimation assumes that the dispersion is **constant** across **all genes**. It is typically used when there are **not enough replicates** available for a per-condition estimation.
- This method estimates a **single dispersion** value for **all genes**, which is based on the overall variation observed in the dataset.

Per-Condition (“Treatment-wise”)

- The per-condition dispersion estimation calculates a **separate dispersion value** for **each gene**, taking into account the biological and technical variation specific to each condition.
- This method requires **replicates** for each condition to obtain reliable estimates.

Generalized linear model (1)



$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j . q_{ij}$$

$$\log_2(q_{ij}) = \sum_r x_j . \beta_{ir}$$

Log2 Fold Change **conditions**

coefficient

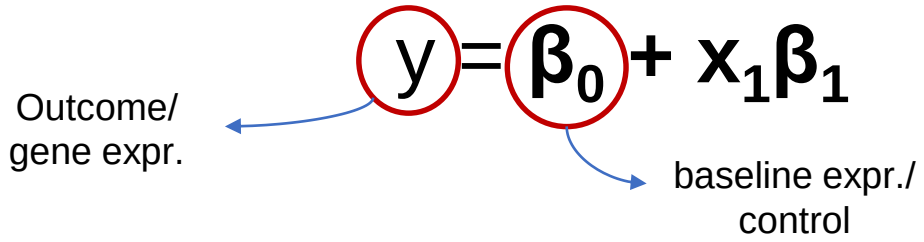
treated **untreated**

1 design factor

design factor

$$y = \beta_0 + x_1\beta_1$$

Generalized linear model (2)



Condition
U = 0
T = 1

Treated

- $y = \beta_0 + 1.\ \beta_1$
- $B_1 = y - \beta_0$
  $= \log(y) - \log(\beta_0)$
  $= \log$ (out.expr./ control)
  = treated / control

Untreated

- $y = \beta_0 + 0.\ \beta_1$
- $y = \beta_0$

Applied to 2 design factors

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_1\beta_2$$

Hypothesis testing

- Null hypothesis: no differential expression between two sample groups (LogFoldChange =0).
- Based on a thought experiment, this hypothesis can be assumed without observing any data.
- A statistical test is utilized to determine if based on the observed data, the null hypothesis is true.

Wald test

- With DESeq2, Wald test is the default used for hypothesis testing when comparing two groups.
- The Wald test is a test of hypothesis usually performed on parameters that have been estimated by maximum likelihood.
- DESeq2 implements the Wald test by:
  - ❖ Taking the **LFC** and dividing it by its standard error, resulting in a **z-statistic**
  - ❖ The z-statistic is compared to a standard normal distribution, and a **p-value** is computed reporting the probability that a z-statistic at least as extreme as the observed value would be selected at random
  - ❖ If the p-value is small we reject the null hypothesis and state that there is evidence against the null (i.e. the gene is differentially expressed).

DESeq2 result exploration

	A	B	C	D	E	F	G	H
1	Gene ID	Gene Name	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
2	YMR174C	PAI3	6660.51528	-6.939791406	0.21536036	-29.5077109	0	0
3	YBR072W	HSP26	61686.7482	-8.392251633	0.27534178	-28.3547649	0	0
4	YMR175W	SIP18	22124.4775	-9.794759216	0.34443418	-26.7388072	0	0
5	YBR117C	TKL2	5075.66676	-9.569236819	0.39288104	-22.8675752	0	0
6	YGR256W	GND2	4915.64491	-7.18222056	0.31102526	-21.2112048	0	0
7	YGL121C	GPG1	4571.80927	-5.079030659	0.21888846	-20.5311442	0	0
8	YPL223C	GRE1	3316.82570	-8.196330393	0.38714273	-19.6602690	0	0
9	YMR090W	YMR090W	6768.21961	-4.003823106	0.21470563	-15.9233037	0	0
10	YKR076W	ECM4	2072.42378	-4.498724879	0.25102169	-15.5911819	0	0
11	YDL130W-A	STF1	3910.89868	-3.103476042	0.16187961	-15.5577094	0	0
12	YIL136W	OM45	13216.3199	-4.855610968	0.28795838	-14.8306531	0	0
13	YDL223C	HBT1	1132.82533	-5.368560875	0.33076685	-14.4620320	0	0
14	YKL107W	YKL107W	277.999346	-5.658392633	0.35600776	-14.2507921	0	0
15	YFL014W	HSP12	134528.310	-7.104074298	0.46490892	-14.0222608	0	0
16	YKR049C	FMP46	2804.11353	-3.310053737	0.19718923	-13.8194854	0	0
17	YOR120W	GCY1	7817.10304	-3.71465497	0.23482670	-13.3275087	0	0
18	YBL075C	SSA3	2846.50584	-4.813297683	0.31972027	-13.2249908	0	0
19	YML042W	CAT2	712.249988	-3.779948076	0.24253887	-13.1729321	0	0
20	YBL049W	MOH1	672.121539	-4.524306334	0.30311778	-12.9959589	0	0
21	YDR032C	PST2	14441.9083	-2.723462882	0.16492755	-12.9660736	0	0
22	YGL156W	AMS1	2223.19047	-3.902025146	0.25665764	-12.9239290	0	0
23	YGR043C	NQM1	14459.1562	-6.91156118	0.49697797	-12.7300634	0	0
24	YMR118C	SHH3	619.406776	-5.163101809	0.36111034	-12.6778472	0	0
25	YBL002W	HTB2	3347.47347	3.15729999	0.20578603	12.4998763	0	0
26	YMR107W	SPG4	2910.40421	-6.879894688	0.50753365	-12.4029107	0	0
27								

- **`baseMean`**: mean of normalized counts for all samples
- **`log2FoldChange`**: log2 fold change
- **`lfcSE`**: standard error
- **`stat`**: Wald statistic
- **`pvalue`**: Wald test p-value
- **`padj`**: BH adjusted p-values

When will p-values be NA ?

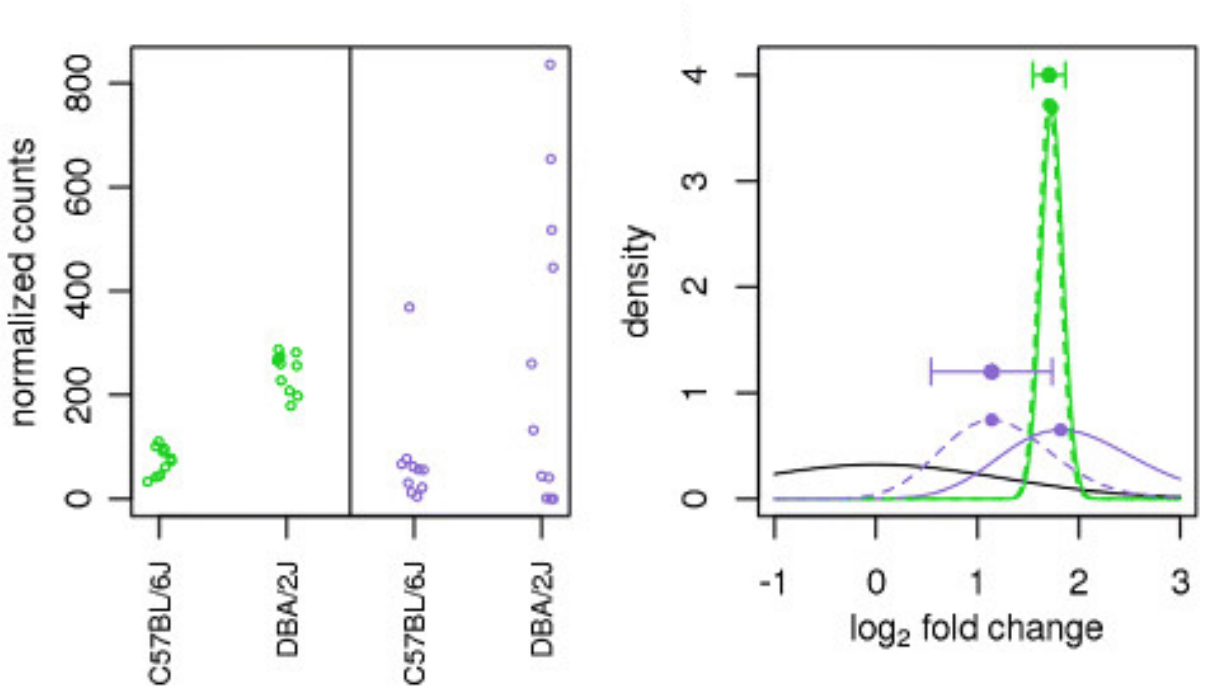
Note: p-values set to **NA**

1. If within a row, all samples have **zero counts**, the baseMean column will be zero, and the log2 fold change estimates, p-value and adjusted p-value will all be set to NA.
2. If a row contains a sample with an **extreme count outlier** then the p-value and adjusted p-value will be set to NA. These outlier counts are detected by Cook's distance.
3. If a row is filtered by automatic **independent filtering**, for having a **low mean normalized count**, then only the adjusted p-value will be set to NA.

***What is independent filtering?** This is a low mean threshold that is empirically determined from your data, in which the fraction of significant genes can be increased by reducing the number of genes that are considered in the multiple testing.

Shrunken log2 foldchanges (LFC)

- To generate more precise log2 foldchange estimates, DESeq2 allows for the shrinkage of the LFC estimates toward zero when the information for a gene is low, include:
  - Low counts
  - High dispersion values



NOTE: Shrinking the log2 fold changes will not change the total number of genes that are identified as significantly differentially expressed

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8

Gene Ontology in RNAseq

Enrichment analysis

Introduction

- Differential expression analysis is *univariate* – each gene is tested on its own.
- This does not reflect the underlying biology because genes work in *conjunction*, not in isolation.
- The univariate approach expects significant changes in single gene. Moderate effects in many related genes cannot, by definition be identified as statistically significant.
- The goal of an enrichment analysis is to test for a group of related genes, called **gene sets**, and test whether the genes within these sets are enriched for differentially expression.

Gene sets

1. Gene ontology (GO)
2. Kyoto Encyclopedia of Genes and Genomes (KEGG)
3. Reactome
4. Molecular Signatures Database (MSigDB)

Gene ontology (GO)

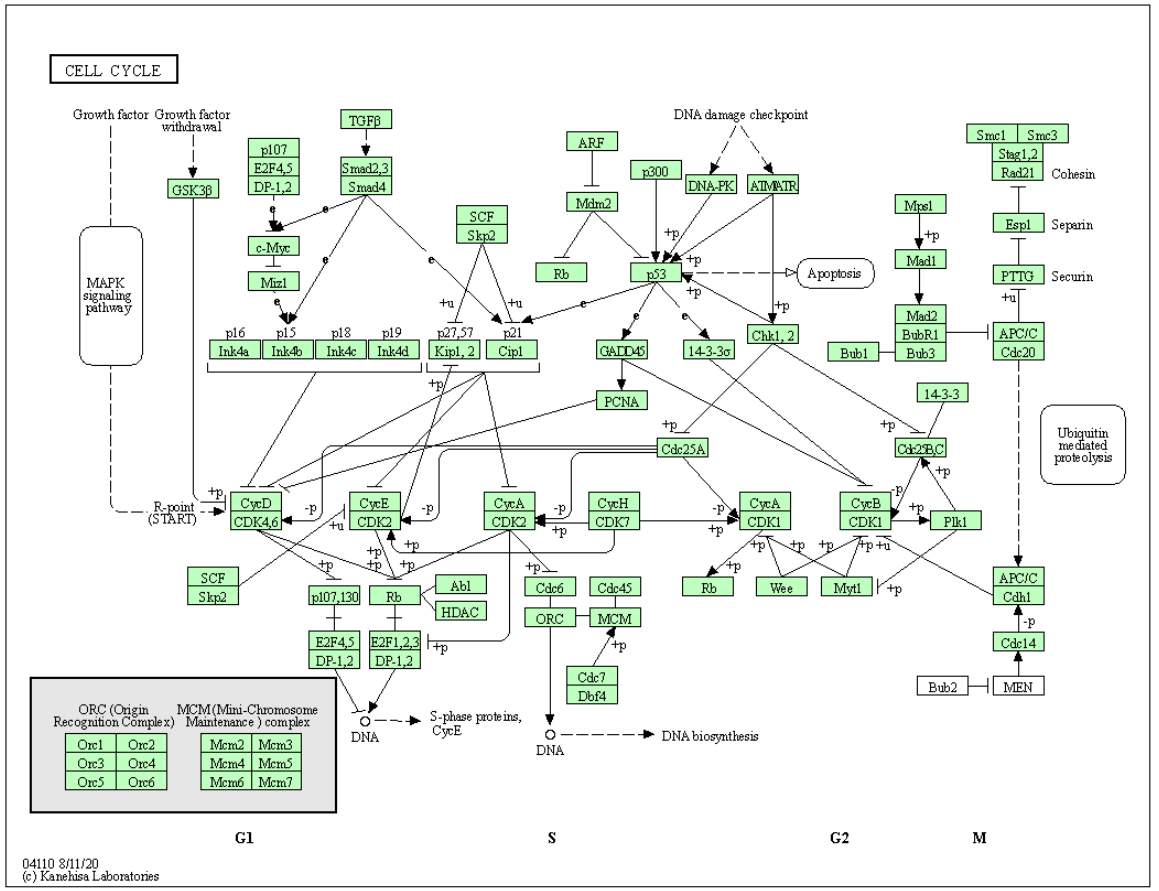
- The Gene Ontology (Ashburner et al.2000) defines GO terms.
- These terms are based on a controlled vocabulary and relations that define the directed links between terms -> define a hierarchy between GO terms.
- These terms have been classified into 3 categories, called namespaces:
  - **Molecular Function** (MF): molecular activities of gene products
  - **Cellular Component** (CC): where gene products are active
  - **Biological Process** (BP): pathways and larger processes made up of the activities of multiple gene products

http://geneontology.org/

Kyoto Encyclopedia of Genes and Genomes (KEGG)

- KEGG pathway is a collection of manually drawn and curated pathway maps representing current knowledge of the molecular interaction, reaction and relation networks.

1. Metabolism
2. Genetic information processing
3. Environmental information processing
4. Cellular processes
5. Organismal systems
6. Human diseases
7. Drug development.

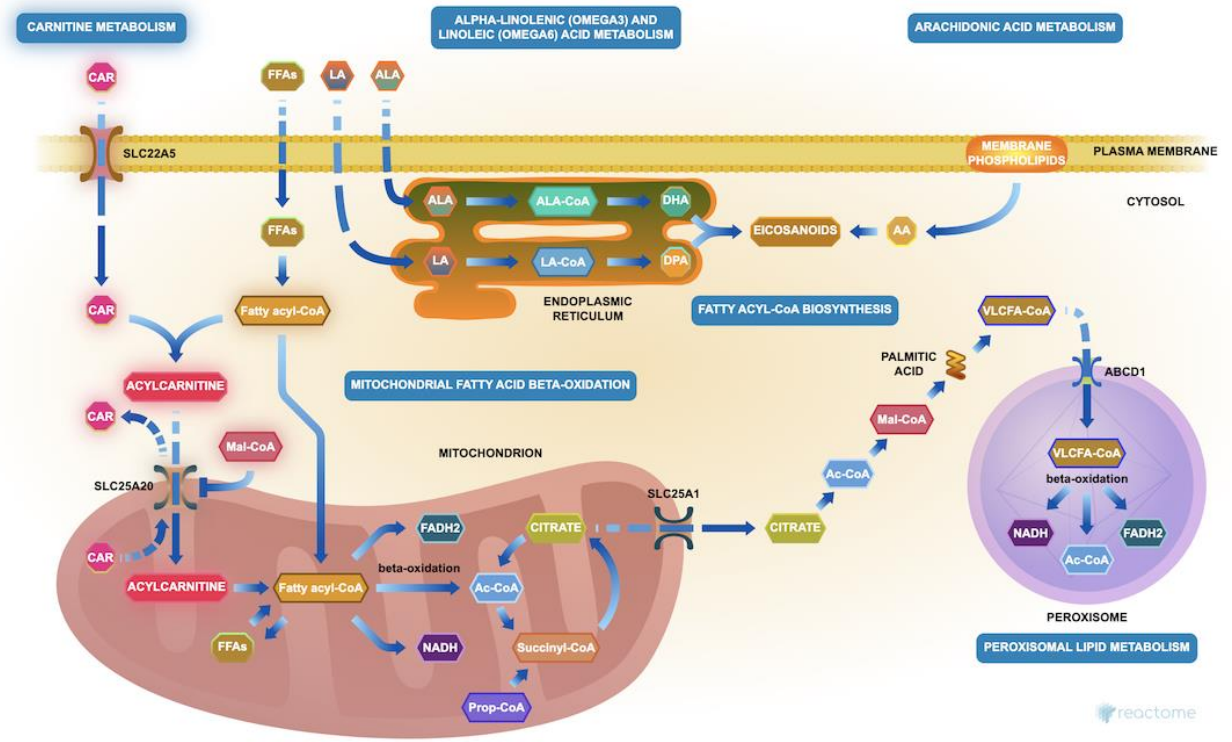


KEGG pathway for cell cycle

https://www.genome.jp/kegg-bin/show_pathway?hsa04110

Reactome

- Alike KEGG patway, Reactome is a free, open-source, curated and peer-reviewed pathway database.



https://reactome.org/

Molecular Signatures Database (MSigDB)

- MSigDB is a collection of annotated gene sets for use with GSEA software. The MSigDB gene sets are divided into 9 collections:
  - Hallmark gene sets (H) are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
  - Positional gene sets (C1) for each human chromosome and cytogenetic band.
  - Curated gene sets (C2) from online pathway databases, publications in PubMed, and knowledge of domain experts.
  - Regulatory target gene sets (C3) based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.
  - Computational gene sets (C4) defined by mining large collections of cancer-oriented microarray data.
  - Ontology gene sets (C5) consist of genes annotated by the same ontology term.
  - Oncogenic signature gene (C6) sets defined directly from microarray gene expression data from cancer gene perturbations.
  - Immunologic signature gene sets (C7) defined directly from microarray gene expression data from immunologic studies.
  - Cell type signature gene sets (C8) curated from cluster markers identified in single-cell sequencing studies of human tissue.

Over representation analysis (ORA)

- Over Representation Analysis (ORA) (Boyle et al. 2004) is a widely used approach to determine whether known biological functions or processes are over-represented (= enriched) in an experimentally-derived gene list, *e.g.* a list of differentially expressed genes (DEGs).
- The *p*-value can be calculated by **hypergeometric distribution**.

Over representation analysis (ORA)

- The example used to describe the distribution is an urn contain \(N\) marbles of two colours. There are \(K\) green and and \(N-K\) red marbles in the urn. Drawing a green marble is defined as success, and a red marble failure. Using the formula above, we can compute the probability to draw \(k\) green marbles from the urn.
- In the frame of an enrichment analysis (Rivals et al.2007) we use the following formulation to calculate a probabiliy that we have more green marbles than we would expect by change, i.e. there to be an over representation of green marbles.

Performing ORA

- To perform an over representation analysis, we thus need to define:
  - among all the genes (called the universe), which ones are differentially expressed (DE);
  - among all the genes, which ones are part of the gene set of interest.

	GO	**not_GO**
DE	n	p
not_DE	m	q

- Fisher's exact (or hypergeometric test) that will test whether we can identify a statistically enrichment of DE genes in the GO category.

Gene set enrichment analysis (GSEA)

- Gene set enrichment analysis refers to a broad family of tests.
- The major advantage of GSEA approaches is that they don't rely on defining DE genes.
- The first step is to **order** the genes of interest based on the statistics used.
- Specify a list of interest ordered genes for analysis.
- Genes are ranked based on their phenotypes. Given apriori defined **set of gene *S*** (e.g., genes sharing the same *DO* category), the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout the **ranked gene list (*L*)** or primarily found at the top or bottom.

Performing GSEA

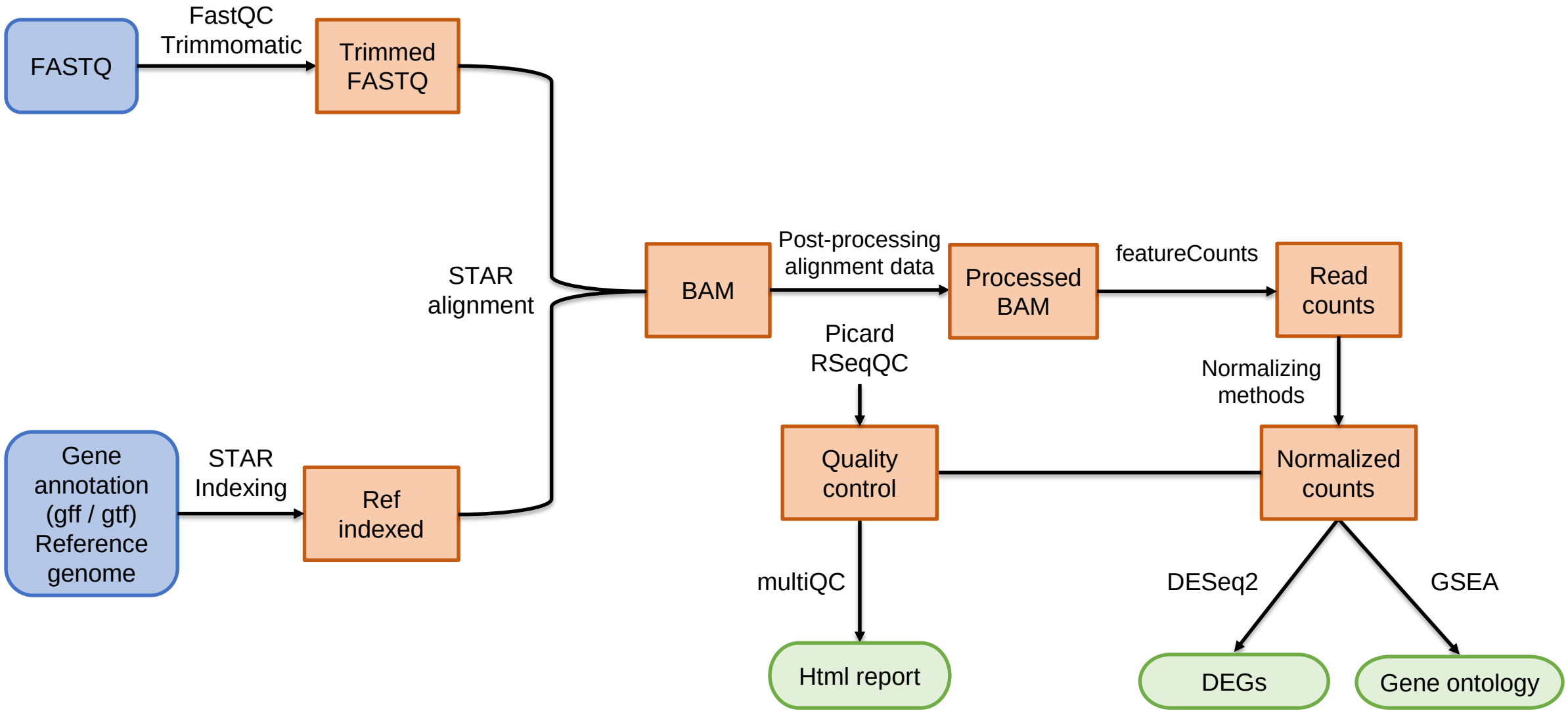
- There are three key elements of the GSEA method:

1. Calculation of an Enrichment Score.
   - The enrichment score (*ES*) represents the degree to which a set *S* is over-represented at the top or bottom of the ranked list *L*.
2. Esimation of Significance Level of *ES*
   - The *p*-value of the *ES* is calculated using a permutation test
3. Adjustment for Multiple Hypothesis Testing
   - When the entire gene sets are evaluated, the estimated significance level is adjusted to account for multiple hypothesis testing and also *q*-values are calculated for FDR control.

Comparison: ORA vs GSEA

	ORA	GSEA
Conceptual Approach	starts with a predefined set of genes or gene sets	does not rely on predefined gene sets
Input Data	a list of differentially expressed genes (DEGs) or genes of interest	a **ranked** list of genes based on a specific metric, such as differential expression scores or correlation values
Statistical Analysis	hypergeometric test or Fisher's exact test	Kolmogorov-Smirnov-like running sum statistic
Interpretation of Results	provides **p-values** or other statistical measures indicating the **significance of the enrichment** for each tested gene set	calculates an **enrichment score** for each gene set, which represents the degree to which the gene set is **overrepresented** at the top or bottom of the ranked gene list

Summary RNA sequencing analysis workflow



Reference

- https://en.bio-protocol.org/en/bpdetail?id=4161&type=0
- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE227381
- https://bioconductor.org/packages/release/bioc/html/DESeq2.html
- https://rseqc.sourceforge.net/
- https://www.researchgate.net/figure/Gene-Set-Enrichment-Analysis-GSEA-for-Gene-Ontology-GO-and-Kyoto-Encyclopedia-of_fig5_343311093
- https://www.gsea-msigdb.org/gsea/index.jsp


THANK
YOU