Differential Gene Expression (based on RNA-seq)

Laboratory Course Bioinformatics 2025

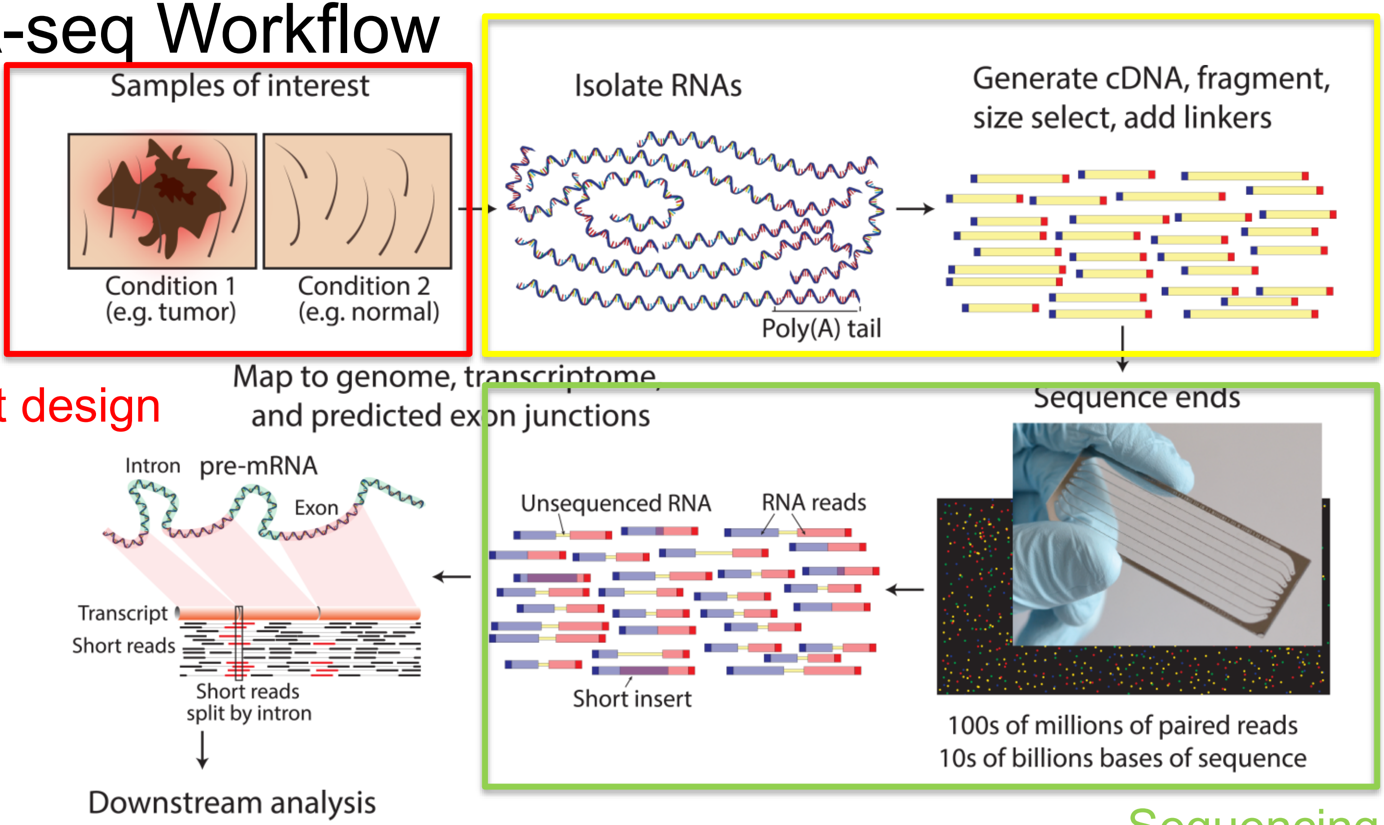
Dr. techn. Veronika Schusterbauer

Sunday, November 23, 2025

Background

Sample preparation

RNA-seq Workflow



Experiment design

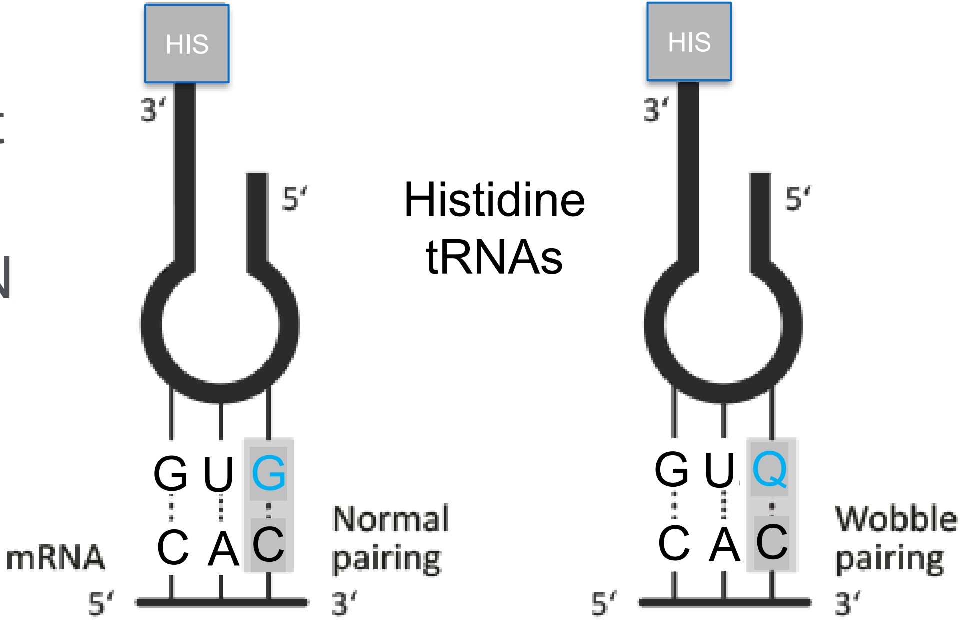
Sequencing

https://doi.org/10.1371/journal.pcbi.1004393

How does queuosine affect metal homeostasis in *E. coli*?

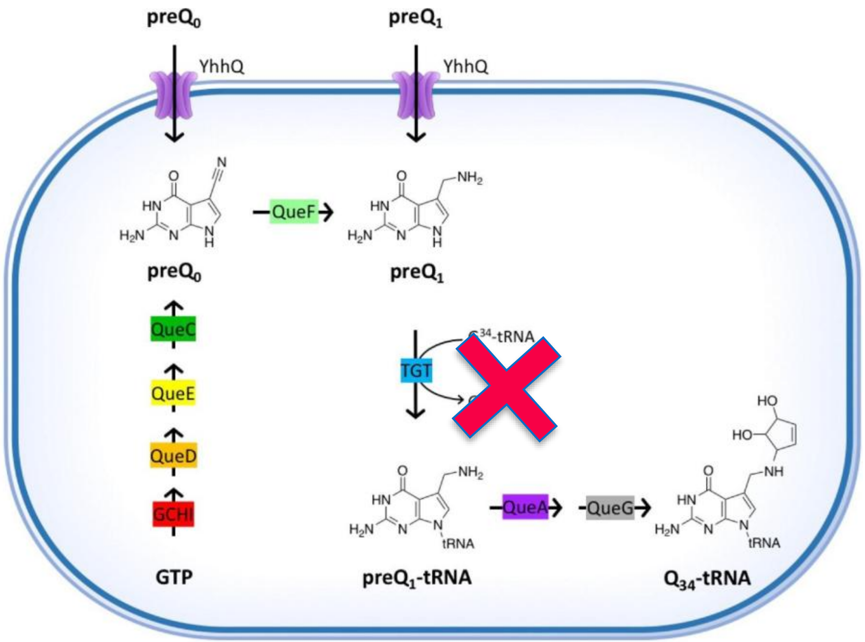
Queuosine:

- Rare modified nucleoside that is present in certain tRNAs
- At the wobble position of GUN anticodon
- Only synthesized by bacteria
  - Might be a vitamin



https://academic.oup.com/metallomics/article/14/9/mfac065/6692928

Experiment Design



WT strain: *Escherichia coli* K-12 MG1655

TGT protein -> replace G with Q in tRNA

Knockout TGT gene -> **Q_{34} tRNA deficient** strains

Add Nickel ($2\ \mathrm{mM}\ NiCl_2{\cdot}6H_2O$) to medium

Factor	Strain	Media
Conditions	WT	LB
	TGT	LB+Ni

-> 4 groups in total

Library preparation

Only Sequence the RNA of interest!
-> Remember ~90% of RNA is ribosomal RNA

polyA selection:

Amplify mature mRNAs only (by oligodT affinity)
Only for eukaryotic organisms

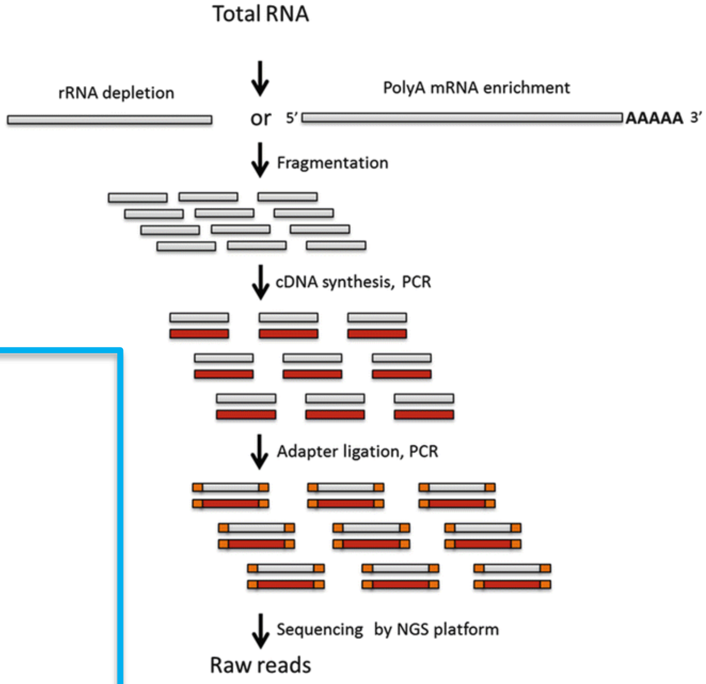
rRNA depletion:

More suitable for low quality / fragmented mRNA
Allows for analysis of non-coding RNAs

- tRNAs
- Short and long non-coding RNAs

!! Kits are species specific !!
Usually more expensive

RNeasy Mini Kit +
Ribo-Zero Magnetic Kit



https://link.springer.com/protocol/10.1007/978-1-4939-2697-8_9/figures/4

Library Prep/ Sequencing Considerations

Read Depth:

40 million reads per sample

- More depth needed for lowly expressed genes (regulatory genes)
- Detecting low fold differences need more depth (because of less variance)

Read Length:

75 bp paired end, Illumina

- The longer the length the more likely to map uniquely
- Paired read help in mapping and junctions
- PacBio and ONT allow sequencing of full transcripts and splicing isoforms

Strand specific Protocols

? -> not strand specific

- Give clearer results -> especially important for densely packed genomes

Replicates

3 “biological” replicates

- Detecting subtle differences in expression needs more replicates
- Detecting novel genes or alternate iso-forms need more replicates

Increasing depth, length, and/or replicates and strand specific protocols increase costs

Data retrieval

RNA-seq reads (FASTQ, Bioproject PRJNA751097):

NCBI:

- SRA (Sequence Read Archive)
- Https://www.ncbi.nlm.nih.gov/Traces/study/

EMBL-EBI:

- ENA (European Nucleotide Archive)
- https://www.ebi.ac.uk/ena/browser/home
- Pro: Allows direct download of FASTQ files

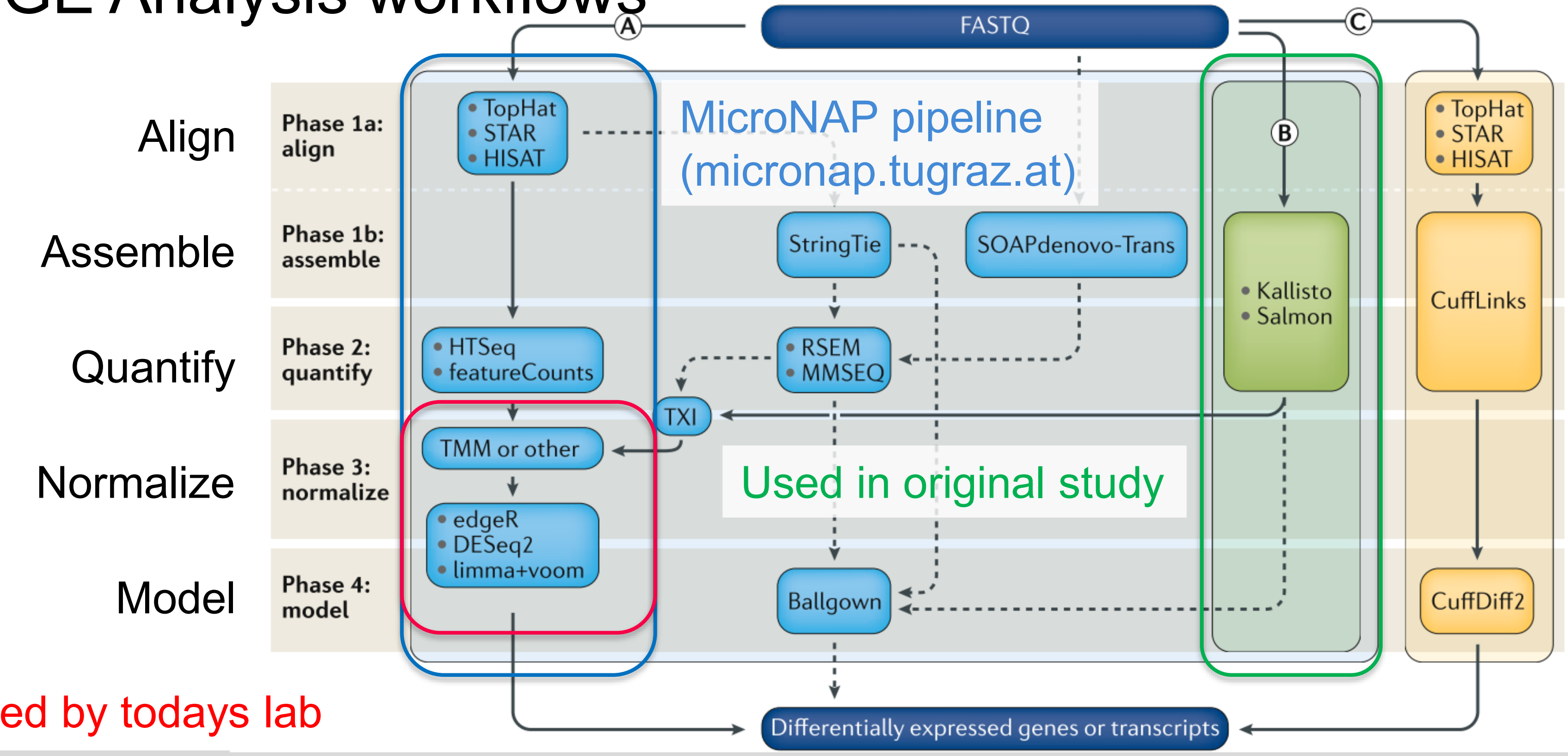
Reference (FASTA + GFF/GTF):

NCBI Datasets for *E. coli* K12 MG1655:

- (Now combines Taxonomy/Genome/Assembly archives)
- https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000005845.2/

Bioinformatics Analysis

DGE Analysis workflows



MicroNAP pipeline (micronap.tugraz.at)

Used in original study

Covered by todays lab

Mapping (provided by MicroNAP)

STAR

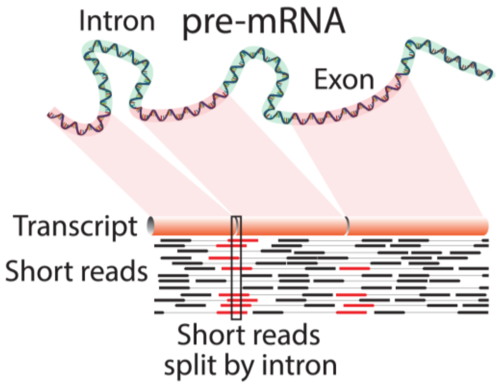
- Splice aware aligner (Considers introns)
- Can indicate new fusions in genes

Input:

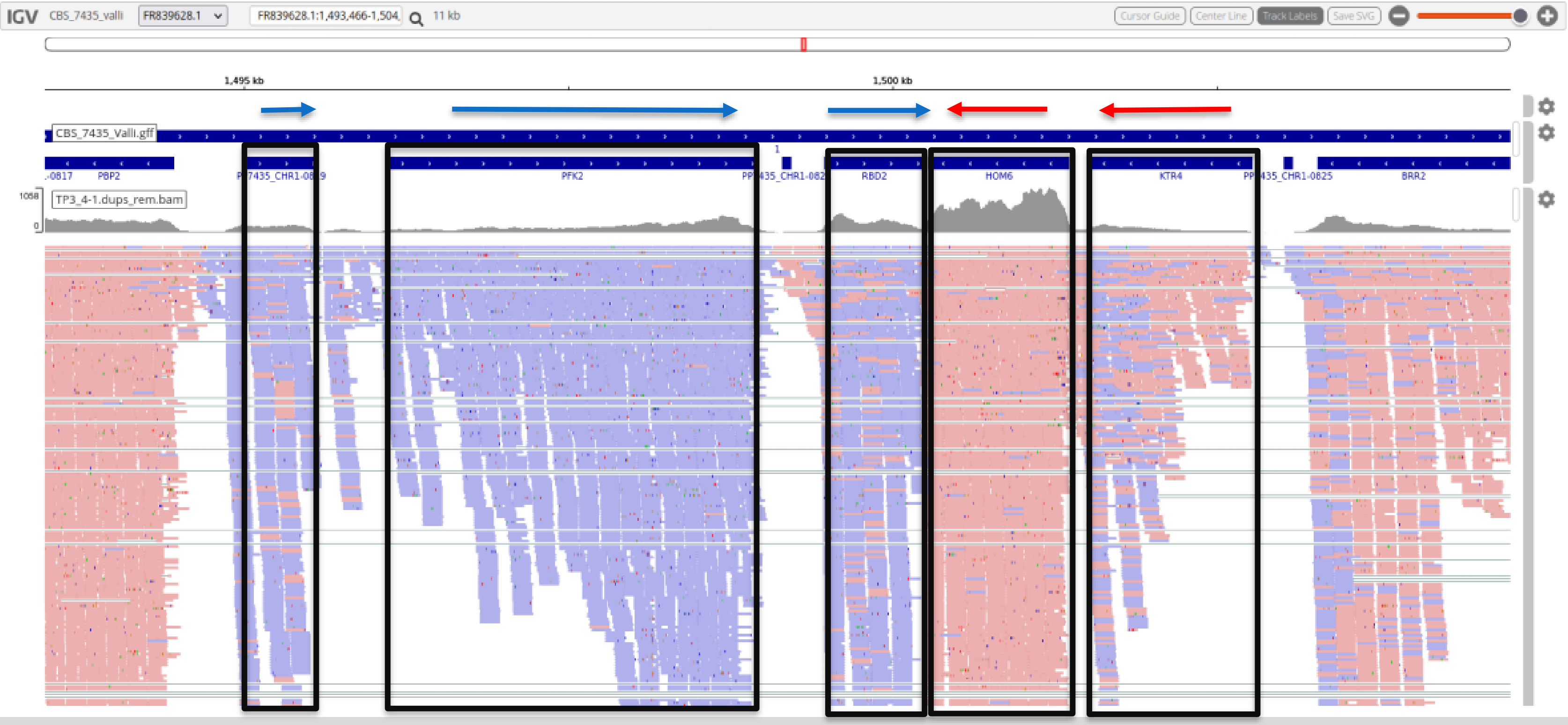
- Reference & Annotation
- FASTQ files of raw reads

Output:

- Reads mapped to reference (BAM file)



Read Mapping – RNA Sequencing



Counting Reads (provided by MicroNAP)

Htseq-count:

- Counts reads mapped to genes or other features

Input:

- Mapped Reads (BAM)
- Reference Genome & Annotation

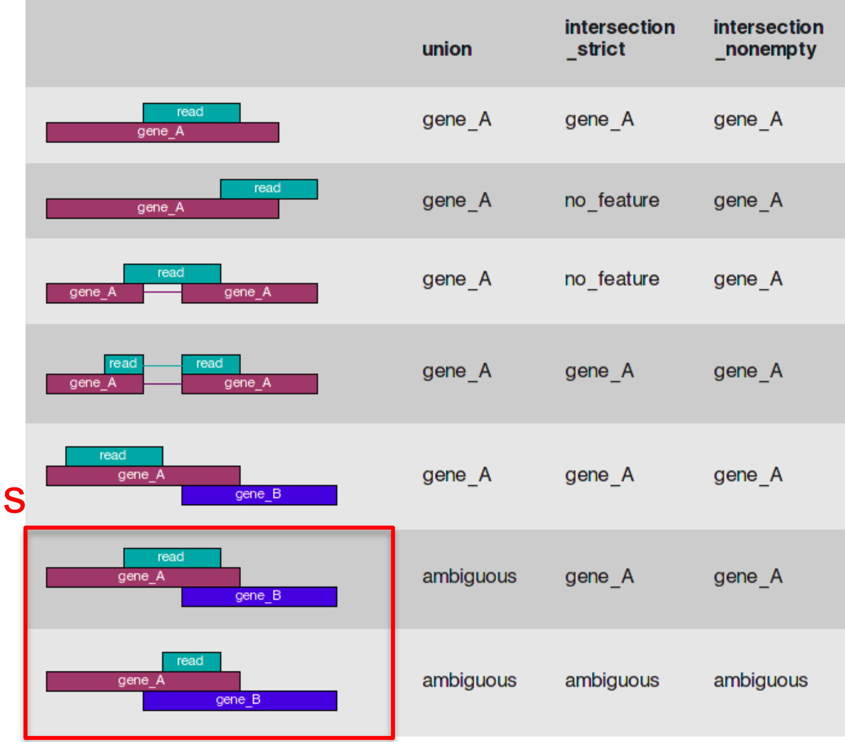
Output:

- Table of read counts (TSV)

Standard setting:

- Ambiguous reads not counted
- Multimapped reads not counted

ambiguous



	union	intersection _strict	intersection _nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

Count normalization

To account for different amounts of Illumina reads per sample (sequencing depth)

Trimmed Mean of M-values (TMM)

- Used for statistical testing between conditions (edgeR & Deseq2)
- Assumption: Most genes are expressed similarly

Makes read counts per gene comparable across samples

(But not different genes within one sample)

- For comparing genes use: RPKM/FPKM or TPM

Model Fitting and Statistical Testing

Design Matrix:

- Table that lists different factors for each sample

sample	strain	treatment	group
sample1	WT	none	WT.none
sample2	WT	none	WT.none
sample3	WT	Nickel	WT.Nickel
sample4	WT	Nickel	WT.Nickel
sample5	TGT	none	TGT.none
sample6	TGT	none	TGT.none

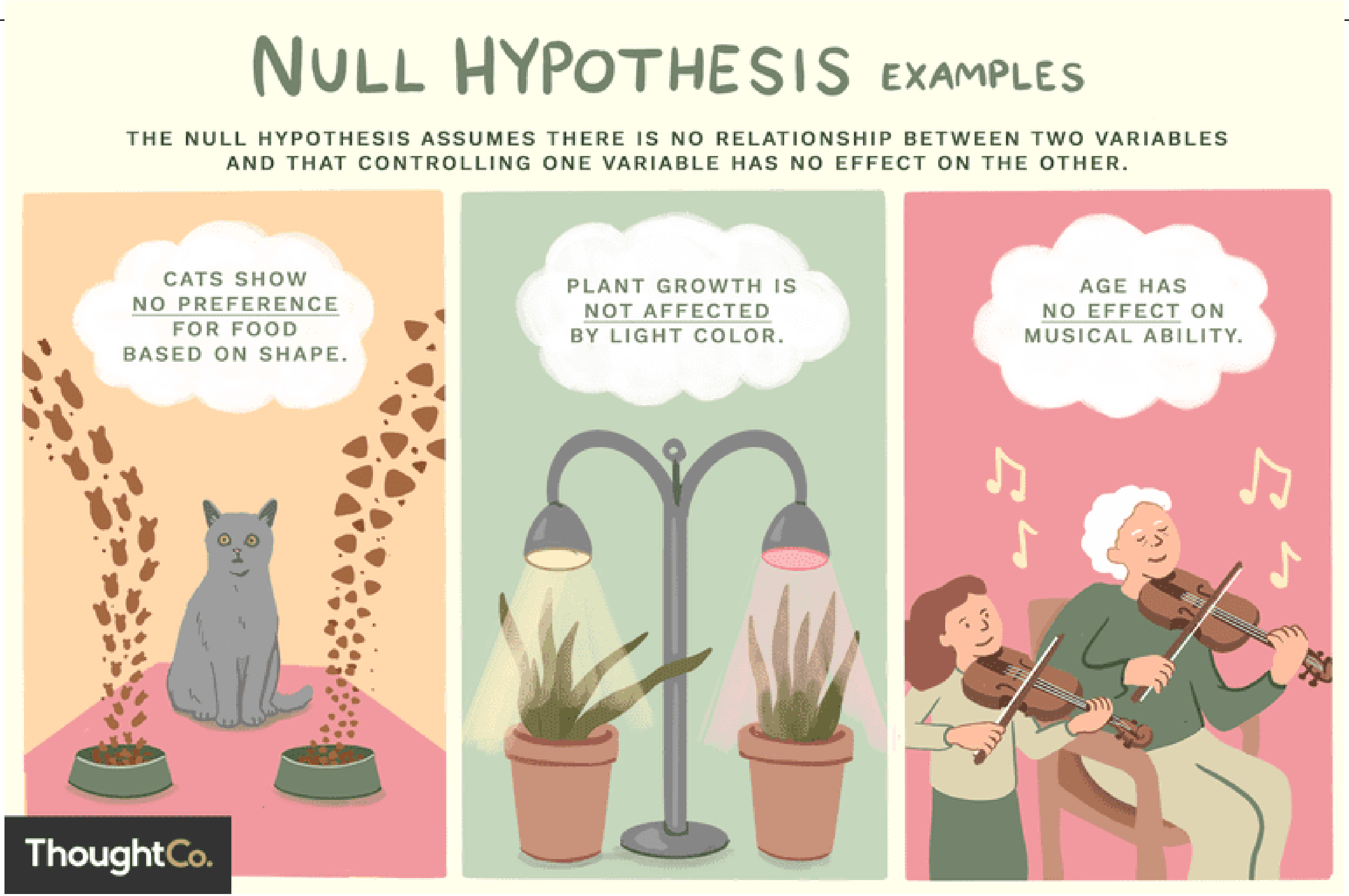
Model Fitting and Statistical Testing

Contrast:

- Formula that defines which groups / factors we want to compare
  - Simplest form: group B – **group A**

  - Effect of Nickel on wildtype E. coli:
    - WT.Nickel – **WT.none**
  - Effect of mutation on wildtype E. coli:
    - TGT.none – **WT.none**

The **reference (group A)** is always **deducted from** the deviant state **(group B)**
→ positive LFC means gene is higher expressed in group B than in the group A



NULL HYPOTHESIS EXAMPLES

THE NULL HYPOTHESIS ASSUMES THERE IS NO RELATIONSHIP BETWEEN TWO VARIABLES AND THAT CONTROLLING ONE VARIABLE HAS NO EFFECT ON THE OTHER.

CATS SHOW NO PREFERENCE FOR FOOD BASED ON SHAPE.
PLANT GROWTH IS NOT AFFECTED BY LIGHT COLOR.
AGE HAS NO EFFECT ON MUSICAL ABILITY.
ThoughtCo.

https://www.thoughtco.com/thmb/Re2LtGKysci0lpqcuYSCvdaHjp8=/768x0/filters:no_upscale():max_bytes(150000):strip_icc()/null-hypothesis-examples-609097_FINAL-100262e70b70426fb0633304eb2f49f4.png

Model Fitting and Statistical Testing

Hypothesis is tested for each gene separately

H0:

- No difference in gene expression between groups.

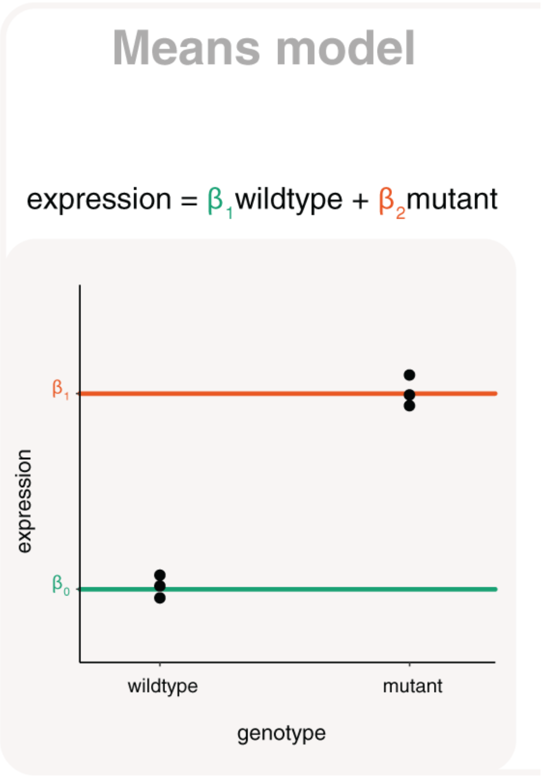
H1:

- There is a difference in gene expression between groups.

p-value:

- Definition: *the probability of obtaining the observed results, assuming that the null hypothesis is true.*
- ~ The probability that what we are **seeing/sampling**, happened by chance
  (rather than an actual **difference in population mean**)

If p-value < cutoff, we reject H0 and accept H1



Means model

expression = $\beta_1$wildtype + $\beta_2$mutant

More Info:
https://biocorecrg.github.io/CRG_Bioinformatics_for_Biologists/differential_gene_expression.html

Multiple testing correction

As more attributes are compared, differences due solely to chance become more likely!

Well known problem for genomic data

- 10,000s genes/transcripts
- 100,000s exons

→ We need to adjust our initial p-value, taking into account how many statistical tests we performed

→ This results in an **adjusted p-value**
(sometimes **also called q-value or false discovery rate (FDR))**

(Log) Fold Change

Expression level WT = E_{wt}

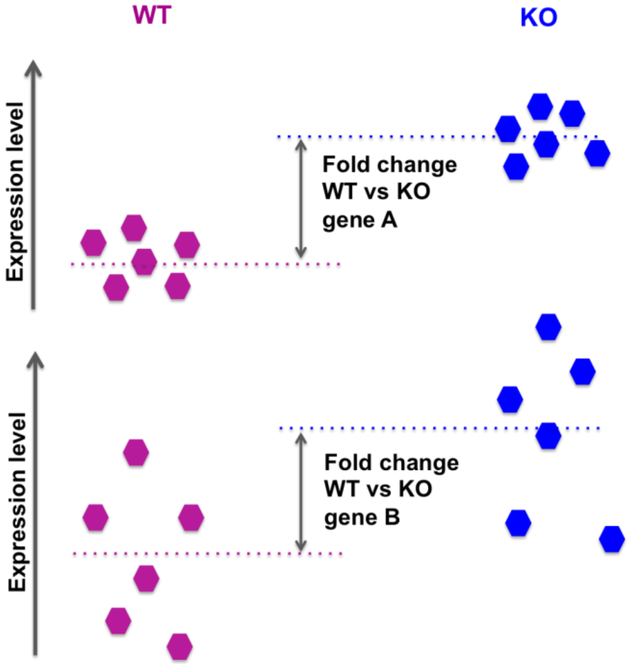
Expression level KO = E_{ko}

Fold Change = FC

Log Fold Change = LFC

$$\mathbf{FC = E_{ko} / E_{wt} = 2^{LFC}}$$

$$\mathbf{LFC} = \log_2(FC) = \log_2(E_{ko} / E_{wt}) = \mathbf{\log_2(E_{ko}) - \log_2(E_{wt})}$$



LFC is the same,
although the variance is not

https://biocorecrg.github.io/CRG_Bioinformatics_for_Biologists/differential_gene_expression.html

Differentially expressed genes (DEGs)

Definition: *A gene is declared differentially expressed if a* ***difference*** *or change observed* ***in read counts or expression*** *levels between two experimental conditions* ***is statistically significant*** **[1].**

Can be further restricted by cutoffs, but it is not strictly necessary.

Most common cutoffs:

- Adjusted p-value < 0.05
- LFC > 1 (or FC > 2)
- LFC < -1 (or FC < ½)

[1] Anjum et. al, J Comput Biol. 2016, doi: 10.1089/cmb.2015.0205

Quality Control

Quality Control (provided by MicroNAP)

FastQC

- Quality metrics for **raw FastQ** files
- adapter sequences, low-quality reads, uncalled bases (Ns)

Qualimap

- Quality metrics for mapped reads
- Detect unmapped reads (-> contamination ?)

MultiQC

- Summarizes QC results from multiple tools and multiple samples

Principal Component Analysis (PCA)

- Principal component analysis (PCA) simplifies the complexity in high-dimensional data while retaining trends and patterns.
- It does this by transforming the data into fewer dimensions, which act as summaries of features.
- Each Principal Component (PC) is a linear combination of all variables.
- PC1 maximizes the variance.

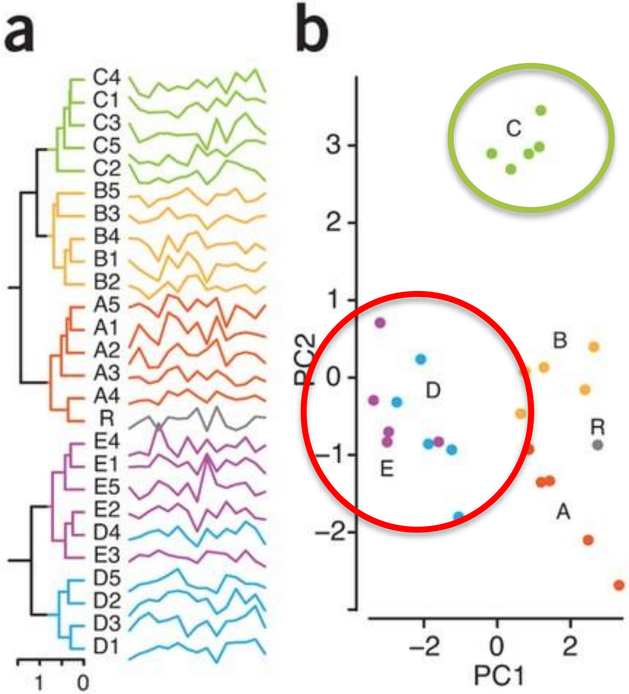
- Quality Control for DGE: Optimally, the replicates of a sample cluster together

Principal Component Analysis (PCA)

a) Raw signal per sample

b) PCA of all samples:

- Group C clusters nicely
- D & E not clearly separable



https://www.nature.com/articles/nmeth.4346/figures/3

Lab 09: Differential gene expression

Create a new folder for lab09 & lab19

Go to the Teachcenter (MOL923UF Laboratory Course Bioinformatics) and download:

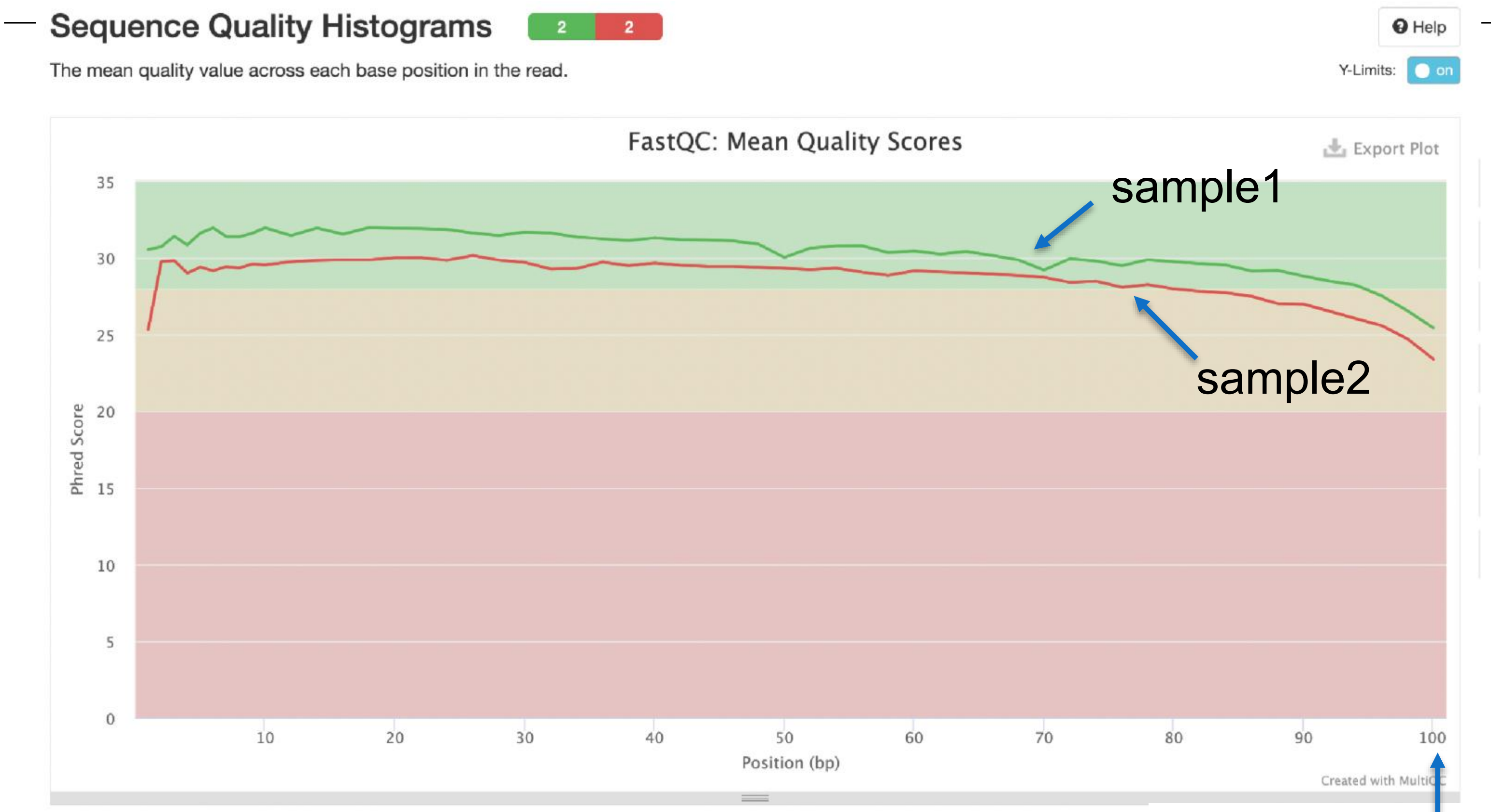
- Annotated Raw Counts (Counts_raw.tsv)
- MultiQC Report (multiqc_report.html)
- Lab 09 Instructions (2025_lab09_Vsc.html)

- Answer the questions in the instructions in a separate report

Further Reading

https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/

https://www.nature.com/collections/qghhqm/pointsofsignificance



Sequence Quality Histograms

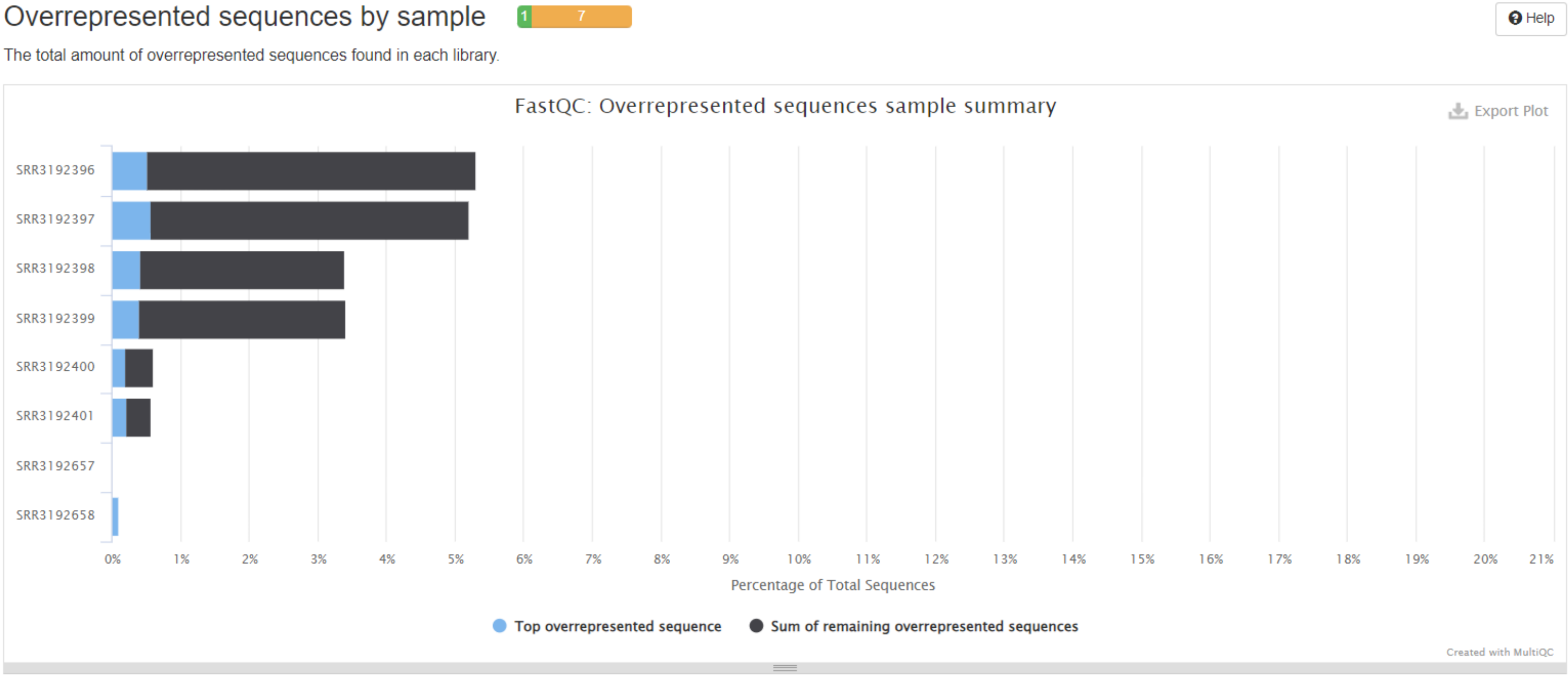
The mean quality value across each base position in the read.

FastQC: Mean Quality Scores

sample1

sample2

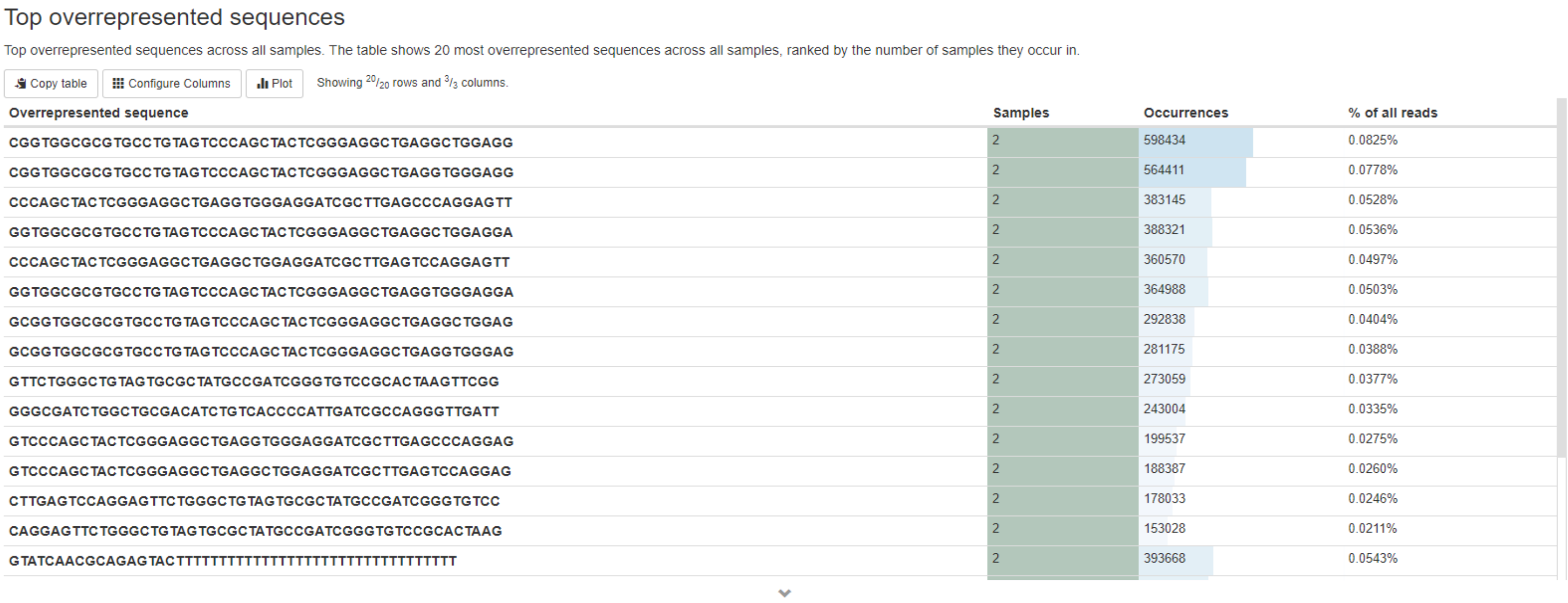
Read length



Overrepresented sequences by sample

The total amount of overrepresented sequences found in each library.

- Adapter contamination
- Presence of rRNA



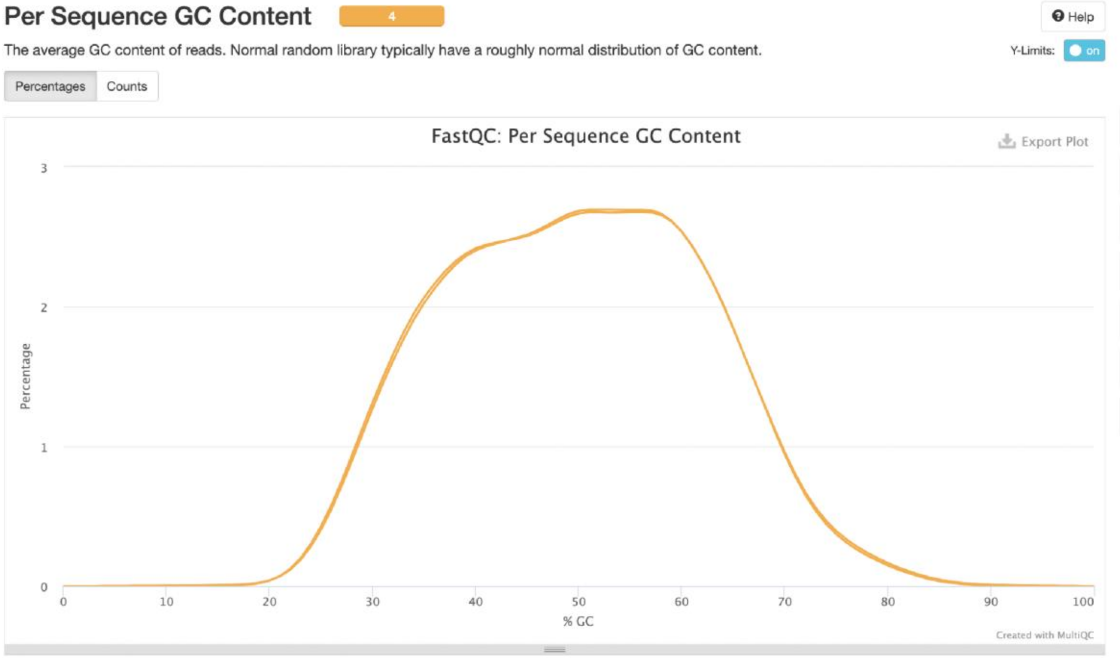
Top overrepresented sequences

Top overrepresented sequences across all samples. The table shows 20 most overrepresented sequences across all samples, ranked by the number of samples they occur in.

Showing $^{20}/_{20}$ rows and $^{3}/_{3}$ columns.

Overrepresented sequence	Samples	Occurrences	% of all reads
CGGTGGCGCGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCTGGAGG	2	598434	0.0825%
CGGTGGCGCGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGTGGGAGG	2	564411	0.0778%
CCCAGCTACTCGGGAGGCTGAGGTGGGAGGATCGCTTGAGCCCAGGAGTT	2	383145	0.0528%
GGTGGCGCGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCTGGAGGA	2	388321	0.0536%
CCCAGCTACTCGGGAGGCTGAGGCTGGAGGATCGCTTGAGTCCAGGAGTT	2	360570	0.0497%
GGTGGCGCGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGTGGGAGGA	2	364988	0.0503%
GCGGTGGCGCGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCTGGAG	2	292838	0.0404%
GCGGTGGCGCGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGTGGGAG	2	281175	0.0388%
GTTCTGGGCTGTAGTGCGCTATGCCGATCGGGTGTCCGCACTAAGTTCGG	2	273059	0.0377%
GGGCGATCTGGCTGCGACATCTGTCACCCCATTGATCGCCAGGGTTGATT	2	243004	0.0335%
GTCCCAGCTACTCGGGAGGCTGAGGTGGGAGGATCGCTTGAGCCCAGGAG	2	199537	0.0275%
GTCCCAGCTACTCGGGAGGCTGAGGCTGGAGGATCGCTTGAGTCCAGGAG	2	188387	0.0260%
CTTGAGTCCAGGAGTTCTGGGCTGTAGTGCGCTATGCCGATCGGGTGTCC	2	178033	0.0246%
CAGGAGTTCTGGGCTGTAGTGCGCTATGCCGATCGGGTGTCCGCACTAAG	2	153028	0.0211%
GTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	2	393668	0.0543%

Can be used to determine which genomic regions cause problems



Per Sequence GC Content

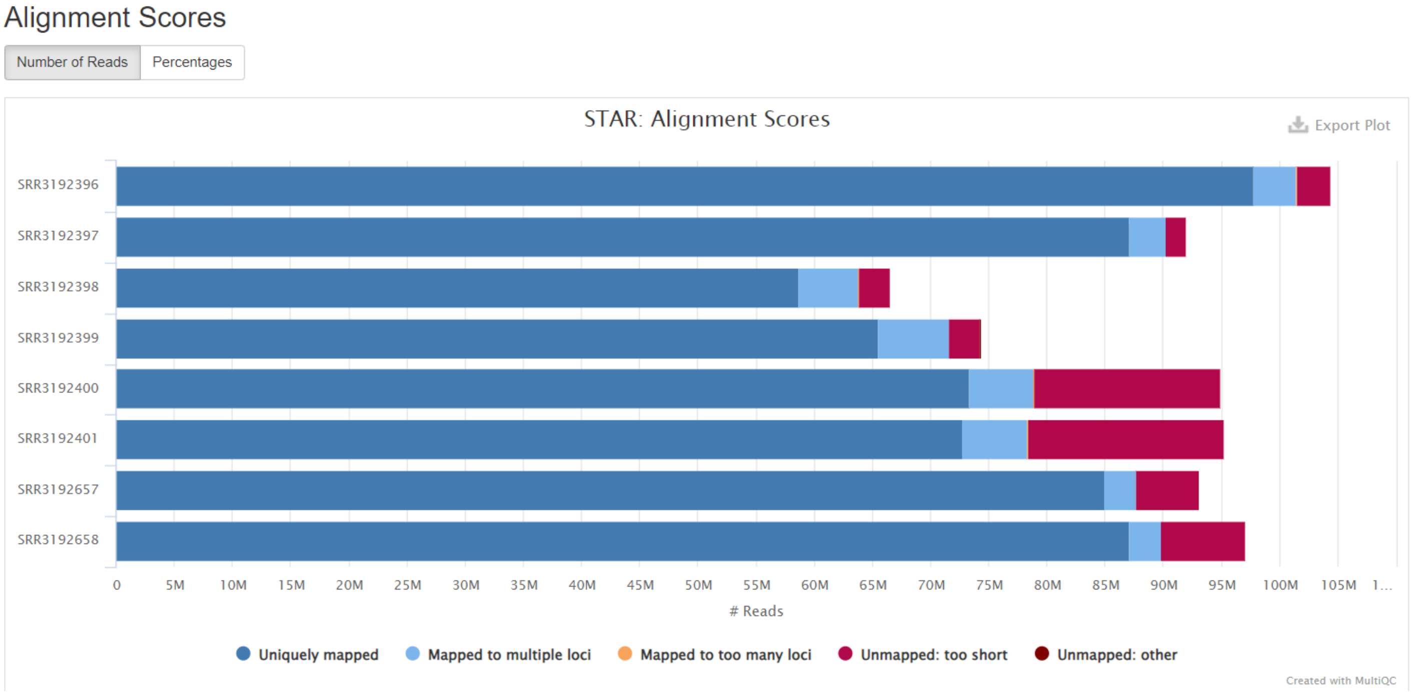
The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

Y-Limits: on

Harvard Chan Bioinformatics Core

A shift/difference to expectations could be caused by a bias in library prep

A second maximum could indicate contamination from another species



Alignment Scores

Multimapped reads stem from repetitive regions in the genome:
e.g.: rRNA, transposons, other homologous genes

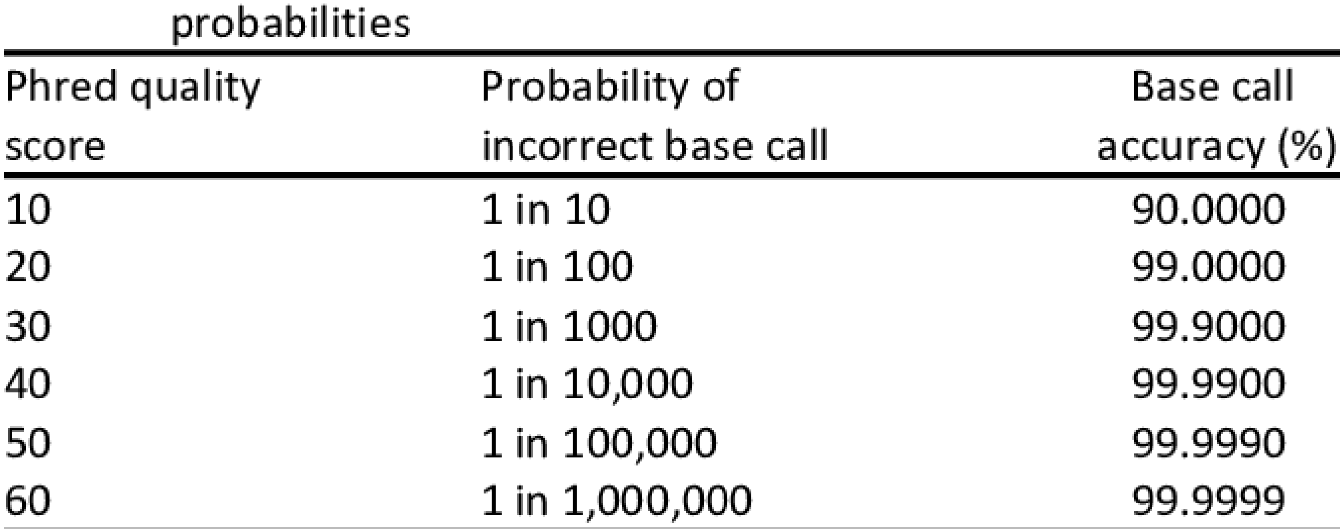
https://multiqc.info/examples/rna-seq/multiqc_report.html

FastQ format

- Standard format for NGS reads
  - Illumina, PacBio, Oxford Nanopore Technologies (ONT)

- Similar to multi-fasta
- Additional quality value for each base
  - Phred Score (Q), relates to error probability of a base (P):
    - $Q = -10 * \log_{10}(P)$
    - Q ranges from 0-60

Phred Score



probabilities

Phred quality score	Probability of incorrect base call	Base call accuracy (%)
10	1 in 10	90.0000
20	1 in 100	99.0000
30	1 in 1000	99.9000
40	1 in 10,000	99.9900
50	1 in 100,000	99.9990
60	1 in 1,000,000	99.9999