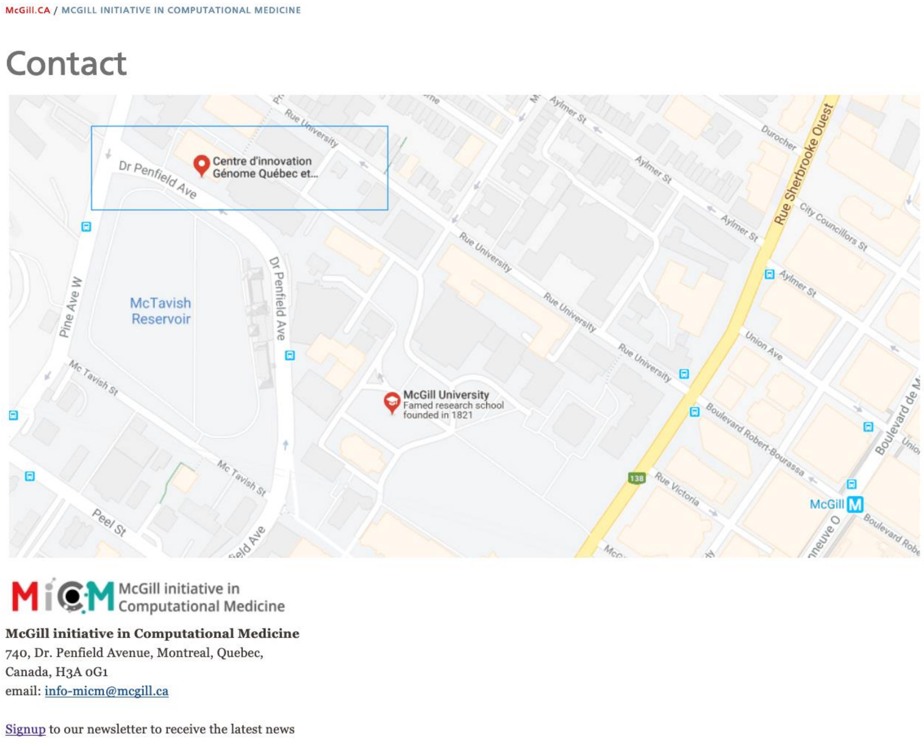
RNA-seq Quantification

December 1st, 2022

Audrey Baguette
Rached Alkallas
Georgette Femerling

Mission : aims to deliver inter-disciplinary research programs and empower the use of data in health research and health care delivery



McGILL.CA / MCGILL INITIATIVE IN COMPUTATIONAL MEDICINE

Contact

McGill initiative in Computational Medicine

McGill initiative in Computational Medicine
740, Dr. Penfield Avenue, Montreal, Quebec,
Canada, H3A 0G1
email: info-micm@mcgill.ca

Signup to our newsletter to receive the latest news

https://www.mcgill.ca/micm

Overview

- Introduction (5 min)
  - What does bulk RNA-seq measure?
  - What are the limitations of bulk RNA-seq?
- Overview of the preprocessing steps (35 min)
  - From FASTQ files to raw read counts: what does each step mean?
  - Fastqc
  - Galaxy
  - Quality Control report: what should be flagged?
  - Hands on: run a QC analysis and interpret the results (15 min)

Overview

- Normalization (15 min)
  - Why do we normalize?
  - Common normalization techniques
  - Normalization in differential gene expression analysis
  - Hands on: identify the appropriate model depending on the analysis (5 min)

LUNCH BREAK

Introduction

What does bulk RNA-seq measure?

- (Generally) un-targeted sequencing
- Transcripts produced
- Uses a population to increase sensitivity
- Captures mature mRNA (polyA) or mature and immature transcripts (total RNA)
- Can identify different splicing events

Protocol

- RNA is isolated. DNA is depleted with DNAse
- RNA is selected or depleted
  - Selecting for polyA
  - Depleting rRNA
  - Isolation of specific transcripts (probes, size)
- RNA is converted in completmentary DNA. Reverse transcription into cDNA improves the stability and prepares for PCR.
- cDNA is then fragmented, size-selected and sequenced

What are the limitations of bulk RNA-seq?

- Average of a population
- Does not capture all transcripts
- Genes with a high output overshadow low-output genes
- # mRNA ≠ # proteins

Overview of the preprocessing steps

From FASTQ files to raw read counts: what does each step mean?

1. QC and trimming
2. Alignment
3. Quality of the alignment
4. Counting reads in each gene
5. Optionally, create a wig file

GenPipes (https://genpipes.readthedocs.io/en/latest/user_guide/pipelines/gp_rnaseq.html)

From FASTQ files to raw read counts: what does each step mean?

1. QC and trimming

(fastqc, timmomatic)

We want to measure the quality of sequencing

Reads must be trimmed to remove adapters and low-quality bases

Low-quality reads are filtered out

For paired end, paired reads and unique reads are separated

From FASTQ files to raw read counts: what does each step mean?

2. Alignment

(STAR, BWA)

Finding where transcripts fall on the genome

The genome of reference may change the results

The reads can be aligned to the transcriptome instead of the genome

From FASTQ files to raw read counts: what does each step mean?

3. Quality of the alignment

(Picard)

Quantity of uniquely mapped reads

Duplicated reads (coming from single fragment, created during PCR)

From FASTQ files to raw read counts: what does each step mean?

4. Counting reads in each gene

(HT Seq Count)

Sort the aligned reads

Count the number of reads falling in each gene (or known isoform)

Output the results in a matrix

rows are genes/isoforms

column(s) are raw read counts

From FASTQ files to raw read counts: what does each step mean?

5. Optionally, create a wig file

(Wiggle)

Visualize the repartition of reads along the genome

Estimation of transcription levels

Fastqc

- Command line
- One report per file
- Stats can be merged with MultiQC

Galaxy

What if I can't access a server?

- Web-based (https://usegalaxy.org/)
- Interactive
- Also includes samtools, bedtools, tools for long reads, lift-over tool, alignment, DEG etc.
- Limited storage (~6Gb?)
- Needs to create an account

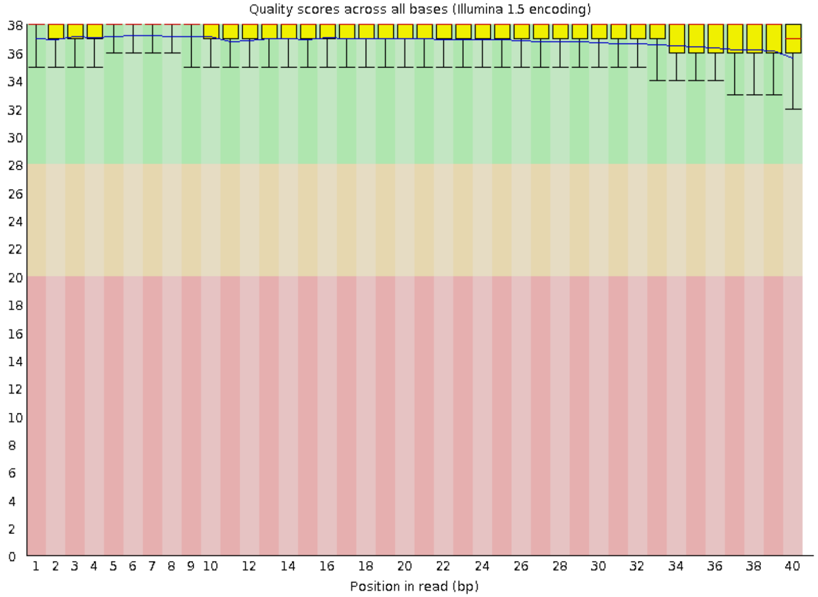
Quality Control report: what should be flagged?

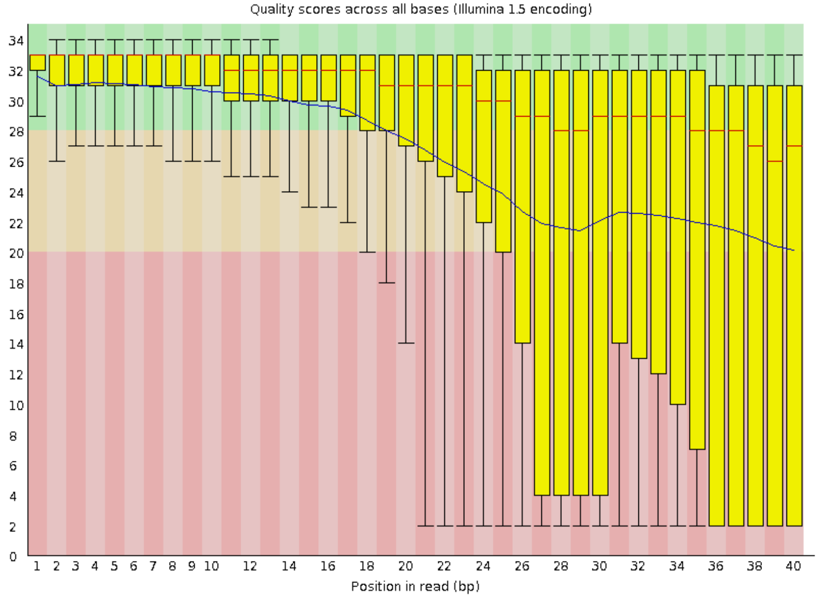
- Read length
- Base quality
- Repeated k-mers
- Base content
- Adapter content

Per base sequence quality

Warning: lower quartile < 10 OR median < 25

Fail: lower quartile < 5 OR median < 20





https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Per base sequence quality

REMINDER:

Why is the quality decreasing over read length?
What does a Q20 mean?

Because as bases get added, there is more change to slip and have a shift.

Q20 -> 10^(20/-10) -> 0.01 -> 1% errors -> about 1 wrong base per read (length is 100 bp)

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Per tile sequence quality

Heatmap

Deviation from the mean quality score for each flowcell tile, for a given base

Blue: mean tile >= mean all tiles

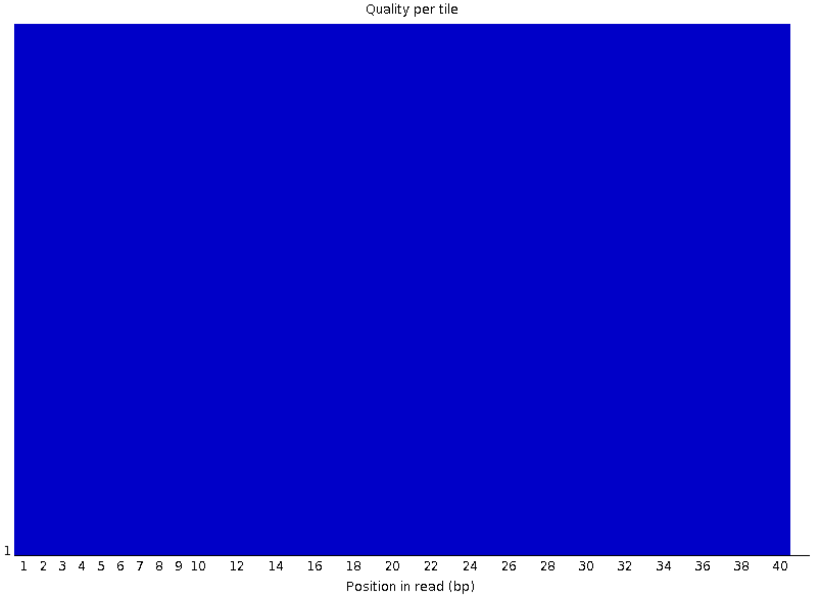
Red: mean tile < mean all tiles

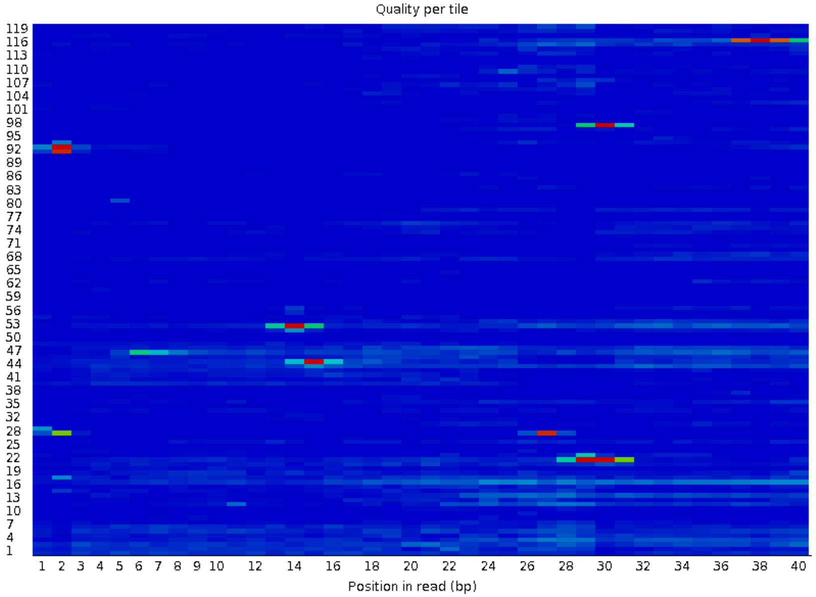
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html

Per tile sequence quality

Warning: mean tile < mean all tiles -2

Fail: mean tile < mean all tiles -5 ●



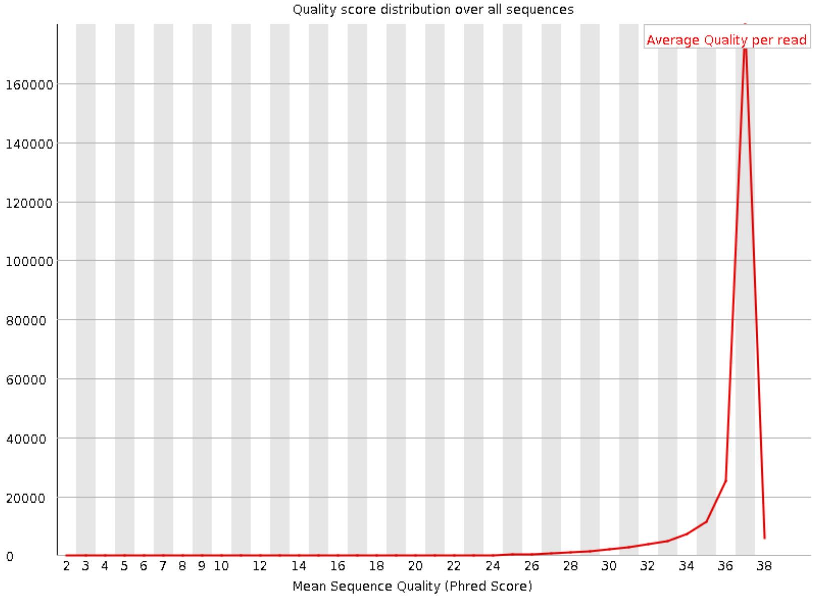


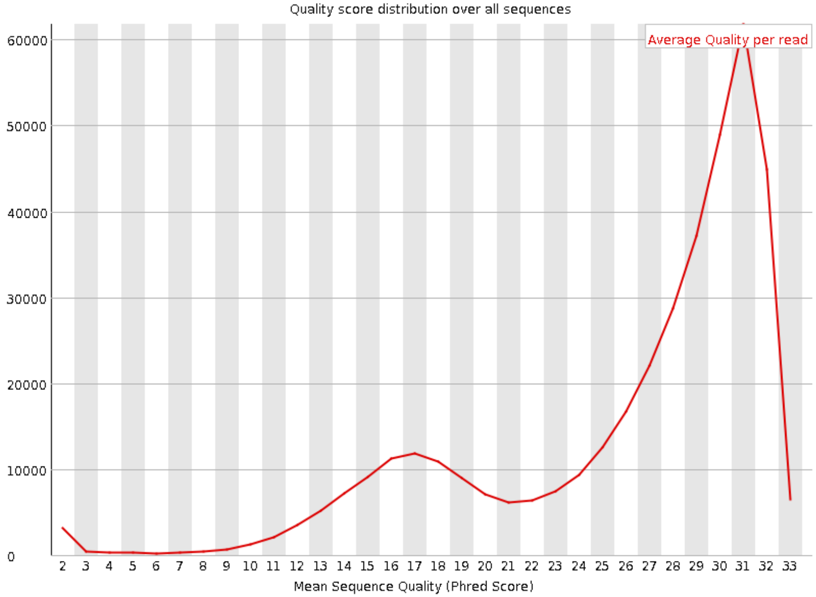
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Per sequence quality scores

Warning: most frequent score < 27 (0.2%)

Fail: most frequent score < 20 (1%)





https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/3%20Per%20Sequence%20Quality%20Scores.html

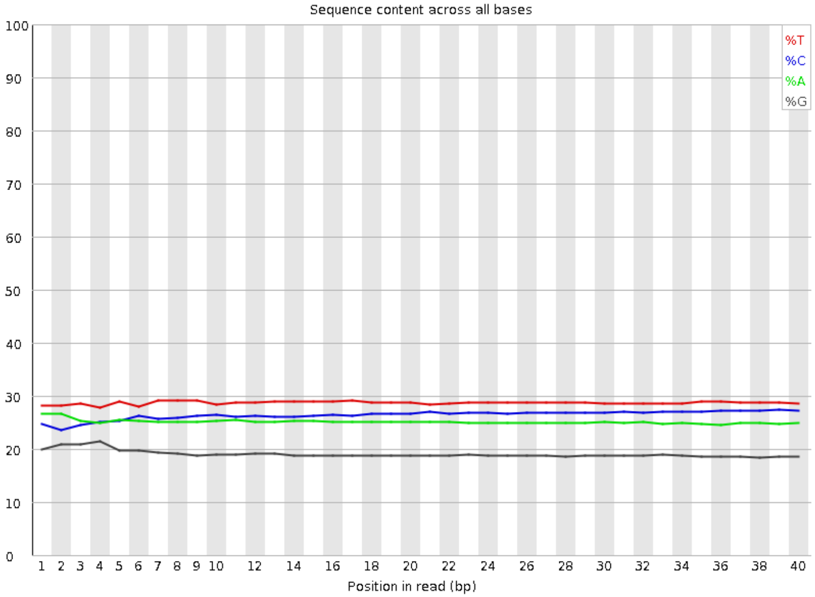
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

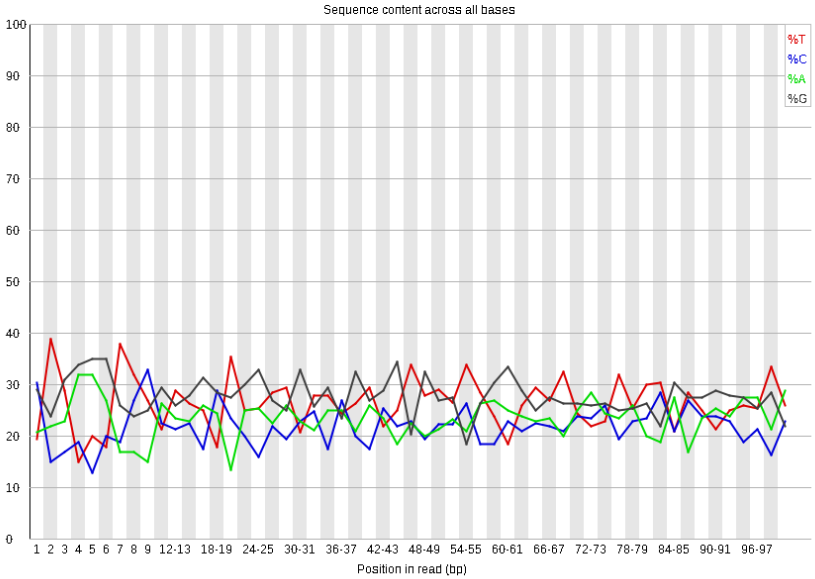
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Per base sequence content

Warning: difference in abundance > 10%

Fail: difference in abundance > 20% ●





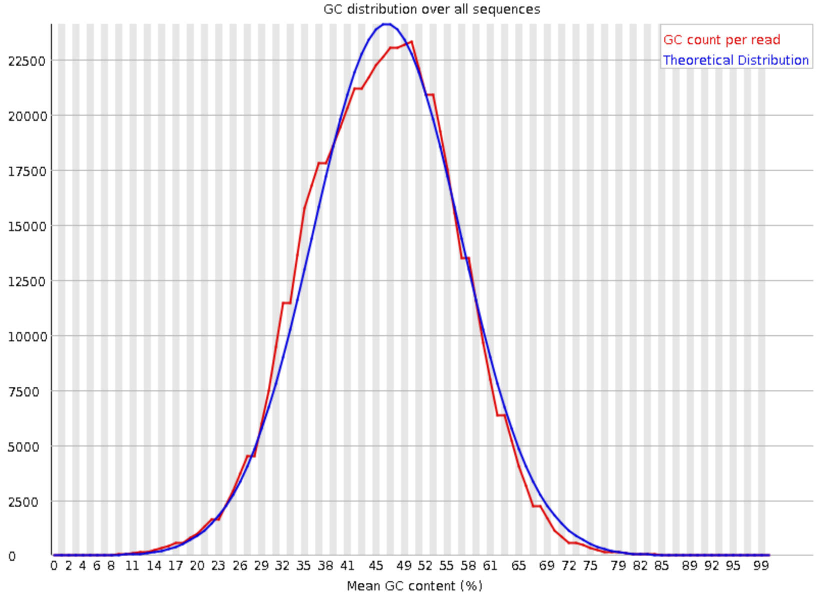
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html

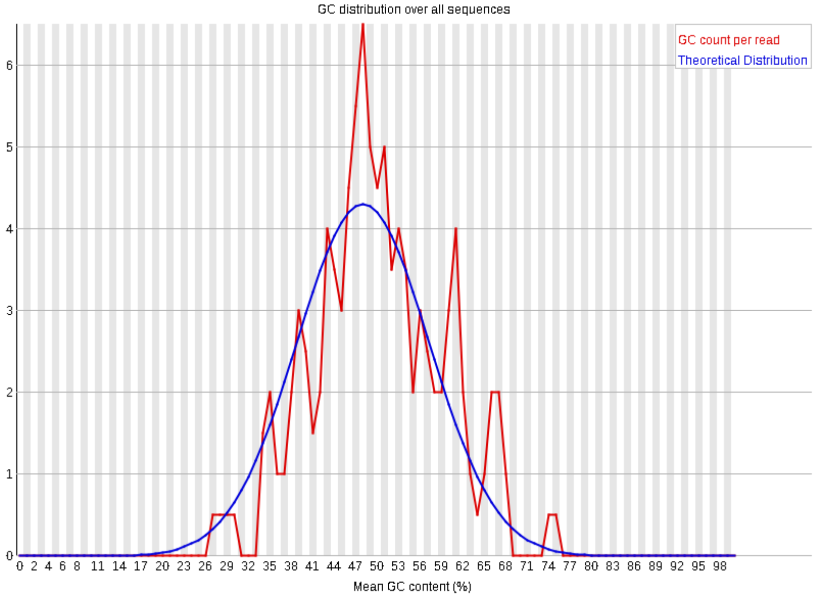
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

Per sequence GC content

Warning: deviation from normal in > 15% reads

Fail: deviation from normal in > 20% reads ●



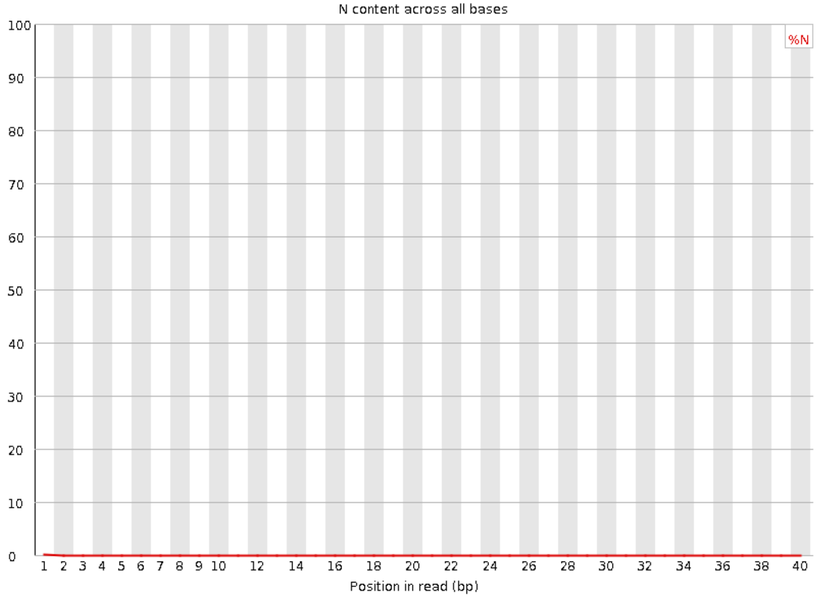


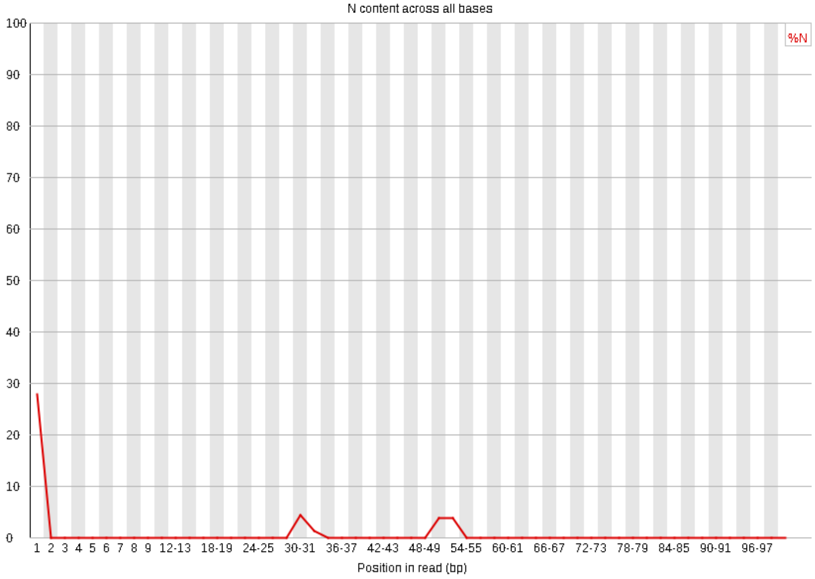
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

Per base N content

Warning: N content > 5% at any position

Fail: N content > 20% at any position ●



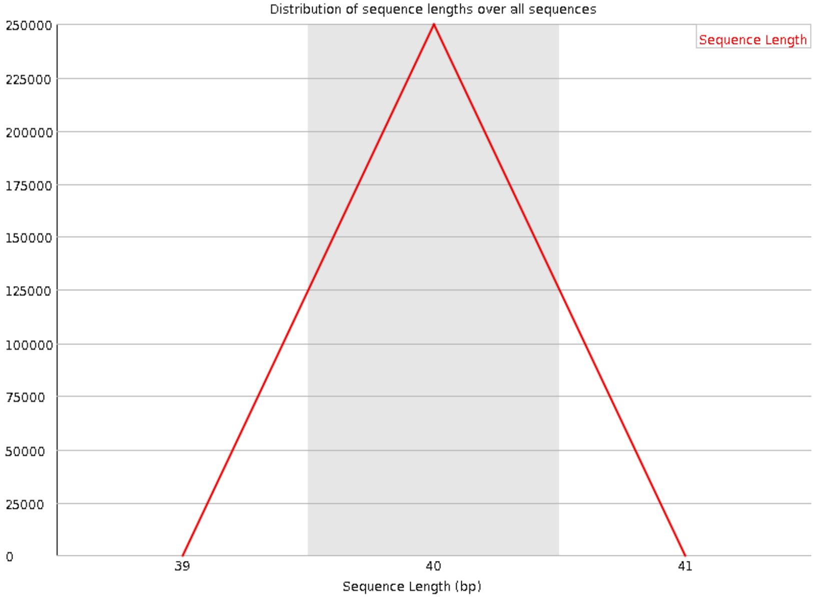


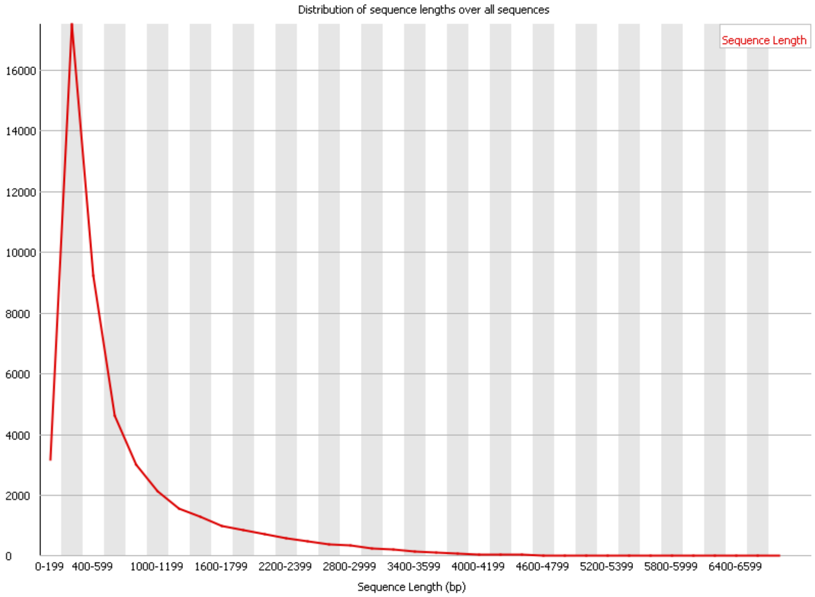
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/6%20Per%20Base%20N%20Content.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

Sequence length distribution

Warning: all sequences are not the same length

Fail: at least one 0 length





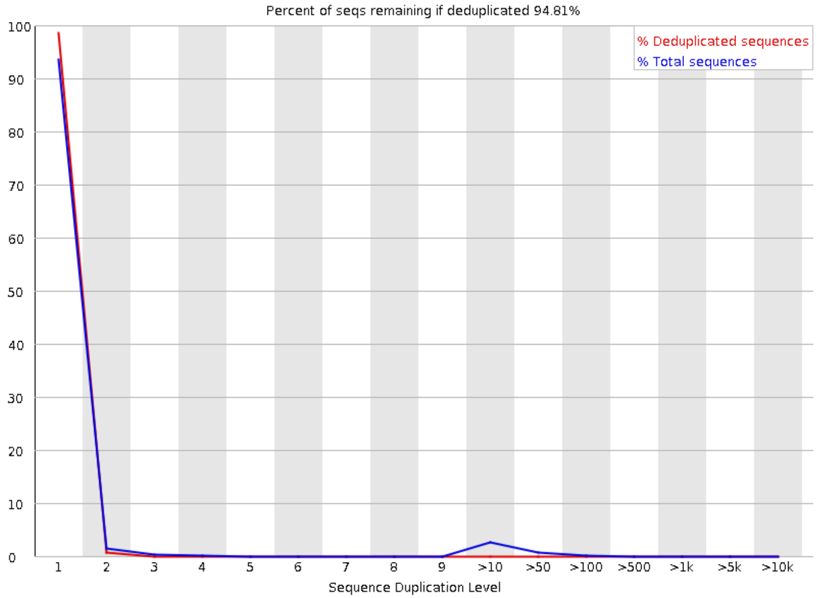
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

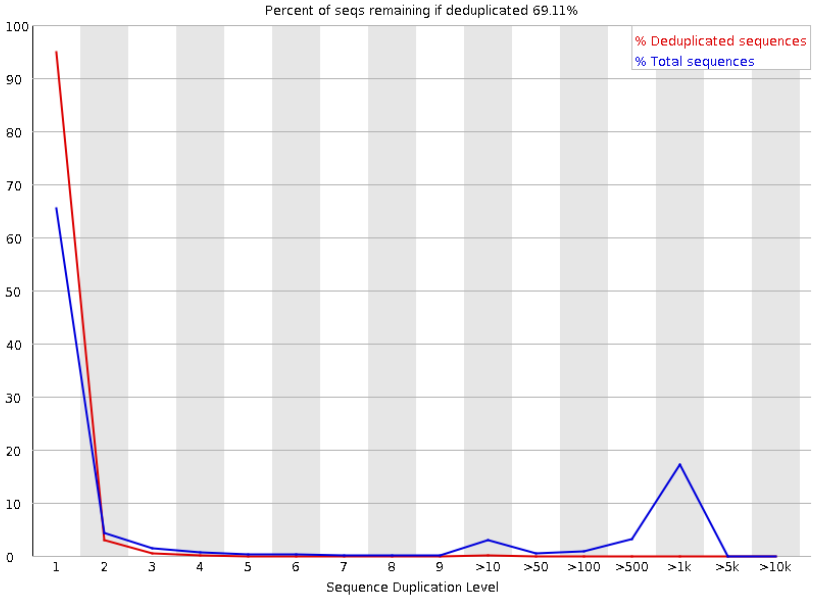
Sequence duplication levels

Warning: non-unique sequences > 20%

Fail: non-unique sequences > 50%

●





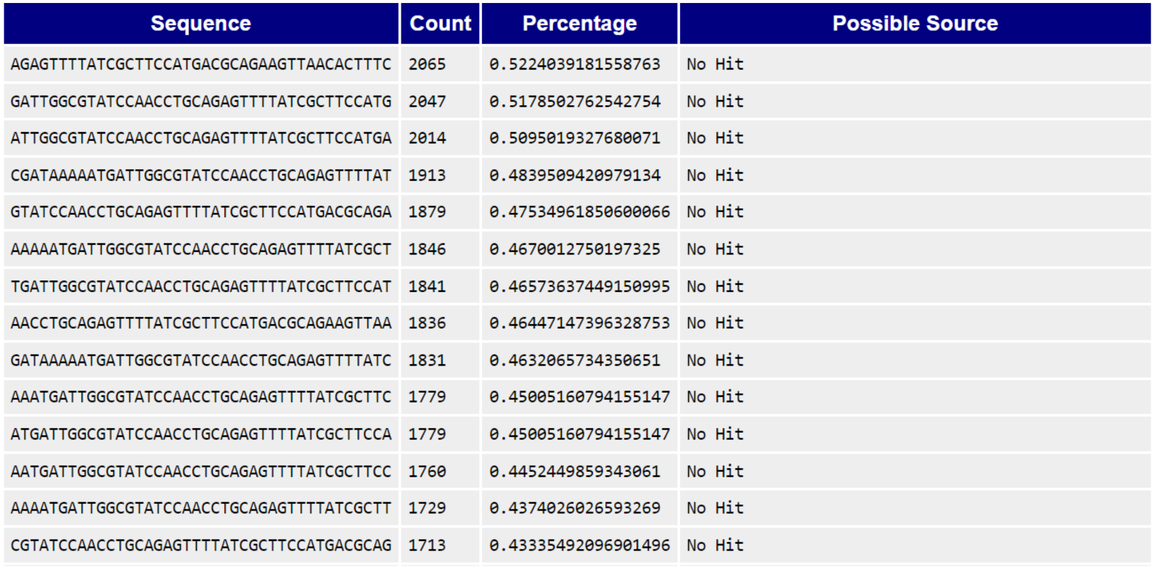
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/8%20Duplicate%20Sequences.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Overrepresented sequences

Warning: sequence found in > 0.1% of total

Fail: sequence found in > 1% of total

None



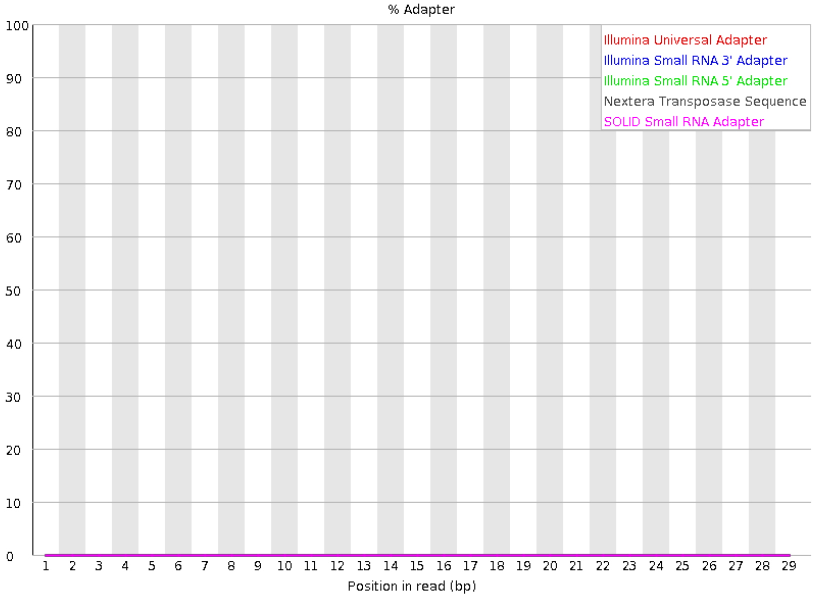
Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/9%20Overrepresented%20Sequences.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Adapter content

Warning: adapter in > 5%

Fail: adapter in > 10%

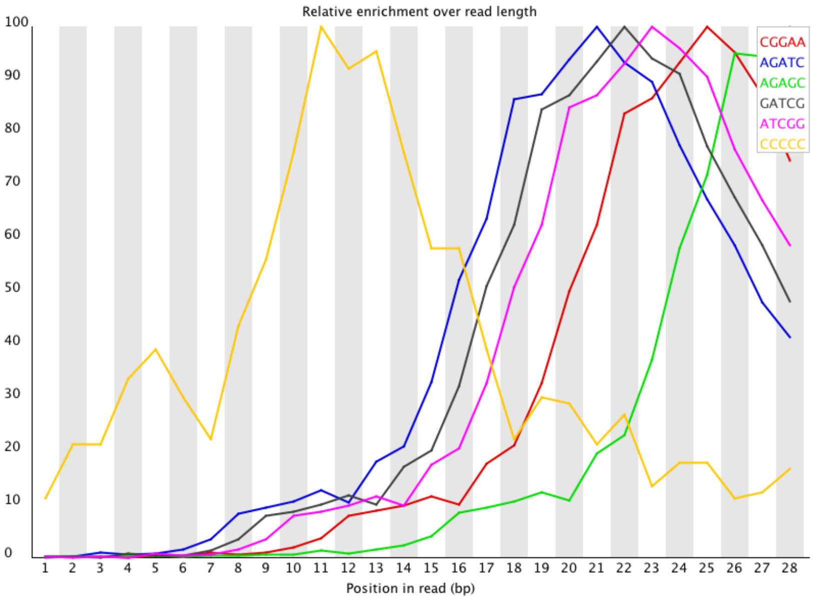


https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/10%20Adapter%20Content.html
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

K-mer content

Warning: kmer imbalanced with p-value < 0.01

Fail: kmer imbalanced with p-value < 10^-5



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/10%20Adapter%20Content.html

Hands on (15 min)
Run a QC analysis and interpret the results

Run fastqc on the 6 available fastq files in the Data folder.

```
module load fastqc/0.11.9
fastqc file1 file2 …
```

Optional: merge the results with multiqc

```
module load mugqic/MultiQC/1.12
multiqc .
```

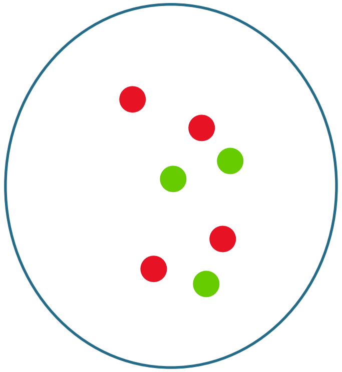
What is your interpretation?

Normalization

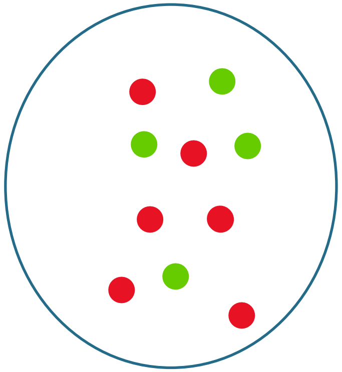
Why do we normalize?

Make things comparable

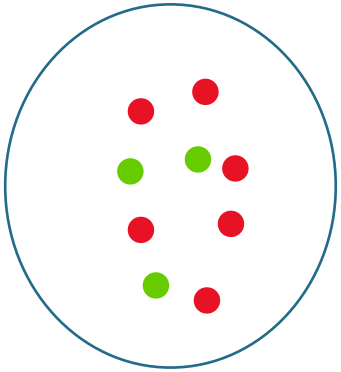
Initially, transform into a normal (gaussian) distribution (now any relevant distribution)



Red : 4
Green : 3



Red : 6
Green : 4

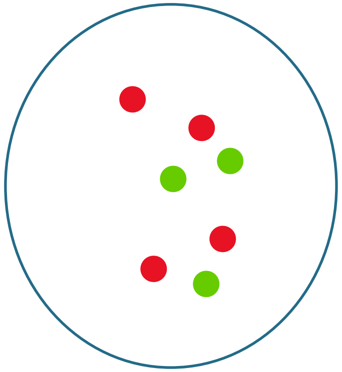


Red : 6
Green : 3

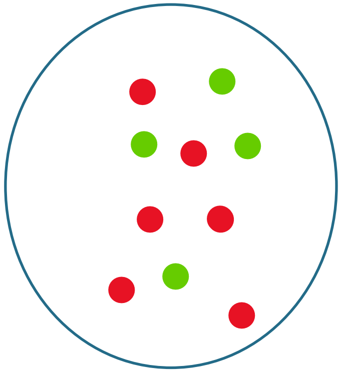
Why do we normalize?

Make things comparable

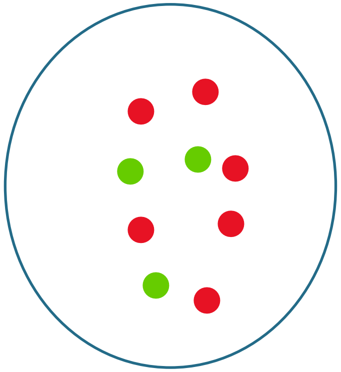
Initially, transform into a normal (gaussian) distribution (now any relevant distribution)



Red : 4/7 ≈ 57%
Green : 3/7 ≈ 43%



Red : 6/10 = 60%
Green : 4/10 = 40%



Red : 6/9 ≈ 67%
Green : 3/9 ≈ 33%

Common normalization techniques

- RPM (reads per million)

  # reads mapped to a gene * 10^6 / library size (total # mapped reads)

  -> gene length not considered

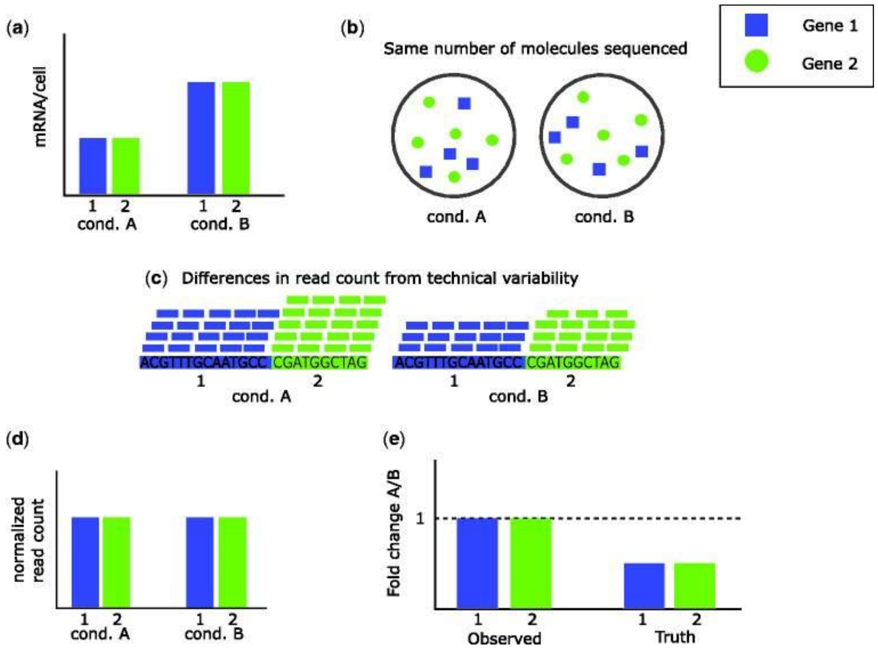
- RPKM/FPKM (reads/fragments per kb per million)

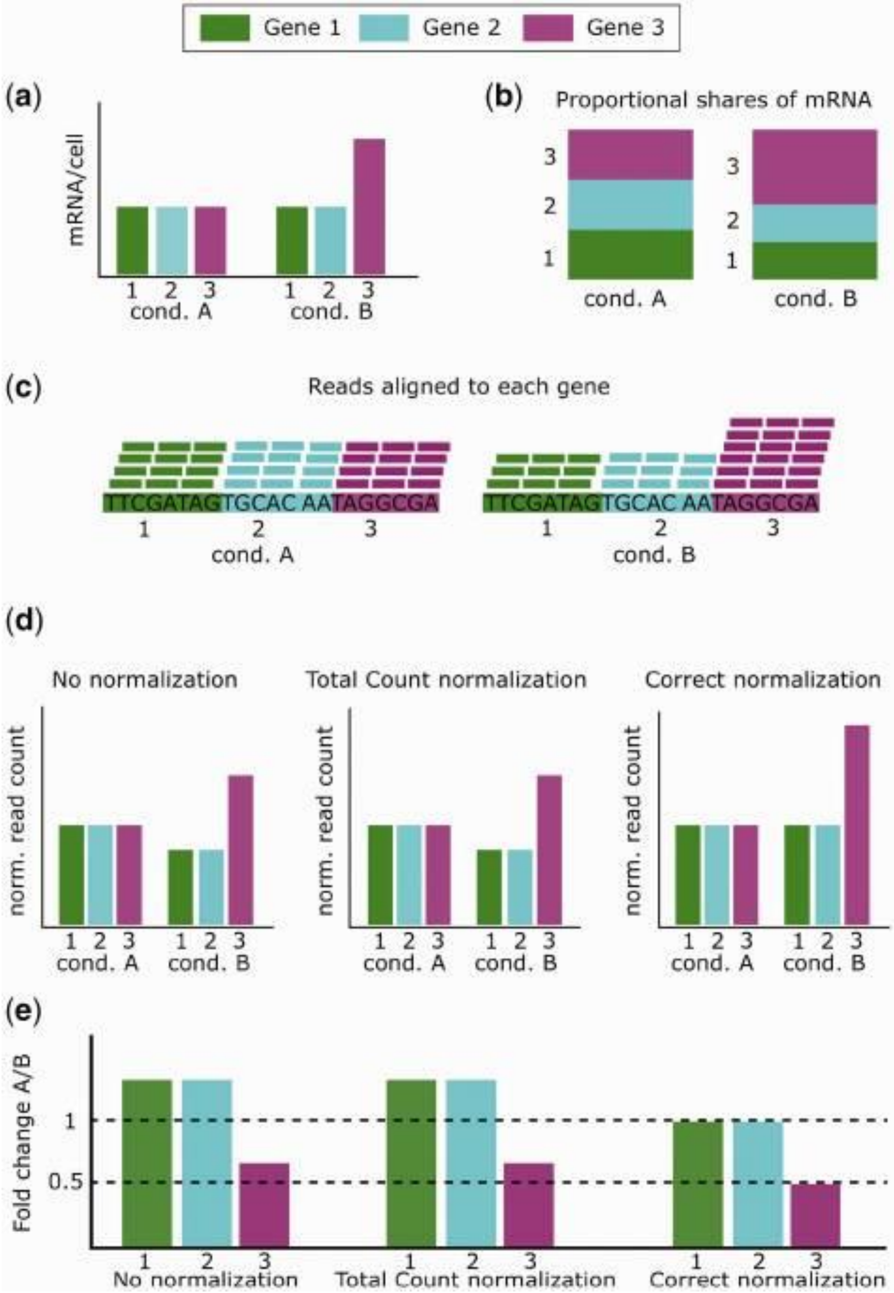
  # reads mapped to a gene *10^6 / (library size * gene length in kb)

- TPM (transcripts per million)

  # reads mapped to a gene / gene length in kb

Normalization issues…



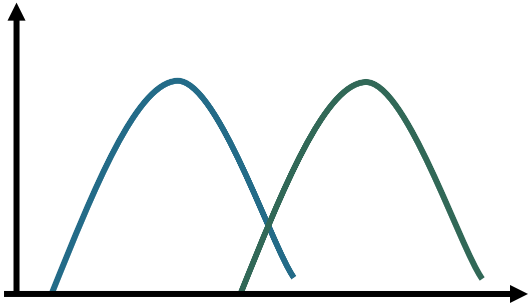


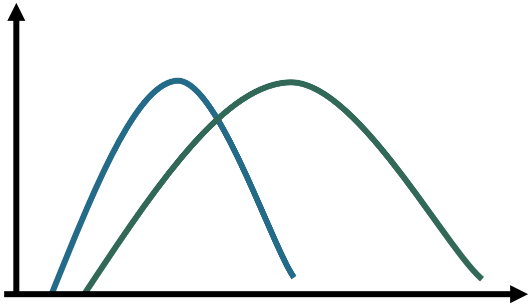
Evans, Ciaran et al. "Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions." *Briefings in bioinformatics* vol. 19,5 (2018): 776-792. doi:10.1093/bib/bbx008
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6171491/

Using replicates

Technical or biological?

- Introducing variance





- Stats

  Fold change (magnitude)

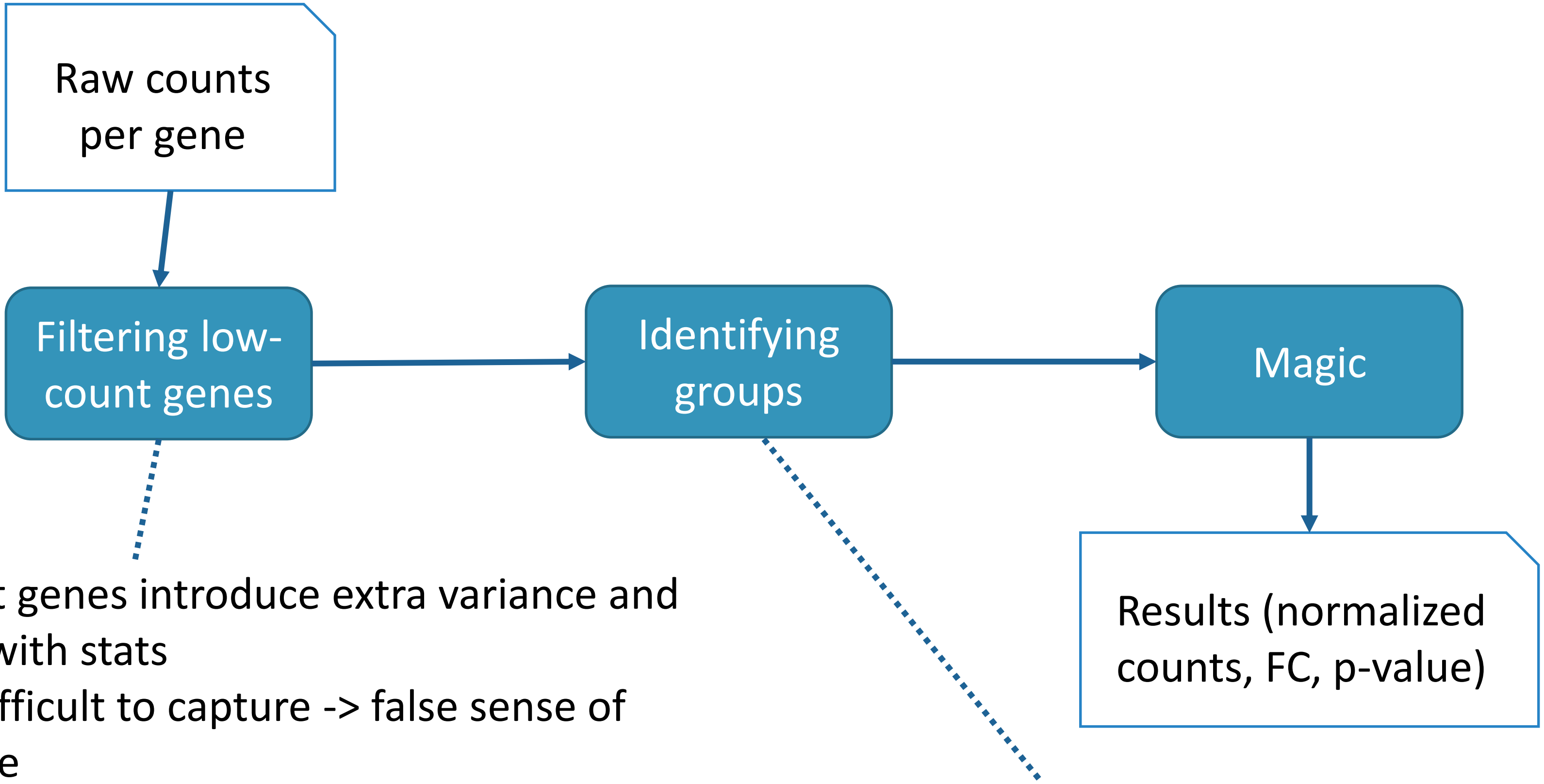
  P-value (significance)

- Assumptions on how genes should behave

  The counts for sample i, gene j follow a negative binomial distribution

  H_0: no gene is differentially expressed

Normalization in differential gene expression analysis


Raw counts per gene
Filtering low-count genes
Identifying groups
Magic
Results (normalized counts, FC, p-value)

Low count genes introduce extra variance and interfere with stats
May be difficult to capture -> false sense of abundance
*May have poor biological meaning
Not suitable for miRNAs and tRNAs

What is the control?
What is/are the treatment(s)?

Hands on (5 min)
Identify the appropriate normalization depending on the analysis

What are the characteristics to consider in a DEG?

Hint: Should those elements be considered?

- Gene length
- Coverage
- GC content
- …

Hands on
Identify the appropriate normalization depending on the analysis

Cares about:

- Relationship between conditions
- Sequencing depth/lib size
- Sample-specific effects

Does not care bout:

- Counts themselves
- Gene length
- GC content

Lunch break

Overview

- EdgeR (45 min)
  - Steps to do a differential gene expression analysis
  - Choosing the appropriate fitting and testing function
  - Extracting the results
  - Hands on: produce a DEG analysis, with the appropriate fitting and testing functions (10 min)
- Interpreting common plots (10 min)
  - Volcano plots
  - MA plots
  - Hands on: identify the most interesting genes in the plots (5 min)

EdgeR

Steps to do a differential gene expression analysis

Assumption: more than half of genes are NOT DE

1. Filter low count genes
2. Calculate normalization factors
3. Identify groups to compare
4. Estimate the dispersion
5. Fit and test the appropriate model
6. Get DEGs

edgR – calcNormFactors()

- Scale the lib size to minimize log-FC between samples. It avoids under-sampling problems when a few genes make up the majority of reads
- Trimmed mean of M-values (TMM)

  Pairwise comparison of two samples, for all genes, what is the relation between the # of counts of a gene and the total number of counts of the sample?

  Robinson, Mark D, and Alicia Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data.” Genome biology vol. 11,3 (2010): R25. doi:10.1186/gb-2010-11-3-r25

  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2864565/
- Effective lib size = lib size * scaling factor

  -> scaling factor < 1: high counts on few genes, lib size is reduced, upscaling of counts

  -> scaling factor > 1: lib size is increased, downscaling of counts

edgeR – Biological Coefficient of Variation

- Relationship between the observed counts y_{gi} and the fraction π_{gi} of the total reads N_i

$$E(y_{gi}) = \mu_{gi} = N_i \pi_{gi}.$$

- Assumption: y_{gi} for repeated sequencing of the same sample ~ Poisson distribution

$$\text{var}(y_{gi}) = E_\pi\left[\text{var}(y|\pi)\right] + \text{var}_\pi\left[E(y|\pi)\right] = \mu_{gi} + \phi_g \mu_{gi}^2.$$

Dividing both sides by μ_{gi}^2 gives

$$\text{CV}^2(y_{gi}) = 1/\mu_{gi} + \phi_g.$$

Technical CV^2 + biological CV^2
(CV: coefficient of variation)

edgeR – Biological Coefficient of Variation

- As sequencing depth → ∞, TCV should ↓ and BCV should remain stable
- Assumption: true gene abundance ~ Gamma distribution

  -> read counts ~ negative binomial probability law
- BCV approximated as $\sqrt{\Phi_g}$, where Φ_g is the dispersion

edgeR – estimateDisp()

- Common dispersion estimate

  All genes have the same dispersion

- Trended dispersion estimate

  Smooth function dependent on gene counts

  Common dispersion for bins of genes with similar expression -> loess curve

- Genewise/tagwise dispersion estimate

  Genes have their own dispersion, dependent on gene length, sequence, expression level and/or function

  Compromise between common dispersion and fully-individual dispersion by using a weighted likelihood empirical Bayes approach

  Difficult (low-count) genes' dispersion will tend towards common dispersion

edgeR – exactTest()

- Quasi negative binomial
- Variant that better captures gene-specific BCV and TCV

$$\mathrm{var}(y_{gi}) = \underbrace{\sigma_g^2}_{\text{Gene-specific}}(\mu_{gi} + \underbrace{\phi}_{\text{Global parameter}}\mu_{gi}^2),$$

- Empirical Bayes approach to compensate # replicates
- Reliable
- Only 1 factor

edgeR – Generalized linear models

- 1 or more factors (more complex experiments)
- Linear models for non-normal data
- Mean-variance relationship
- Given μ_{gi} and Φ_g, we can deduce the variance

$$\mathrm{var}(y_{gi}) = E_\pi\left[\mathrm{var}(y|\pi)\right] + \mathrm{var}_\pi\left[E(y|\pi)\right] = \mu_{gi} + \phi_g\mu_{gi}^2.$$

- Quasi-likelihood F-test: glmQLFit() -> glmQLFTest()

  Reflects uncertainty in estimating the dispersion, more robust and reliable error rate with small number of replicates

- Likelihood ratio test: glmFit() -> glmLRT()

edgeR – summary

Preparation (all steps are required)	DGEList()	Create the right object
	filterByExpr()	Remove interference from low-count genes
	calcNormFactors()	Get the effective lib size
	model.matrix()	Identify groups
	estimateDisp()	Get a dispersion estimate (required for variance estimate)
Fitting and testing (chose one)	exactTest()	1 factor only Pairwise comparison
	glmQLFit() + glmQLFTest()	2+ factors
	glmFit() + glmLRT()	No replicates?
	topTags()	Get DEGs

Choosing the appropriate function(s)

	Analysis 1	Analysis 2	Analysis 3
Condition1_rep1	1		1
Condition1_rep2	1		1
Condition2_rep1	2		2
Condition2_rep2	2		2
Condition3_rep1		1	
Condition4_rep1		2	
Condition5_rep1			3
Condition5_rep2			3

- exactTest()
- glmQLFit() + glmQLFTest()
- glmFit() + glmLRT()

Choosing the appropriate function(s)

	Analysis 1	Analysis 2	Analysis 3
Condition1_rep1	1		1
Condition1_rep2	1		1
Condition2_rep1	2		2
Condition2_rep2	2		2
Condition3_rep1		1	
Condition4_rep1		2	
Condition5_rep1			3
Condition5_rep2			3

1. exactTest()

3. glmQLFit() + glmQLFTest()

2. glmFit() + glmLRT()

Resources

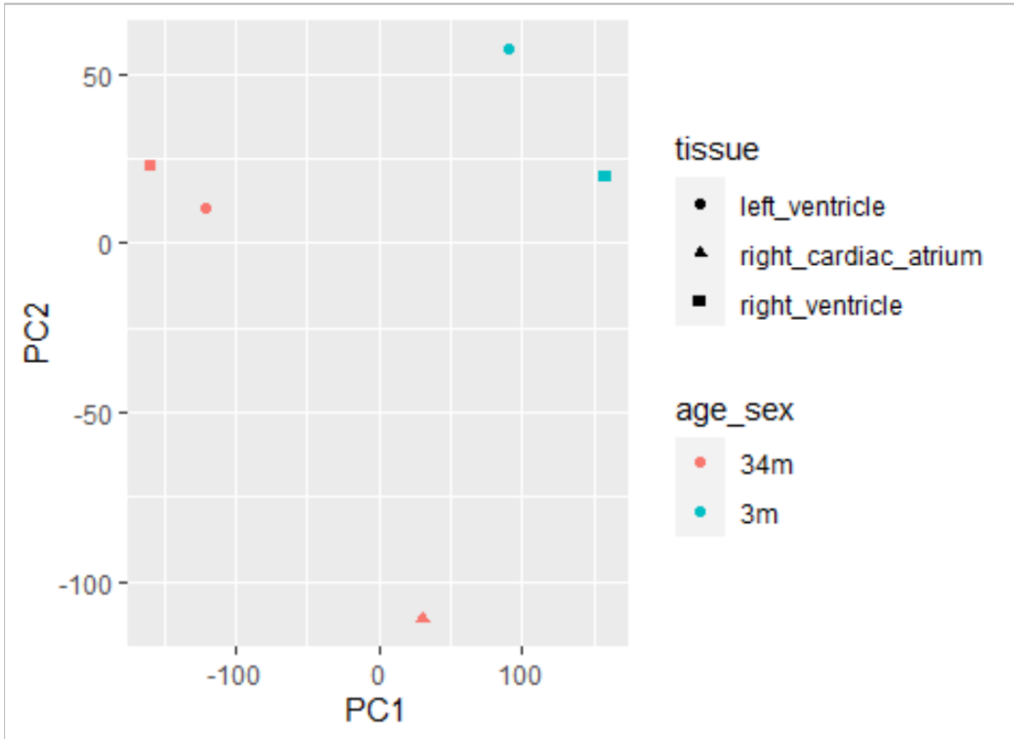
- https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

- Robinson, Mark D et al. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” *Bioinformatics (Oxford, England)* vol. 26,1 (2010): 139-40. doi:10.1093/bioinformatics/btp616
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/

- McCarthy, Davis J et al. “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.” *Nucleic acids research* vol. 40,10 (2012): 4288-97. doi:10.1093/nar/gks042
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3378882/

Hands on (15 min)
Produce a DEG analysis, with the appropriate fitting and testing functions

1. Given the following PCA, what can we accurately compare? Are there outliers we should exclude?

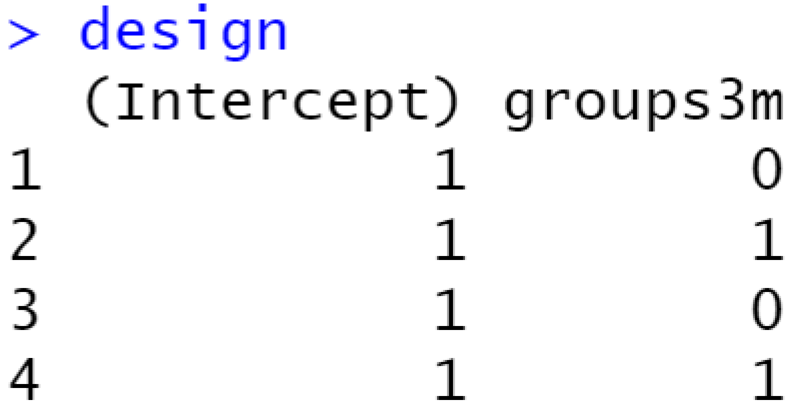


Hands on (15 min)
Produce a DEG analysis, with the appropriate fitting and testing functions

2. Load the files, prepare the groups, filter by expression

34m = control

3m = condition



```
> design
  (Intercept) groups3m
1           1        0
2           1        1
3           1        0
4           1        1
```

Hands on (15 min)
Produce a DEG analysis, with the appropriate fitting and testing functions

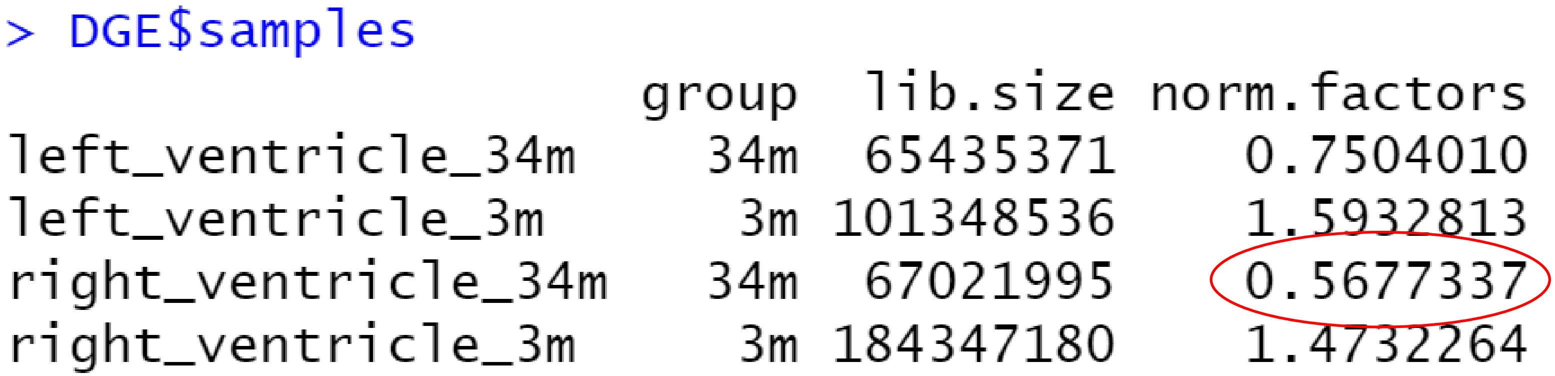
3. Calculate the normalization factors and estimate the dispersion

What sample(s) had the strongest problem of having a few genes monopolizing the reads?

Hands on (15 min)
Produce a DEG analysis, with the appropriate fitting and testing functions

3. Calculate the normalization factors and estimate the dispersion

What sample(s) had the strongest problem of having a few genes monopolizing the reads?



```
> DGE$samples
                     group  lib.size norm.factors
left_ventricle_34m     34m  65435371    0.7504010
left_ventricle_3m       3m 101348536    1.5932813
right_ventricle_34m    34m  67021995    0.5677337
right_ventricle_3m      3m 184347180    1.4732264
```

Hands on (15 min) Produce a DEG analysis, with the appropriate fitting and testing functions

4. Fitting and testing

Choose the appropriate function(s)

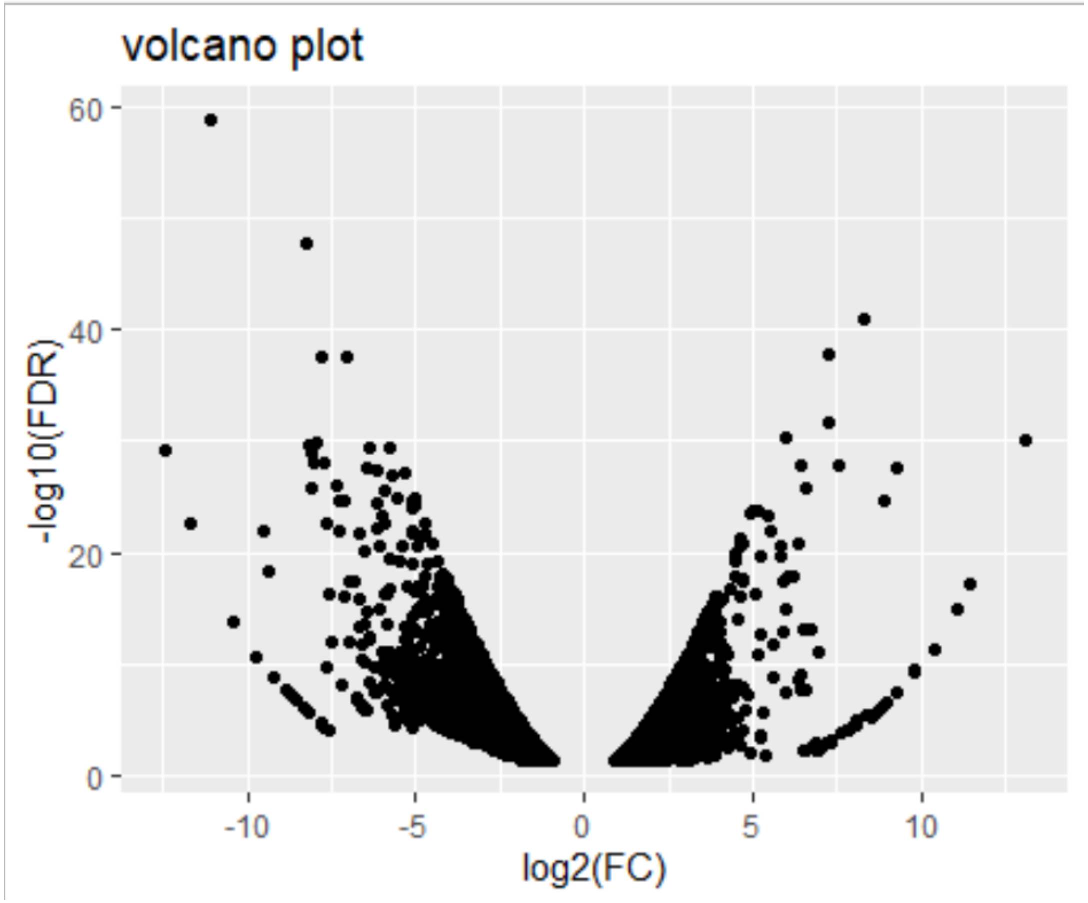
- exactTest(DGE)
- glmQLFit(DGE, design) + glmQLFTest(fit)
- glmFit(DGE, design) + glmLRT(fit)

Optional: test the other functions and compare the results

Interpreting common plots

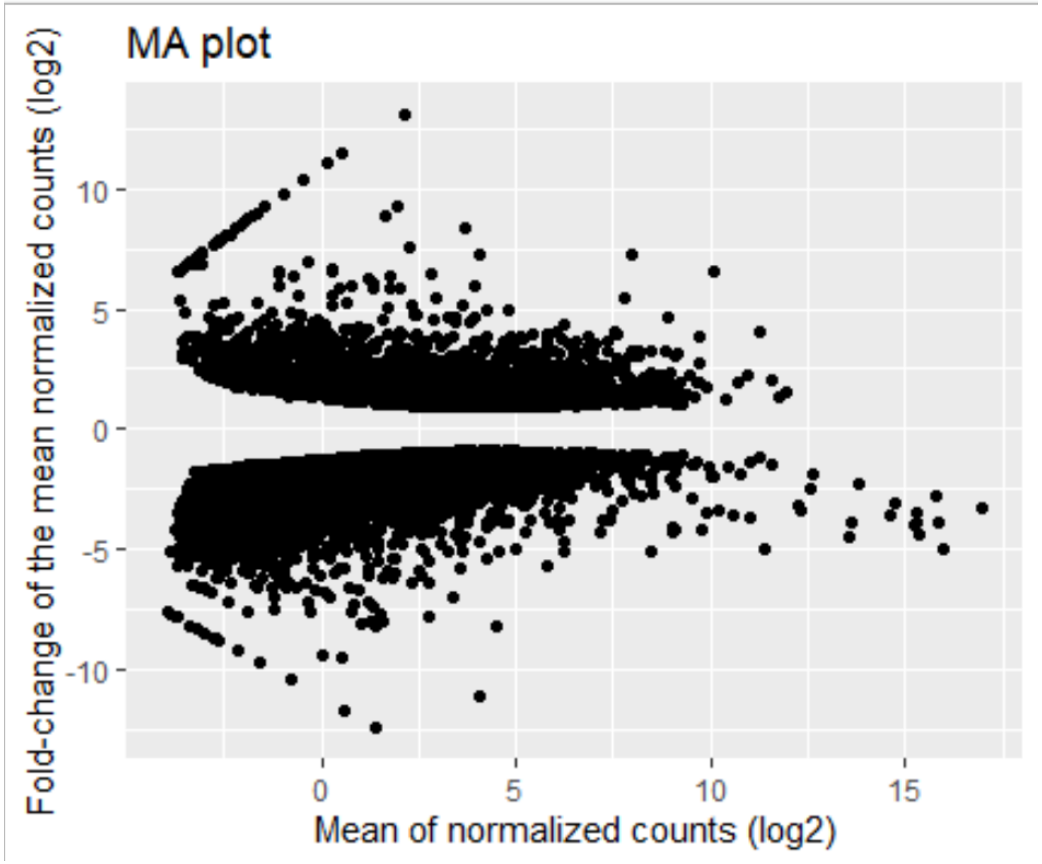
Volcano plots

- X axis: Fold change (log2)
- Y axis: pvalue/FDR (-log10)
- Goal: See what genes have a high FC AND high significance



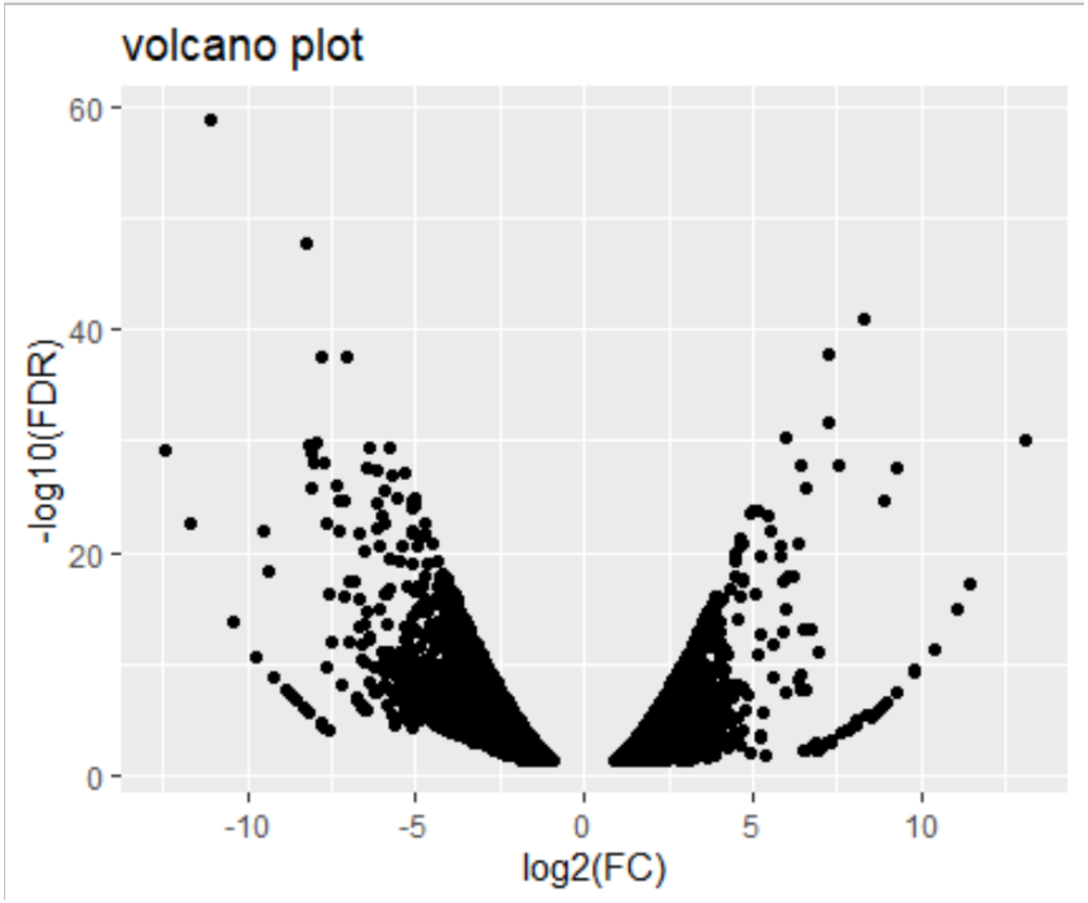
MA plots

- X axis: mean (log) of the normalized counts (logCPM)
- Y axis: log2 FC
- Goal: Differentiate high FC, low counts and high FC, high counts



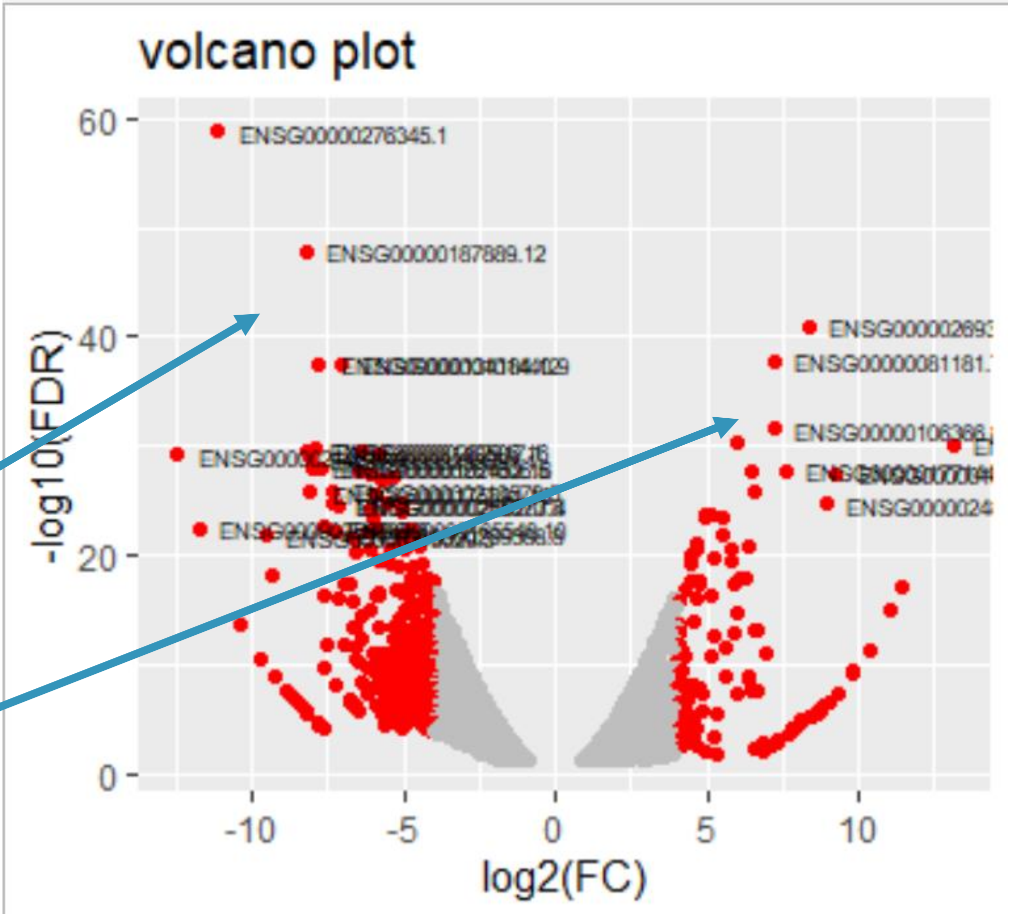
Hands on (5 min)
Identify the most interesting genes in the plots

- X axis: Fold change (log2)
- Y axis: pvalue/FDR (-log10)
- Goal: See what genes have a high FC AND high significance



Hands on
Identify the most interesting genes in the plots

- Values of interest are at the top left (downregulation) and top right (upregulation)

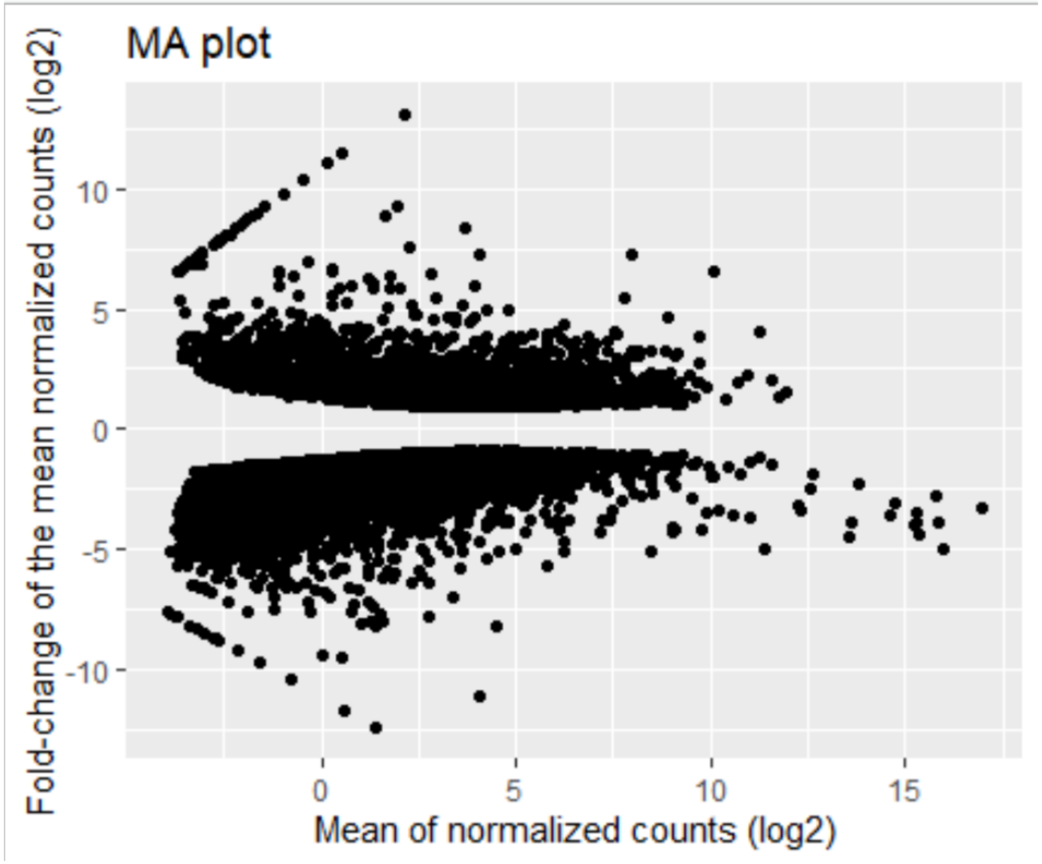


control > condition
34m > 3m

condition > control
3m > 34m

Hands on (5 min)
Identify the most interesting genes in the plots

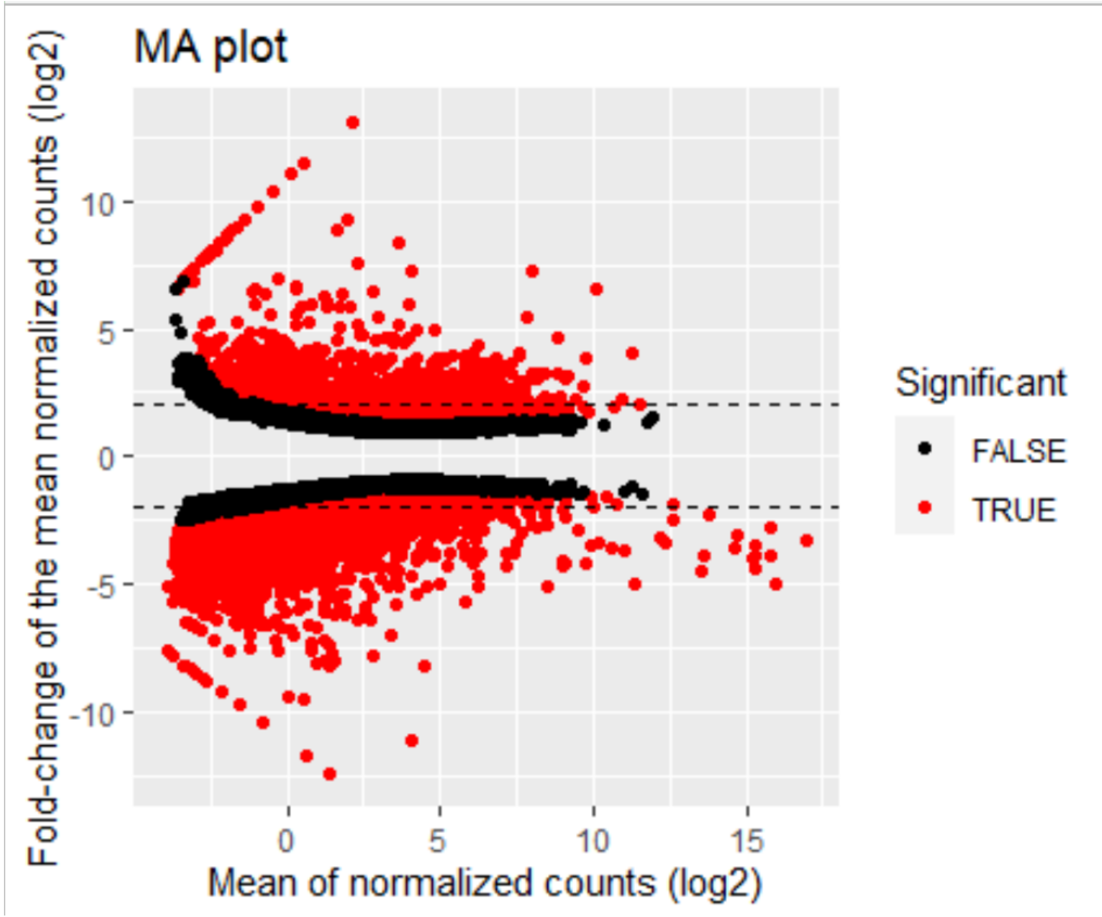
- X axis: mean (log) of the normalized counts (logCPM)
- Y axis: log2 FC
- Goal: Differentiate high FC, low counts and high FC, high counts



McGill initiative in Computational Medicine

Hands on
Identify the most interesting genes in the plots

- Values of interest are at the top right (high expression, high FC) and bottom right (high expression, low FC)


MA plot
Fold-change of the mean normalized counts (log2)
Mean of normalized counts (log2)
Significant
● FALSE
● TRUE