Analysis of RNA-seq data

Céline Keime
keime@igbmc.fr

Analysis of RNA-seq data

Quality analysis

↓

Mapping

↓

Gene expression quantification

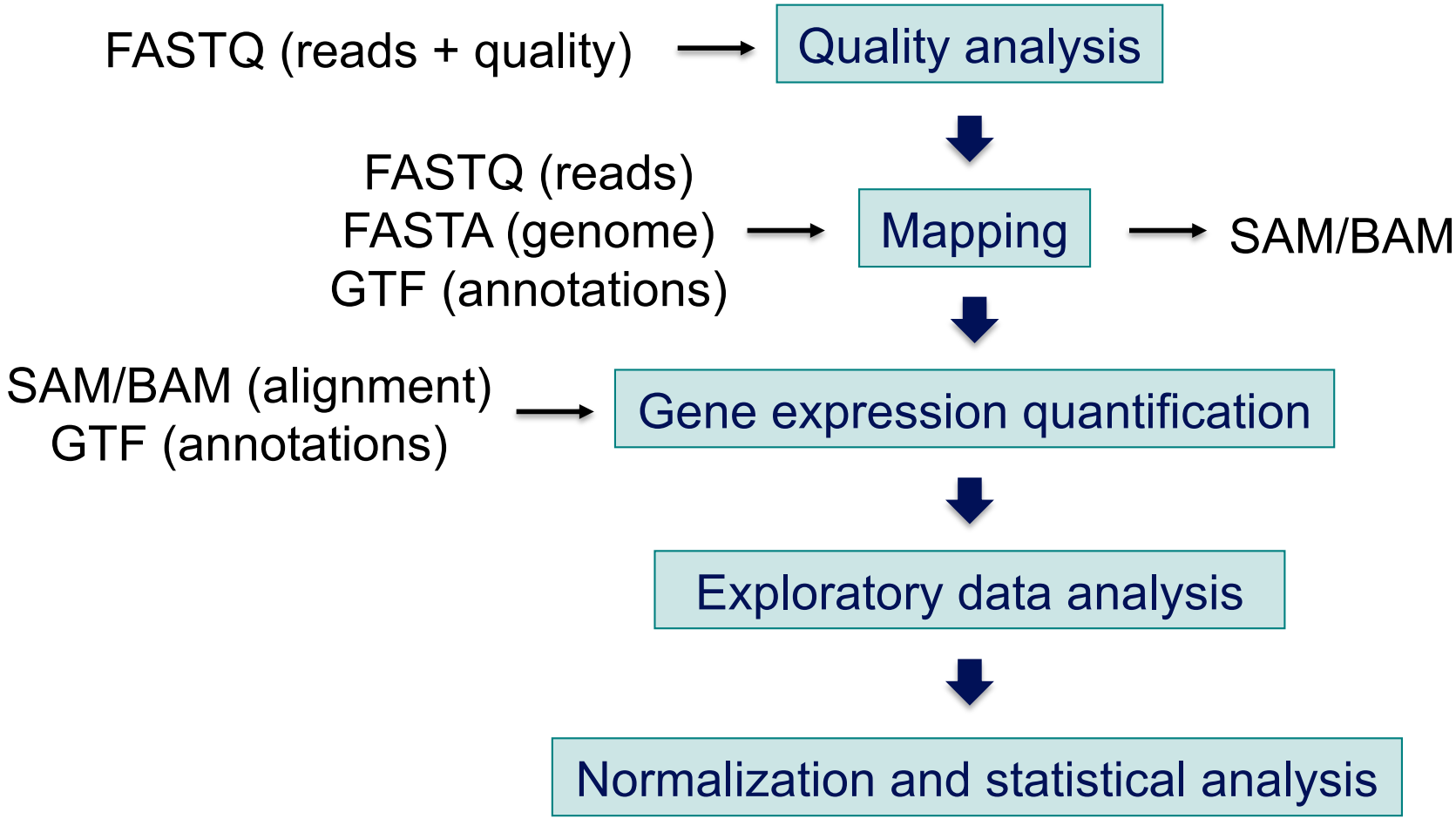
↓

Exploratory data analysis

↓

Normalization and statistical analysis

Analysis of RNA-seq data



Analysis of RNA-seq data

Quality analysis

↓

Mapping

↓

Gene expression quantification

↓

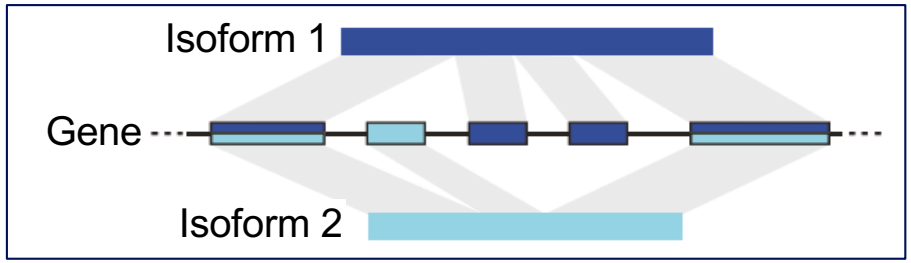
Exploratory data analysis

↓

Normalization and statistical analysis

Gene-level quantification

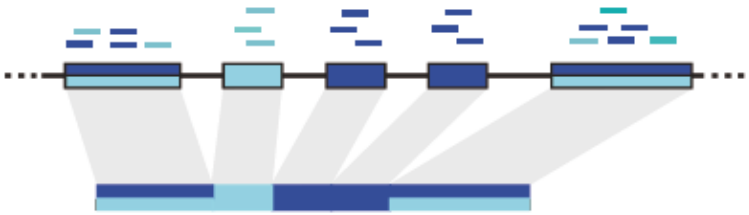
- How to summarize expression level of genes with several isoforms ?



Garber et al., Nature methods 2011; 8(6):469-77

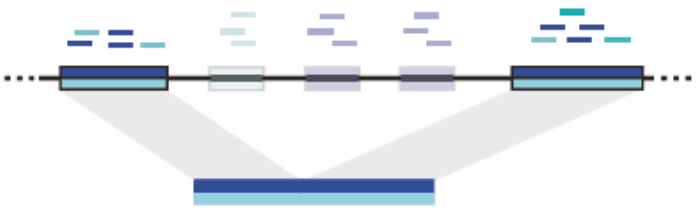
- Exon-union method

  Count reads mapped to all exons from all isoforms of the gene



- Exon-intersection method

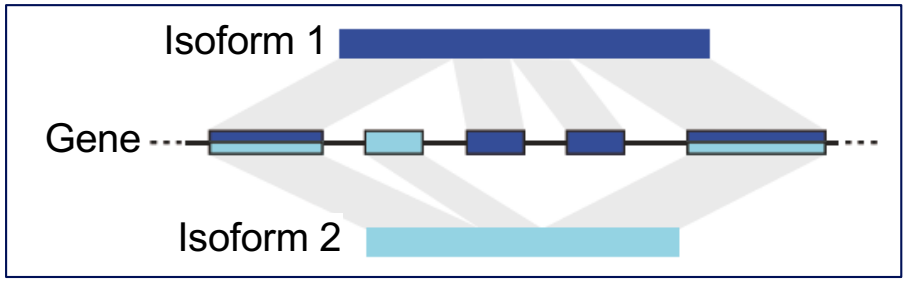
  Count only reads mapped to its constitutive exons



➔ reduce power for differential expression analysis

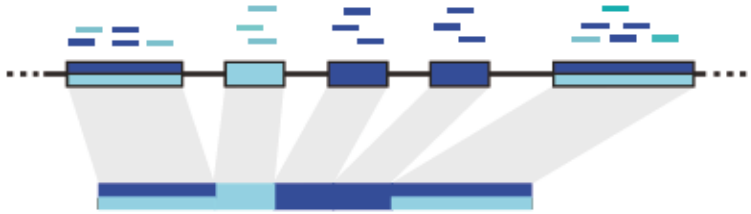
Gene-level quantification

- How to summarize expression level of genes with several isoforms ?



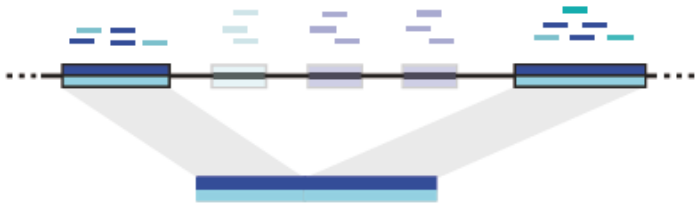
- Exon-union method

  Count reads mapped to all exons from all isoforms of the gene



- Exon-intersection method

  Count only reads mapped to its constitutive exons



➔ reduce power for differential expression analysis

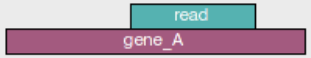
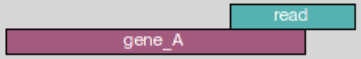
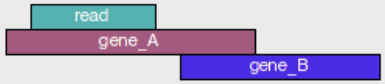
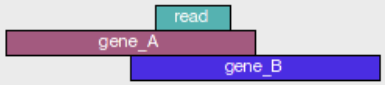
Gene-level quantification : HTSeq-count

Anders et al., Bioinformatics 2015;31(2):166-9

- How to deal with multiple aligned reads ?
  - Multi-mapped reads are discarded rather than counted for each feature
    - Because the primary intended use case for htseq-count is differential expression analysis
    - i.e. comparison of the expression of the same gene across samples
  - Why ?
    - Consider 2 genes with multiple aligned reads on these genes
    - Discard multiple aligned reads
      - → undercount the total output of these 2 genes
      - But the expression ratio between conditions will still be correct
      - Because we discard the same fraction of reads in all samples
    - If we counted these reads for both genes
      - → differential expression analysis might find false positives
      - Even if only one of the gene is differentially expressed
      - Multi-mapped reads would be counted for both genes
      - Gives the wrong appearance that both genes are differentially expressed

Gene-level quantification : HTSeq-count

- How to deal with overlapping features ?

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

https://htseq.readthedocs.io/en/latest/count.html

HTSeq-count

- Input
  - Alignment file (SAM/BAM)
  - Annotation file (GFF/GTF) **with the same chromosome names** as in the alignment file
- Options

Mode → cf. previous slide

Union

Mode to handle reads overlapping more than one feature. (--mode)

Stranded →

Yes

Specify whether the data is from a strand-specific assay. **Be sure to choose the correct value** (see help for more information). (--stranded)

Reverse for a directional protocol that generates reads in the opposite strand as the transcribed one
No for a non-directional protocol

Minimum alignment quality

10

Skip all reads with alignment quality lower than the given minimum value. (--minaqual)

Feature type

exon

Feature type (3rd column in GFF file) to be used. All features of other types are ignored. The default, suitable for RNA-Seq and Ensembl GTF files, is exon. (--type)

ID Attribute

gene_id

GFF attribute to be used as feature ID. Several GFF lines with the same feature ID will be considered as parts of the same feature. The feature ID is used to identity the counts in the output table. All features of the specified type MUST have a value for this attribute. The default, suitable for RNA-Seq and Ensembl GTF files, is gene_id. (--idattr)

OK for Ensembl
cf. next slide

Ensembl GTF file

→ **Feature type :** Feature type (3rd column in GFF file) to be used. All features of other types are ignored. The default, suitable for RNA-Seq and Ensembl GTF files, is exon.

3rd column

Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
2	ensembl_havana	gene	227813842	227817564	.	+	.	
2	havana	transcript	227813842	227817564	.	+	.	
2	havana	exon	227813842	227813987	.	+	.	
2	havana	CDS	227813912	227813987	.	+	0	
2	havana	start_codon	227813912	227813914	.	+	0	
2	havana	exon	227815457	227815568	.	+	.	
2	havana	CDS	227815457	227815568	.	+	2	

→ **ID Attribute :** GFF attribute to be used as feature ID. Several GFF lines with the same feature ID will be considered as parts of the same feature. The feature ID is used to identity the counts in the output table. All features of the specified type MUST have a value for this attribute. The default, suitable for RNA-Seq and Ensembl GTF files, is gene_id.

gene_id "ENSG00000115009"; gene_version "11"; transcript_id "ENST00000409189"; transcript_version "7"; exon_number "1"; gene_name "CCL20"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000133189"; havana_gene_version "3"; transcript_name "CCL20-001"; transcript_source "havana"; transcript_biotype "protein_coding"; …

Exercise : quantification of gene expression using HTSeq-count on Galaxy

- Launch HTSeq-count to quantify gene expression on siLuc2_1000000 sample
- Inputs
  - Alignment file you obtained with STAR on siLuc2_1000000.fastq.gz
  - Annotations : Ensembl release 105 GTF file : Homo_sapiens.GRCh38.105.chr.gtf.gz (already imported)

Exercise : quantification of gene expression using HTSeq-count on Galaxy

htseq-count - Count aligned reads in a BAM file that overlap features in a GFF file (Galaxy Version 0.9.1)

Aligned SAM/BAM File

Alignment file you obtained with STAR on siLuc2_1000000

8: RNA STAR on data 5 and data 4: mapped.bam

GFF File

5: Homo_sapiens.GRCh38.105.chr.gtf.gz

Mode

Union

Mode to handle reads overlapping more than one feature. (--mode)

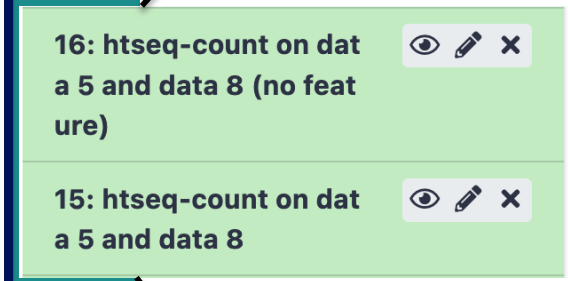
Stranded

Reverse

Specify whether the data is from a strand-specific assay. **Be sure to choose the correct value** (see help for more information). (--stranded)

HTSeq-count on Galaxy

- Output
  - A tabulated text file providing



Category	RNA STAR on data 5 and data 4: mapped.bam
__no_feature	67657
__ambiguous	32425
__too_low_aQual	0
__not_aligned	13608
__alignment_not_unique	450475

the number of reads not assigned to genes

the number of alignments not taken into account

  - A tabulated text file containing the number of reads assigned to each gene

Geneid	RNA STAR on data 5 and data 4: mapped.bam
ENSG00000000003	31
ENSG00000000005	0
ENSG00000000419	95
ENSG00000000457	18
ENSG00000000460	55
ENSG00000000938	0
ENSG00000000971	3
ENSG00000001036	66

HTSeq-count

- Results on siLuc2_1000000

  1. Among uniquely mapped reads, what is the proportion of assigned, no feature and ambiguous reads ?

  → What is the number of uniquely mapped reads ?

  → What is the number of no feature reads ? Calculate the corresponding proportion

  → What is the number of ambiguous reads ? Calculate the corresponding proportion

  → Calculate the proportion of assigned reads

HTSeq-count

- Results on whole dataset
  - Gene quantification results on the whole dataset are available in "NGS data analysis training Strasbourg" history
  - Summary of quantification results

Sample name	% of assigned reads	% of no feature reads	% of ambiguous reads
siLuc2	88.22	7.95	3.83
siLuc3	87.61	8.62	3.77
siMitf3	88.91	7.43	3.65
siMitf4	89.32	6.98	3.70

Transcript-level quantification

- Some reads cannot be assigned unequivocally to a transcript

- Alexa-seq (Griffith et al. Nature methods 2010)

  Counts only reads that map uniquely to a single isoform ➔ Fails for genes that do not contain unique exons from which to estimate isoform expression

- Cufflinks (Trapnell et al. Nature Biotechnology 2010) ; MISO (Katz et al. Nature Methods 2010) ; RSEM (Li et al. BMC Bioinformatics 2011)
  - Construct a likelihood function that models the sequencing process
  - Calculate isoforms abundance estimates that best explain the reads observed in the experiment
- Alignment-free methods
  - Salmon (Patro et al. Nature methods 2017) ; kallisto (Bray et al. Nature Biotechnology 2016)
  - Search which transcript has generated the read
    - Where the read aligns is not necessary
  - Ultra-fast methods

Analysis of RNA-seq data

Quality analysis

↓

Mapping

↓

Gene expression quantification

↓

Exploratory data analysis

↓

Normalization and statistical analysis

→ Launch exploratory data analysis, normalization and statistical analysis on Galaxy

Exercise : SARTools

- SARTools
  - R package dedicated to differential analysis of RNA-seq data
  - Allows to
    - Generate descriptive and diagnostic graphs
    - Run differential analysis with DESeq2 or edgeR package
    - Export the results into tab-delimited files
    - Generate a report
  - Does not replace DESeq2 or edgeR but simply provides an environment to use some of their functionalities

→ **We will use SARTools with DESeq2**

Exercise : SARTools

- Input files for SARTools
  - A zip file containing raw counts files
  - A design file describing the experiment

label	files	group
s1c1	count_file_sample1_cond1.txt	cond1
s2c1	count_file_sample2_cond1.txt	cond1
s1c2	count_file_sample1_cond2.txt	cond2
s2c2	count_file_sample2_cond2.txt	cond2

    - Design file for the analysis we would like to perform :

label	files	group
siLuc2	siLuc2_htseq.txt	siLuc
siLuc3	siLuc3_htseq.txt	siLuc
siMitf3	siMitf3_htseq.txt	siMitf
siMitf4	siMitf4_htseq.txt	siMitf

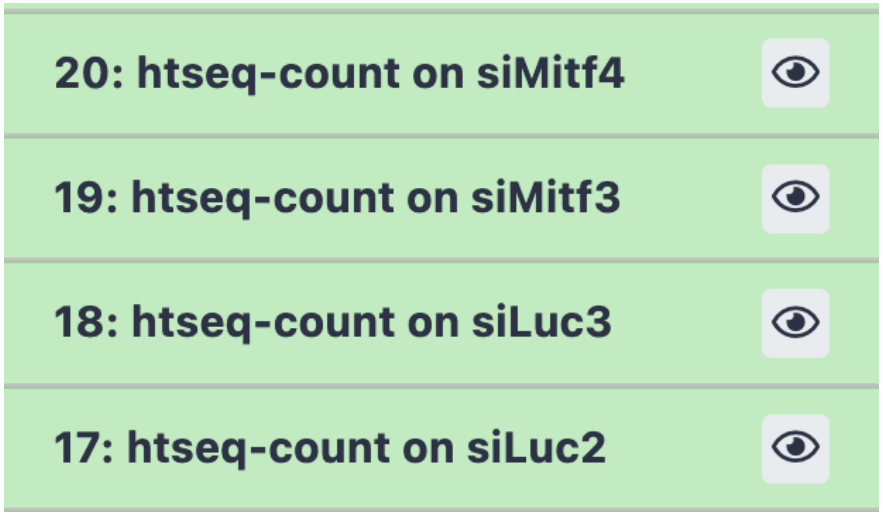
→ **These files can be prepared using the tool "Preprocess files for SARTools"**

Exercise : SARTools

- Launch statistical analysis using SARTools DESeq2
  1. Import raw count files obtained on the whole dataset
     - 17 : htseq-count on siLuc2
     - 18 : htseq-count on siLuc3
     - 19 : htseq-count on siMitf3
     - 20 : htseq-count on siMitf4
  2. Prepare files for SARTools using **Preprocess files for SARTools**
  3. Launch **SARTools DESeq2**

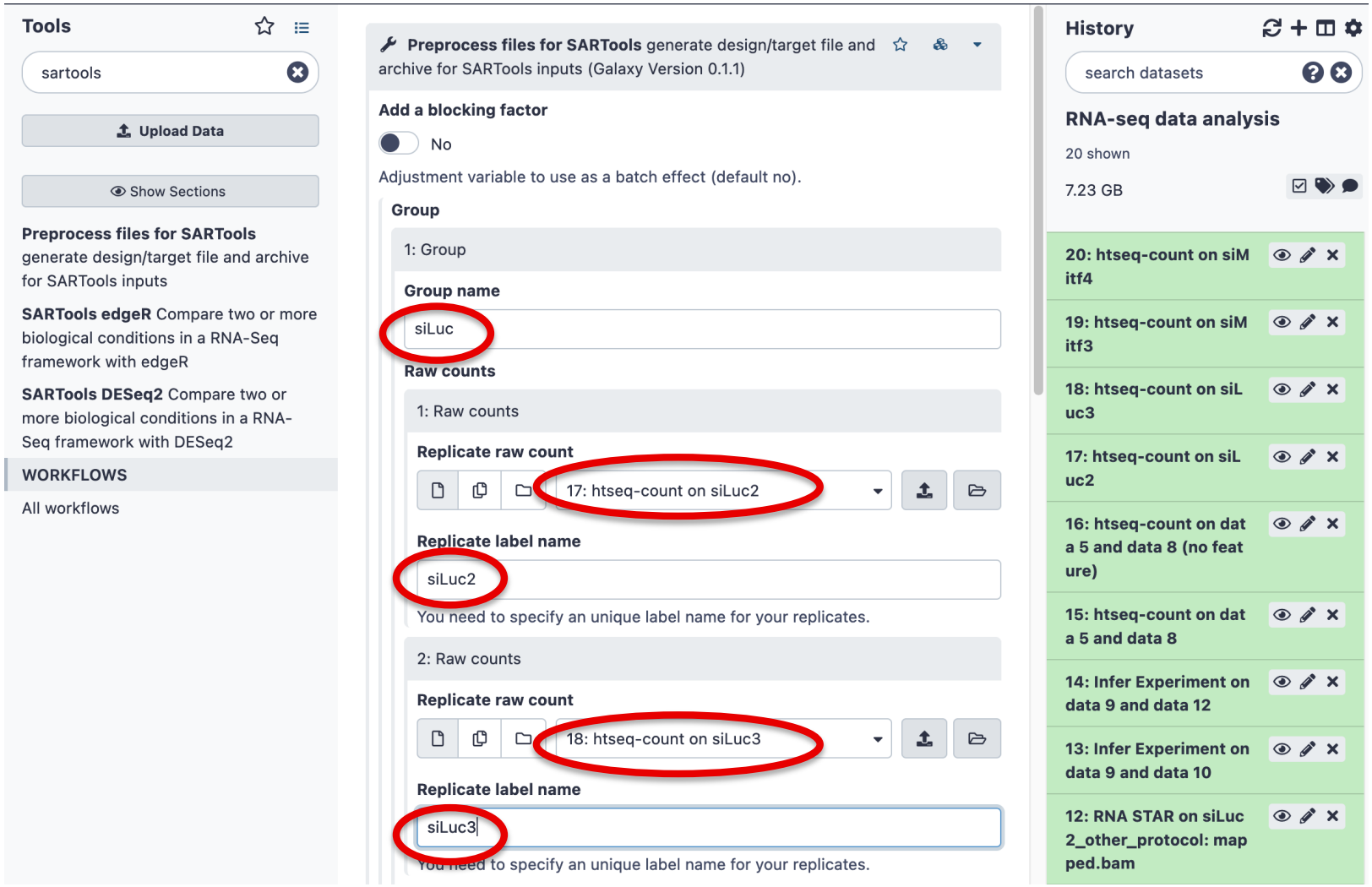
Exercise
1. Import raw counts files

- Import all counts tables that have been obtained with HTSeq-count on the whole dataset (datasets 17 to 20) :



Exercise
2. Prepare files for SARTools

- Use the tool **Preprocess files for SARTools**



Exercise
2. Prepare files for SARTools

2: Group

Group name

siMitf

Raw counts

1: Raw counts

Replicate raw count

19: htseq-count on siMitf3

Replicate label name

siMitf3

You need to specify an unique label name for your replicates.

2: Raw counts

Replicate raw count

20: htseq-count on siMitf4

Replicate label name

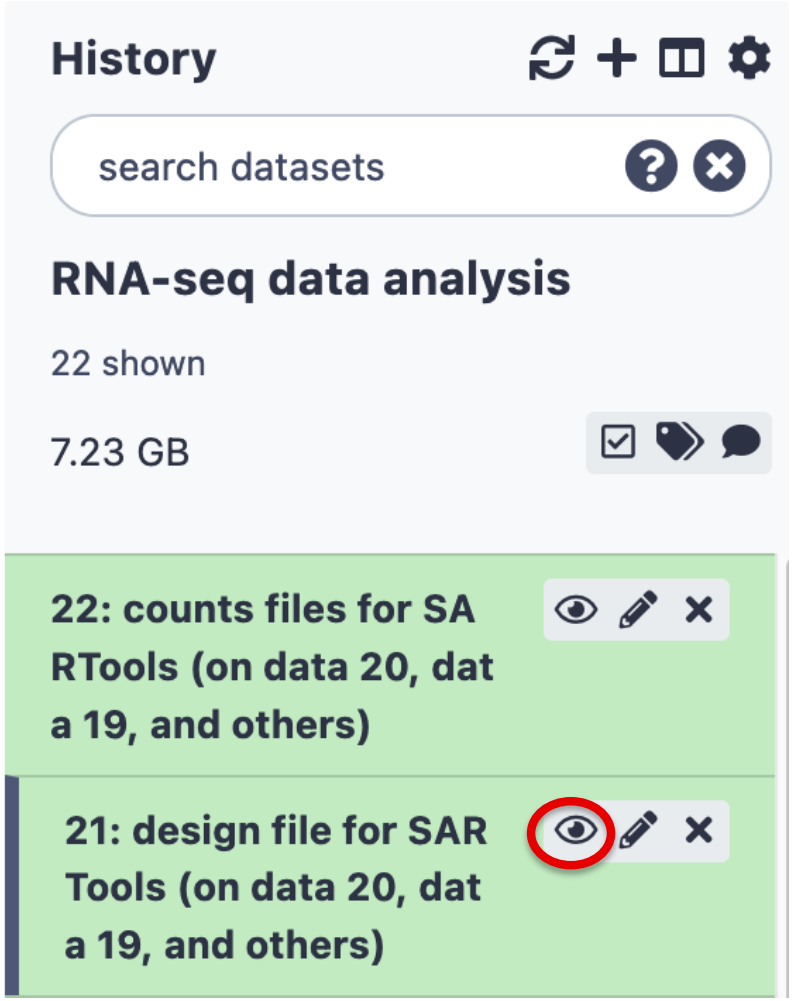
siMitf4

You need to specify an unique label name for your replicates.

+ Insert Raw counts

Exercise
2. Prepare files for SARTools : results

label	files	group
siLuc2	dataset_2432739.dat	siLuc
siLuc3	dataset_2432741.dat	siLuc
siMitf3	dataset_2432743.dat	siMitf
siMitf4	dataset_2432745.dat	siMitf



Exercise
3. Launch SARTools DESeq2

SARTools DESeq2 Compare two or more biological conditions in a RNA-Seq framework with DESeq2 (Galaxy Version 1.7.3+galaxy0)

Name of the project used for the report

Analysis_siMitf_siLuc without space

No space allowed. (--projectName)

Name of the report author

Galaxy

No space allowed. (--author)

Design / target file

21: design file for SARTools (on data 20, data 19, and others)

See the help section below for details on the required format. (--targetFile)

Zip file containing raw counts files

22: counts files for SARTools (on data 20, data 19, and others)

See the help section below for details on the required format. (--rawDir)

Names of the features to be removed

alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual

Separate the features with a comma, no space allowed. More than once can be specified. Specific HTSeq-count information and rRNA for example. Default are 'alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual'. (--featuresToRemove)

Factor of interest

group

Biological condition in the target file. Default is 'group'. (--varInt)

Reference biological condition

siLuc

SARTools results

- Figures

Galaxy Tool SARTools_DESeq2

Run at 14/04/2023 14:28:59

Figures available for downloading

Output File Name (click to view)	Size
MAPlot.png	1.0 MB
PCA.png	112.7 KB
barplotNull.png	61.5 KB
barplotTotal.png	60.9 KB
cluster.png	37.4 KB
countsBoxplots.png	95.5 KB
densplot.png	143.8 KB
diagSizeFactorsHist.png	125.1 KB
diagSizeFactorsTC.png	99.9 KB
dispersionsPlot.png	354.4 KB
majSeq.png	91.7 KB
pairwiseScatter.png	249.5 KB
rawpHist.png	44.7 KB
volcanoPlot.png	182.8 KB

SARTools results

- Tables

Galaxy Tool SARTools_DESeq2

Run at 14/04/2023 14:28:59

Tables available for downloading

Output File Name (click to view)	Size
siMitfvssiLuc.complete.txt	6.1 MB
siMitfvssiLuc.down.txt	521.9 KB
siMitfvssiLuc.up.txt	587.0 KB

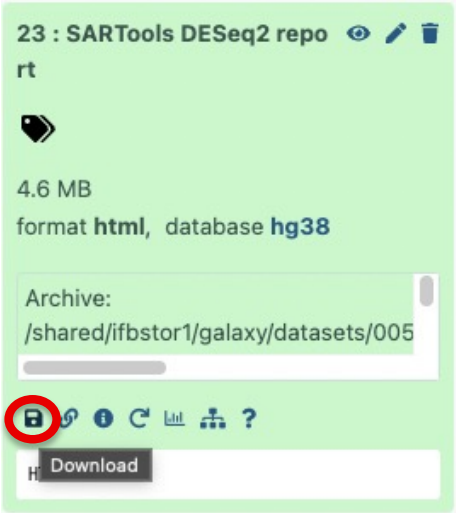
→ All genes

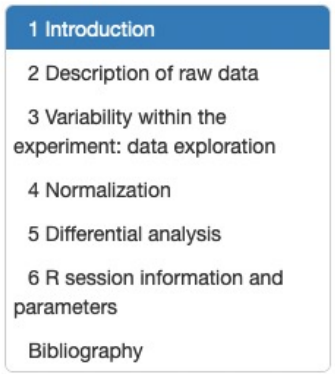
→ Only significant down-regulated genes (i.e. less expressed in siMitf than in siLuc)

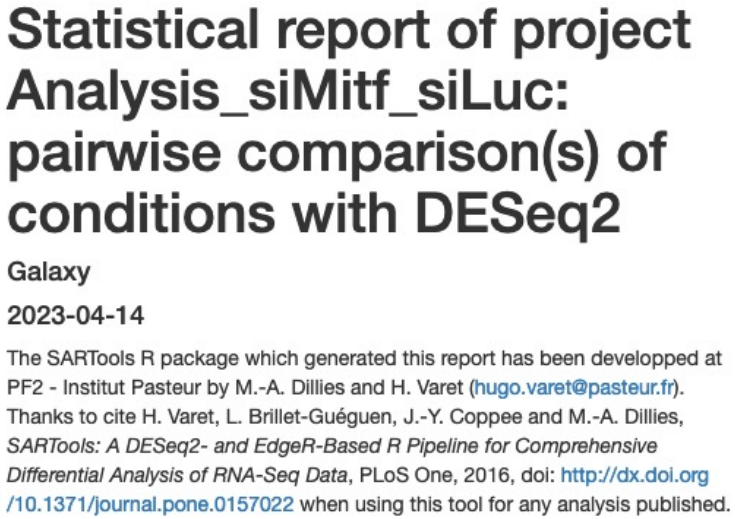
→ Only significant up-regulated genes (i.e. more expressed in siMitf than in siLuc)

SARTools results

- Report
  - Provides details about the methodology, the different steps and results
  - Displays all figures produced + a summary of differential analysis results







Statistical report of project Analysis_siMitf_siLuc: pairwise comparison(s) of conditions with DESeq2

Galaxy

2023-04-14

The SARTools R package which generated this report has been developped at PF2 - Institut Pasteur by M.-A. Dillies and H. Varet (hugo.varet@pasteur.fr). Thanks to cite H. Varet, L. Brillet-Guéguen, J.-Y. Coppee and M.-A. Dillies, *SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data*, PLoS One, 2016, doi: http://dx.doi.org/10.1371/journal.pone.0157022 when using this tool for any analysis published.

SARTools results

- Report
  - Description of raw data

Table 1: Data files and associated biological conditions.

label	files	group
siLuc2	dataset_2432739.dat	siLuc
siLuc3	dataset_2432741.dat	siLuc
siMitf3	dataset_2432743.dat	siMitf
siMitf4	dataset_2432745.dat	siMitf

For this project, there are 61487 features in the count data table.

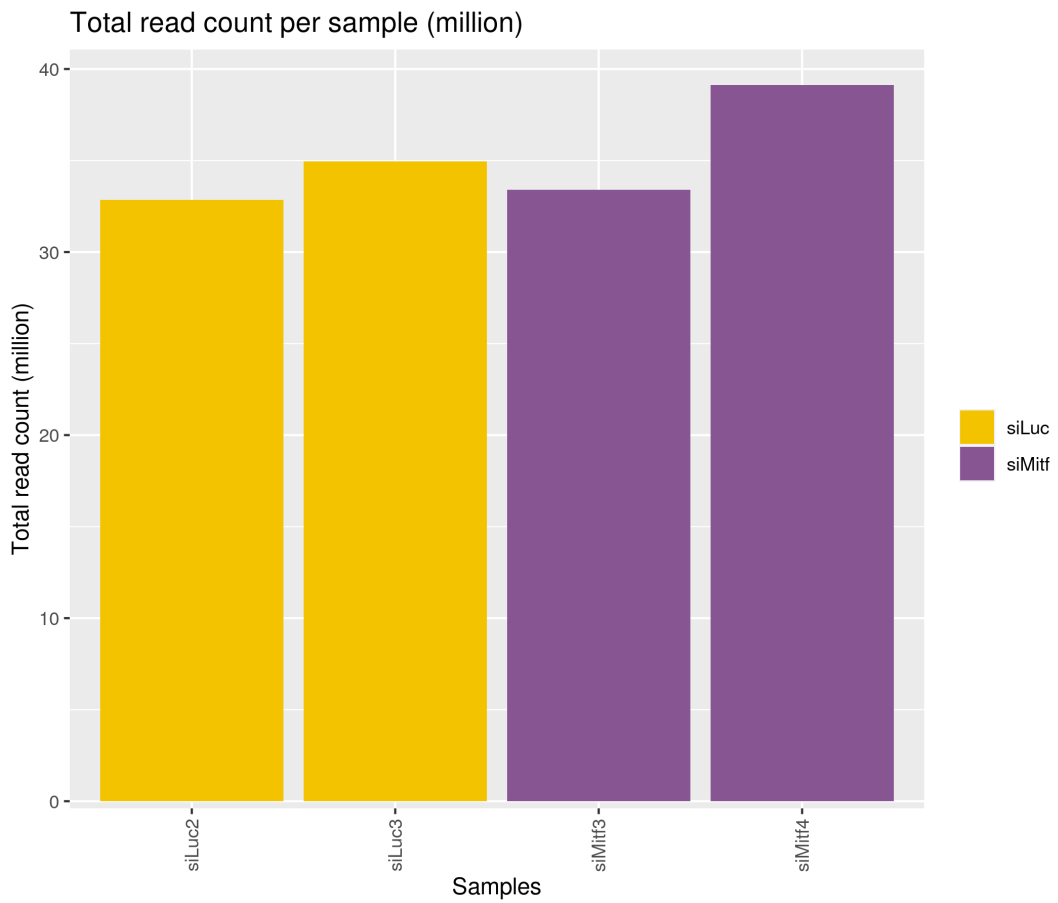
Table 2: Partial view of the count data table.

	siLuc2	siLuc3	siMitf3	siMitf4
ENSG00000000003	1271	1358	1282	1366
ENSG00000000005	0	0	0	0
ENSG00000000419	3700	3960	3760	3910
ENSG00000000457	655	647	636	755
ENSG00000000460	2442	2764	1449	1731
ENSG00000000938	0	0	0	0

Table 3: Summary of the raw counts.

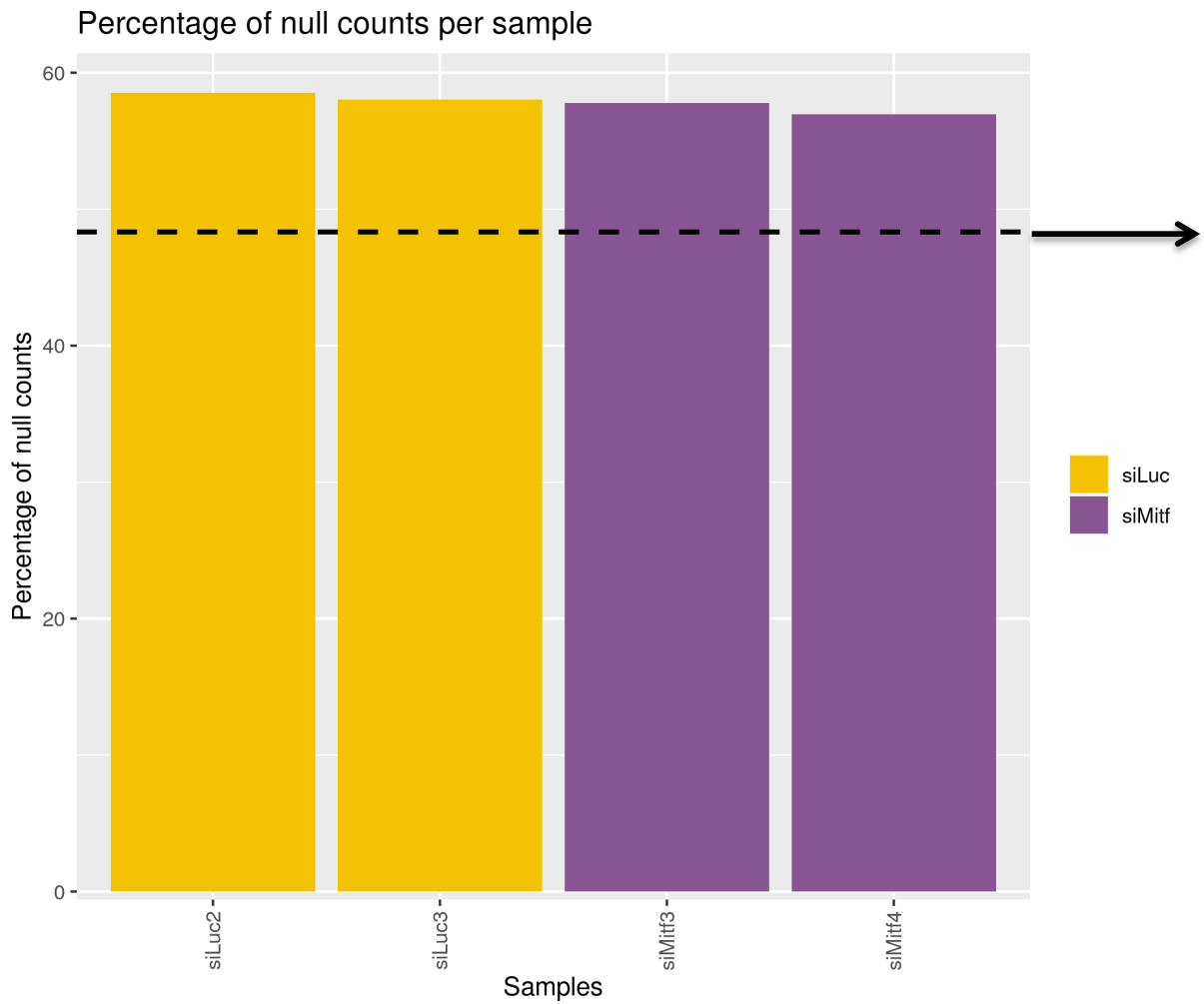
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
siLuc2	0	0	0	534	13	283474
siLuc3	0	0	0	569	14	276224
siMitf3	0	0	0	543	13	335630
siMitf4	0	0	0	636	15	380273

Total read count per sample



Different between samples, as expected → normalization needed
More difficult when major differences between samples

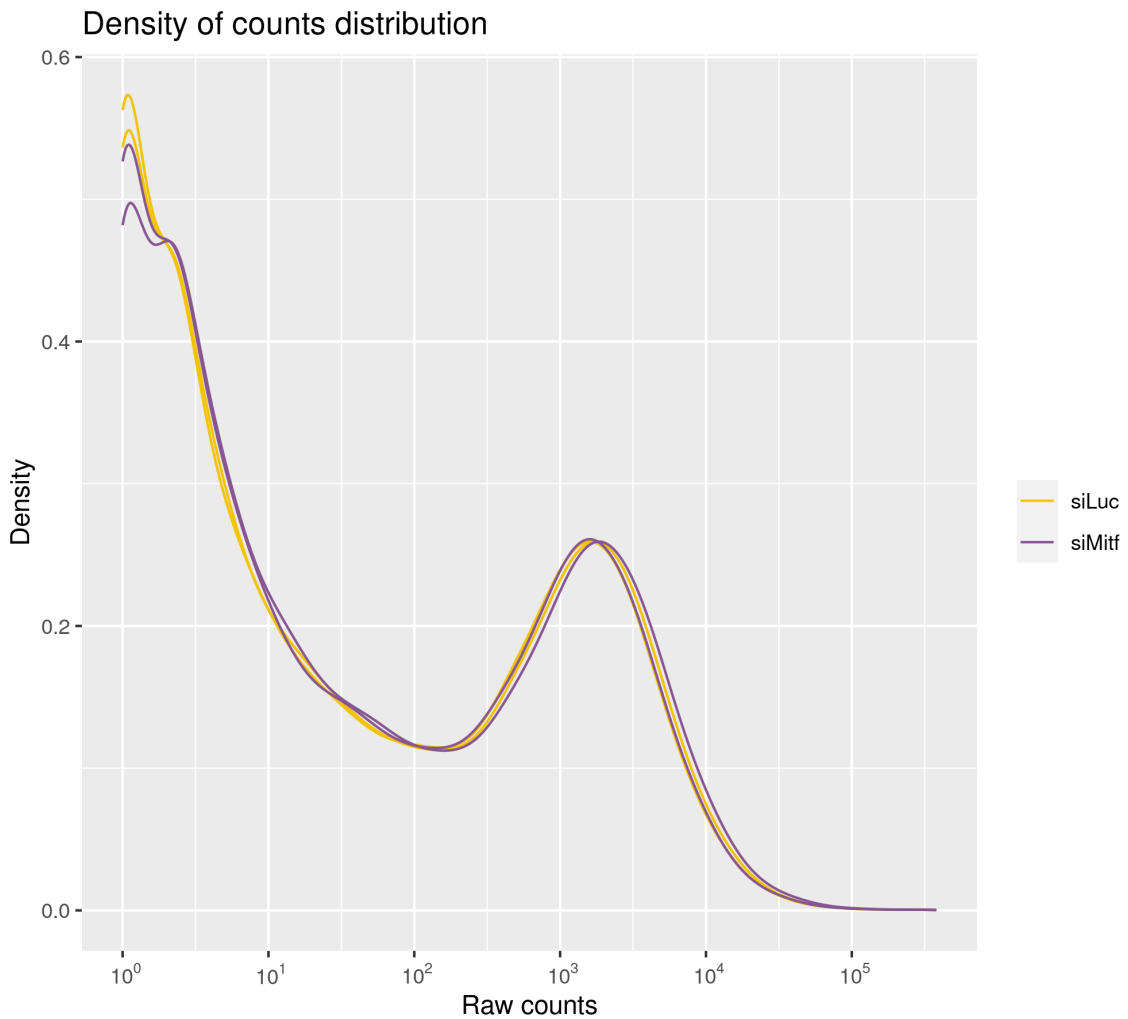
Proportion of null counts per sample



Proportion of genes with null counts in all samples
→ Such genes are left in the data but not taken into account in the analysis (results=NA in the result file)

We expect this proportion to be similar between samples

Density distribution of read counts



We expect replicates to have similar distributions

Proportion of reads from most expressed genes

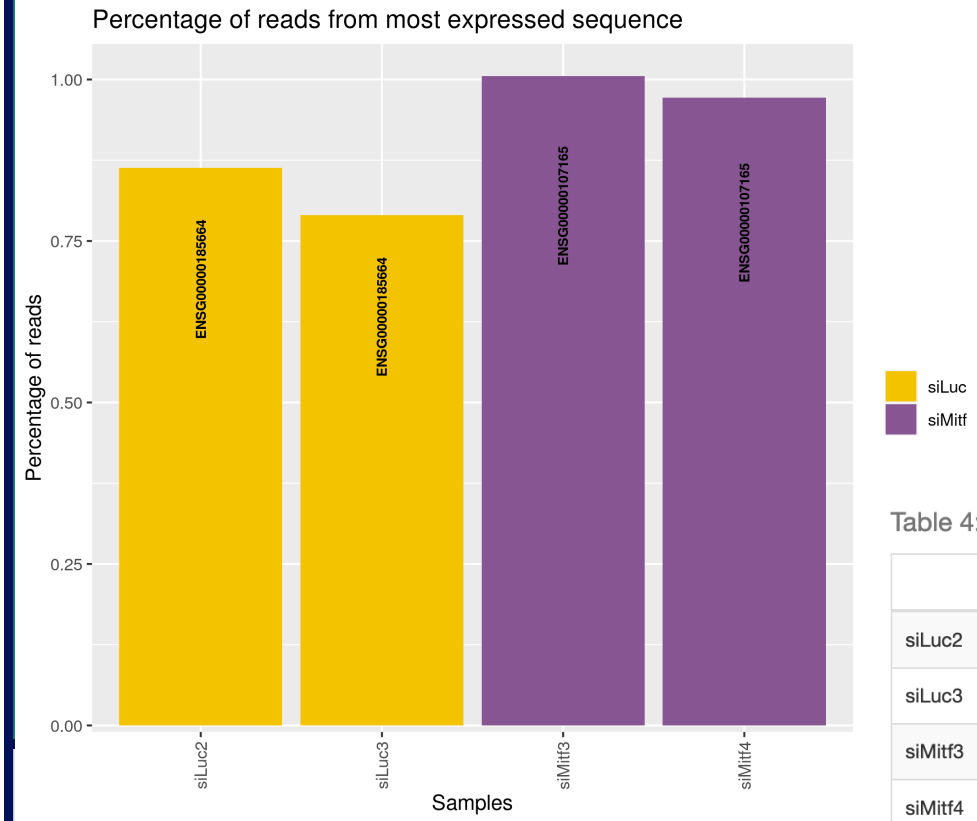


Table 4: Percentage of reads associated with the sequences having the highest counts.

	ENSG00000185664	ENSG00000198886	ENSG00000198804	ENSG00000210082	ENSG00000107165
siLuc2	0.86	0.78	0.76	0.75	0.68
siLuc3	0.79	0.72	0.71	0.71	0.62
siMitf3	0.71	0.84	0.93	0.69	1.00
siMitf4	0.73	0.91	0.91	0.69	0.97

We expect these high count features to be the same across replicates

Analysis of RNA-seq data

Quality analysis

↓

Mapping

↓

Gene expression quantification

↓

Exploratory data analysis

↓

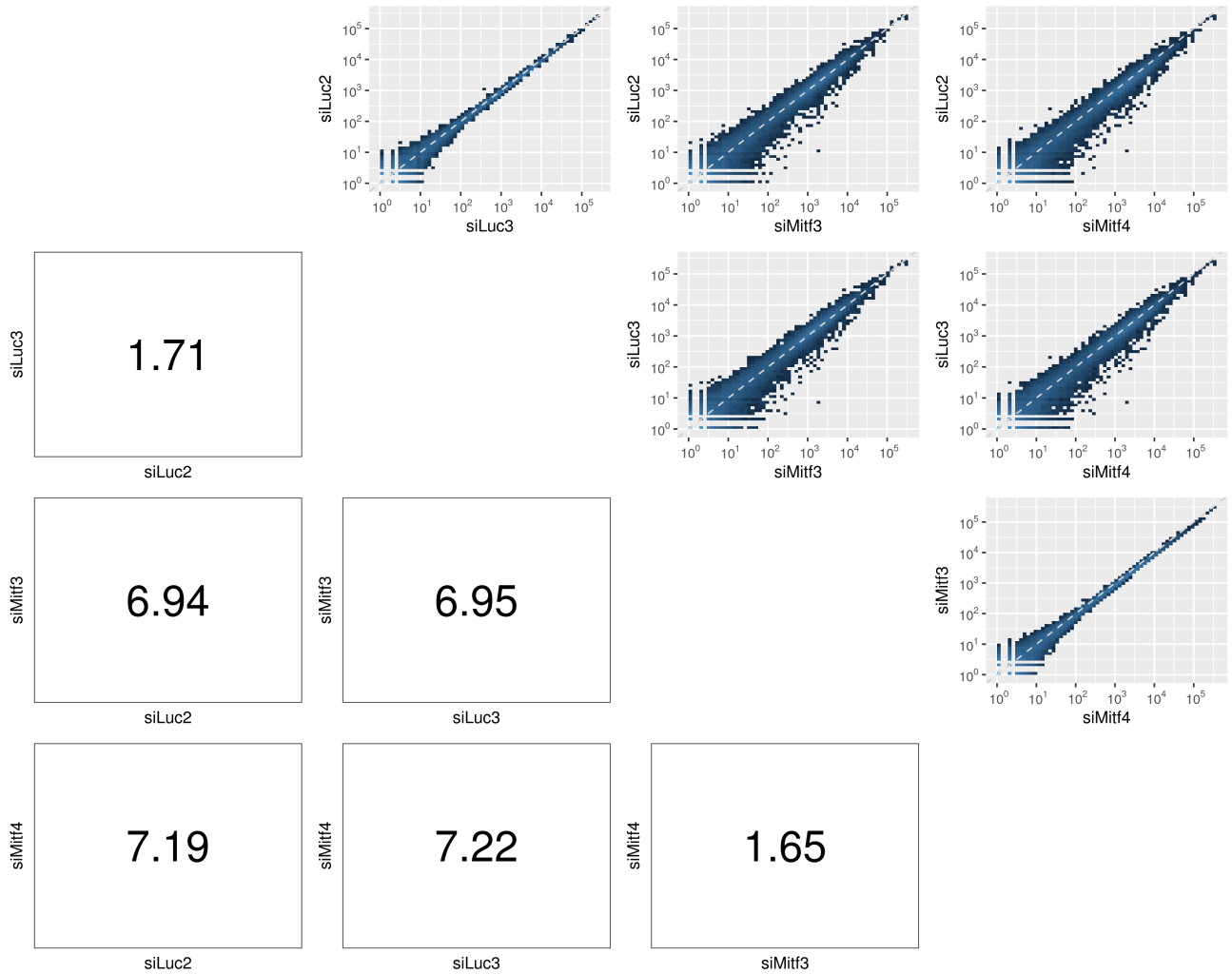
Normalization and statistical analysis

Exploration and visualization of data

- Essential step before any analysis
- Allows data quality assessment and control
- Eventually leads to remove data with insufficient quality

Pairwise comparison of samples

Pairwise scatter plot



SERE values

We expect replicates to have correlated read counts

SERE coefficient

- Simple Error Ratio Estimate (Schulze et al. BMC Genomics 2012;13:524)

$$\text{SERE} = \frac{\text{Observed standard deviation between two samples}}{\text{Value that would be expected from an ideal experiment}}$$

  - SERE = 0 ➔ sample duplication
  - SERE = 1 ➔ technical replication
  - SERE > 1 ➔ biological variation
  - SERE ↑ ➔ Similarity ↓

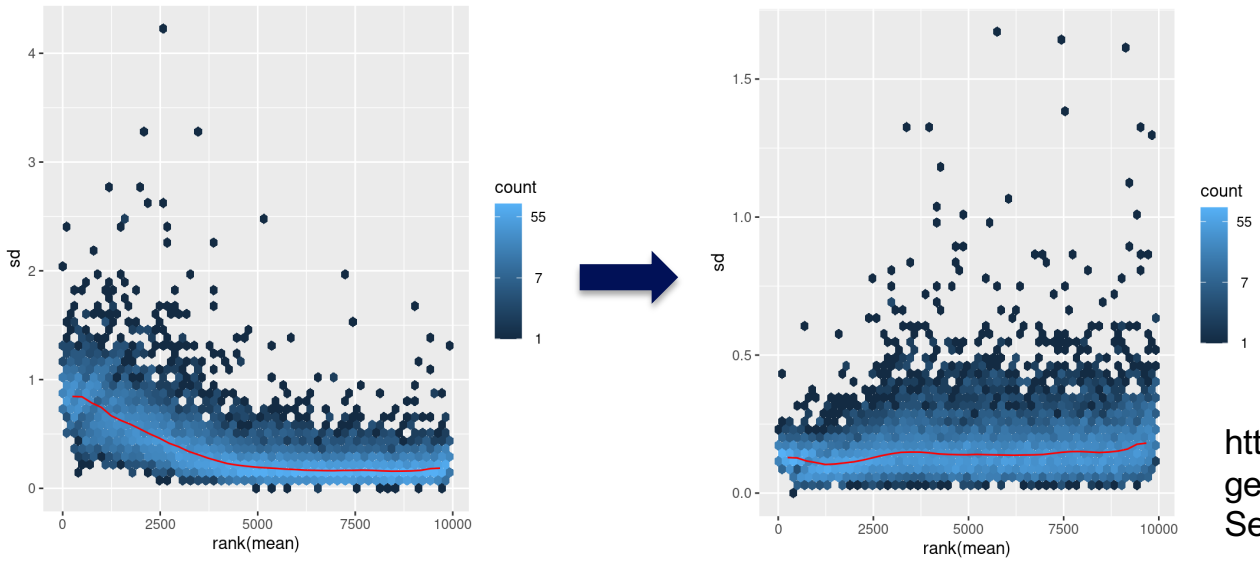
Data transformation

- Many methods for exploratory data analysis (clustering, PCA) work best for data that generally have the same range of variance at different ranges of mean values
- However this is not the case for RNA-seq data
- To avoid that results are dominated by a few highly variable genes

→ Remove the dependence of the variance on the mean :

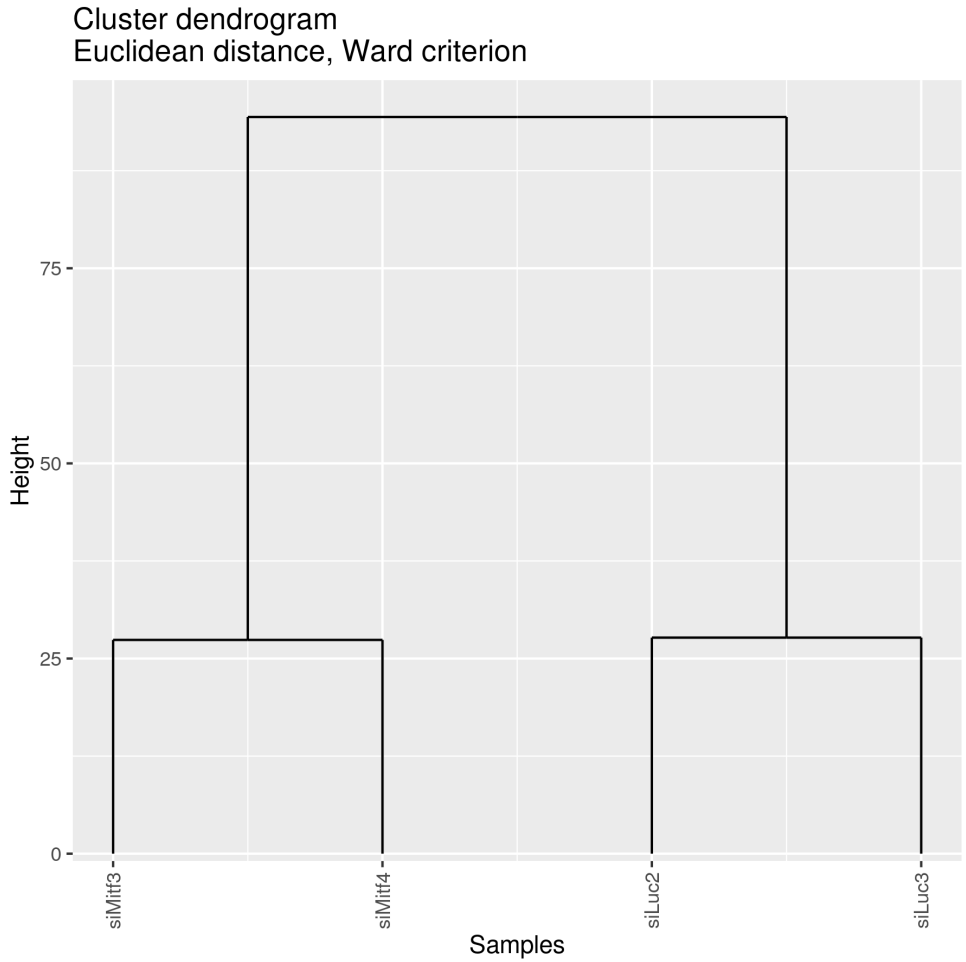
VST (variance-stabilizing transformation ; Anders et al. Genome Biology 2010, 11:106)

→ Only for exploratory data analysis !



https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Samples clustering

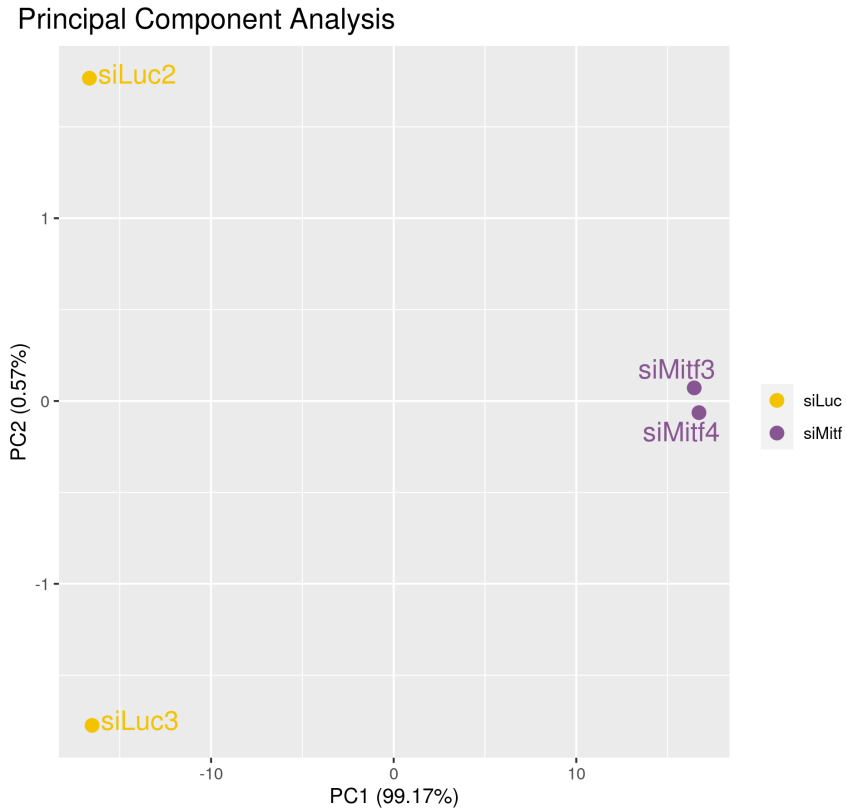


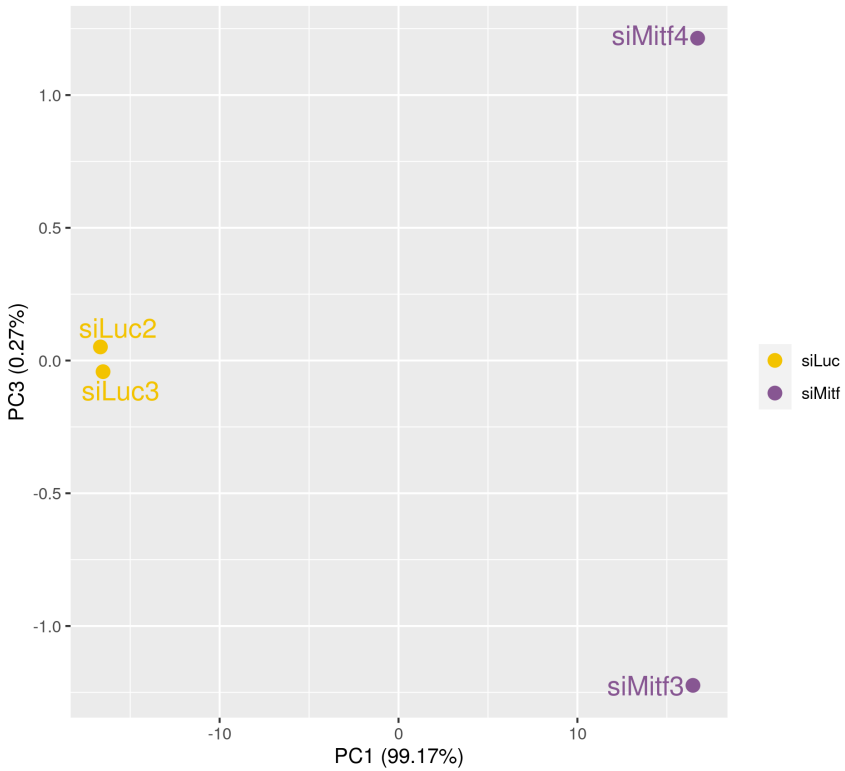
Obtained from VST-transformed data

We expect this dendrogram to group replicates and separate biological conditions

Principal Component Analysis

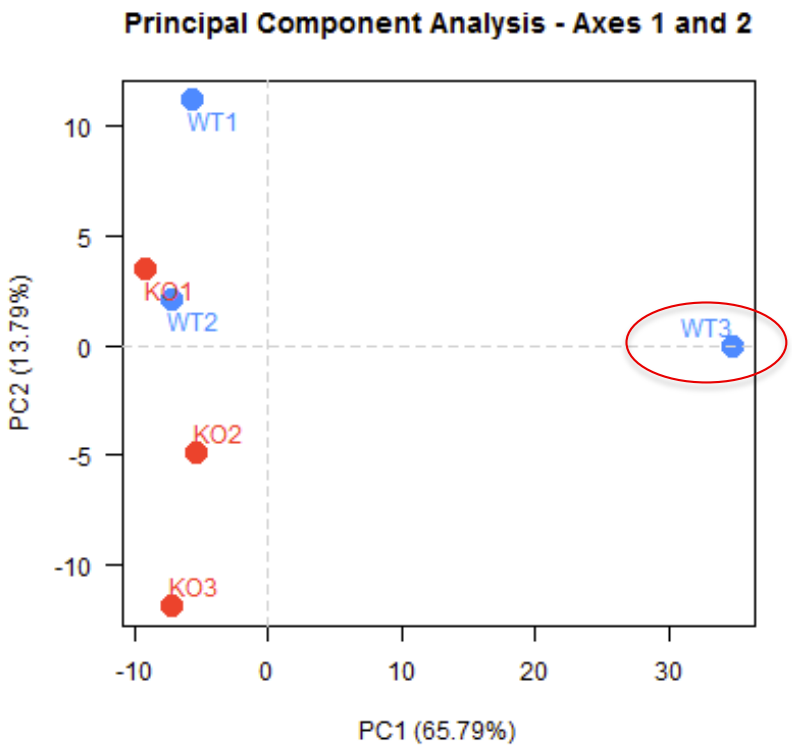
Obtained from VST-transformed data



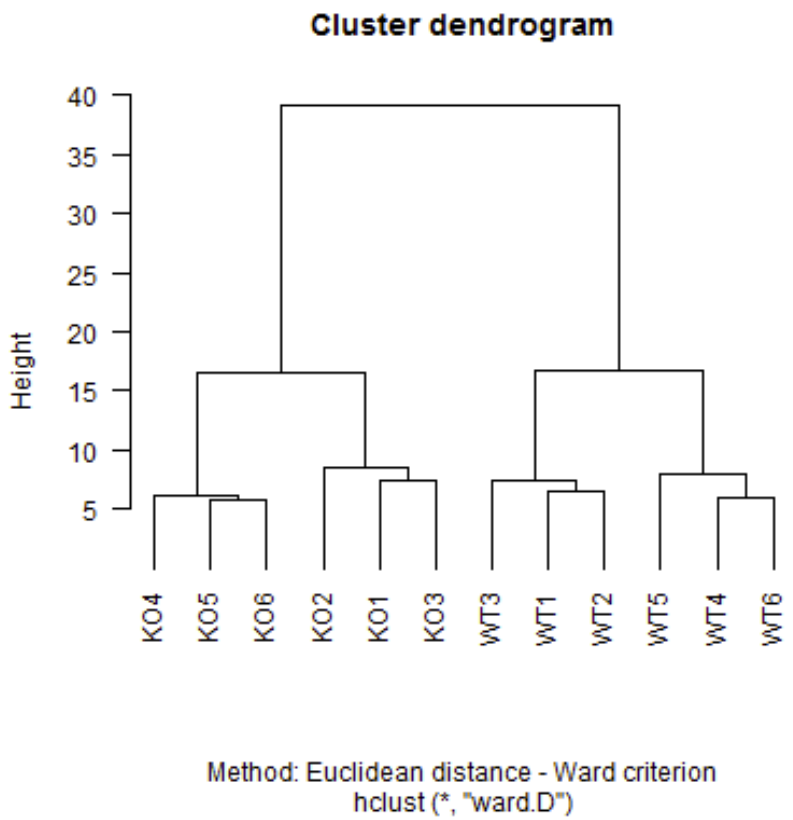


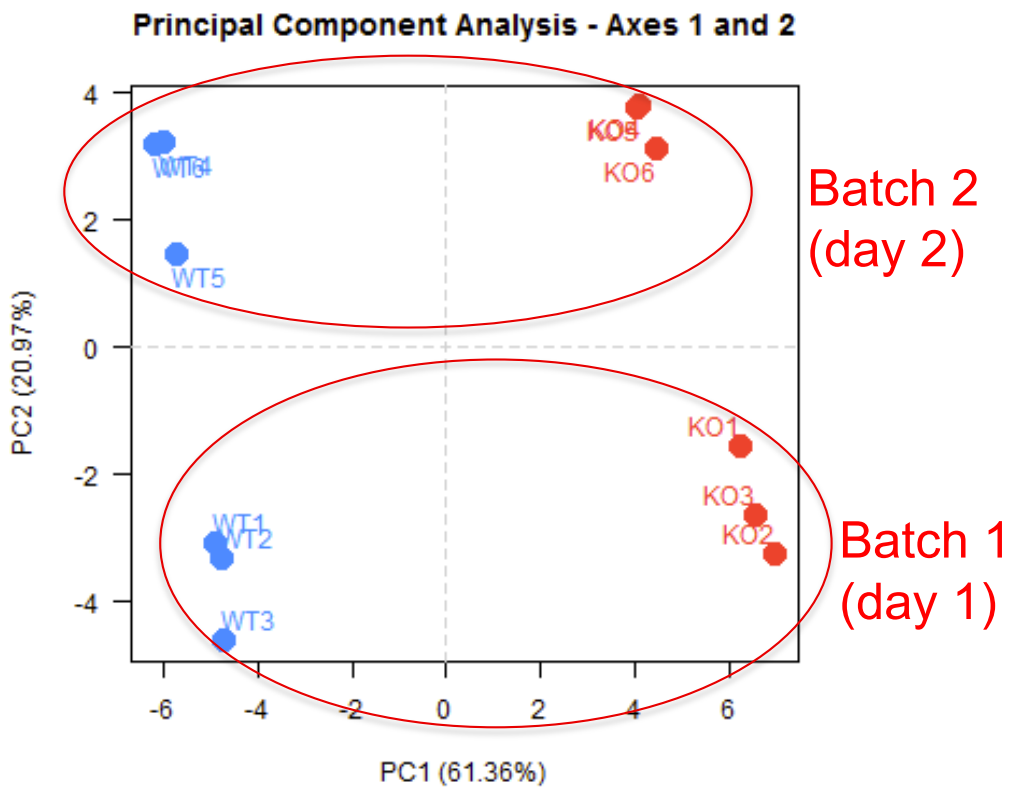
The first principal component is expected to separate samples from the different biological conditions
(i.e. corresponds to the main source of variance in the data)

Data exploration on another dataset : outlier sample



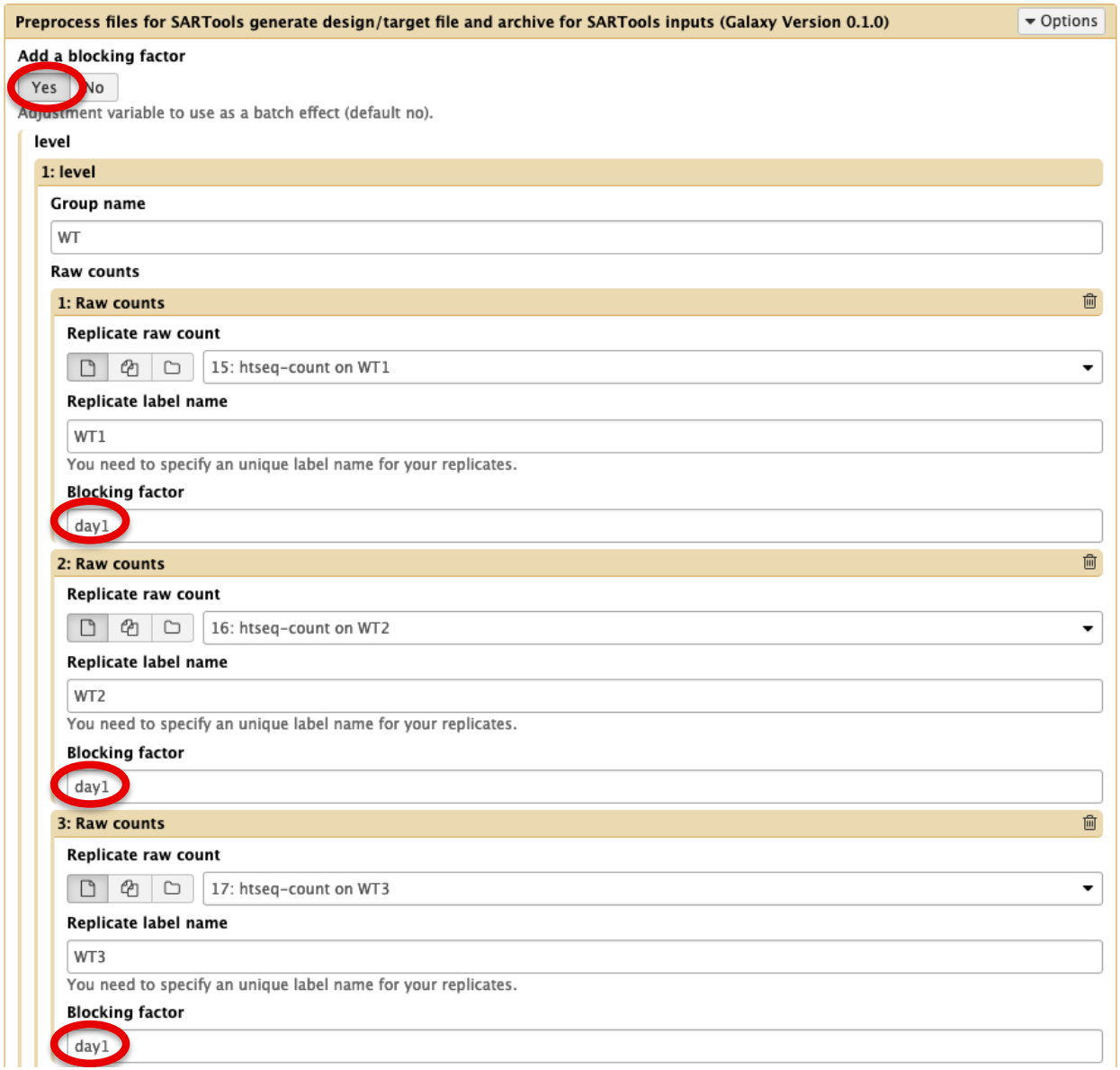
Data exploration on another dataset : batch effect



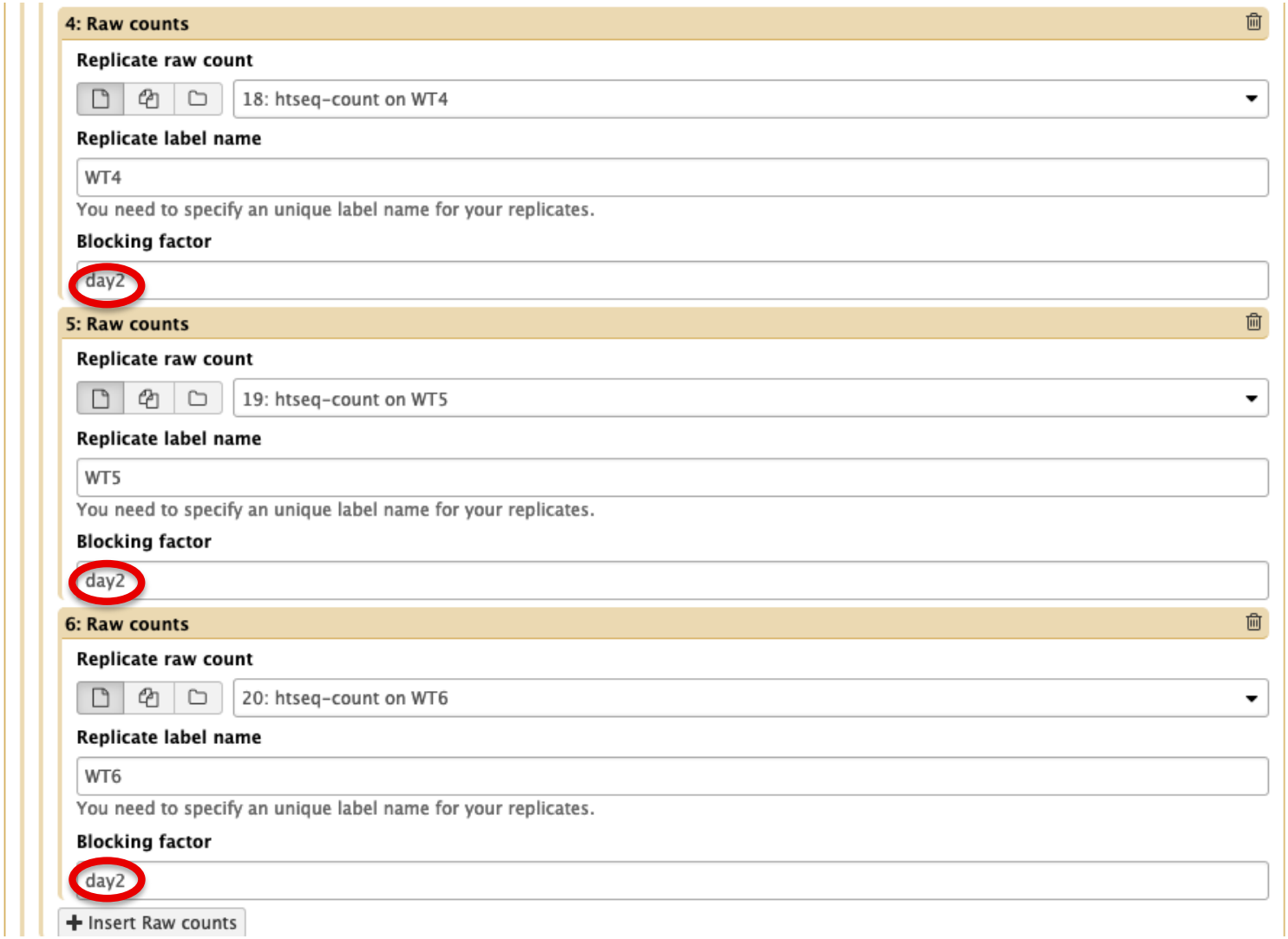


→ Take into account this batch effect in statistical analysis

Take into account batch effect in SARTools
1. Preprocess files for SARTools



Take into account batch effect in SARTools
1. Preprocess files for SARTools



Take into account batch effect in SARTools
1. Preprocess files for SARTools

2: level

Group name

KO

Raw counts

1: Raw counts

Replicate raw count

21: htseq-count on KO1

Replicate label name

KO1

You need to specify an unique label name for your replicates.

Blocking factor

day1

2: Raw counts

Replicate raw count

22: htseq-count on KO2

Replicate label name

KO2

You need to specify an unique label name for your replicates.

Blocking factor

day1

3: Raw counts

Replicate raw count

23: htseq-count on KO3

Replicate label name

KO3

You need to specify an unique label name for your replicates.

Blocking factor

day1

Take into account batch effect in SARTools
1. Preprocess files for SARTools

4: Raw counts

Replicate raw count

24: htseq-count on KO4

Replicate label name

KO4

You need to specify an unique label name for your replicates.

Blocking factor

day2

5: Raw counts

Replicate raw count

25: htseq-count on KO5

Replicate label name

KO5

You need to specify an unique label name for your replicates.

Blocking factor

day2

6: Raw counts

Replicate raw count

26: htseq-count on KO6

Replicate label name

KO6

You need to specify an unique label name for your replicates.

Blocking factor

day2

+ Insert Raw counts

+ Insert level

Execute

Take into account batch effect in SARTools
1. Preprocess files for SARTools

- Design file :

1	2		3	4
label	files		group	batch
WT1	dataset_	dat	WT	day1
WT2	dataset_	dat	WT	day1
WT3	dataset_	dat	WT	day1
WT4	dataset_	dat	WT	day2
WT5	dataset_	dat	WT	day2
WT6	dataset_	dat	WT	day2
KO1	dataset_	dat	KO	day1
KO2	dataset_	dat	KO	day1
KO3	dataset_	dat	KO	day1
KO4	dataset_	dat	KO	day2
KO5	dataset_	dat	KO	day2
KO6	dataset_	dat	KO	day2

Take into account batch effect in SARTools
2. SARTools DESeq2

Name of the project used for the report

Example_batch

(-P, --projectName)

Name of the report author

keime

(-A, --author)

Design / target file

27: design file for SARTools (on data 26, data 25, and others)

(-t, --targetFile) See the help section below for details on the required format.

Zip file containing raw counts files

28: counts files for SARTools (on data 26, data 25, and others)

(-r, --rawDir) See the help section below for details on the required format.

Have you a header in your count files ?

No

The tool needs no header in the input files, so if there is an header, select yes, and it removes it during the processing.

Names of the features to be removed

alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual

(-F, --featuresToRemove) Separate the features with a comma, no space allowed. More than once can be specified. Specific HTSeq-count information and rRNA for example. Default are 'alignment_not_unique,ambiguous,no_feature,not_aligned,too_low_aQual'.

Factor of interest

group

(-v, --varInt) Biological condition in the target file. Default is 'group'.

Reference biological condition

WT

(-c, --condRef) Reference biological condition used to compute fold-changes, must be one of the levels of 'Factor of interest'.

Advanced Parameters

Show

Add a blocking factor

Yes No

(-B, --batch) Adjustment variable to use as a batch effect. Default: unchecked if no batch effect needs to be taken into account.

Blocking factor value

batch

Must be a column of the target file

Analysis of RNA-seq data

Quality analysis

↓

Mapping

↓

Gene expression quantification

↓

Exploratory data analysis

↓

Normalization and statistical analysis

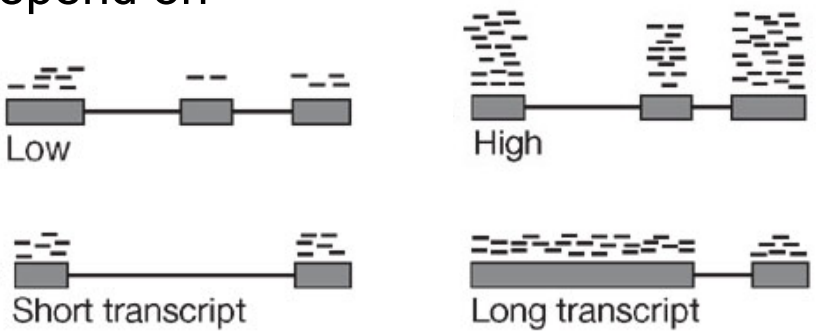
Normalization : why ?

- To compare RNA-seq libraries
  - with different sizes, eg :

Sample name	Total number of reads
siLuc2	43,672,265
siLuc3	46,565,834
siMitf3	43,985,979
siMitf4	51,348,313

- To compare the expression level of several genes within a library

  Indeed read counts depend on
  - Expression level
  - Gene length
  - Library size



Different normalization methods

- Based on distribution adjustment
  - Total read count
    - Motivation

      Higher library size ➔ higher counts
    - Method

      Divide counts by total number of reads
  - Upper quartile (Bullard et al. BMC Bioinformatics 2010;11,94), Median
    - Motivation

      Total read count is strongly dependent on a few highly expressed transcripts
    - Method

      Divide counts by the upper quartile/median of the counts different from 0
  - Quantile (Bolstad et al. Bioinformatics 2003; 19:185–93)
    - Assumption

      Read counts have identical distribution across libraries
    - Method

      Count distributions are matched between libraries

Different normalization methods

- Take into account gene/transcript length
  - RPKM (Mortazavi et al. Nat Methods 2008;5:621–8), FPKM
  - **R**eads (**F**ragments) per **K**ilobase per **M**illion mapped reads
  - Assumption
    - Read counts =f(expression level, gene length, library size)
  - Method
    - Divide counts by gene length (kb) and total number of reads (million)
  - Allows to compare expression levels between genes

Different normalization methods

- Based on the “effective library size” concept
  - Assumption
    - Most genes are not differentially expressed
  - 2 methods
    - Trimmed Mean of M values (Robinson et al. Genome Biol. 2010;11:R25)
    - DESeq normalization (Anders et al. Genome Biol. 2010;11:R106)

Which normalization method to choose ?

- Comparison on 4 real and 1 simulated dataset
- Summary of comparison results

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	–	+	+	–	–
UQ	++	++	+	++	–
Med	++	++	–	++	–
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	–	+	++	–
RPKM	–	+	+	–	–

- : the method provided unsatisfactory results for the given criterion
+ : satisfactory results
++ : very satisfactory results

(Dillies et al. Brief. Bioinformatics 2013 Nov;14(6):671-83)

DESeq normalization method

	lib1	lib2	lib3	…	lib **j**	lib **n**
gene1	468	475	501			
gene2	45	56	76			
gene3	2576	560	578			
gene4	1678	1798	1867			
…						
gene **i**					x_{ij}	

n : number of samples to compare

xij : number of reads for gene i in sample j

(Anders et al. Genome Biol. 2010;11:R106)

DESeq normalization method

	lib1	lib2	lib3	…	lib **j**	lib **n**
gene1	468	475	501			
gene2	45	56	76			
gene3	2576	560	578			
gene4	1678	1798	1867			
…						
gene **i**					x_{ij}	

n : number of samples to compare

xij : number of reads for gene i in sample j

Normalization factor for library j :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{\nu=1}^{n} x_{i\nu})^{1/n}}$$

➔ Each value is divided by the geometric mean of its row

➔ Normalization factor = median of all these ratios

DESeq normalization method

	lib1	lib2	lib3	mean
gene1	468	475	501	m1=481.1263
gene2	45	56	76	m2=57.64187
gene3	2576	560	578	m3=941.2115
gene4	1678	1798	1867	m4=1779.271

Normalization factor for library j :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{\nu=1}^{n} x_{i\nu})^{1/n}}$$

DESeq normalization method

	lib1	lib2	lib3	mean
gene1	468 / m1	475 / m1	501 / m1	m1=481.1263
gene2	45 / m2	56 / m2	76 / m2	m2=57.64187
gene3	2576 / m3	560 / m3	578 / m3	m3=941.2115
gene4	1678 / m4	1798 / m4	1867 / m4	m4=1779.271

Normalization factor for library j :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{\nu=1}^{n} x_{i\nu})^{1/n}}$$

➔ Underlying idea : non-differentially expressed genes should have similar read counts across samples leading to a ratio of 1

DESeq normalization method

	lib1	lib2	lib3	mean
gene1	468 / m1	475 / m1	501 / m1	m1=481.1263
gene2	45 / m2	56 / m2	76 / m2	m2=57.64187
gene3	2576 / m3	560 / m3	578 / m3	m3=941.2115
gene4	1678 / m4	1798 / m4	1867 / m4	m4=1779.271
median	0.9577858	0.9793598	1.0452989	

normalization factors

Normalization factor for library j :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{\nu=1}^{n} x_{i\nu})^{1/n}}$$

→ Median of these ratios for a library → estimate of the correction factor that should be applied to all read counts of this library

→ Normalized read counts = raw read counts / normalization factor

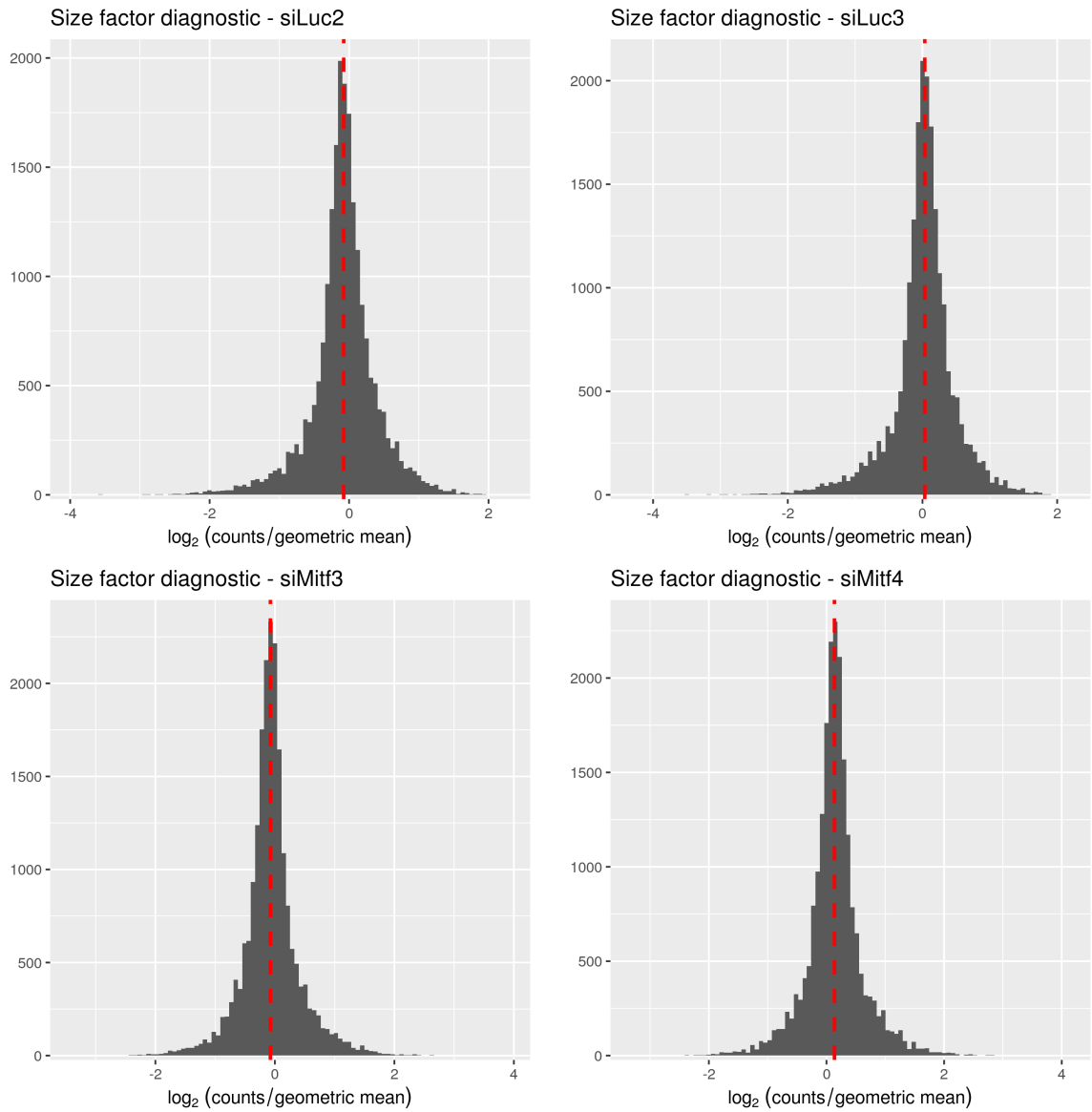
DESeq normalization

- Normalization factors for Mitf dataset :

Table 5: Normalization factors.

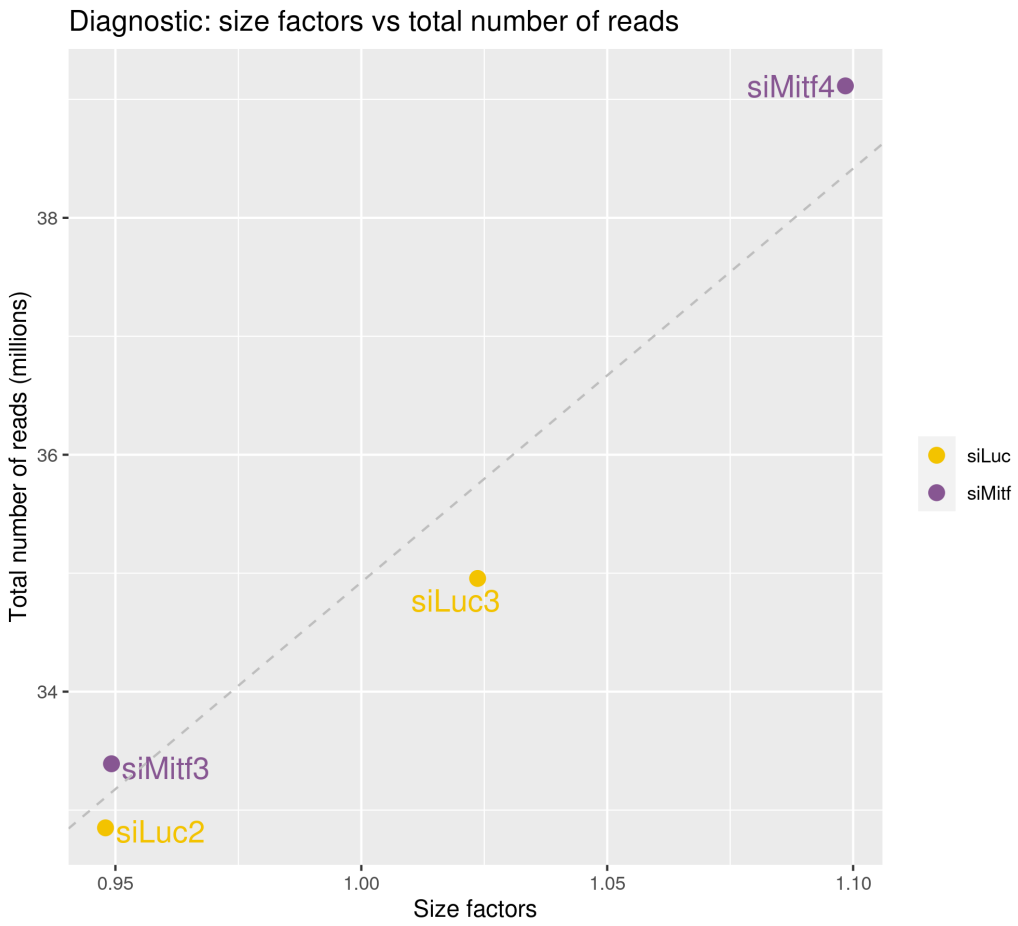
	siLuc2	siLuc3	siMitf3	siMitf4
Size factor	0.95	1.02	0.95	1.1

Diagnostic plot for the estimation of normalization factors



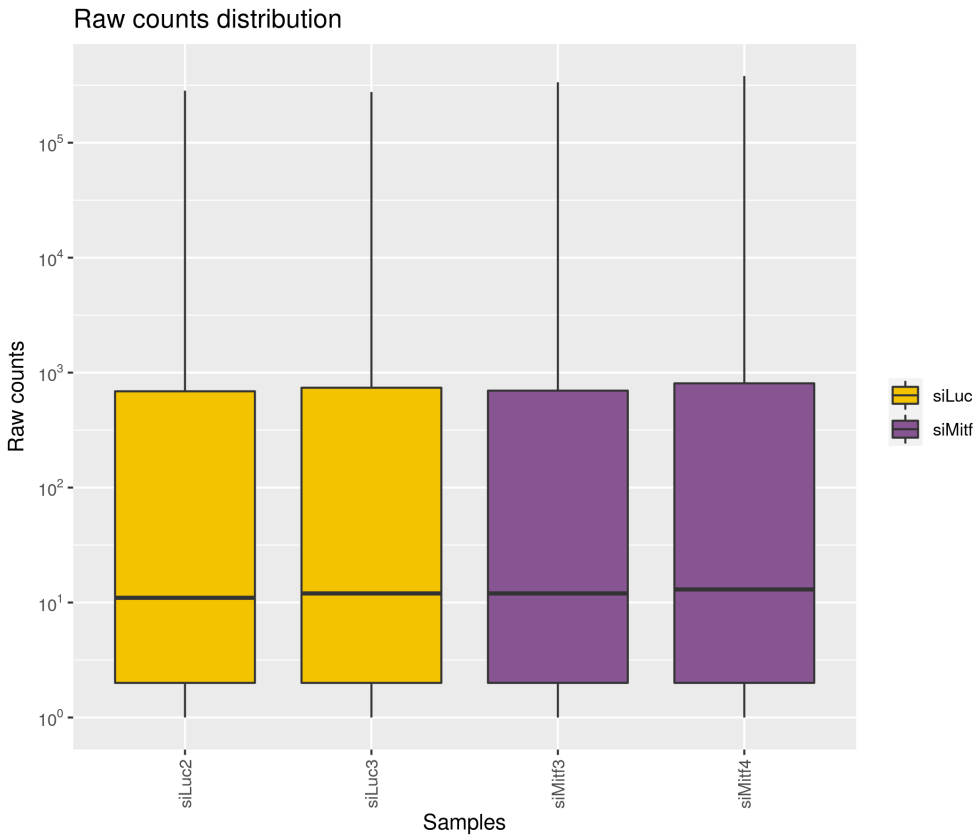
This histogram should be unimodal,
with a clear peak at the value of the size factor (represented in red)

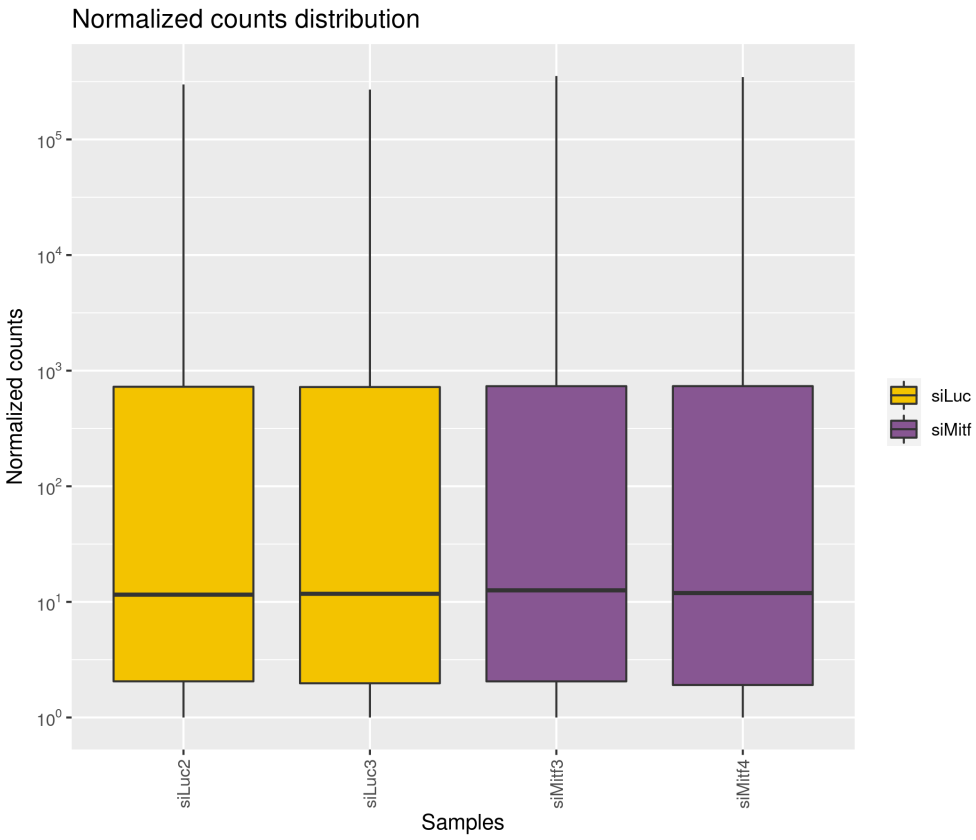
Total number of reads vs size factors



Normalization by total number of reads and DESeq2 size factors is not exactly the same, but very close for this dataset

Boxplots of raw and normalized read counts





We expect normalization to stabilize distributions across samples

Analysis of RNA-seq data

Quality analysis

↓

Mapping

↓

Gene expression quantification

↓

Exploratory data analysis

↓

Normalization and **statistical analysis**

Search for significantly differentially expressed genes

- What is significant differential expression ?
  - The observed difference between conditions is statistically significant i.e. greater than expected just due to random variation

- Microarray vs RNA-seq
  - Microarray
    Fluorescence proportional to expression ➔ continuous data
  - RNA-seq
    Number of reads assigned to a feature (gene, transcript) proportional to expression ➔ count data

- Here we focus on count-based measures of gene expression

Search for significantly differentially expressed genes

- Use only a fold-change ranking ?
  - Do not take variability into account
  - Do not take level of expression into account
  - No control of the false positive rate
- Hypothesis testing
  - For each gene
    - H0 : No gene expression difference between the compared conditions
    - H1 : There is a gene expression difference between the compared conditions
- Steps
  - Choose a statistic
  - Define a decision rule
    - Define a threshold below which we will reject H0

Statistic to search for significantly differentially expressed genes

- Sequencing a library = randomly and independently choose N sequences from the library
  → read counts ~ multinomial distribution

- High number of reads, probability of a read assigned to a given gene small → Poisson approximation
  - Distribution of counts across technical replicates for the majority of genes fit well to a Poisson distribution

    Marioni et al. Genome Research 2008;18(9):1509-17

    Bullard et al. BMC Bioinformatics 2010;11,94

→ Technical replicates ~ Poisson distribution

Statistic to search for significantly differentially expressed genes

- But Poisson distribution : variance = mean

  ➔ Across biological replicates variance > mean for many genes (Anders et al. Genome Biology 2010;11:R106) : overdispersion

  ➔ Negative binomial distribution : a good alternative to Poisson in the case of overdispersion

→ Biological replicates ~ Negative binomial distribution

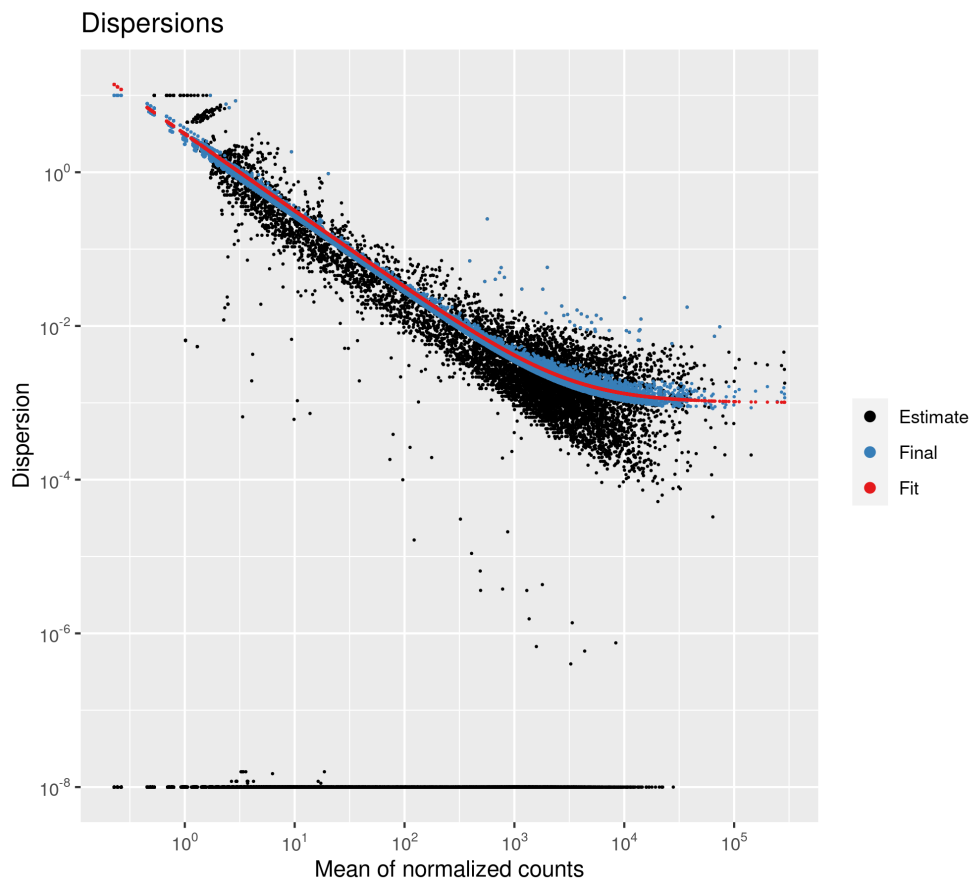
- How to estimate the overdispersion parameter ?
  - Very few replicates ➔ challenging issue
  - DESeq2 (Love et al. Genome Biol. 2014;15:550)

    Shares information across genes to improve the estimation of dispersion

    Assumes that genes of similar average expression strength have similar dispersion

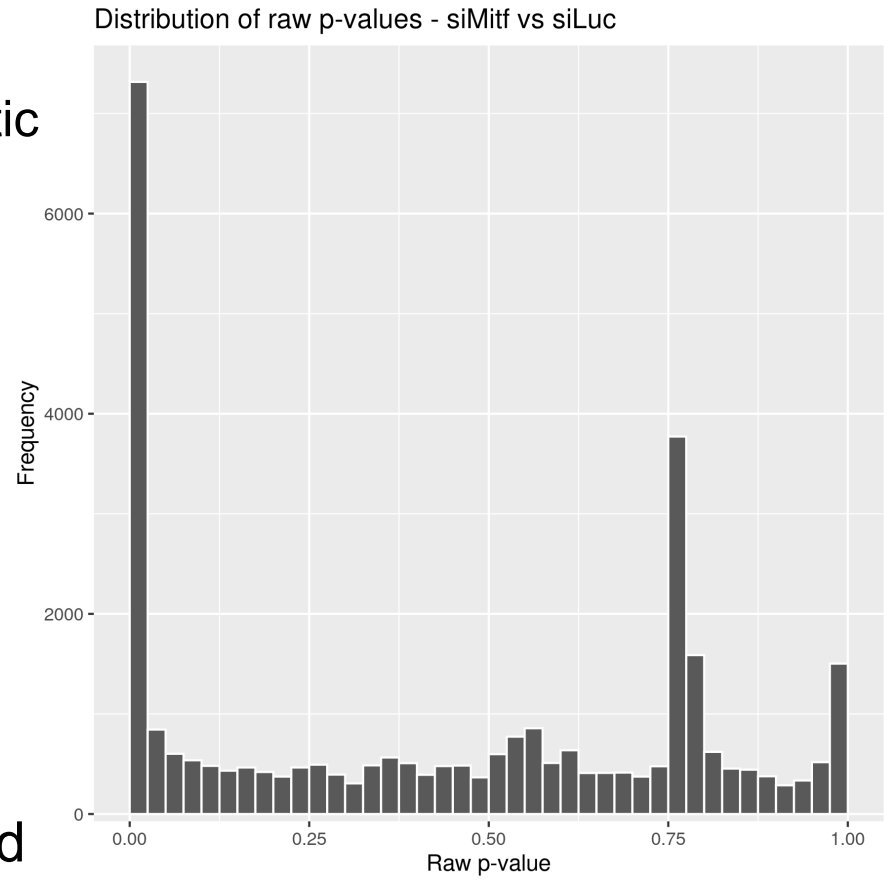
Dispersion plot

- **Black** : gene dispersion values (calculated using only the observed counts)
- **Red** : curve fitted to black dots to capture the overall trend of dispersion-mean dependence
- The red curve is used as a prior mean for a second estimation round, which results in final **blue** values (used during the test)
- Dispersions outliers (**blue**) → for these genes the statistical test is based on the empirical variance to be more conservative



Definition of a decision rule

- p-value
  - Probability of obtaining a statistic at least as extreme as the one that was actually observed, assuming that H0 is true



- Reject H0 if p-value < threshold
  - Common threshold = 0.05
  - ➔ the observed result would be highly unlikely under H0

But be careful : you perform multiple testing !

Multiple testing problem

- To identify significantly differentially expressed genes
  → as many tests as the number of genes (G)
- With a type I error α for each gene
  - we expect to find $G\alpha$ false positives
  - i.e. $G\alpha$ genes declared to be differentially expressed even if there are not
  - e.g. G=30,000 genes α=0.05 → we expect to find 1,500 false positives

  → Important to control the false positive rate when we make a lot of tests
- 2 points of views
  - Individually consider the differentially expressed genes sorted according to a statistic
  - Consider a list of differentially expressed genes, in which we would like to control the false positive rate

  ➔ Use a multiple testing correction

Multiple testing correction methods

- Family-Wise Error Rate (FWER)
  - Probability to have at least one false positive
  - e.g. FWER = 0.05 ➔ 5% chances of having at least one false positive

- Bonferroni method
  - Bonferroni

    $p_{g\_adjusted} = \min(Gp_g, 1)$

    ➔ Each test is performed with a type I error α/G
  - Very conservative method (Ge et al. TEST 2003;12(1):1-77)

Multiple testing correction methods

- False Discovery Rate (FDR)
  - Expected proportion of false positives among genes declared as differentially expressed
  - e.g. FDR = 0.05 ➔ We expect to find 5% of false positives among genes declared as significantly differentially expressed

- Benjamini and Hochberg method

  (Journal of the R. Stat. Soc., Series B 57 (1): 125–133)
  - Calculation of adjusted p-values that allows to control the FDR

How many genes are significantly differentially expressed between siMitf and siLuc (FDR<0.05) ?

Significantly differentially expressed genes

- Number of significantly differentially expressed genes between siMitf and siLuc (FDR<0.05) :

Table 7: Number of up-, down- and total number of differentially expressed features for each comparison.

Test vs Ref	# down	# up	# total
siMitf vs siLuc	3282	3762	7044

→ 7044 significantly differentially expressed genes
  → 3282 genes significantly under-expressed in siMitf vs siLuc
  → 3762 genes significantly over-expressed in siMitf vs siLuc

Independant filtering

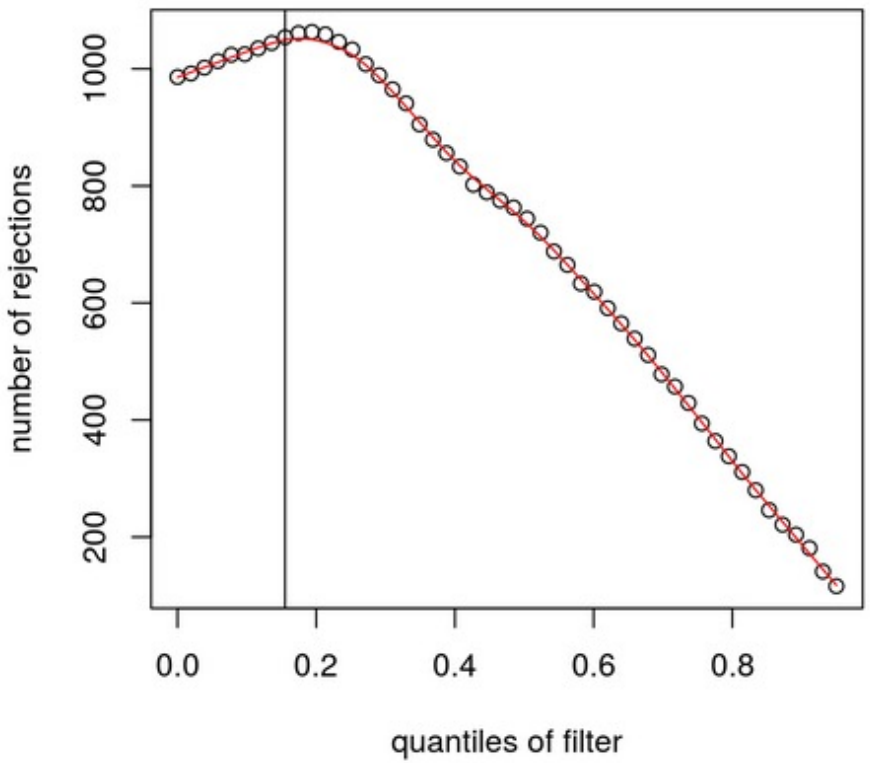
- Goal : filter out those tests from the procedure that have no, or little chance of being significant, without even looking at their test statistic
  - → Results in increased detection power at the same type I error

- Genes with very low counts are not likely to be significantly differentially expressed typically due to high dispersion
  - → DESeq2 defines a threshold on the mean of the normalized counts irrespective of the biological condition
  - → Independent because the information about the variables in the design formula is not used (Love et al. Genome Biol. 2014;15:550)

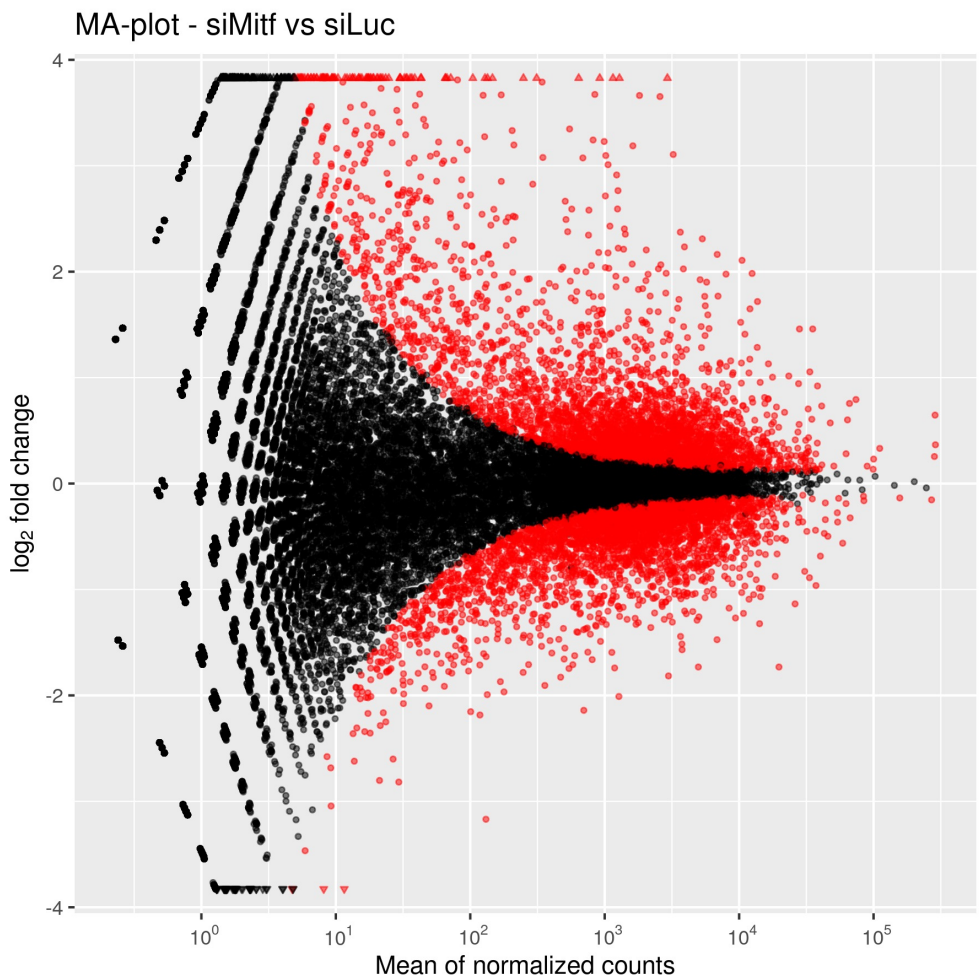
Genes discarded by the independent filtering
→ adjusted p-value = NA in the results table

Independant filtering

- Maximizes the number of rejections
  - adjusted p-value less than a significance level
- over the quantiles of a filter statistic
  - the mean of normalized counts
- Threshold chosen (vertical line)
  - Lowest quantile of the filter for which the number of rejections is within 1 residual standard deviation to the peak of a curve fit to the number of rejections over the filter quantiles:



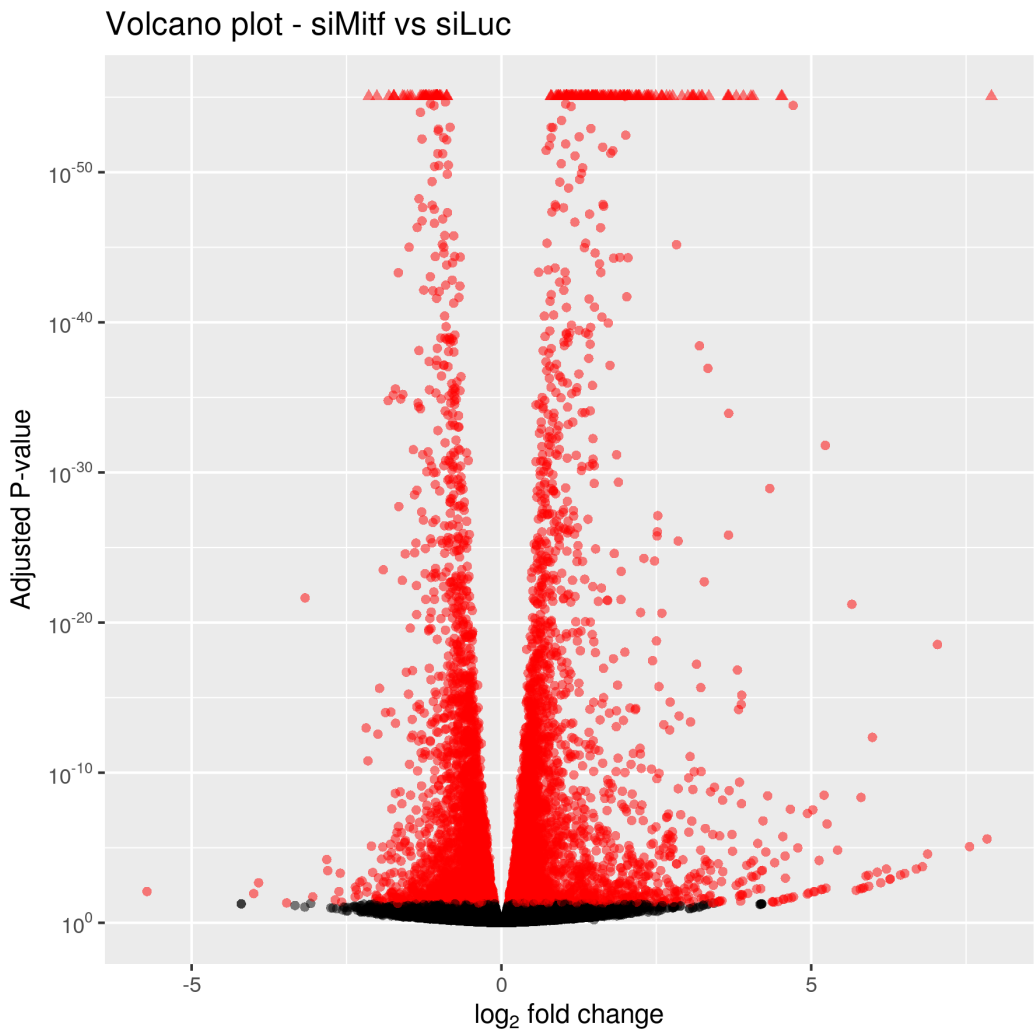
Visualization of significantly differentially expressed genes : MA-plot



Red dots : FDR < 0.05

Triangles : features having a too low/high $\log_2$FC to be displayed on the plot

Visualization of significantly differentially expressed genes : volcano plot



Red dots : FDR < 0.05

Differential analysis results

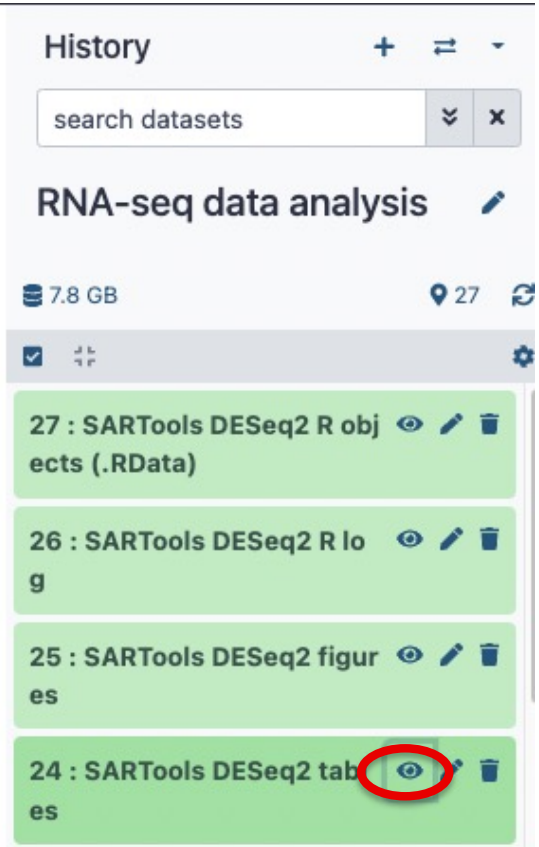
Galaxy Tool SARTools_DESeq2

Run at 14/04/2023 14:28:59

Tables available for downloading

Output File Name (click to view)	Size
siMitfvssiLuc.complete.txt	6.1 MB
siMitfvssiLuc.down.txt	521.9 KB
siMitfvssiLuc.up.txt	587.0 KB

Tabulated text files



- The format of the 3 tables is the same
- Download siMitfvssiLuc.up.txt file
- Open this file with Excel

Differential analysis results

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Id	siLuc2	siLuc3	siMitf3	siMitf4	norm.siLuc2	norm.siLuc3	norm.siMitf3	norm.siMitf4	baseMean	siLuc	siMitf	FoldChange	log2FoldChai	stat	pvalue	padj	dispGeneEst	dispFit	dispMAP	dispersion	betaConv	maxCooks
2	ENSG00000(	4685	5261	18762	22078	4942	5139	19766	20100	12486.79	5040	19933	3.954	1.983	39.51	0	0	2,00E-04	0.0013	0.0011	0.0011	TRUE	NA
3	ENSG00000(	1716	1806	8410	9728	1810	1764	8860	8856	5322.69	1787	8858	4.957	2.31	39.245	0	0	0	0.0016	0.0013	0.0013	TRUE	NA
4	ENSG00000(	3063	3316	12095	13980	3231	3239	12742	12727	7985	3235	12734	3.936	1.977	37.64	0	0	0	0.0014	0.0011	0.0011	TRUE	NA
5	ENSG00000(	309	415	5096	6161	326	405	5369	5609	2927.25	366	5489	14.978	3.905	43.594	0	0	0.0094	0.0021	0.0024	0.0024	TRUE	NA
6	ENSG00000(	3764	4038	15976	18969	3971	3945	16831	17269	10503.85	3958	17050	4.308	2.107	41.423	0	0	0	0.0013	0.0011	0.0011	TRUE	NA
7	ENSG00000(	352	397	4575	5198	371	388	4820	4732	2577.8	380	4776	12.576	3.653	43.641	0	0	0	0.0022	0.0019	0.0019	TRUE	NA
8	ENSG00000(	679	647	5384	6504	716	632	5672	5921	3235.4	674	5796	8.608	3.106	40.262	0	0	0.0027	0.002	0.002	0.002	TRUE	NA
9	ENSG00000(	244	280	3101	3800	257	274	3267	3459	1814.34	266	3363	12.663	3.663	37.5	9.281975425	2.123367903	0	0.0027	0.0025	0.0025	TRUE	NA
10	ENSG00000(	136	151	2266	2714	143	148	2387	2471	1287.26	146	2429	16.692	4.061	34.318	4.255809299	7.788556598	0	0.0035	0.003	0.003	TRUE	NA

→ 1 line per gene, 23 columns

Differential analysis results

siLuc2	siLuc3	siMitf3	siMitf4

- Raw read counts in each sample

norm.siLuc2	norm.siLuc3	norm.siMitf3	norm.siMitf4

- Rounded normalized counts in each sample

baseMean

- Mean of normalized counts over all samples

siLuc	siMitf

- Rounded mean of normalized counts over siLuc or siMitf samples

FoldChange

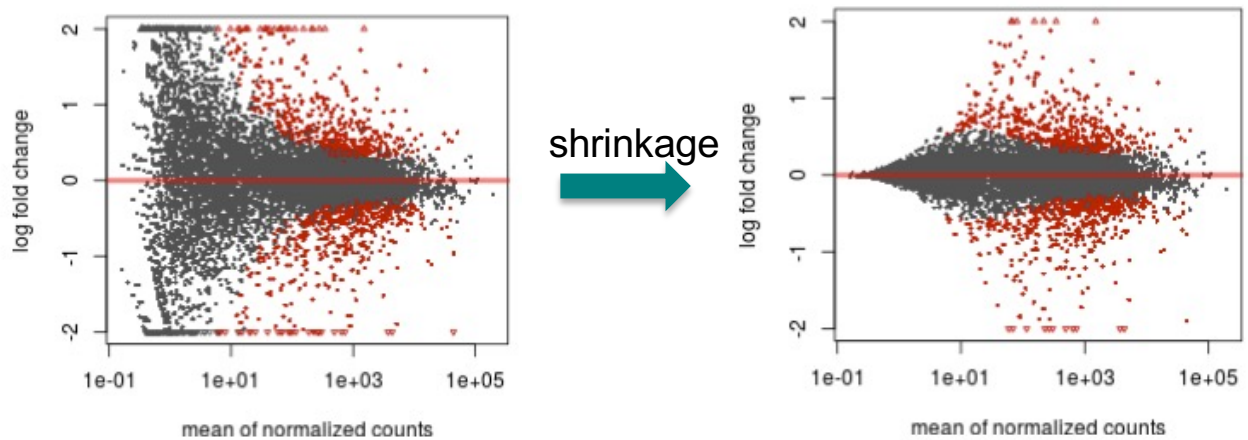
- Expression fold change = $2^{\text{log2FoldChange}}$

log2FoldChange

- log2FoldChange estimated by the generalized linear model
  - Reflects the differential expression between siMitf and siLuc
  - ~0 → similar gene expression in both conditions
  - >0 → over-expressed gene (siMitf > siLuc)
  - <0 → under-expressed gene (siMitf < siLuc)

log2 fold-change (LFC) shrinkage

- To improve stability and interpretability of LFC estimates
- High variance of LFC for genes with low read counts
  - Count data → ratios are inherently noisier when counts are low
- Shrinkage of LFC estimates toward zero
  - Shrinkage is stronger when the information for a gene is low (counts are low or dispersion is high)
  - Avoids that these values, which otherwise would frequently be unrealistically large, dominate the top-ranked LFC
- Shrunken LFC offer a more reproducible quantification of transcriptional differences than standard LFC (Love et al. Genome Biol. 2014;15:550)



Differential analysis results

stat	pvalue	padj

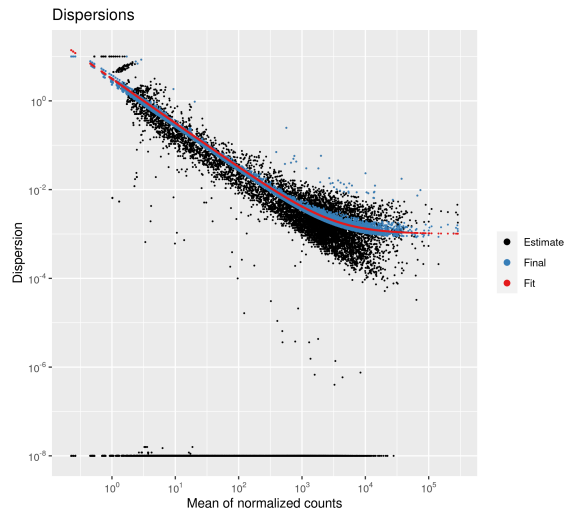
- Statistic, p-value and p-value adjusted for multiple testing

dispGeneEst

- Dispersion parameter estimated from gene counts
  - i.e. black dots on dispersion plot

dispFit

- Dispersion parameter estimated from the model
  - i.e. red dots on dispersion plot



dispMAP

- Maximum *a posteriori* dispersion parameter
  - i.e. blue dots on dispersion plot

dispersion

- Final dispersion parameter used to perform the test
  - i.e. blue dots (with dispersion outliers) on dispersion plot

Differential analysis results

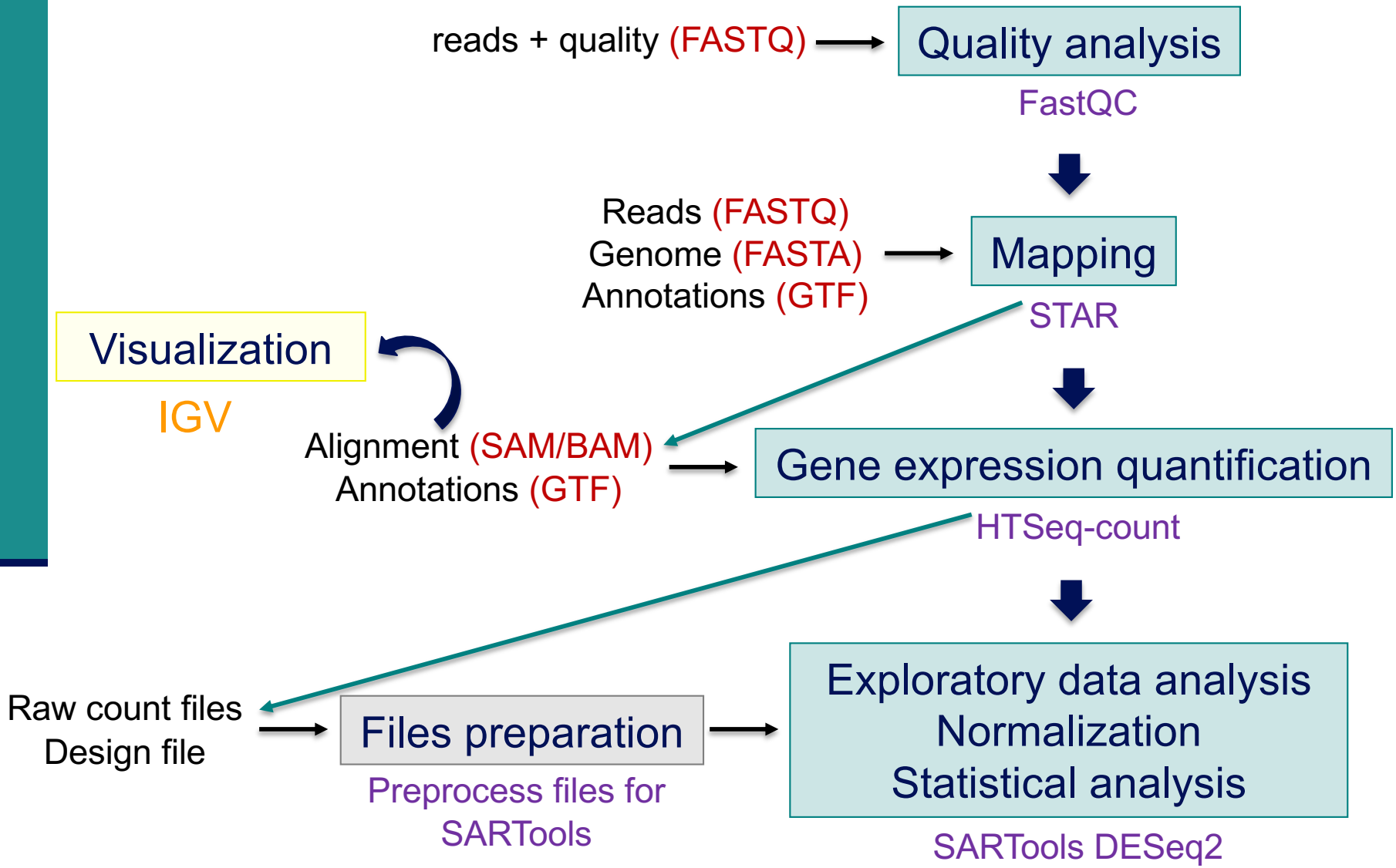
betaConv

- Convergence of the coefficients of the model (True of False)
  - For siMitf project the model converges for all genes

maxCooks

- Maximum Cook's distance of the gene
- If the gene has been detected as a count outlier
  - DESeq2 automatically flags genes which contain a high Cook's distance for samples with 3 or more replicates
    - Therefore = NA for Mitf project
  - Cook's distance
    - Measures of how much a single sample is influencing the fitted coefficients for a gene
    - Large value of Cook's distance is intended to indicate an outlier count

Analysis of RNA-seq data
Files format and Galaxy tools used



RNA-seq data submission to a public data repository

- ArrayExpress (European Nucleotide Archive)
- Gene Expression Omnibus (Sequence Read Archive)
  - How to proceed ?
    https://www.ncbi.nlm.nih.gov/geo/info/seq.html
  - Assembling your submission
    - Metadata spreadsheet : descriptive information about the study
    - Raw data files : FASTQ
    - Processed data files : raw / normalized reads counts
  - Uploading your submission
    - Transfer files to the GEO FTP server
    - Notify GEO and specify when your submission should be released to the public (a private access token can be created for distribution to journal reviewers)