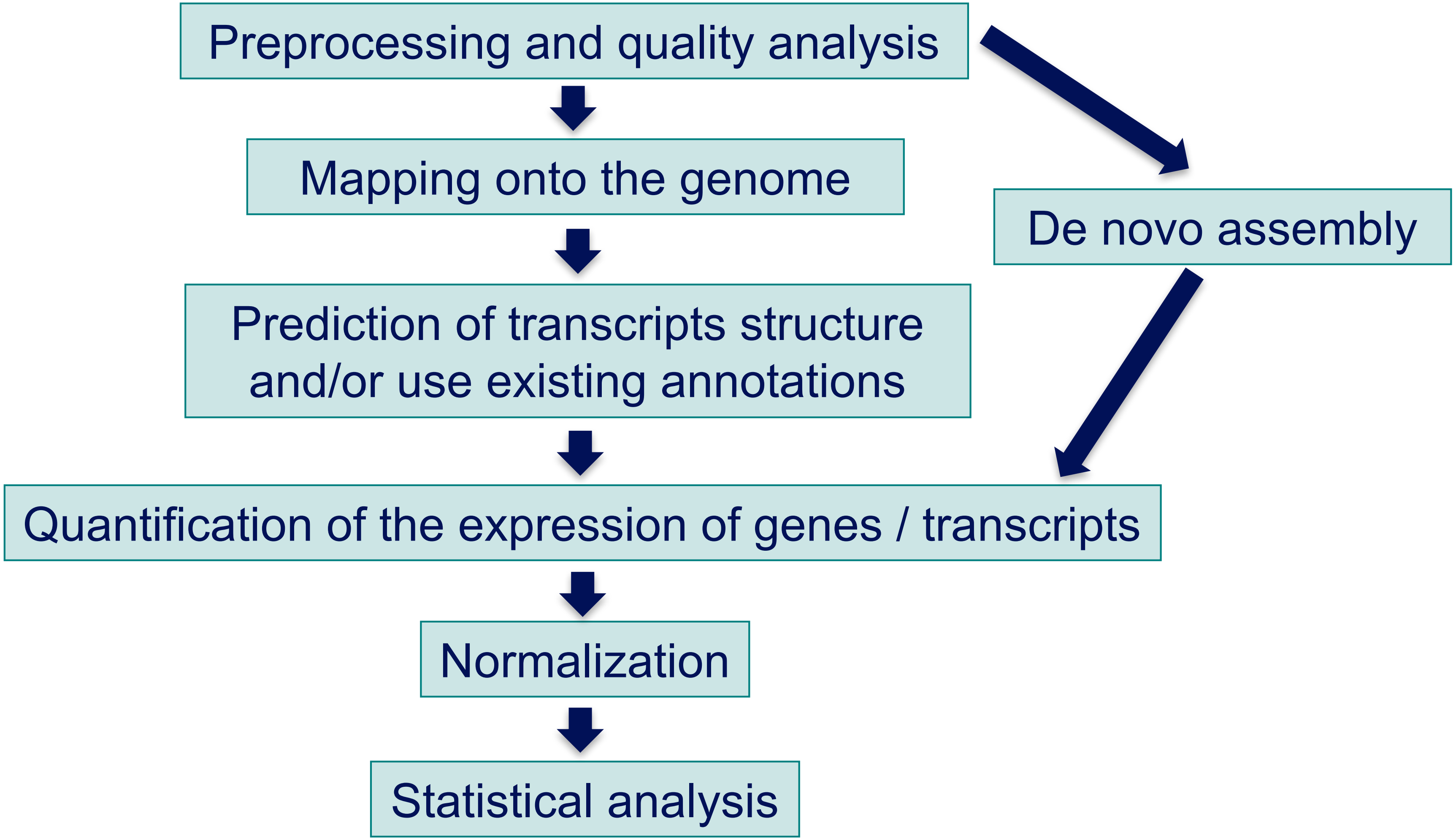
Analysis of RNA-seq data

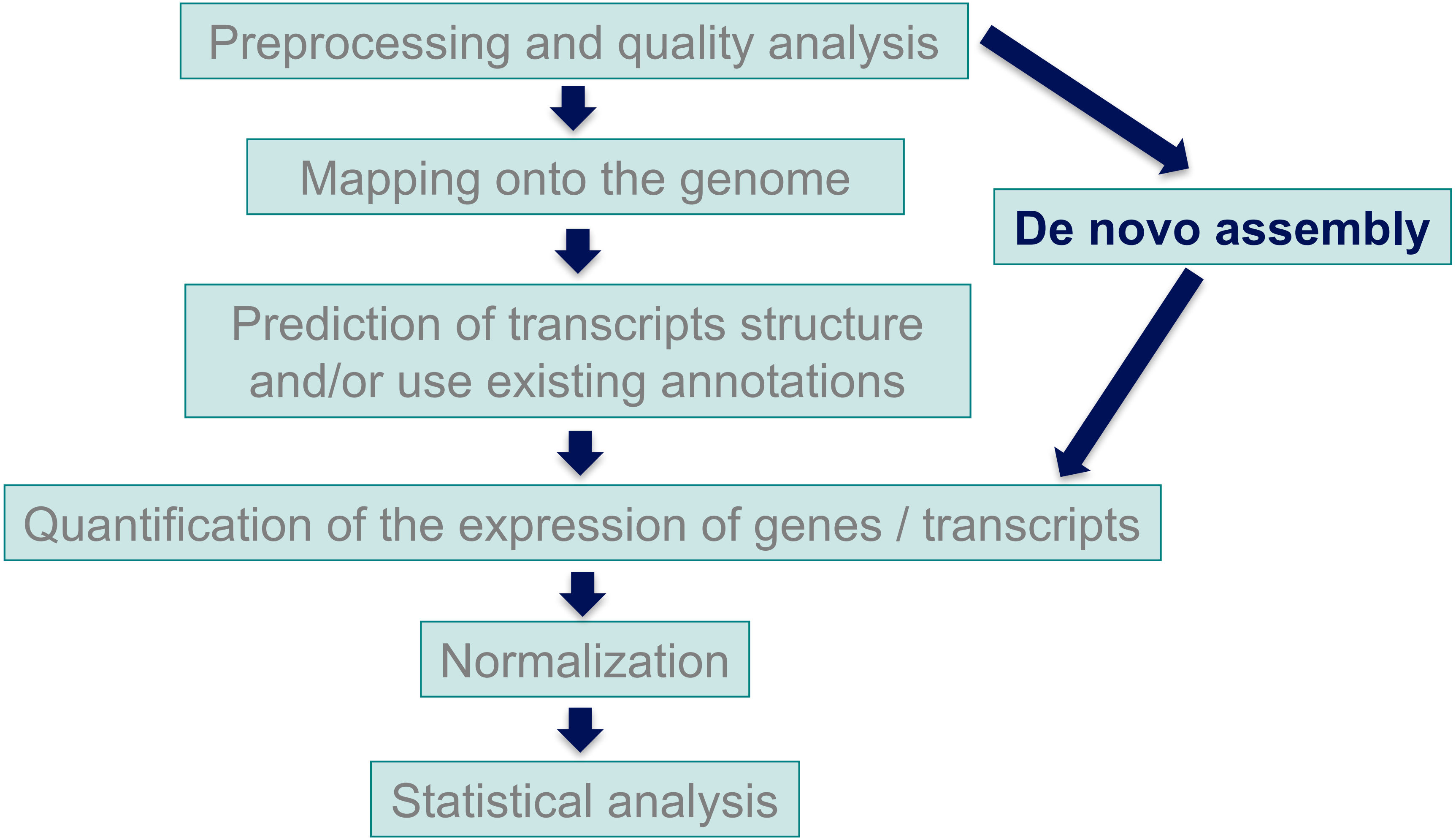
Céline Keime
keime@igbmc.fr

Analysis of RNA-seq data



Preprocessing and quality analysis
Mapping onto the genome
De novo assembly
Prediction of transcripts structure and/or use existing annotations
Quantification of the expression of genes / transcripts
Normalization
Statistical analysis

Analysis of RNA-seq data



Preprocessing and quality analysis

Mapping onto the genome

Prediction of transcripts structure and/or use existing annotations

Quantification of the expression of genes / transcripts

Normalization

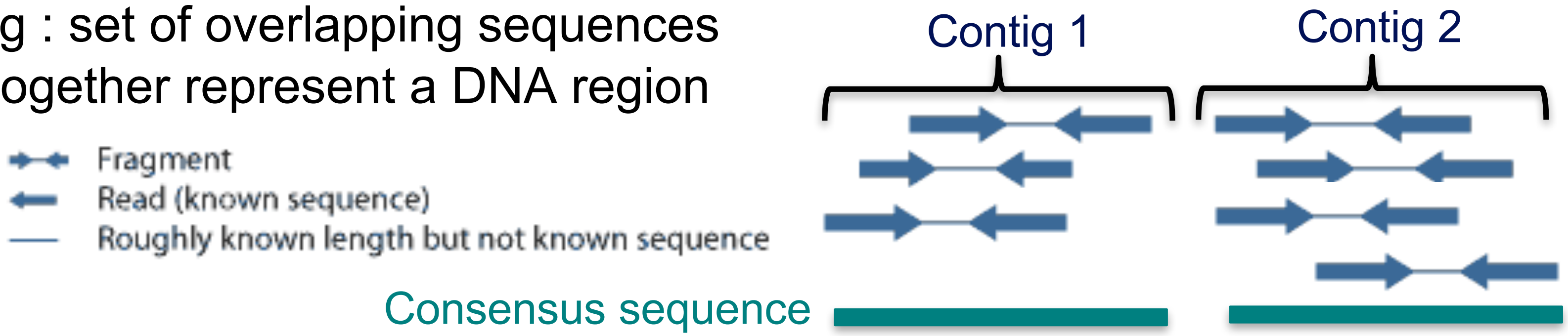
Statistical analysis

De novo assembly

De novo transcriptome assembly

- Purpose
  - Analyse transcriptome on organisms without reference genome
  - Detect chimeric transcripts from chromosomal rearrangements
- Read coverage need to be high enough to build contigs

Contig : set of overlapping sequences that together represent a DNA region



- Challenges (as for genome assembly)
  - Repetitive regions, sequencing errors
- And more challenges specific to transcriptome assembly
  - Transcriptome coverage highly dependent on gene expression
  - Ambiguities in transcriptome assembly due to alternative splicing, alternative promoter usage, alternative polyA, overlapping transcripts

Programs for *de novo* transcriptome assembly

- Different programs
  - Velvet/Oases (Shulz et al. Bioinformatics 2012;28(8):1086-1092)
  - Trans-ABySS (Robertson et al. Nature methods 2010; 7:909–912)
  - Trinity (Haas et al. Nature Protocols 2013; 8:1494–1512)

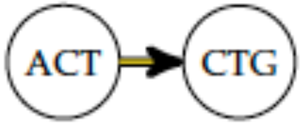
- Comparisons
  - On Illumina data : Zhao et al. (BMC Bioinformatics 2011; 12(14):S2)
  - Which method will perform best is a function of read length, sequencing coverage and transcriptome complexity

De novo transcriptome assembly : general method

- Breaks reads into k-mers (short sub-sequences of length k)

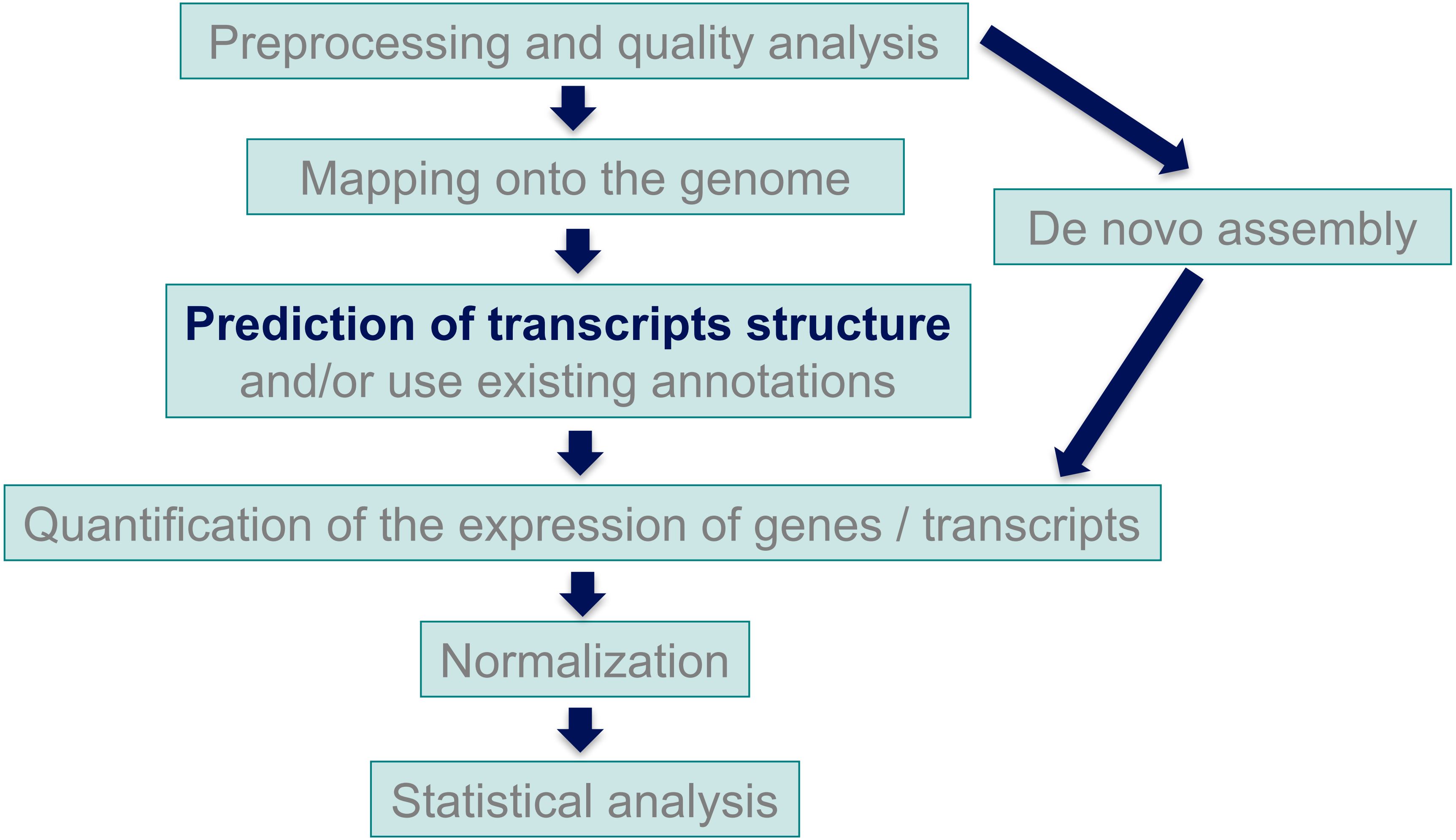
e.g. 1 read = ACTG, k=3 ➔ k-mers = ACT, CTG

- Arranges k-mers into a graph structure (De Brujn graph)
  - Nodes : all sub-sequences of length k present in the sample
  - Arcs : link nodes to represent all sequences present in the sample



- Parse graph in order to create contigs
  - Look at the coverage to decide to follow a path or to remove it in order to avoid sequencing errors
- Choice of k-mer length greatly influence result of the assembly
- Functional annotation of contigs (with Gene Ontology e.g. Blast2GO, screen for Open Reading Frames, for known protein domains, ..)

Analysis of RNA-seq data



Preprocessing and quality analysis

Mapping onto the genome

De novo assembly

Prediction of transcripts structure
and/or use existing annotations

Quantification of the expression of genes / transcripts

Normalization

Statistical analysis

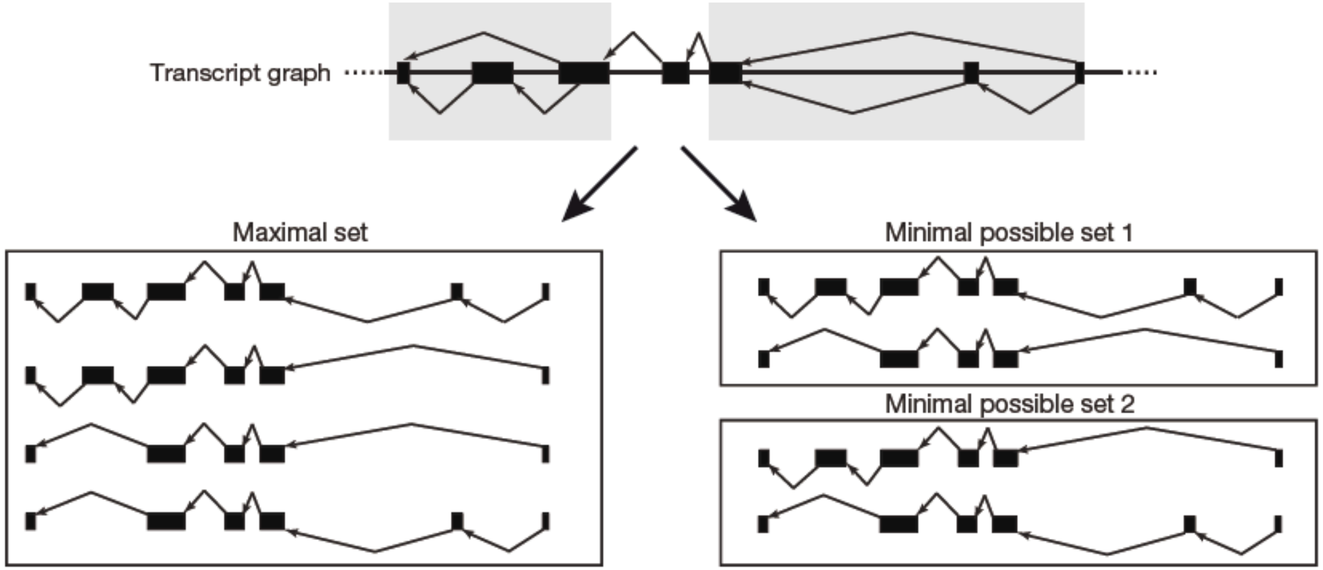
Genome-guided assembly methods

- Use spliced reads to reconstruct the transcriptome

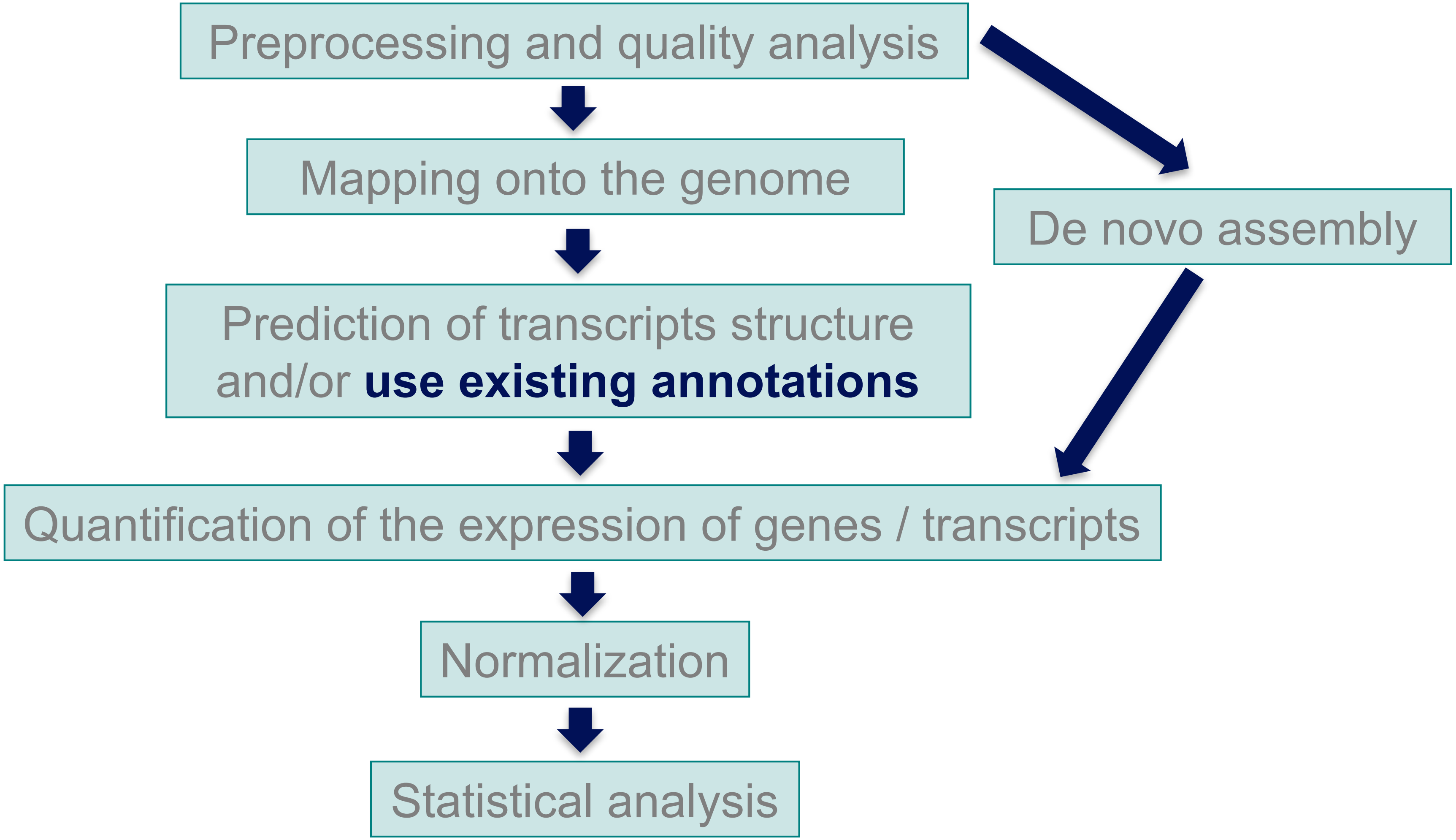
1. Build a transcriptome assembly graph
2. Parse the graph into transcripts (1 path = 1 isoform)

   ➔ Cufflinks reports the minimal number of compatible isoforms i.e. a minimal number of isoforms such that all reads are included in at least one path → uses read coverage to decide which combination of isoforms is most likely to originate from the same RNA

   (Trapnell et al. Nature Biotechnology 2010;28(5):511-5)



Analysis of RNA-seq data



Preprocessing and quality analysis

Mapping onto the genome

De novo assembly

Prediction of transcripts structure and/or **use existing annotations**

Quantification of the expression of genes / transcripts

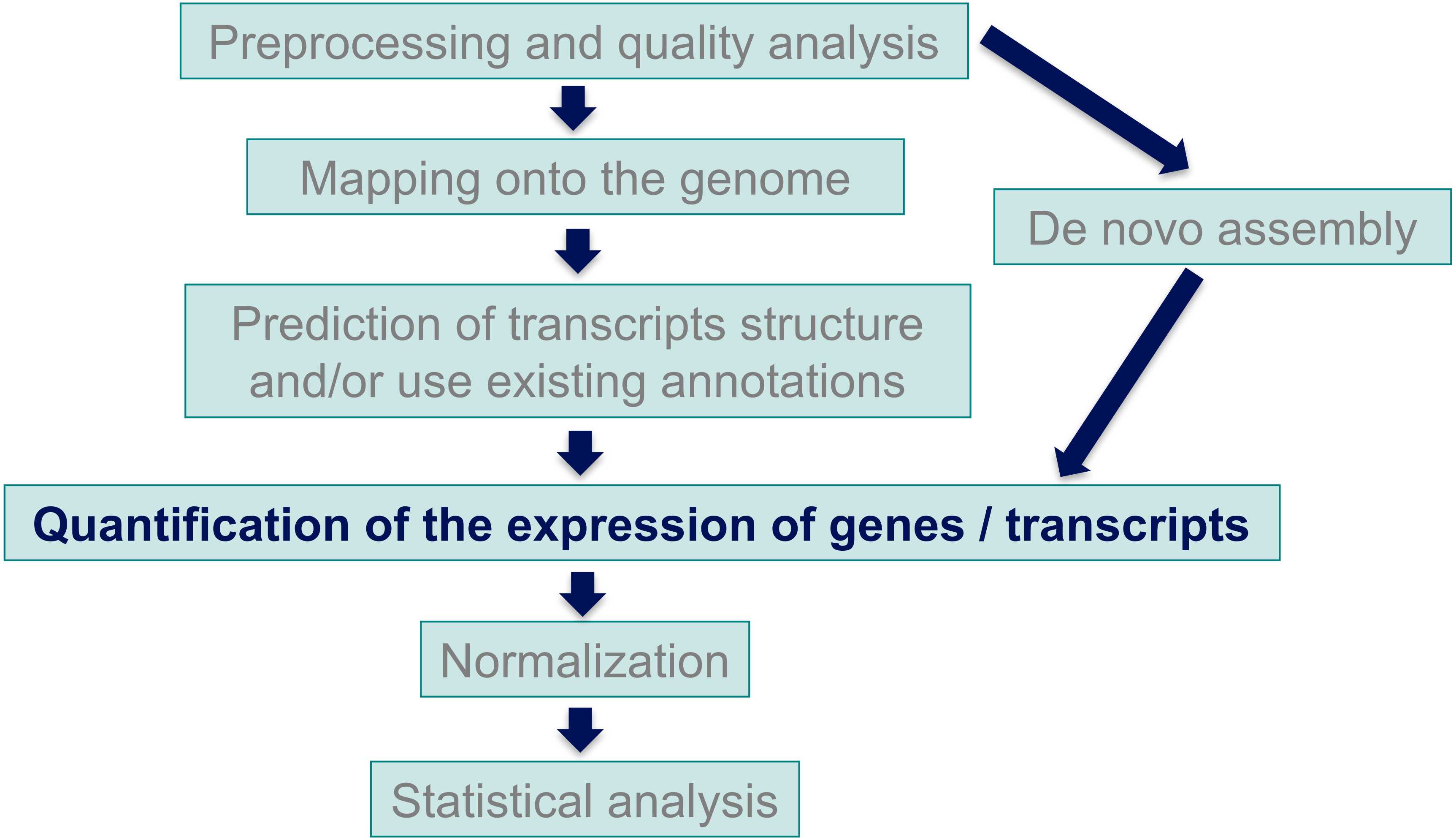
Normalization

Statistical analysis

Genome annotations

- Generally provided in a GTF/GFF file
  - cf. course on read mapping
- Different annotations sources : AceView, Ensembl, UCSC, Refseq…

Analysis of RNAseq data



Preprocessing and quality analysis

Mapping onto the genome

De novo assembly

Prediction of transcripts structure and/or use existing annotations

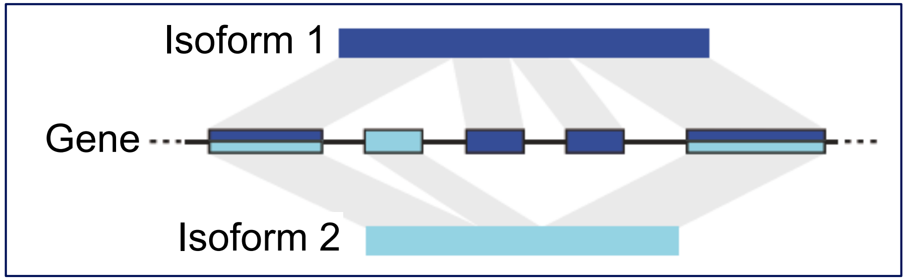
Quantification of the expression of genes / transcripts

Normalization

Statistical analysis

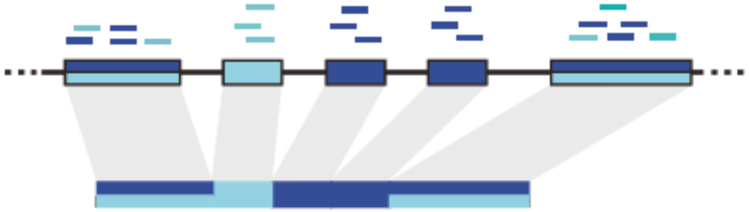
Gene-level quantification

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



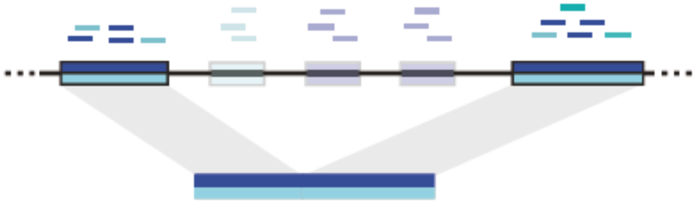
  - Exon-union method

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



  - Exon-intersection method

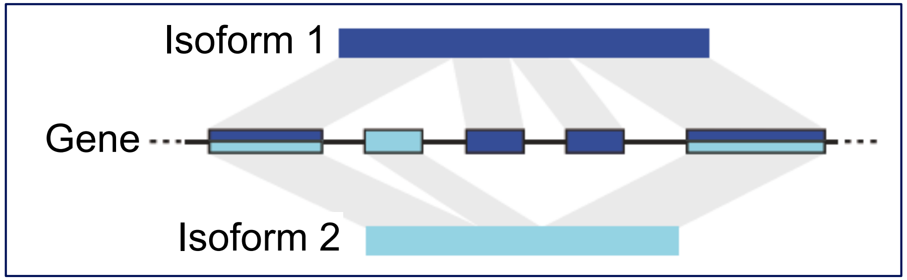
    Count only reads mapped to its constitutive exons



➔ reduce power for differential expression analysis

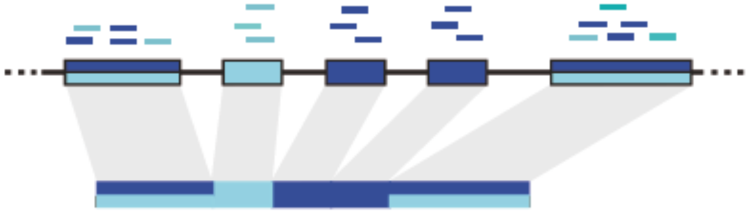
Gene-level quantification

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



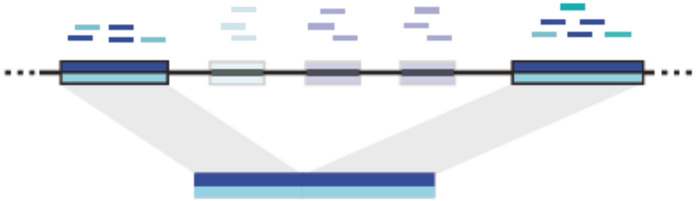
  - Exon-union method

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



  - Exon-intersection method

    Count only reads mapped to its constitutive exons



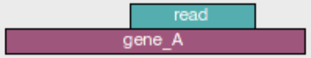
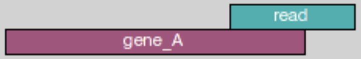
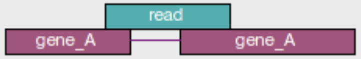
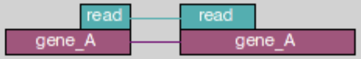
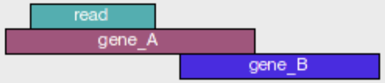
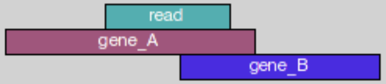
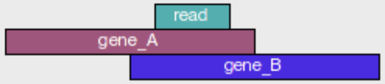
➔ reduce power for differential expression analysis

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, giving the wrong appearance that both genes are differentially expressed

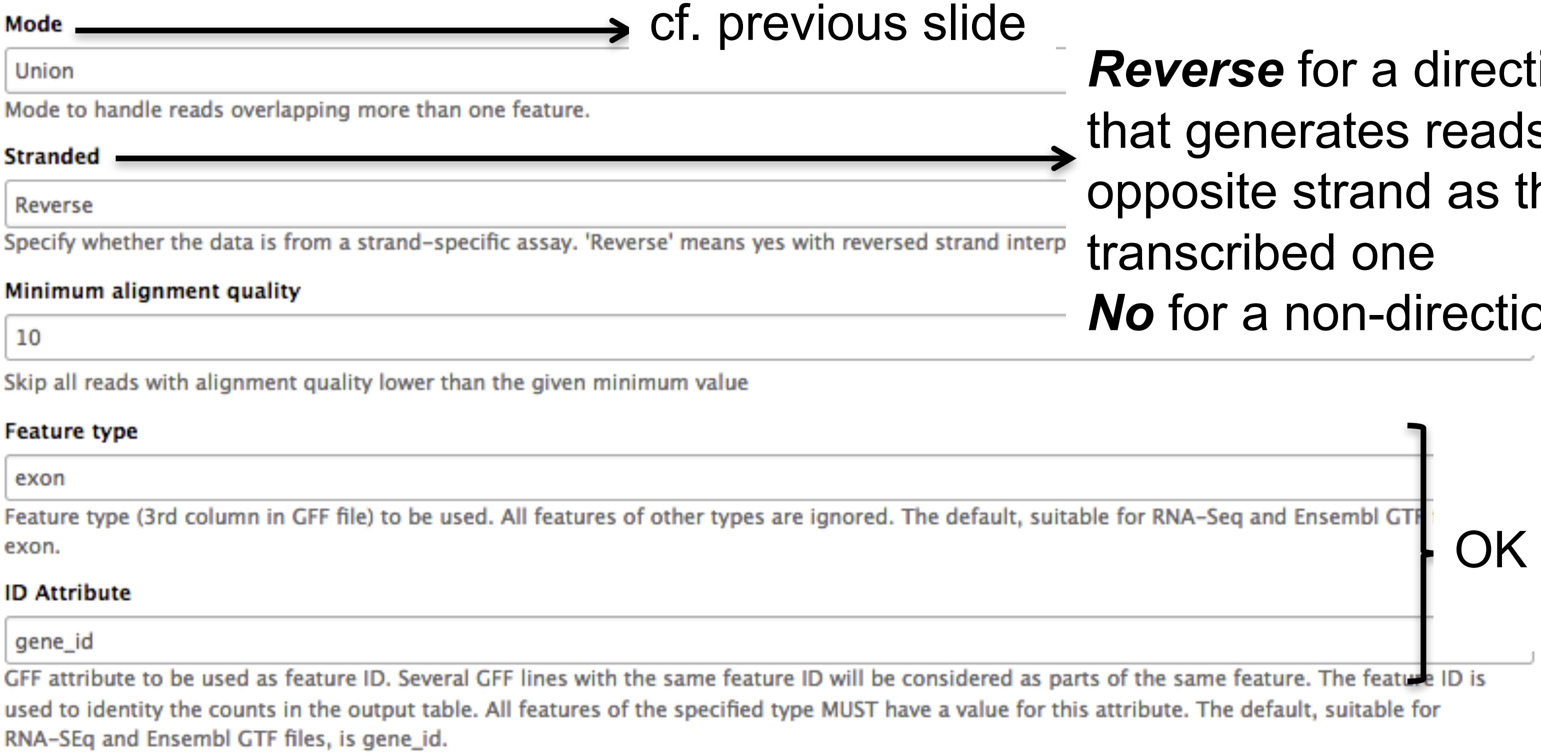
Gene-level quantification : HTSeq-count

- 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

HTSeq-count

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



Mode

Union

Mode to handle reads overlapping more than one feature.

Stranded

Reverse

Specify whether the data is from a strand-specific assay. 'Reverse' means yes with reversed strand interp

Minimum alignment quality

10

Skip all reads with alignment quality lower than the given minimum value

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 exon.

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.

Mode → cf. previous slide

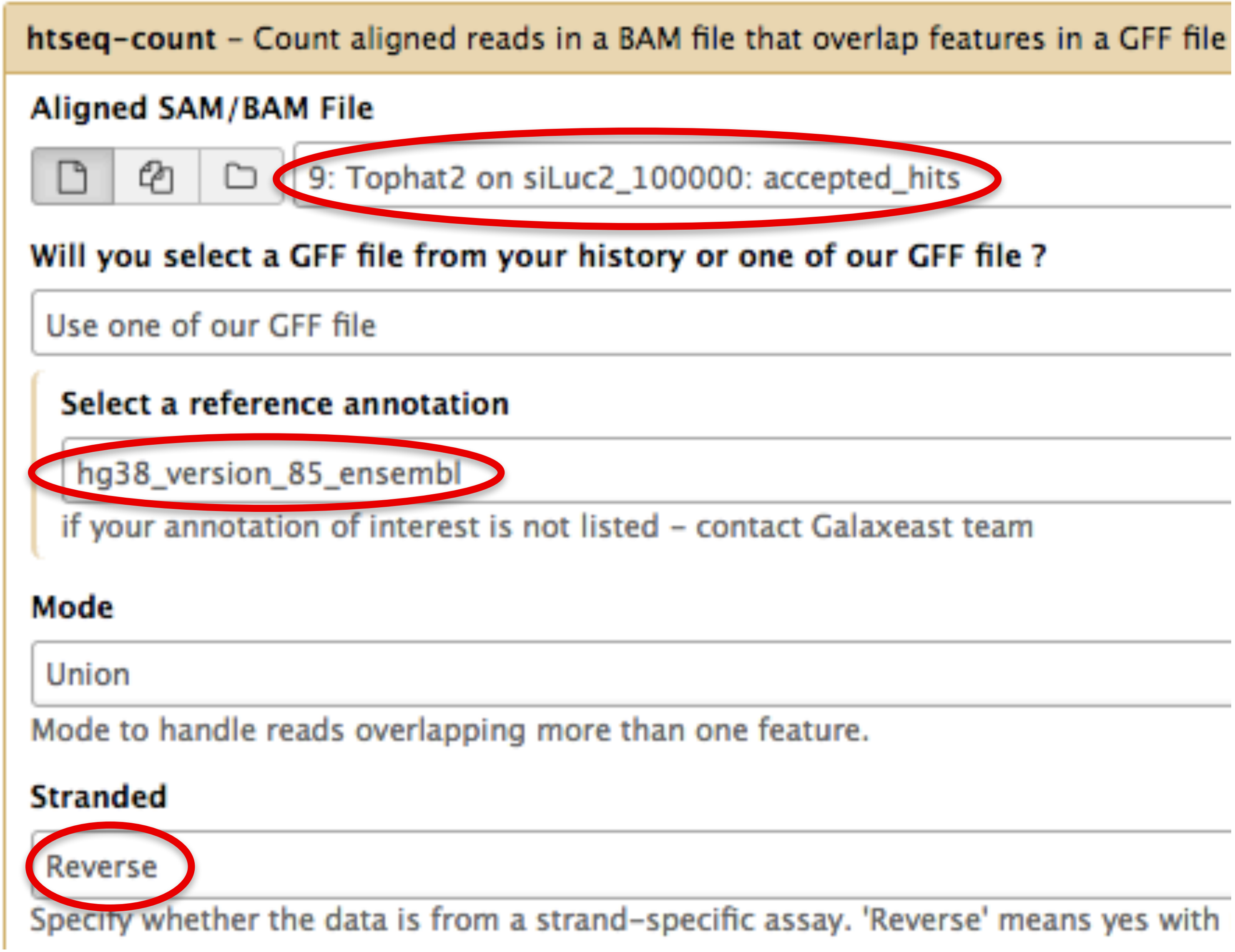
Stranded → ***Reverse*** for a directional protocol that generates reads in the opposite strand as the transcribed one
No for a non-directional protocol

Feature type, ID Attribute: OK for Ensembl

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

- Lauch HTSeq-count to quantify gene expression on siLuc2_1000000 sample
- Inputs
  - Alignment file you obtained with Tophat
  - Ensembl release 85 annotations

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

Aligned SAM/BAM File

9: Tophat2 on siLuc2_100000: accepted_hits

Will you select a GFF file from your history or one of our GFF file ?

Use one of our GFF file

Select a reference annotation

hg38_version_85_ensembl

if your annotation of interest is not listed - contact Galaxeast team

Mode

Union

Mode to handle reads overlapping more than one feature.

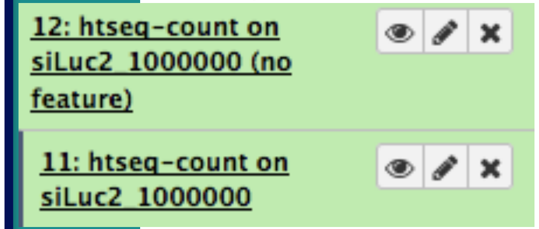
Stranded

Reverse

Specify whether the data is from a strand-specific assay. 'Reverse' means yes with

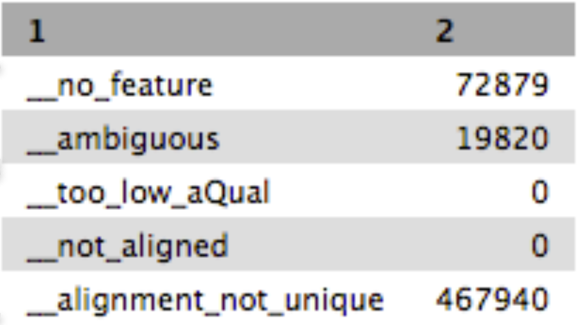
HTSeq-count on GalaxEast

- Output
  - A tabulated text file with
    - the number of reads not assigned to genes
    - The number of alignments not taken into account
  - A tabulated text file with the number of reads assigned to each gene

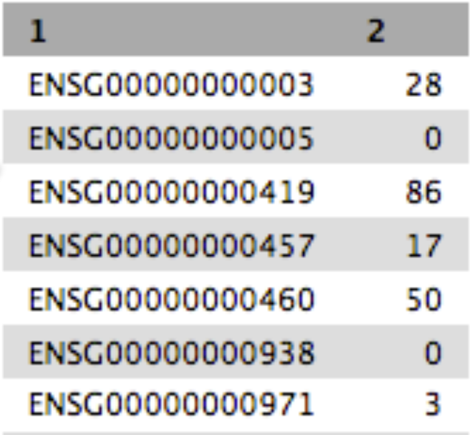


12: htseq-count on siLuc2_1000000 (no feature)

11: htseq-count on siLuc2_1000000



1	2
__no_feature	72879
__ambiguous	19820
__too_low_aQual	0
__not_aligned	0
__alignment_not_unique	467940



1	2
ENSG00000000003	28
ENSG00000000005	0
ENSG00000000419	86
ENSG00000000457	17
ENSG00000000460	50
ENSG00000000938	0
ENSG00000000971	3

HTSeq-count

- Results on siLuc2_1000000

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

  → Calculate the number of uniquely aligned reads

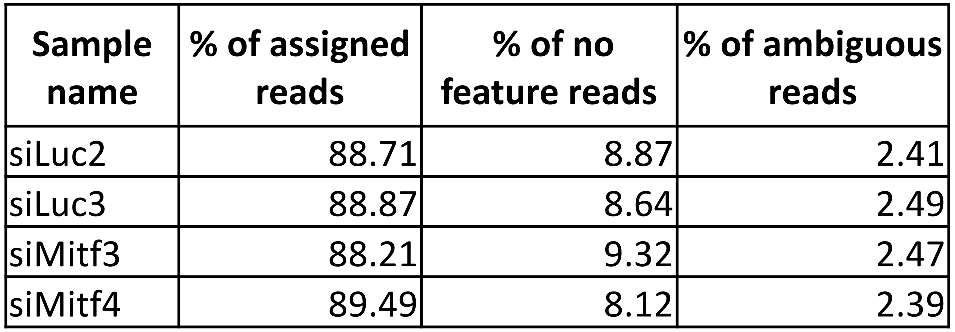
  → What is the number of no feature reads ? Calculate the corresponding proportion

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

  → Calculate the proportion of assigned reads

HTSeq-count

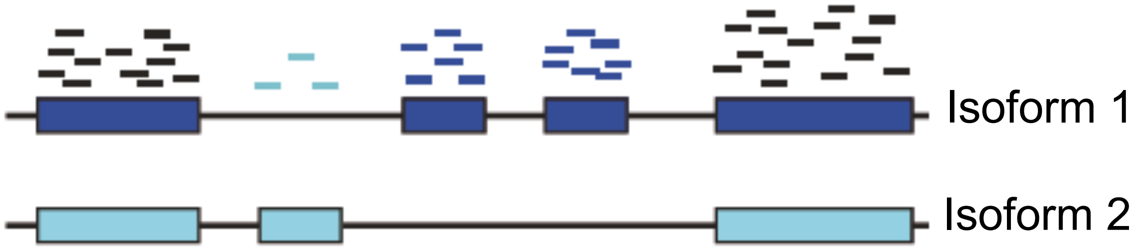
- Results on whole dataset
  - Gene quantification results on the whole dataset are available in
    - Shared Data → Data Libraries → RNAseq → quantification
  - Summary of quantification results



Sample name	% of assigned reads	% of no feature reads	% of ambiguous reads
siLuc2	88.71	8.87	2.41
siLuc3	88.87	8.64	2.49
siMitf3	88.21	9.32	2.47
siMitf4	89.49	8.12	2.39

Transcript-level quantification

- Some reads cannot be assigned unequivocally to a transcript



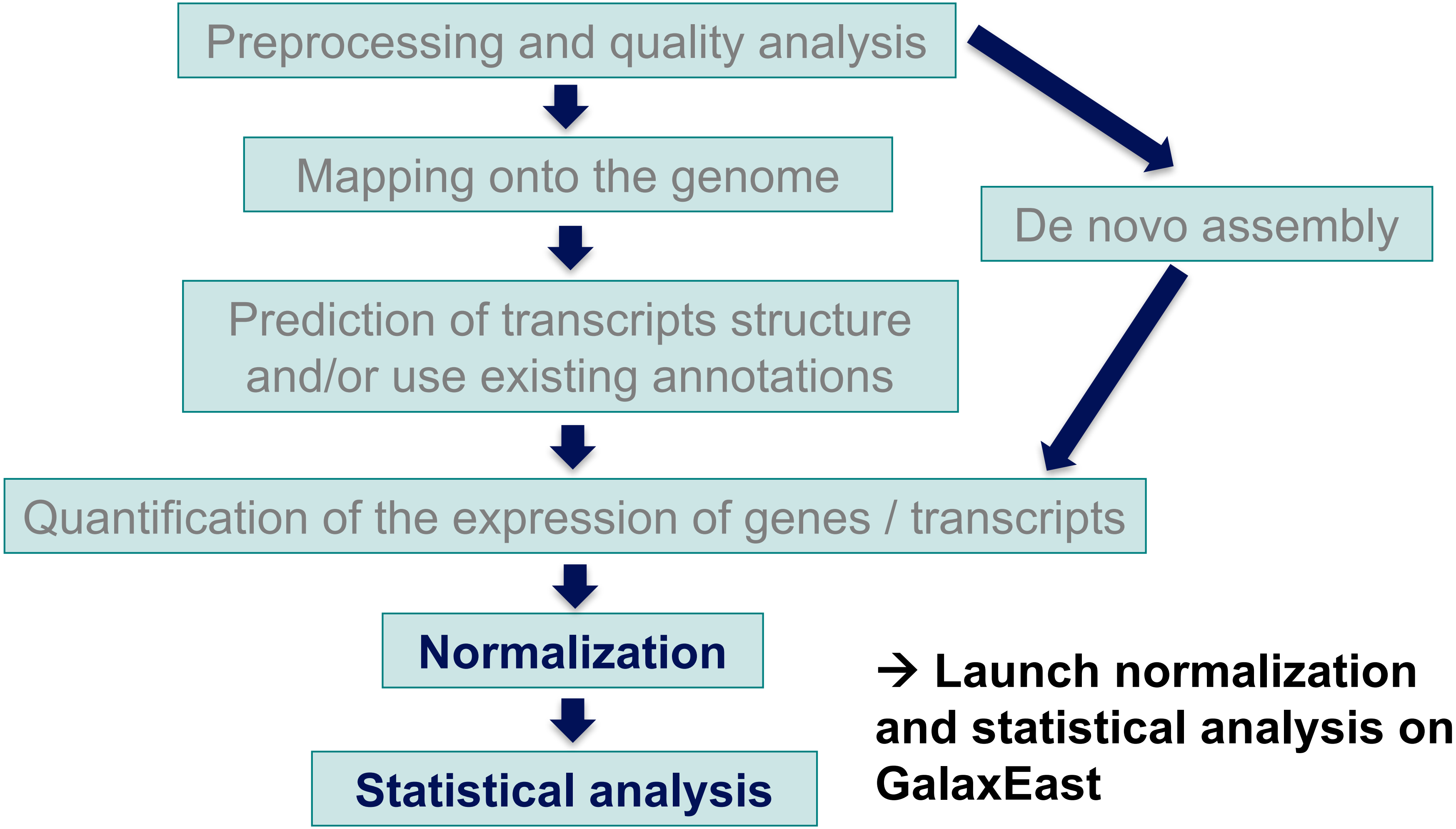
- Alexa-seq (Griffith et al. Nature methods 2010;7(10):843-7)

  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;28(5):511-5)
  MISO (Nature Mathods 2010 Dec;7(12):1009-15)
  - Construct a likelihood function that models the sequencing process
  - Calculate isoforms abundance estimates that best explain reads observed in the experiment

Analysis of RNA-seq data



Preprocessing and quality analysis

Mapping onto the genome

Prediction of transcripts structure and/or use existing annotations

De novo assembly

Quantification of the expression of genes / transcripts

Normalization

Statistical analysis

→ Launch normalization and statistical analysis on GalaxEast

Exercise : statistical analysis using SARTools on GalaxEast

- 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

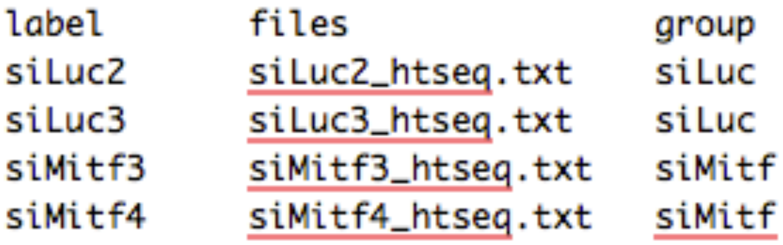
→ **We will use SARTools with DESeq2**

Exercise : statistical analysis using SARTools on GalaxEast

- 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"**

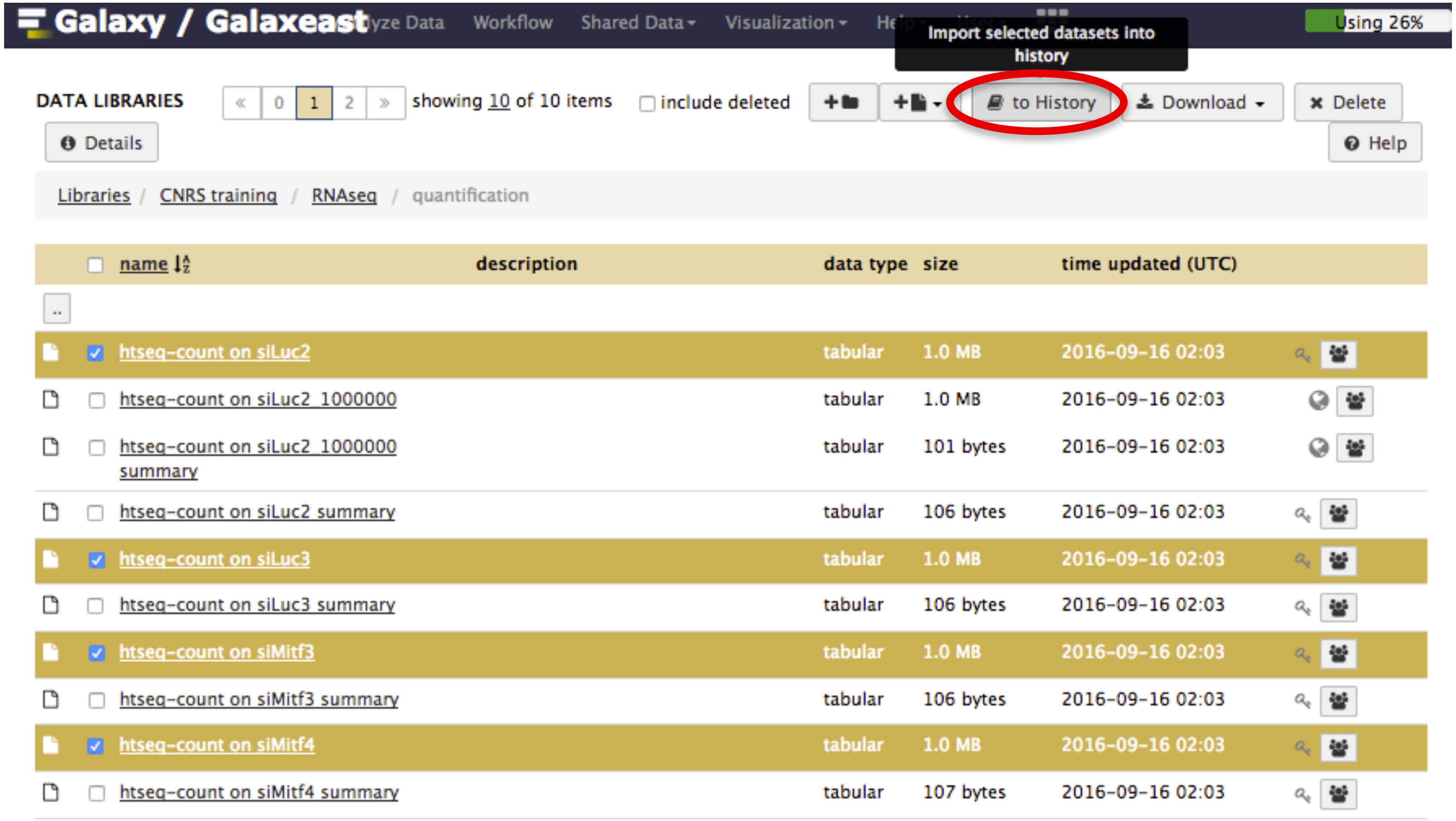
Exercise : statistical analysis using SARTools on GalaxEast

- Launch statistical analysis using SARTools DESeq2
  1. Import raw count files
  2. Prepare 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

Shared Data → Data Libraries → CNRS training → RNAseq → quantification

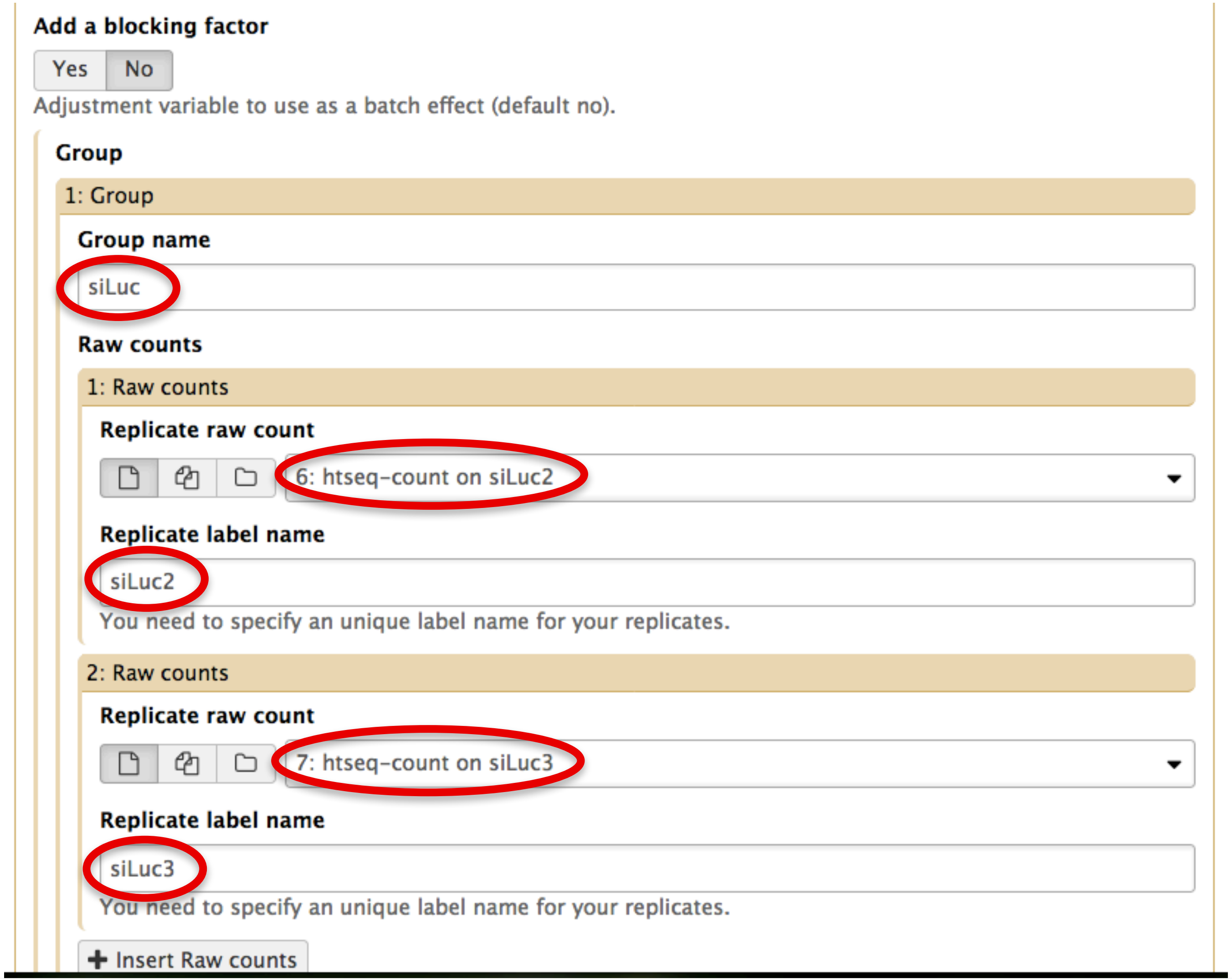


Libraries / CNRS training / RNAseq / quantification

	□	name	description	data type	size	time updated (UTC)
	☑	htseq-count on siLuc2		tabular	1.0 MB	2016-09-16 02:03
	□	htseq-count on siLuc2_1000000		tabular	1.0 MB	2016-09-16 02:03
	□	htseq-count on siLuc2_1000000 summary		tabular	101 bytes	2016-09-16 02:03
	□	htseq-count on siLuc2 summary		tabular	106 bytes	2016-09-16 02:03
	☑	htseq-count on siLuc3		tabular	1.0 MB	2016-09-16 02:03
	□	htseq-count on siLuc3 summary		tabular	106 bytes	2016-09-16 02:03
	☑	htseq-count on siMitf3		tabular	1.0 MB	2016-09-16 02:03
	□	htseq-count on siMitf3 summary		tabular	106 bytes	2016-09-16 02:03
	☑	htseq-count on siMitf4		tabular	1.0 MB	2016-09-16 02:03
	□	htseq-count on siMitf4 summary		tabular	107 bytes	2016-09-16 02:03

Exercise
2. Prepare files for SARTools

- Use the tool “Preprocess files for SARTools”



Add a blocking factor

Yes [✓ No]

Adjustment variable to use as a batch effect (default no).

Group

1: Group

Group name

siLuc

Raw counts

1: Raw counts

Replicate raw count

6: htseq-count on siLuc2

Replicate label name

siLuc2

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

2: Raw counts

Replicate raw count

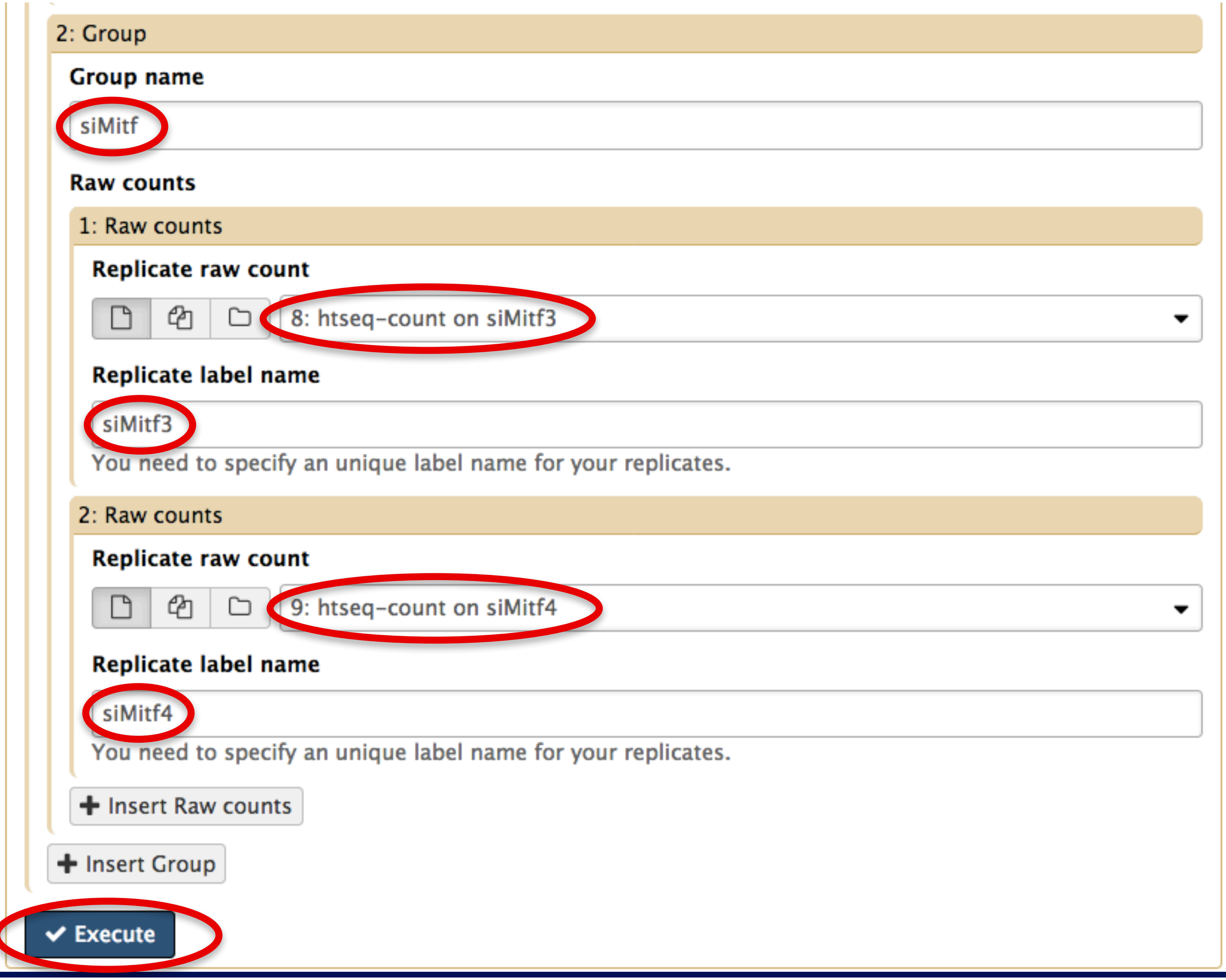
7: htseq-count on siLuc3

Replicate label name

siLuc3

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

Exercise
2. Prepare files for SARTools



2: Group

Group name

siMitf

Raw counts

1: Raw counts

Replicate raw count

8: 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

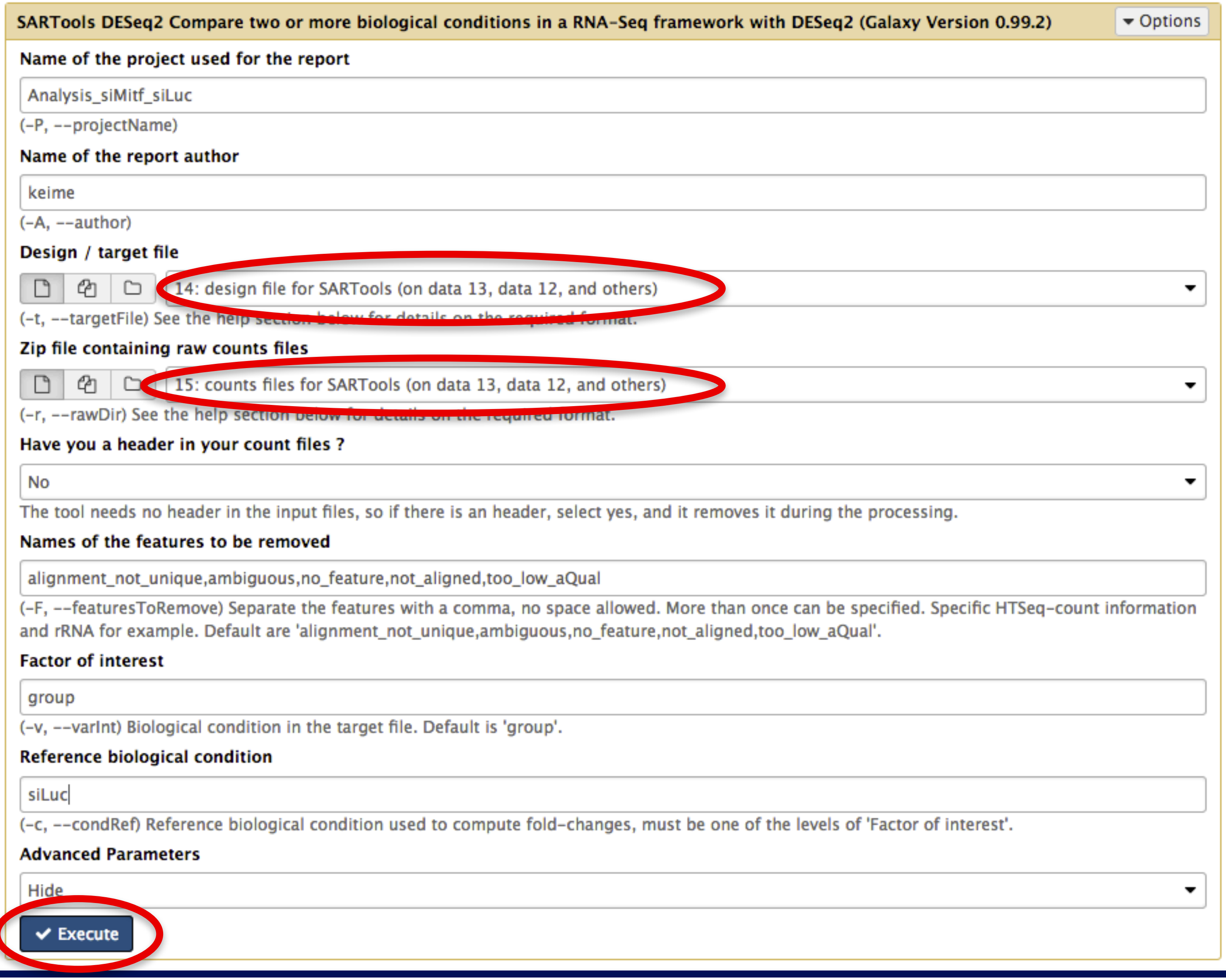
9: htseq-count on siMitf4

Replicate label name

siMitf4

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

Exercise
3. Launch SARTools DESeq2



SARTools DESeq2 Compare two or more biological conditions in a RNA-Seq framework with DESeq2 (Galaxy Version 0.99.2)

Name of the project used for the report

Analysis_siMitf_siLuc

(-P, --projectName)

Name of the report author

keime

(-A, --author)

Design / target file

14: design file for SARTools (on data 13, data 12, and others)

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

Zip file containing raw counts files

15: counts files for SARTools (on data 13, data 12, 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

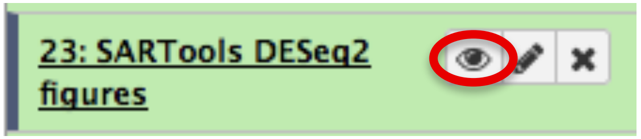
siLuc

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

Advanced Parameters

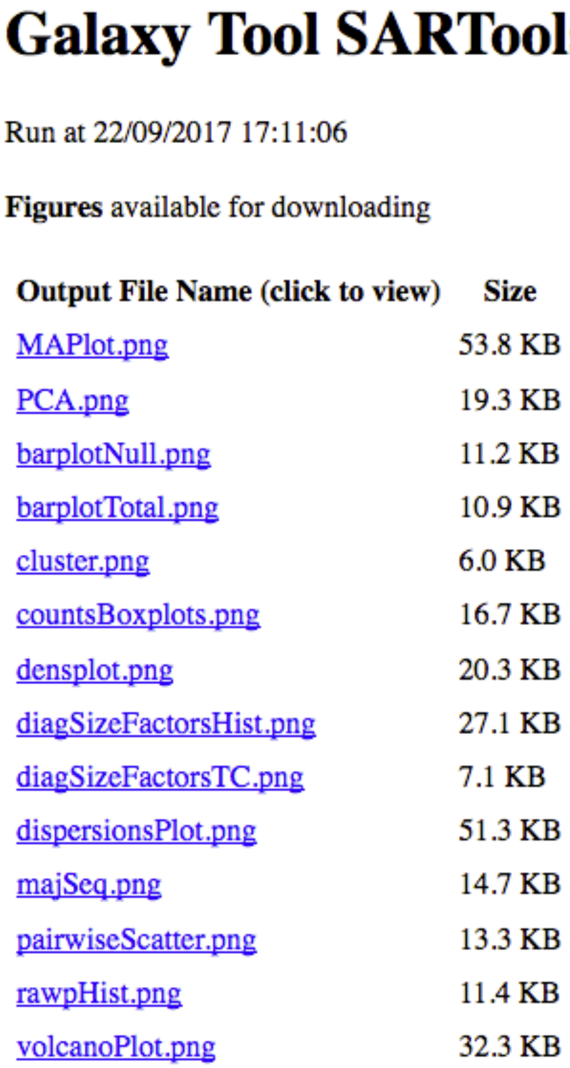
Hide

SARTools results



23: SARTools DESeq2 figures

- Figures



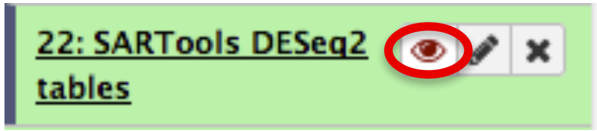
Galaxy Tool SARTools_DESeq2

Run at 22/09/2017 17:11:06

Figures available for downloading

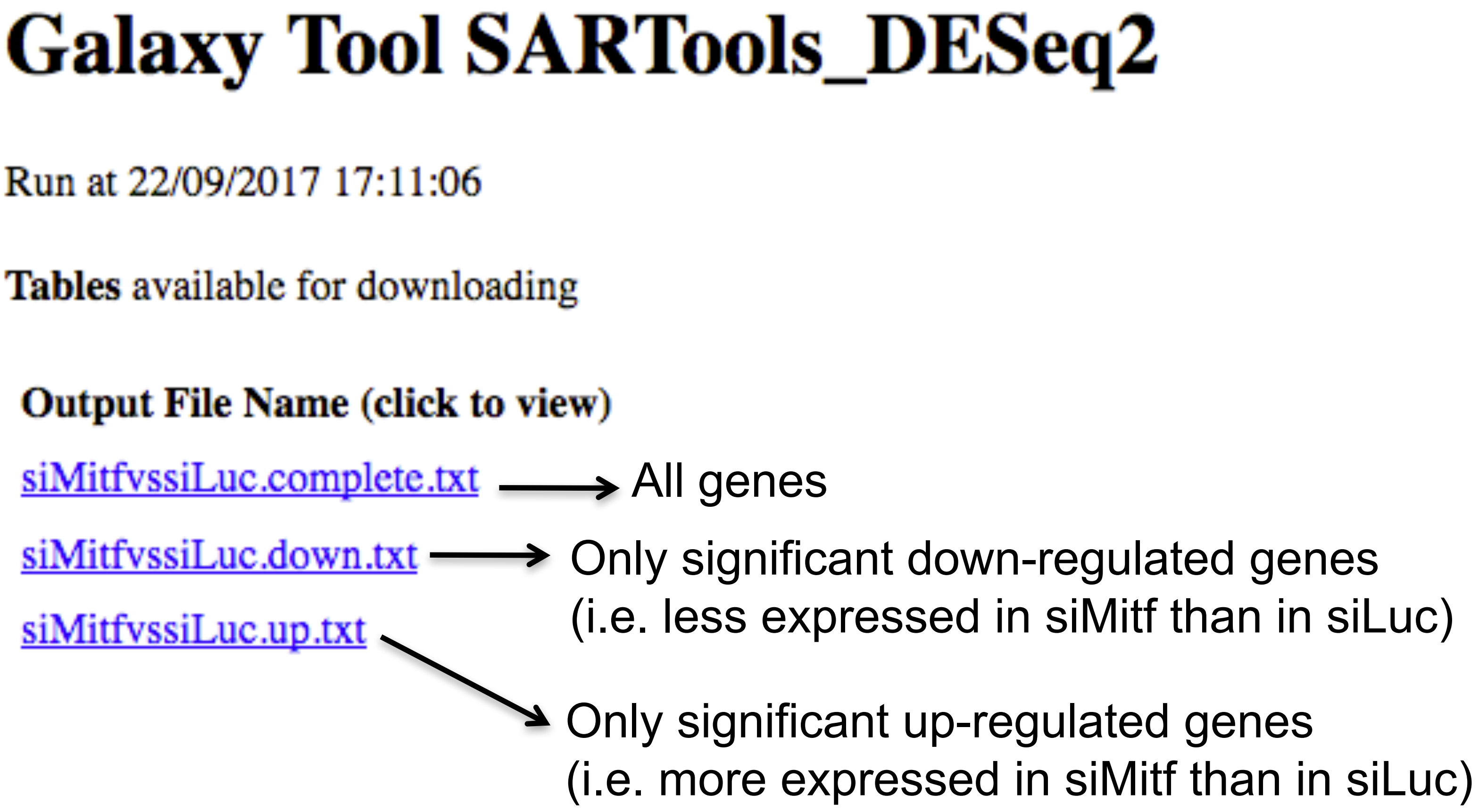
Output File Name (click to view)	Size
MAPlot.png	53.8 KB
PCA.png	19.3 KB
barplotNull.png	11.2 KB
barplotTotal.png	10.9 KB
cluster.png	6.0 KB
countsBoxplots.png	16.7 KB
densplot.png	20.3 KB
diagSizeFactorsHist.png	27.1 KB
diagSizeFactorsTC.png	7.1 KB
dispersionsPlot.png	51.3 KB
majSeq.png	14.7 KB
pairwiseScatter.png	13.3 KB
rawpHist.png	11.4 KB
volcanoPlot.png	32.3 KB

SARTools results



22: SARTools DESeq2 tables

- Tables



Galaxy Tool SARTools_DESeq2

Run at 22/09/2017 17:11:06

Tables available for downloading

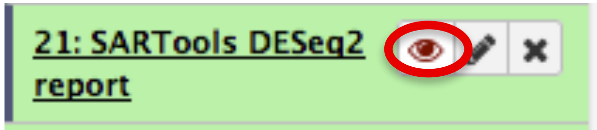
Output File Name (click to view)

siMitfvssiLuc.complete.txt → All genes

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

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

SARTools results



21: SARTools DESeq2 report

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

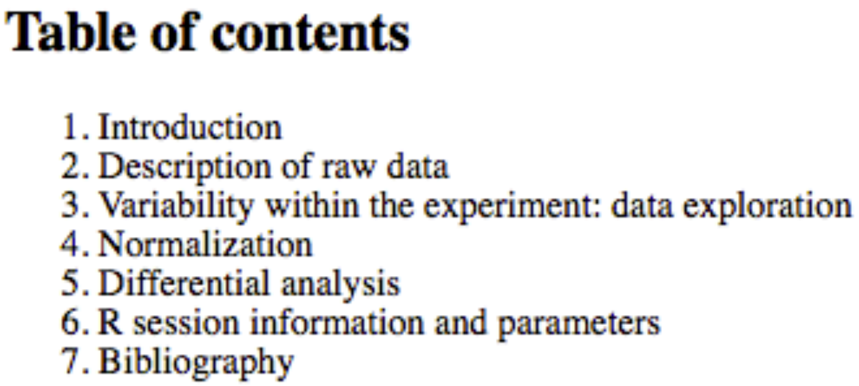


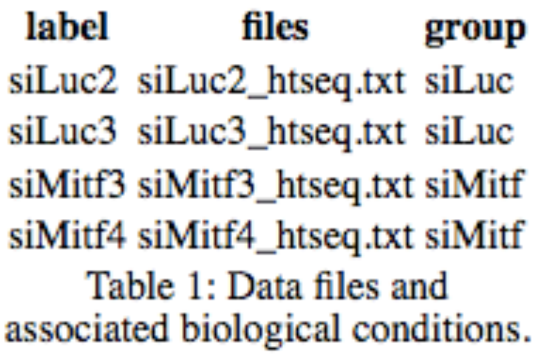
Table of contents

1. Introduction
2. Description of raw data
3. Variability within the experiment: data exploration
4. Normalization
5. Differential analysis
6. R session information and parameters
7. Bibliography

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

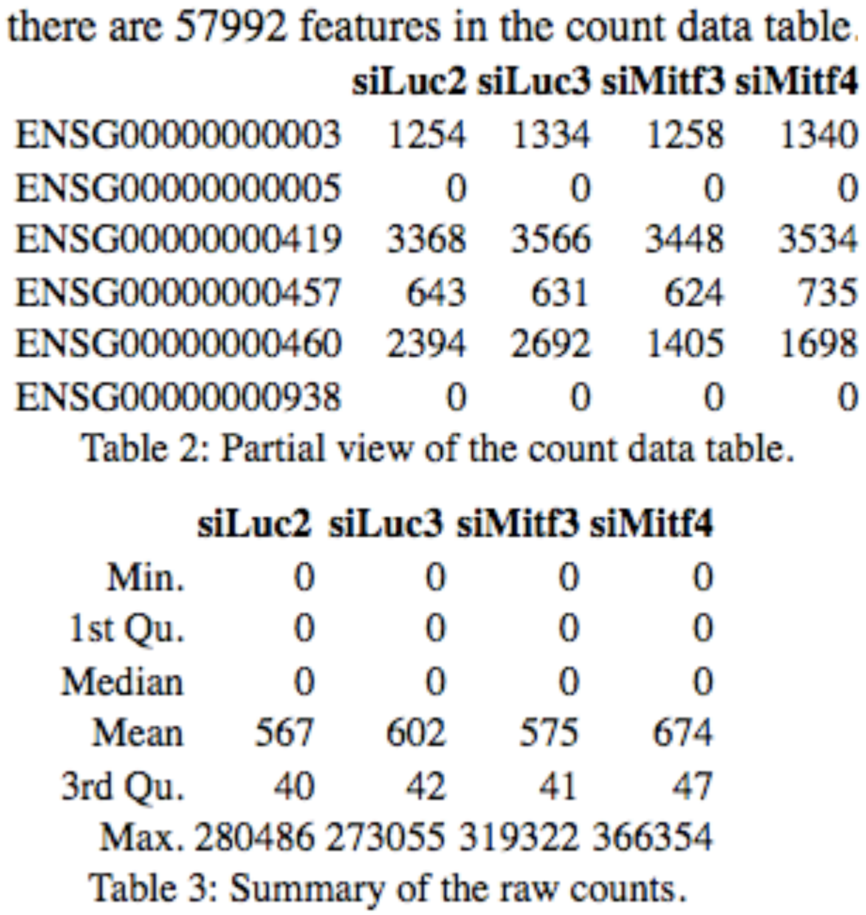
SARTools results

- Report
  - Description of raw data



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

Table 1: Data files and associated biological conditions.



there are 57992 features in the count data table.

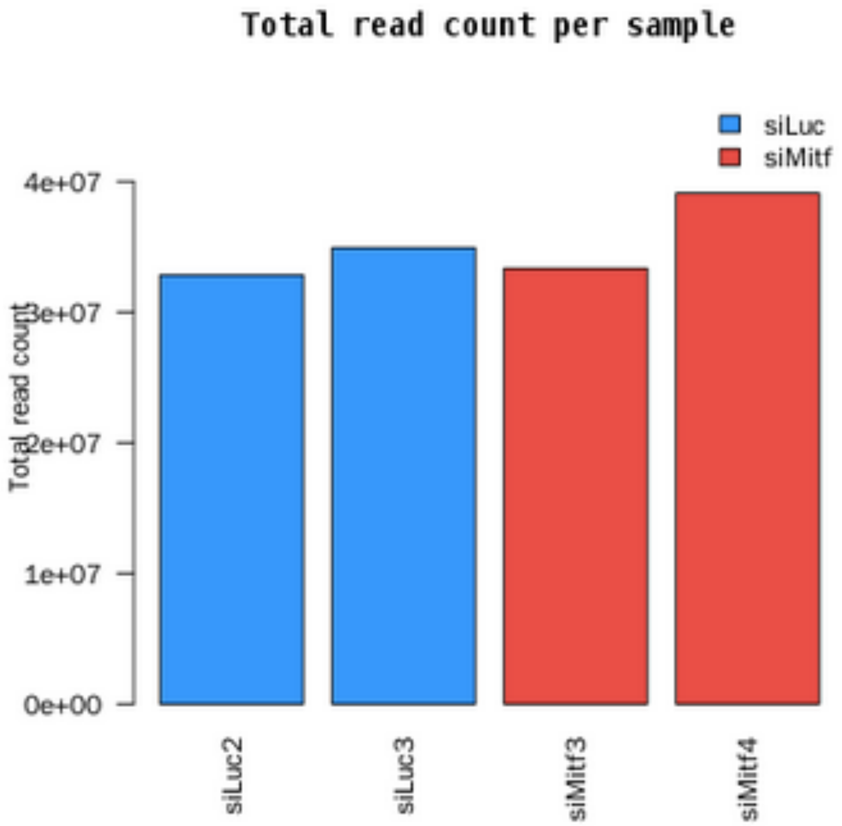
	siLuc2	siLuc3	siMitf3	siMitf4
ENSG00000000003	1254	1334	1258	1340
ENSG00000000005	0	0	0	0
ENSG00000000419	3368	3566	3448	3534
ENSG00000000457	643	631	624	735
ENSG00000000460	2394	2692	1405	1698
ENSG00000000938	0	0	0	0

Table 2: Partial view of the count data table.

	siLuc2	siLuc3	siMitf3	siMitf4
Min.	0	0	0	0
1st Qu.	0	0	0	0
Median	0	0	0	0
Mean	567	602	575	674
3rd Qu.	40	42	41	47
Max.	280486	273055	319322	366354

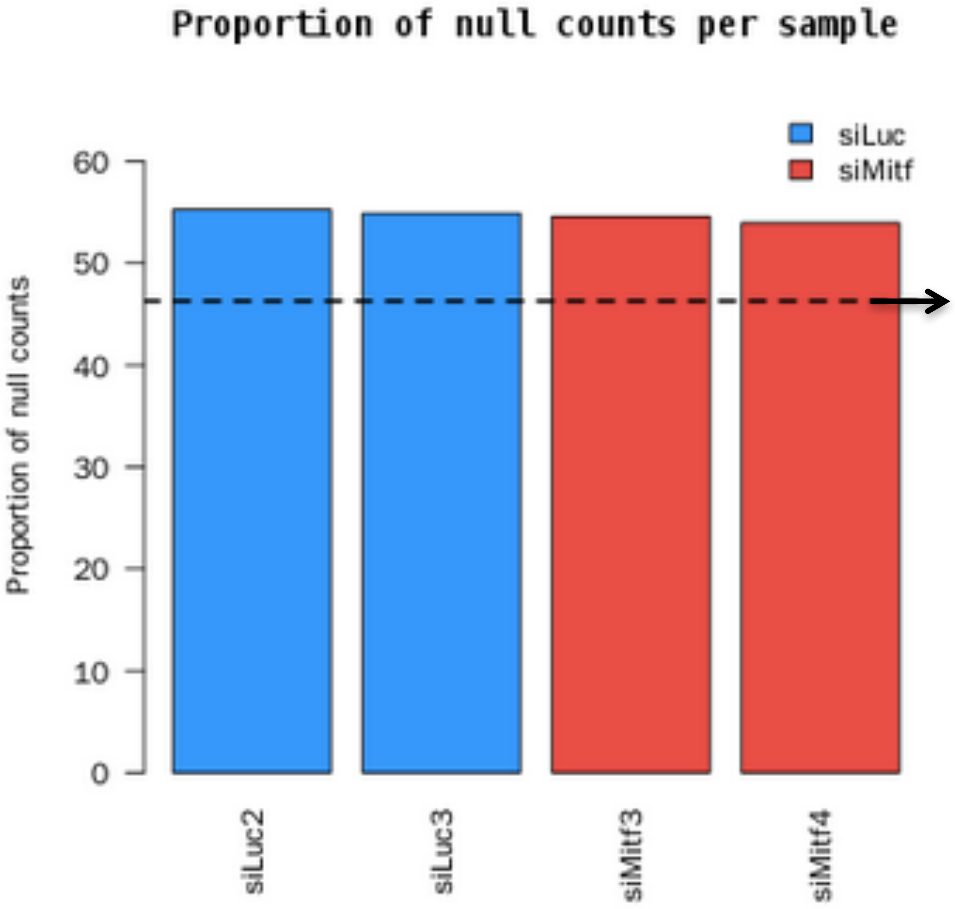
Table 3: Summary of the raw counts.

Total read count per sample



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

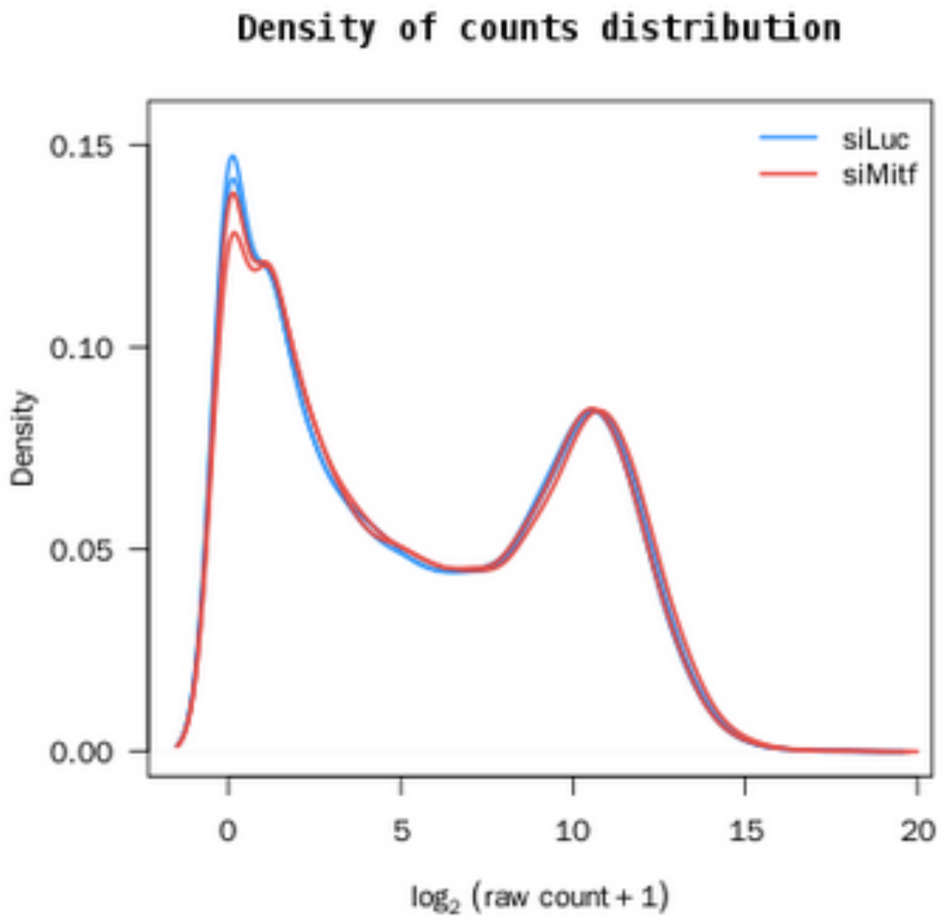
Proportion of null counts per sample



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 results file)

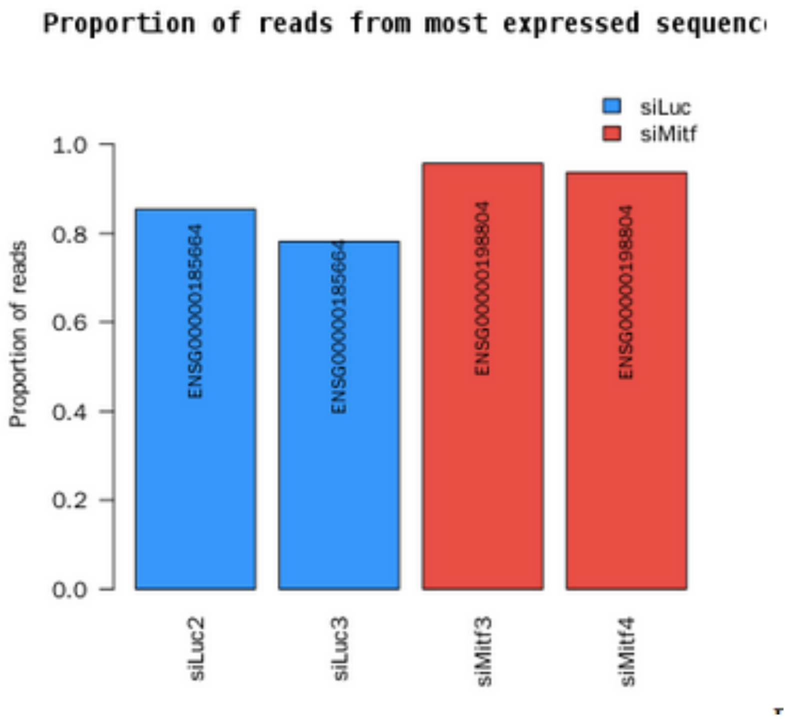
We expect this proportion to be similar between samples

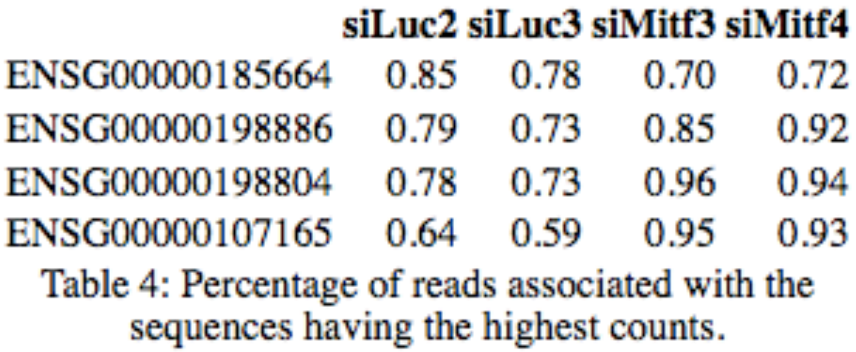
Density distribution of read counts



We expect replicates to have similar distributions

Proportion of reads from most expressed genes





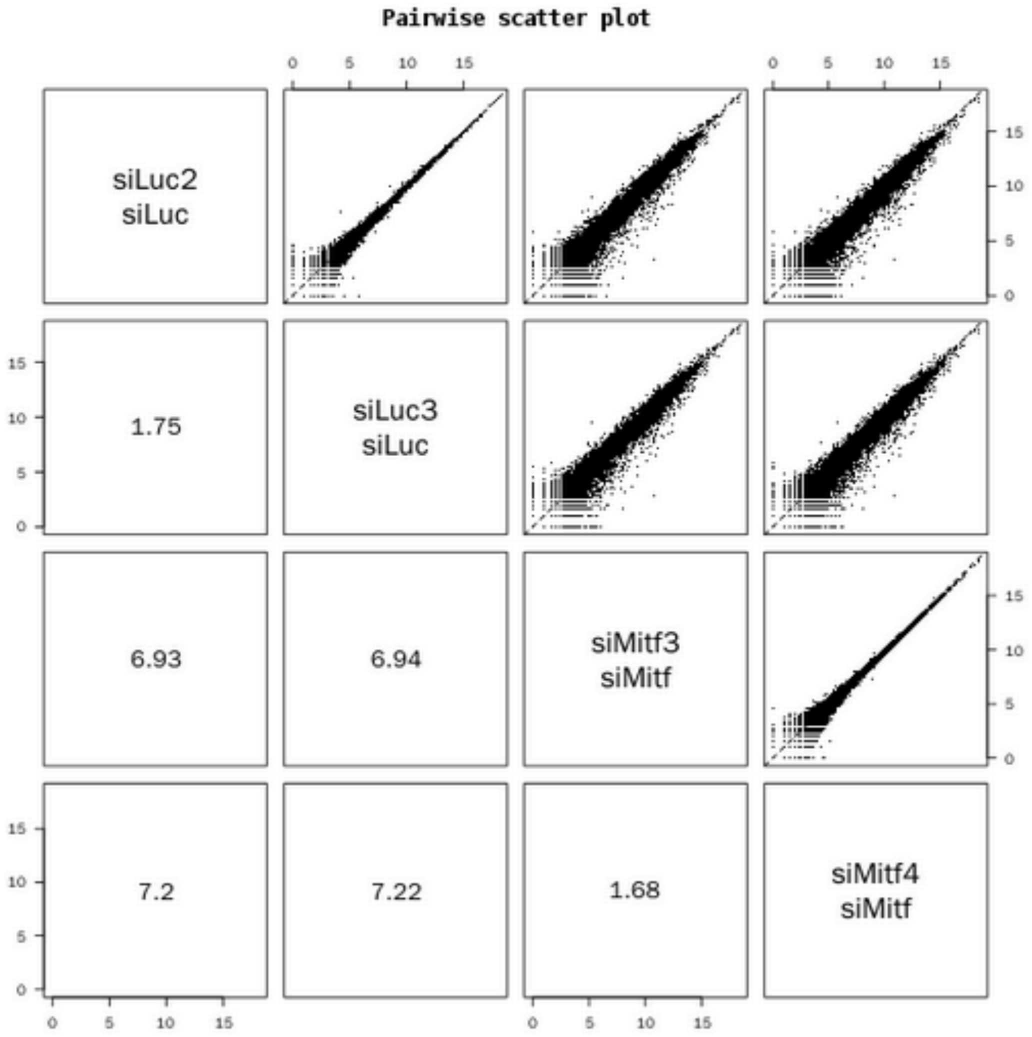
	siLuc2	siLuc3	siMitf3	siMitf4
ENSG00000185664	0.85	0.78	0.70	0.72
ENSG00000198886	0.79	0.73	0.85	0.92
ENSG00000198804	0.78	0.73	0.96	0.94
ENSG00000107165	0.64	0.59	0.95	0.93

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

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

Pairwise comparison of samples

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
- e.g. PCA on RNA-seq data

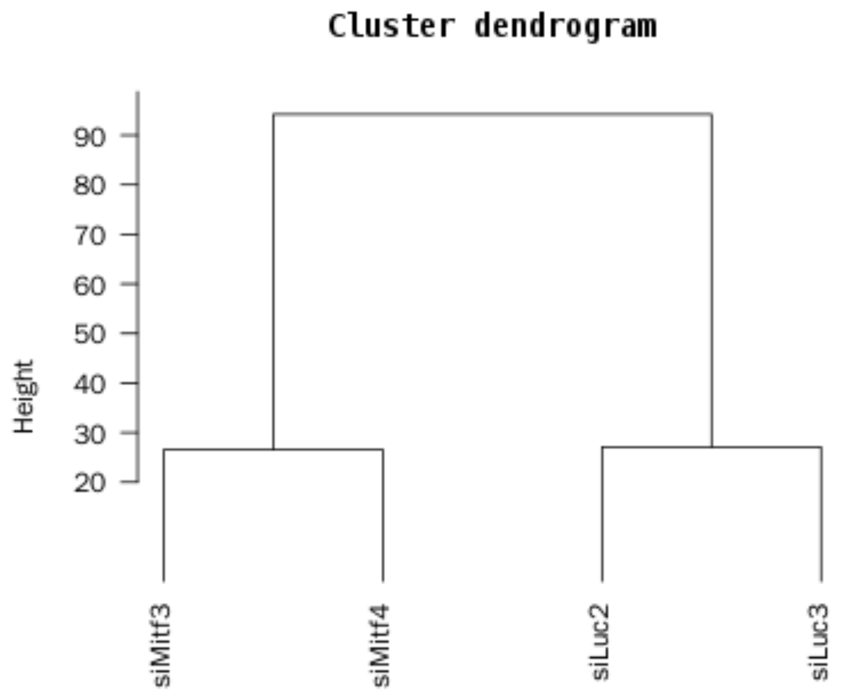
→ result typically depends only on the few most strongly expressed genes because they show the largest absolute differences between samples

- Solution → stabilize variance across the mean
  - VST (variance-stabilizing transformation) : mean-variance relationship estimated from the data (Anders et al. Genome Biology 2010, 11:106)
  - rlog (regularized log-transformation) : fit a generalized linear model from the data, more robust when size factors vary widely (Love et al. Genome Biology 2014, 15:550)

  → Values approximately homoskedastic
  (having constant variance along the range of mean values)

Samples clustering

Obtained from VST-transformed data

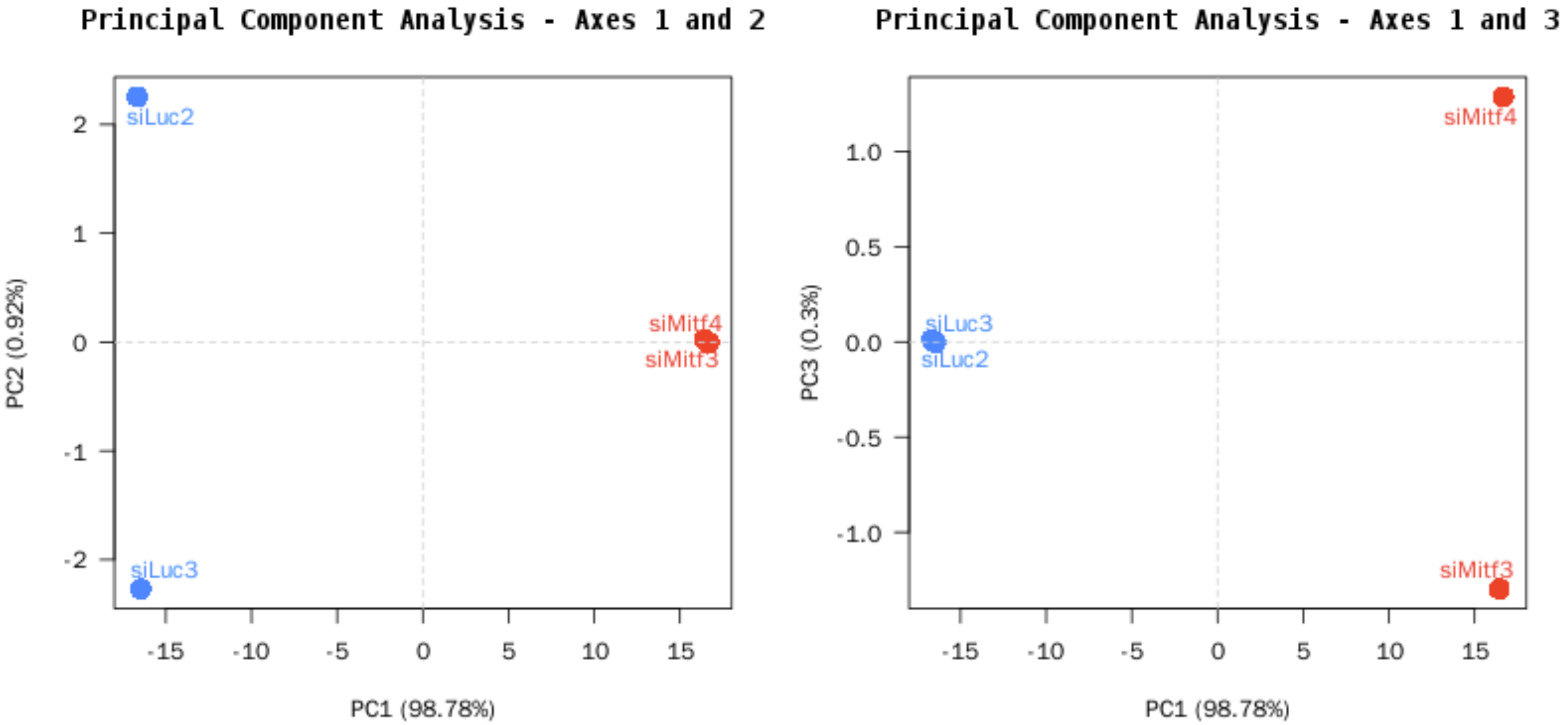


Method: Euclidean distance - Ward criterion
hclust (*, "ward.D")

We expect this dendrogram to group replicates and separate biological conditions

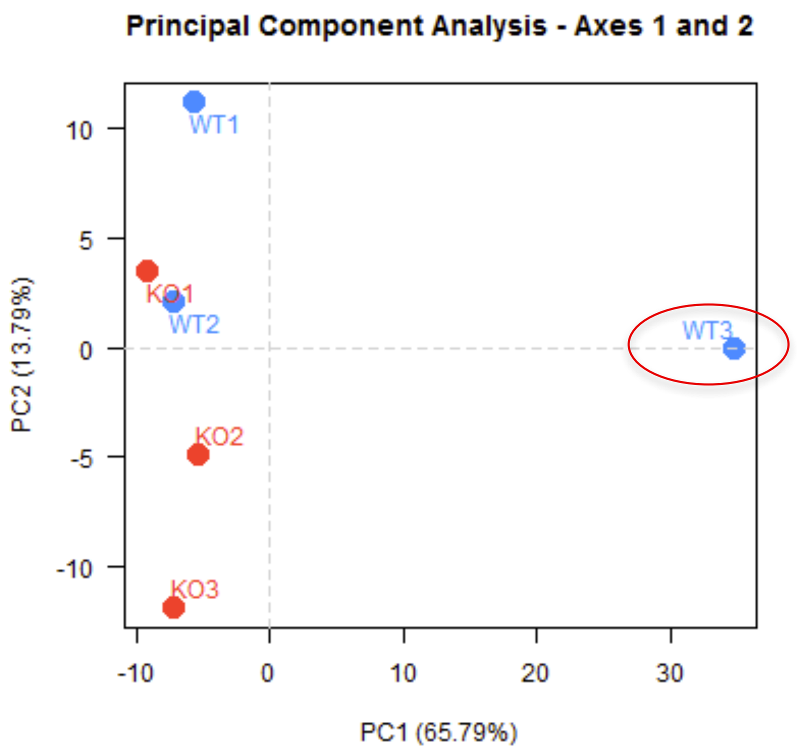
PCA

Obtained from VST-transformed data

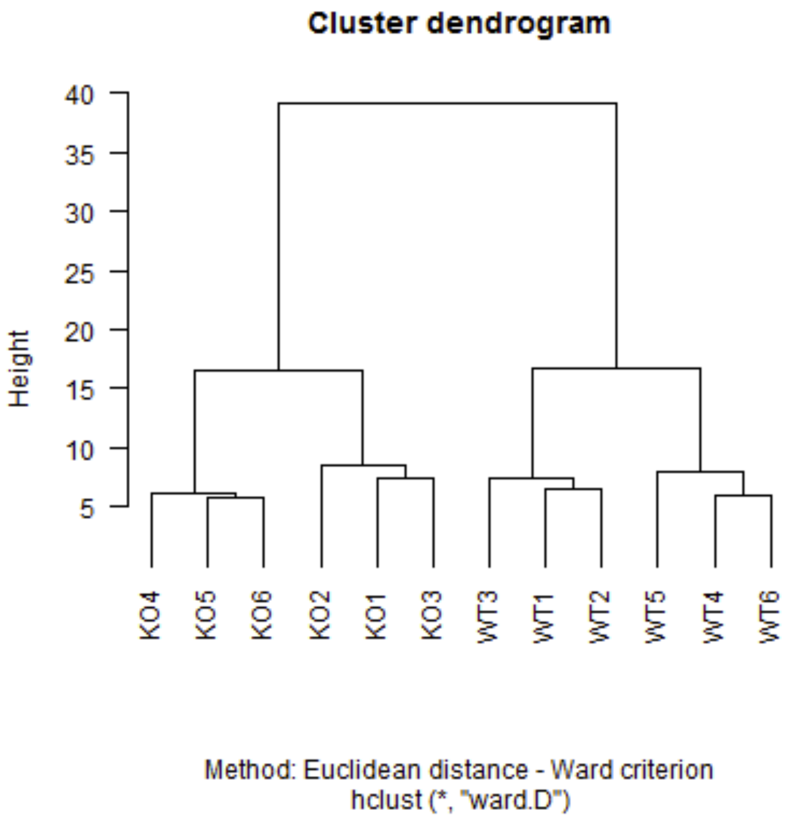


The first principal component is expected to separate samples from the different biological conditions (i.e. corresponds to the main source of variance in the data)

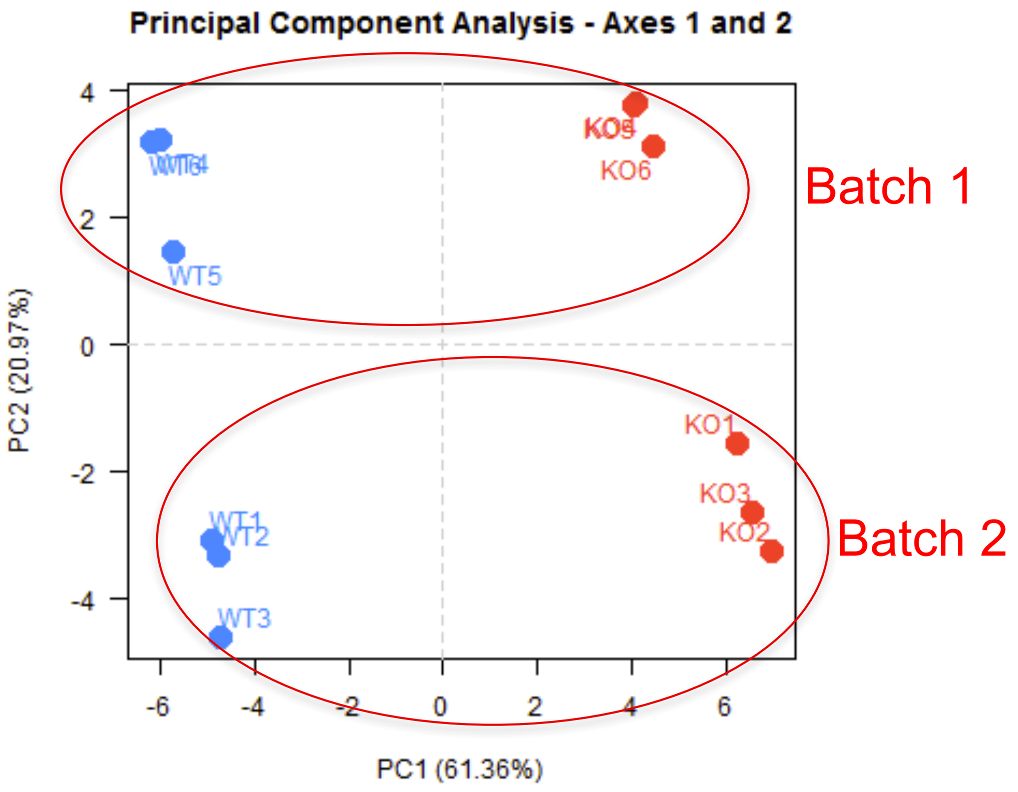
Data exploration on another dataset : outlier sample



Data exploration on another dataset : batch effect



Cluster dendrogram

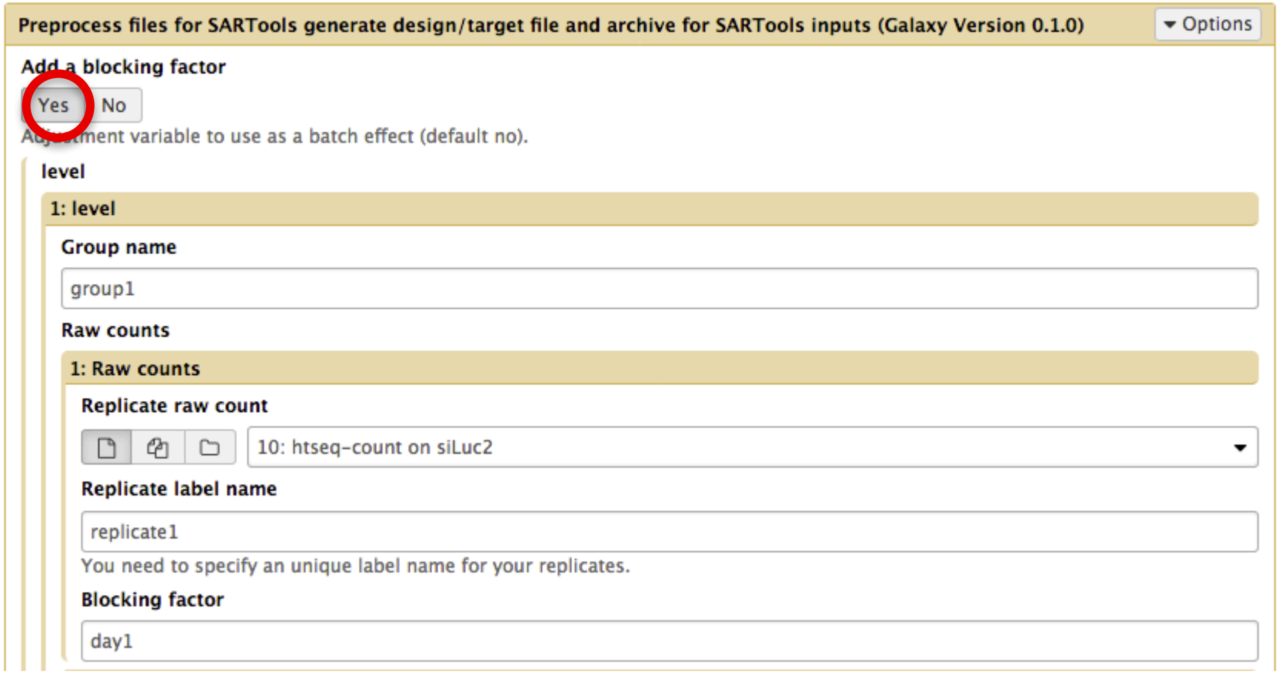


Principal Component Analysis - Axes 1 and 2

→ Take into account this batch effect in statistical analysis

Batch effect

- Preprocess files for SARTools



Preprocess files for SARTools generate design/target file and archive for SARTools inputs (Galaxy Version 0.1.0)

Add a blocking factor

[✓ Yes] No

Adjustment variable to use as a batch effect (default no).

level

1: level

Group name

group1

Raw counts

1: Raw counts

Replicate raw count

10: htseq-count on siLuc2

Replicate label name

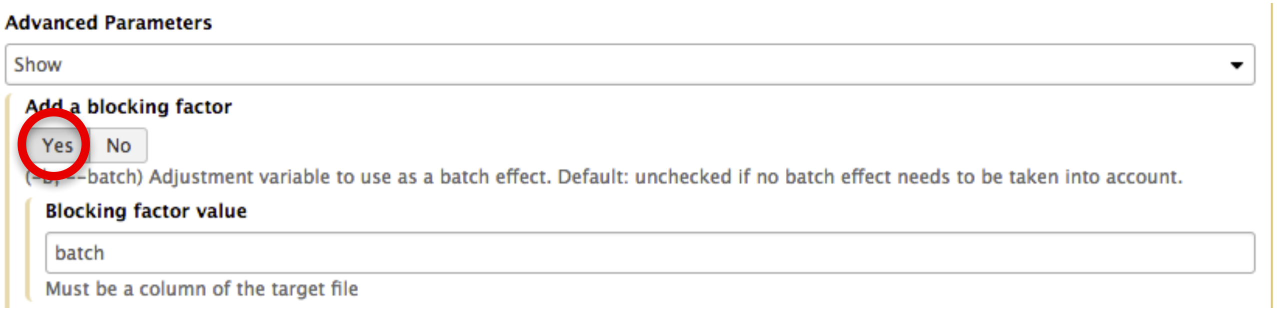
replicate1

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

Blocking factor

day1

- SARTools



Advanced Parameters

Show

Add a blocking factor

[✓ Yes] No

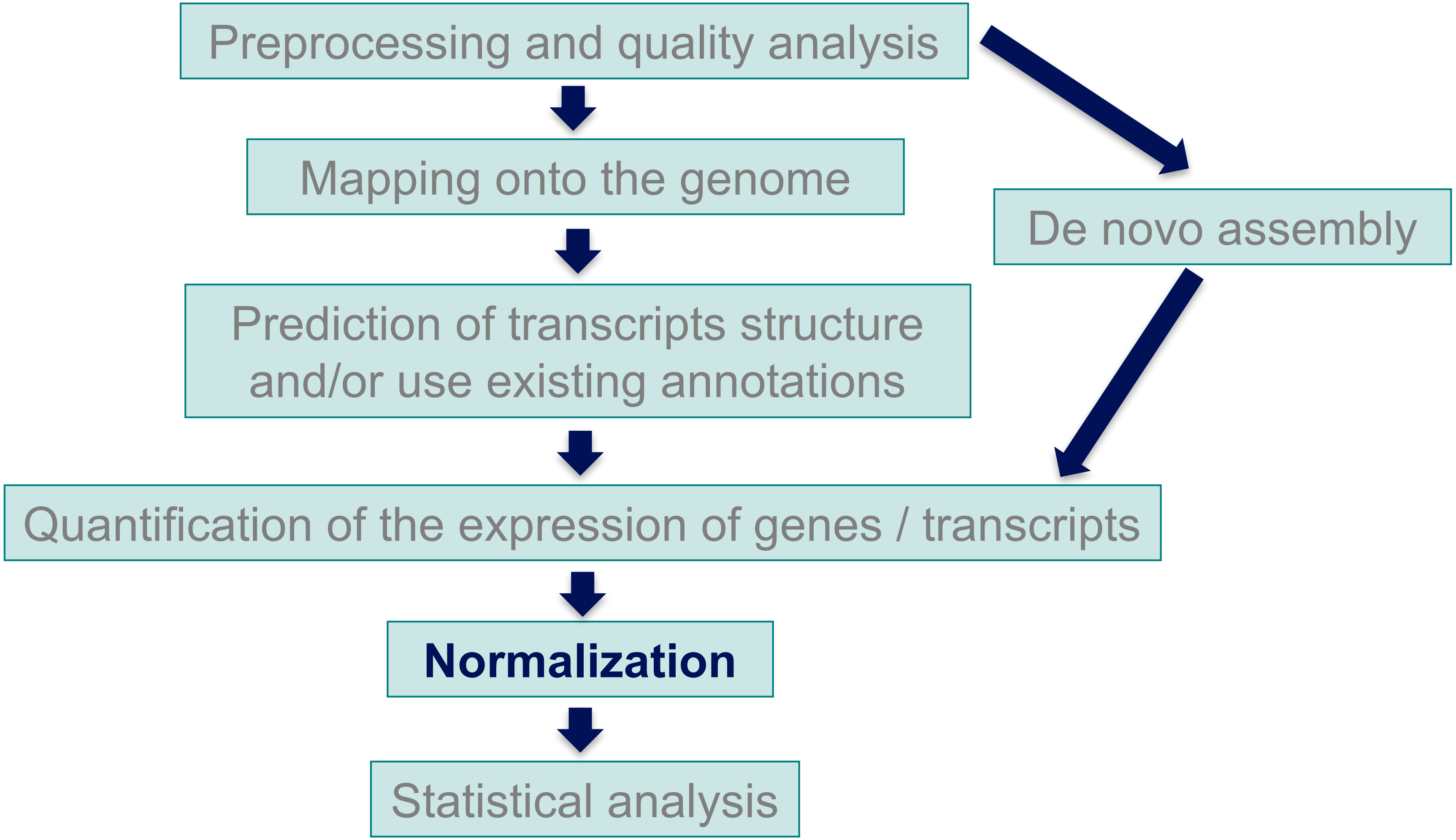
(-b, --batch) Adjustment variable to use as a batch effect. Default: unchecked if no batch effect needs to be taken into account.

Blocking factor value

batch

Must be a column of the target file

Analysis of RNA-seq data


Preprocessing and quality analysis
Mapping onto the genome
De novo assembly
Prediction of transcripts structure and/or use existing annotations
Quantification of the expression of genes / transcripts
Normalization
Statistical analysis

Normalization : why ?

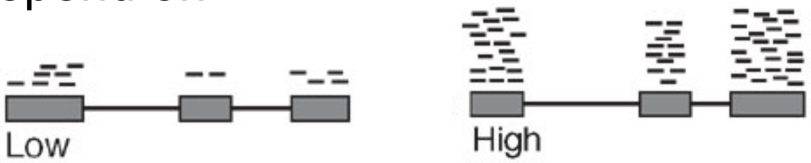
- To compare RNA-seq libraries
  - with different sizes, eg :

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

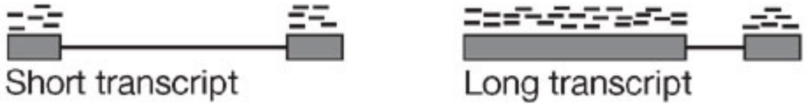
- To compare the expression level of several genes within a library

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

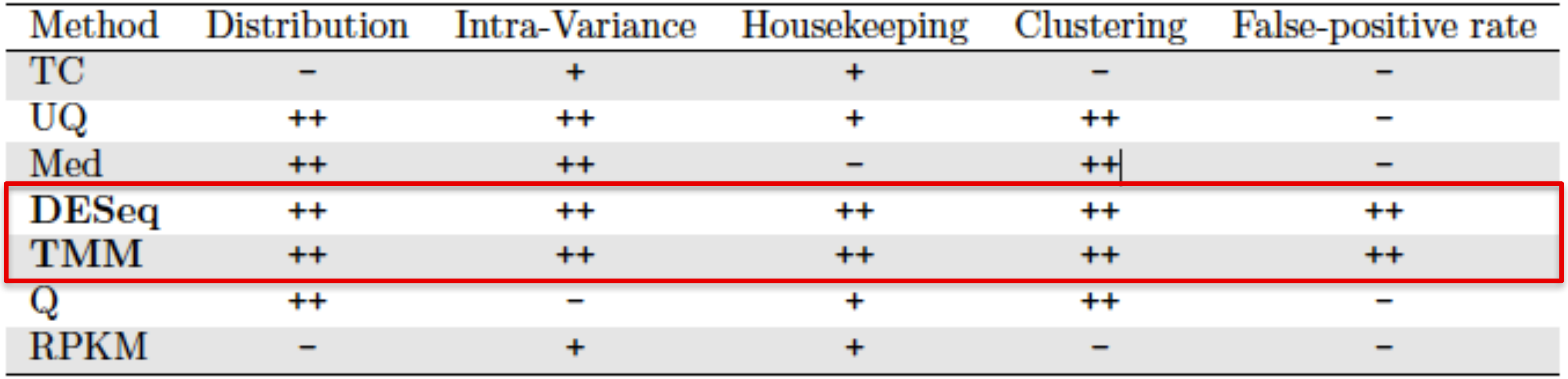
- Take into account gene/transcript length
  - RPKM (Mortazavi et al. Nat Methods 2008;5:621–8), FPKM
  - **R**eads (**F**ragments) per **K**ilobase per **M**illion mapped reads
  - Assumption
    - Read counts =f(expression level, gene length, library size)
  - Method
    - Divide counts by gene length (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**					$\mathbf{x_{ij}}$	

n : number of samples to compare

xij : number of reads
for gene i in sample j

Normalization factor for library j :

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

➔ 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 :

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

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 :

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

➔ Underlying idea : non-differentially expressed genes should have similar read count 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 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	

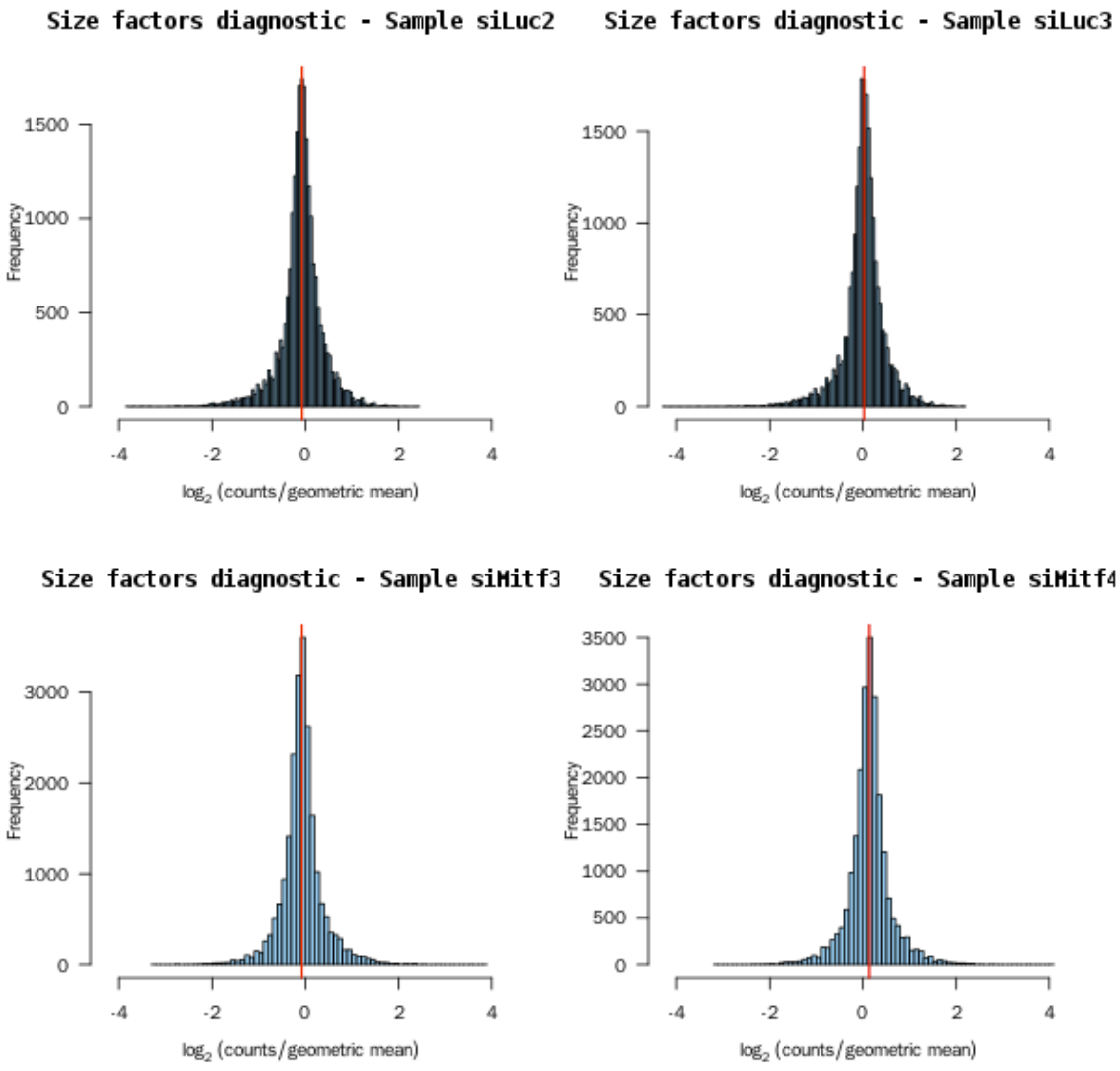
normalization factors

Normalization factor for library j :

$$\hat{s}_j = median_i \frac{x_{ij}}{\left(\prod_{v=1}^{n} x_{iv}\right)^{1/n}}$$

2. What are the values of these normalization factors for Mitf dataset ?

Diagnostic plot for the estimation of normalization factors



Size factors diagnostic - Sample siLuc2

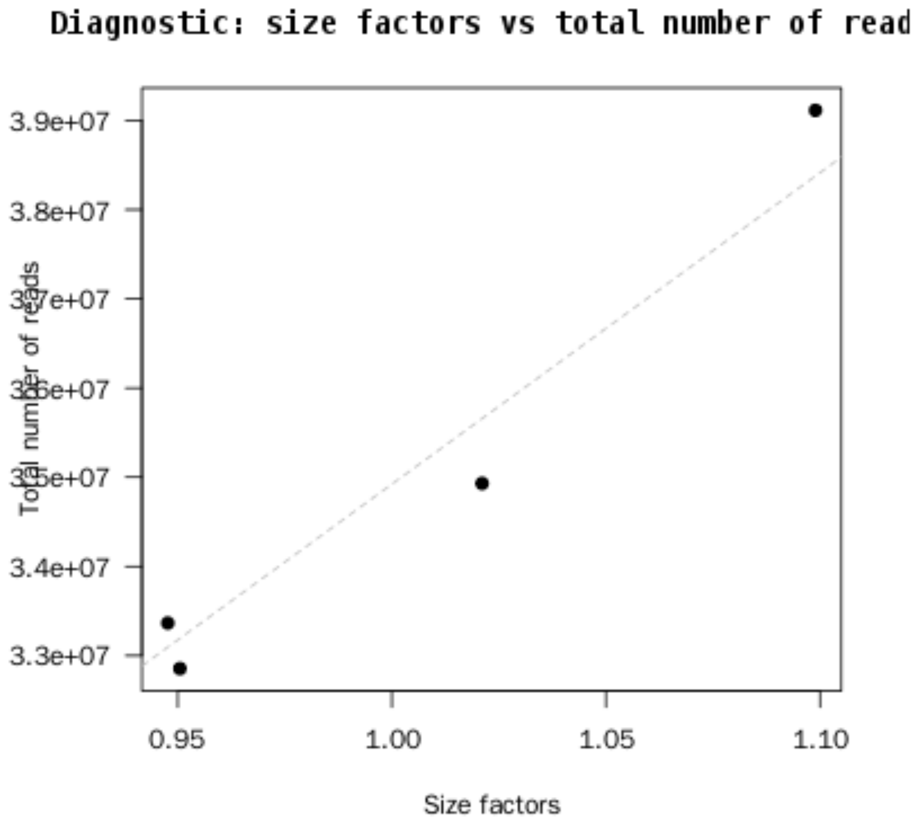
Size factors diagnostic - Sample siLuc3

Size factors diagnostic - Sample siMitf3

Size factors diagnostic - Sample siMitf4

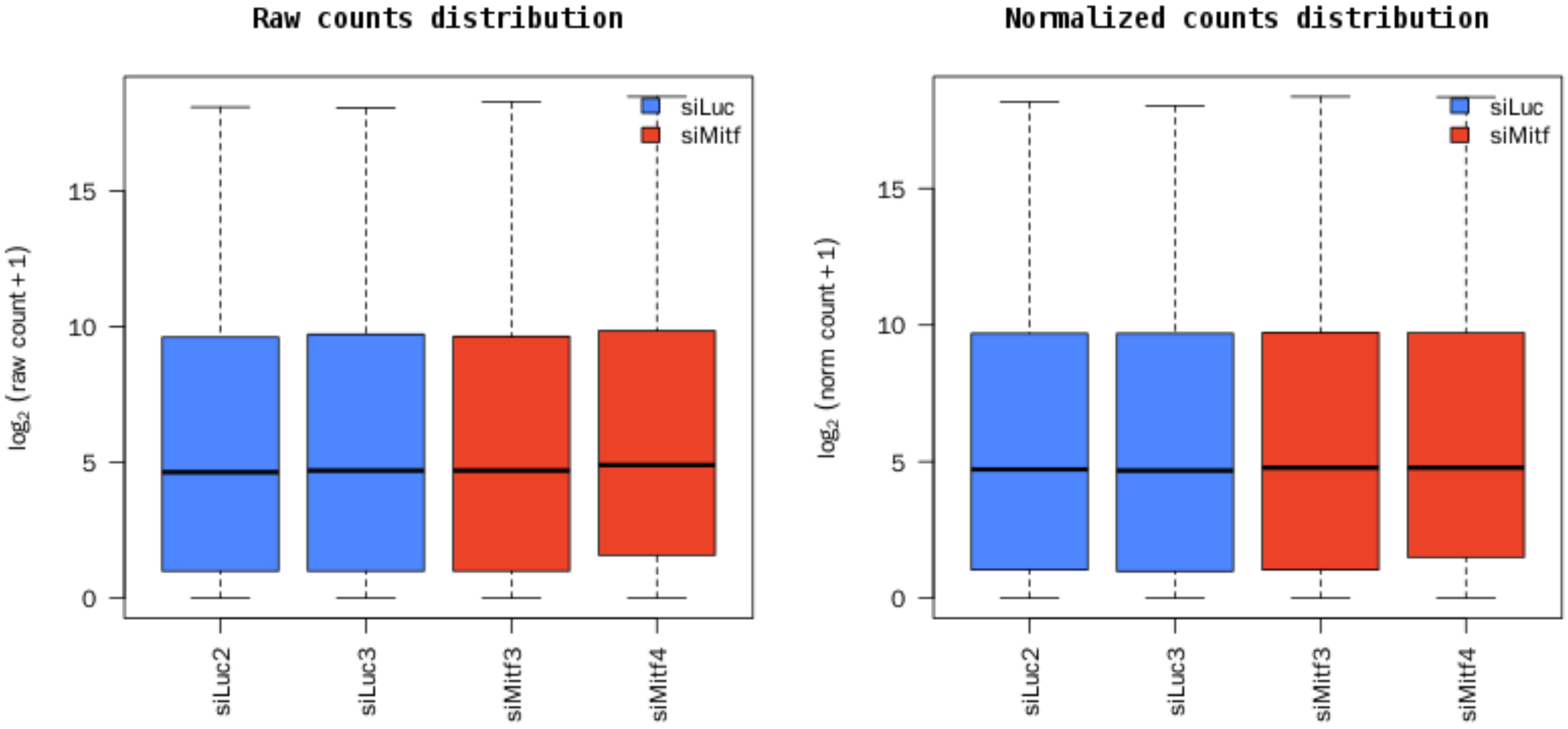
This histogram should be unimodal,
with a clear peak at the value of the size factor (represented in red)

Total number of reads vs size factors



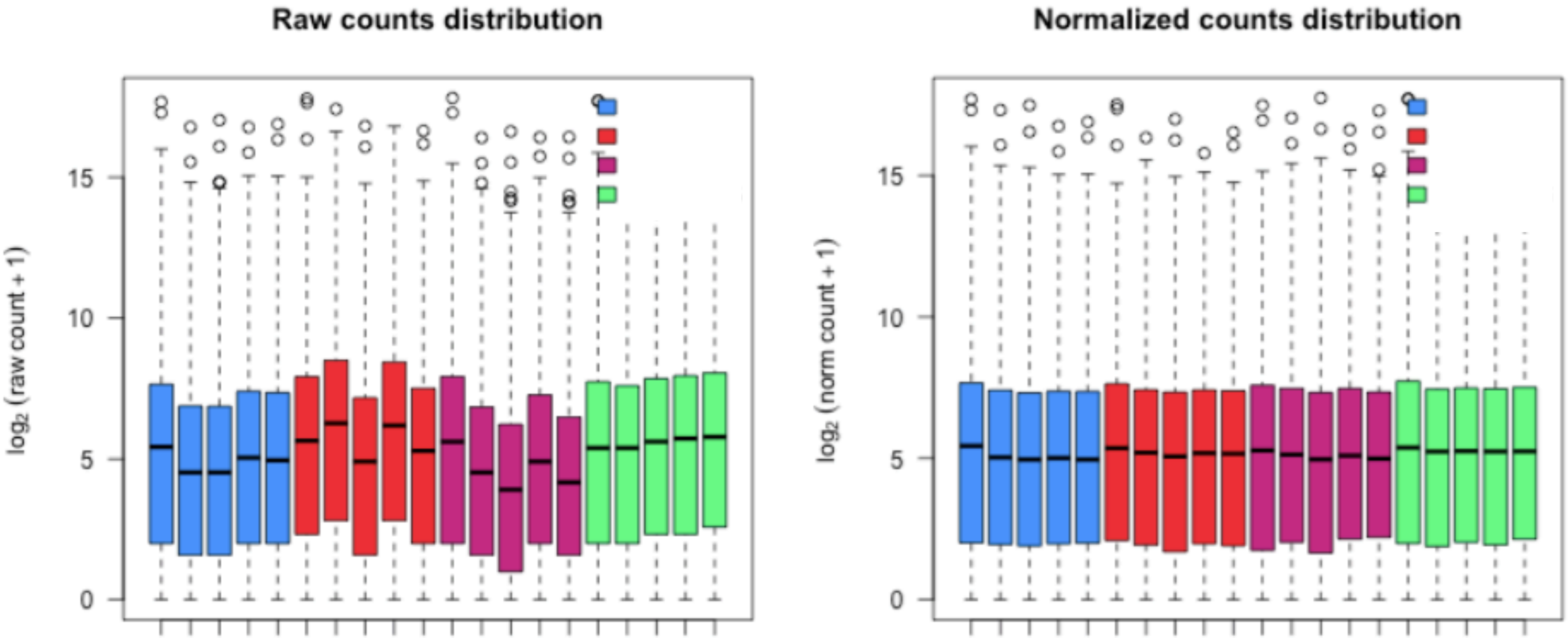
Normalization by total number of reads and DESeq size factors is not exactly the same, but very close for this dataset

Boxplots of raw and normalized read counts

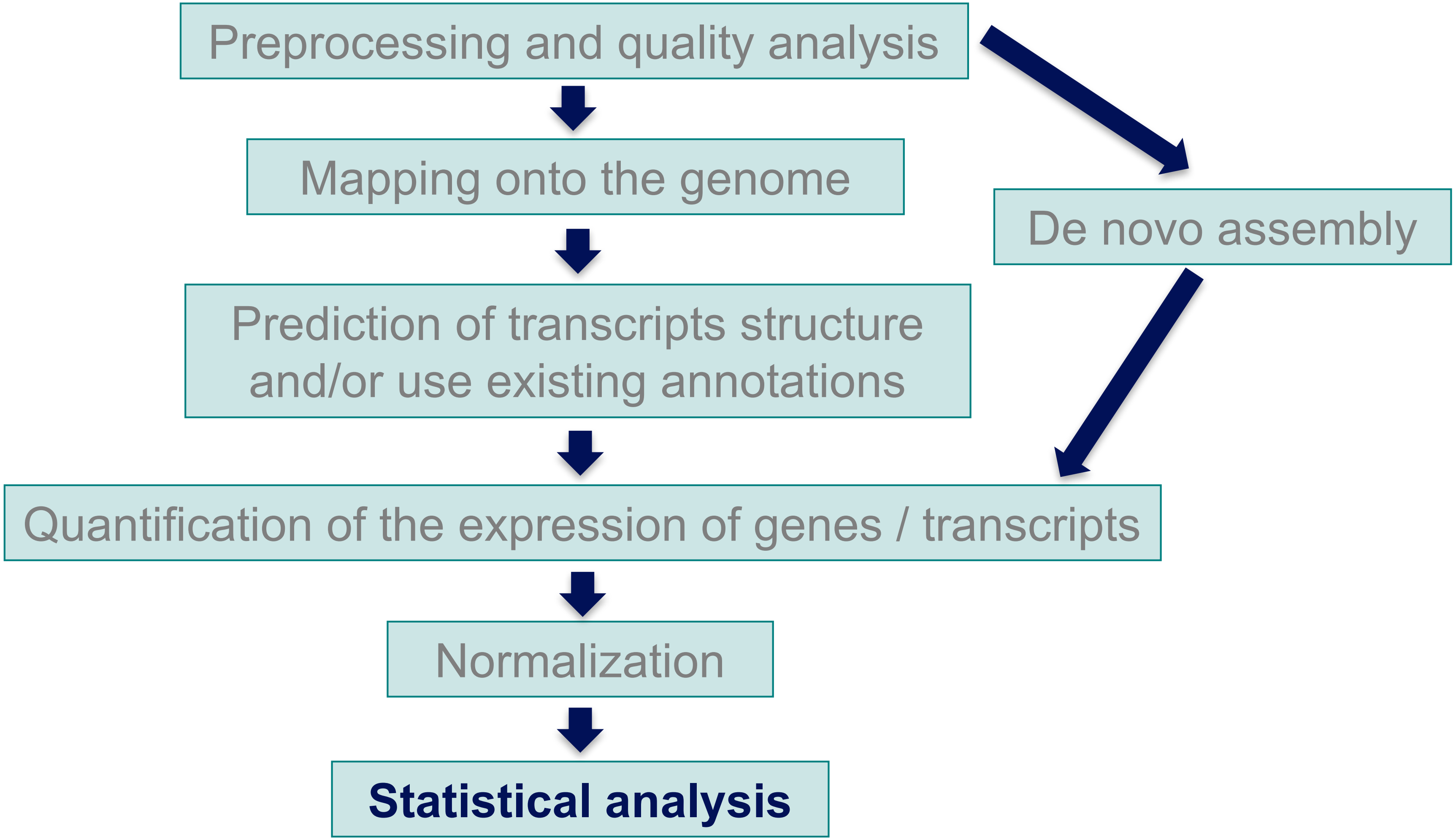


We expect normalization to stabilize distributions across samples

Boxplots of raw and normalized read counts on another dataset



Analysis of RNA-seq data



Preprocessing and quality analysis

Mapping onto the genome

De novo assembly

Prediction of transcripts structure and/or use existing annotations

Quantification of the expression of genes / transcripts

Normalization

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

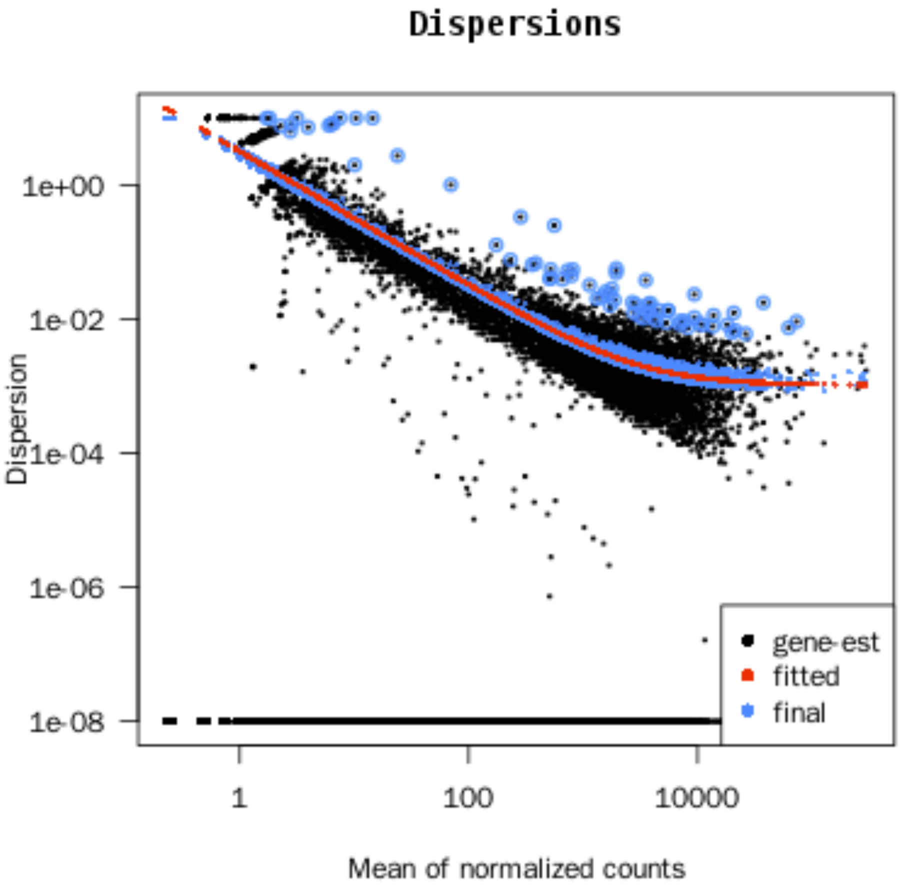
- How to estimate the overdispersion parameter ?
  - Very few replicates ➔ challenging issue
  - DESeq2 (Love et al. Genome Biol. 2014;15:550)

    Shares information across genes to improve the estimation of dispersion

    Assumes that genes of similar average expression strength have similar dispersion

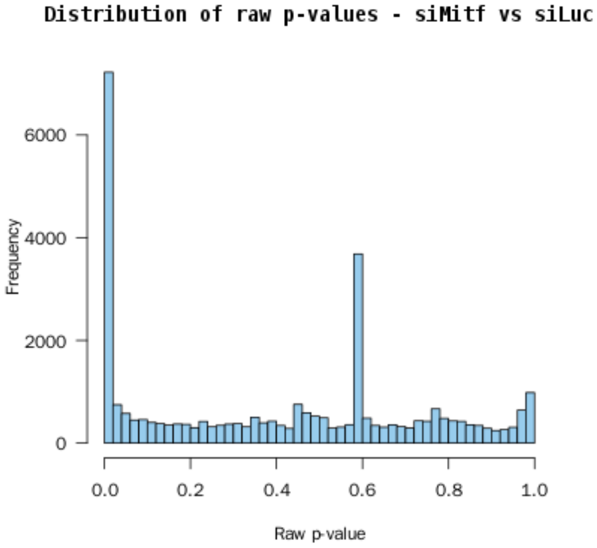
Dispersion plot

- **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)
- **Blue circles** : dispersions outliers → 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

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

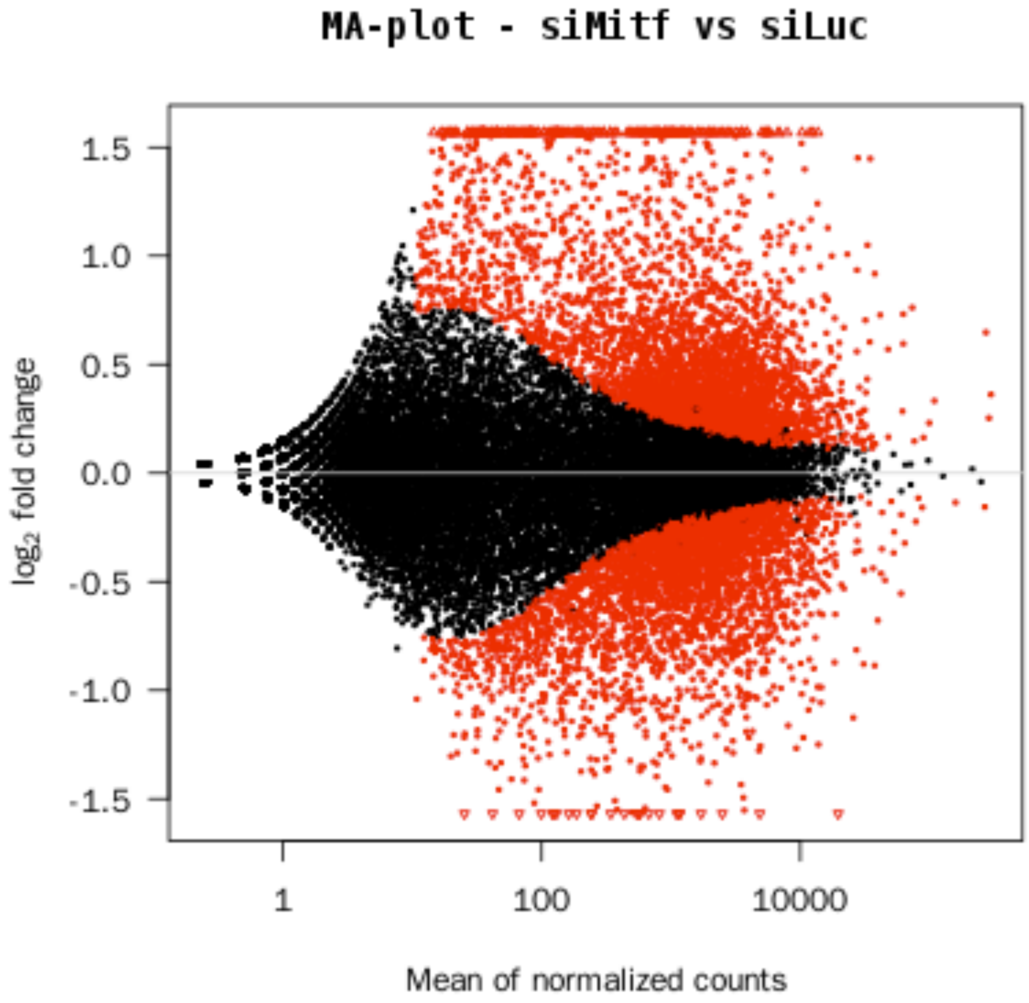
Independant filtering

- 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

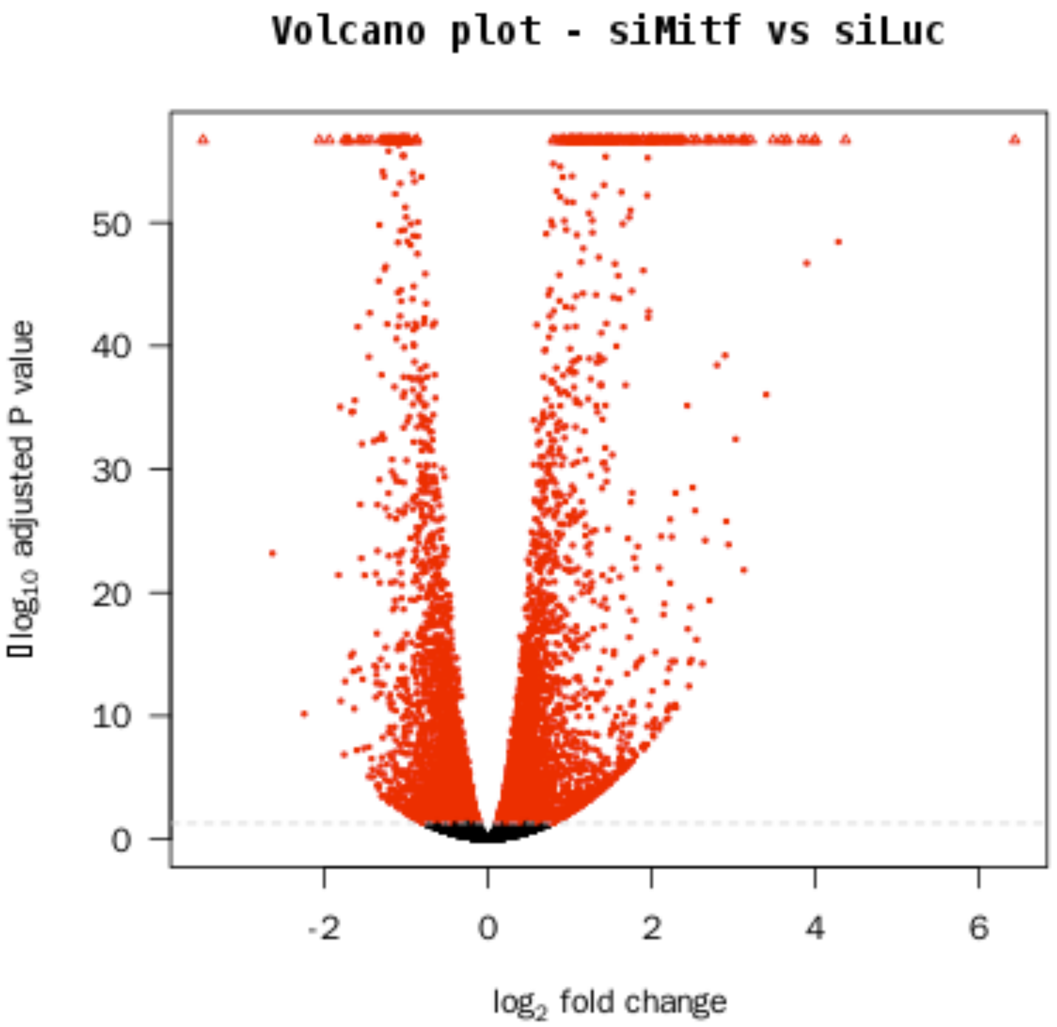
Visualization of significantly differentially expressed genes : MA-plot



Red dots : FDR < 0.05

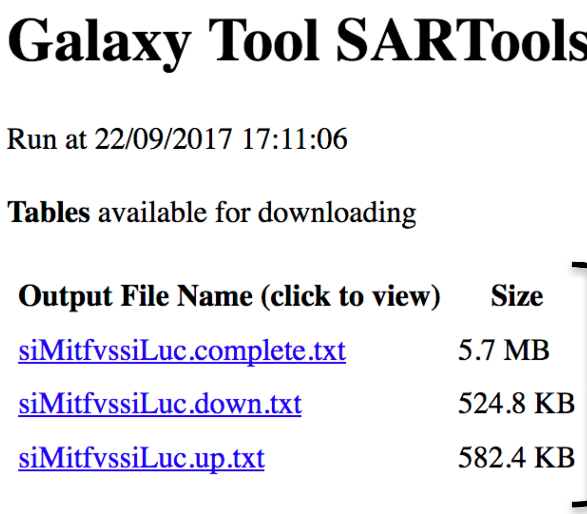
Triangles : features having a too low/high $\log_2$FC to be displayed on the plot

Visualization of significantly differentially expressed genes : volcano plot



Red dots : FDR < 0.05

Differential analysis results



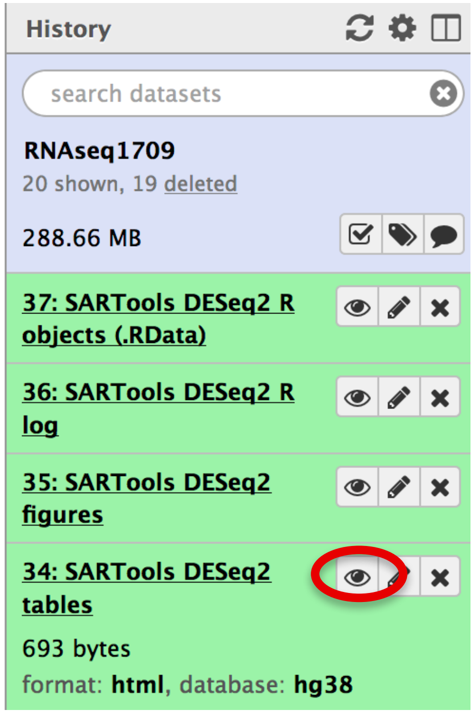
Galaxy Tool SARTools_DESeq2

Run at 22/09/2017 17:11:06

Tables available for downloading

Output File Name (click to view)	Size
siMitfvssiLuc.complete.txt	5.7 MB
siMitfvssiLuc.down.txt	524.8 KB
siMitfvssiLuc.up.txt	582.4 KB

Tabulated text files



RNAseq1709
20 shown, 19 deleted

288.66 MB

37: SARTools DESeq2 R objects (.RData)

36: SARTools DESeq2 R log

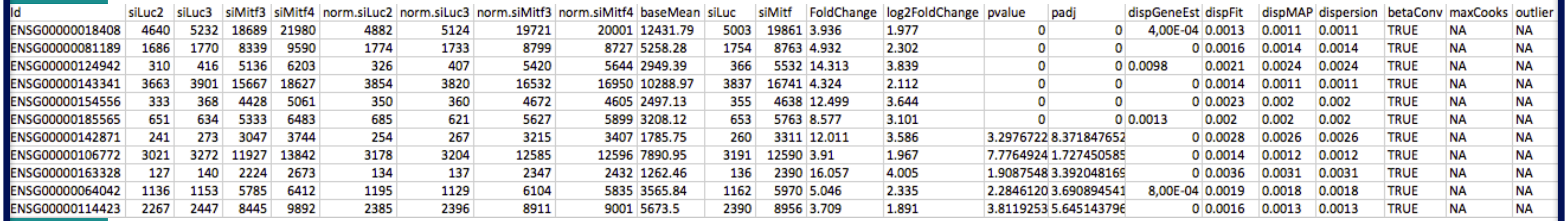
35: SARTools DESeq2 figures

34: SARTools DESeq2 tables

693 bytes
format: **html**, database: **hg38**

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

Differential analysis results

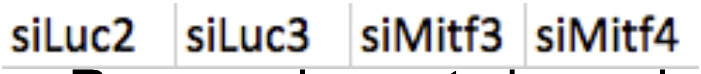


Id	siLuc2	siLuc3	siMitf3	siMitf4	norm.siLuc2	norm.siLuc3	norm.siMitf3	norm.siMitf4	baseMean	siLuc	siMitf	FoldChange	log2FoldChange	pvalue	padj	dispGeneEst	dispFit	dispMAP	dispersion	betaConv	maxCooks	outlier
ENSG00000018408	4640	5232	18689	21980	4882	5124	19721	20001	12431.79	5003	19861	3.936	1.977	0	0	4,00E-04	0.0013	0.0011	0.0011	TRUE	NA	NA
ENSG00000081189	1686	1770	8339	9590	1774	1733	8799	8727	5258.28	1754	8763	4.932	2.302	0	0	0	0.0016	0.0014	0.0014	TRUE	NA	NA
ENSG00000124942	310	416	5136	6203	326	407	5420	5644	2949.39	366	5532	14.313	3.839	0	0	0.0098	0.0021	0.0024	0.0024	TRUE	NA	NA
ENSG00000143341	3663	3901	15667	18627	3854	3820	16532	16950	10288.97	3837	16741	4.324	2.112	0	0	0	0.0014	0.0011	0.0011	TRUE	NA	NA
ENSG00000154556	333	368	4428	5061	350	360	4672	4605	2497.13	355	4638	12.499	3.644	0	0	0	0.0023	0.002	0.002	TRUE	NA	NA
ENSG00000185565	651	634	5333	6483	685	621	5627	5899	3208.12	653	5763	8.577	3.101	0	0	0.0013	0.002	0.002	0.002	TRUE	NA	NA
ENSG00000142871	241	273	3047	3744	254	267	3215	3407	1785.75	260	3311	12.011	3.586	3.2976722	8.371847652	0	0.0028	0.0026	0.0026	TRUE	NA	NA
ENSG00000106772	3021	3272	11927	13842	3178	3204	12585	12596	7890.95	3191	12590	3.91	1.967	7.7764924	1.727450585	0	0.0014	0.0012	0.0012	TRUE	NA	NA
ENSG00000163328	127	140	2224	2673	134	137	2347	2432	1262.46	136	2390	16.057	4.005	1.9087548	3.392048169	0	0.0036	0.0031	0.0031	TRUE	NA	NA
ENSG00000064042	1136	1153	5785	6412	1195	1129	6104	5835	3565.84	1162	5970	5.046	2.335	2.2846120	3.690894541	8,00E-04	0.0019	0.0018	0.0018	TRUE	NA	NA
ENSG00000114423	2267	2447	8445	9892	2385	2396	8911	9001	5673.5	2390	8956	3.709	1.891	3.8119253	5.645143796	0	0.0016	0.0013	0.0013	TRUE	NA	NA

→ 1 line per gene (Id = Ensembl gene id)

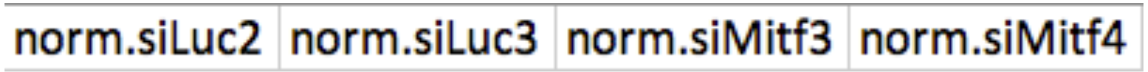
→ 23 columns

Differential analysis results



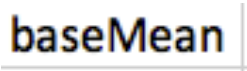
siLuc2	siLuc3	siMitf3	siMitf4

- Raw read counts in each sample



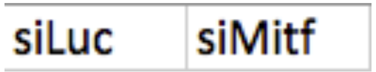
norm.siLuc2	norm.siLuc3	norm.siMitf3	norm.siMitf4

- Rounded normalized counts in each sample



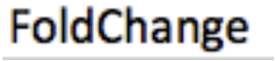
baseMean

- Mean of normalized counts over all samples



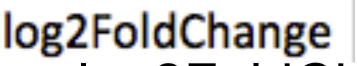
siLuc	siMitf

- Rounded mean of normalized counts over siLuc/siMitf samples



FoldChange

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

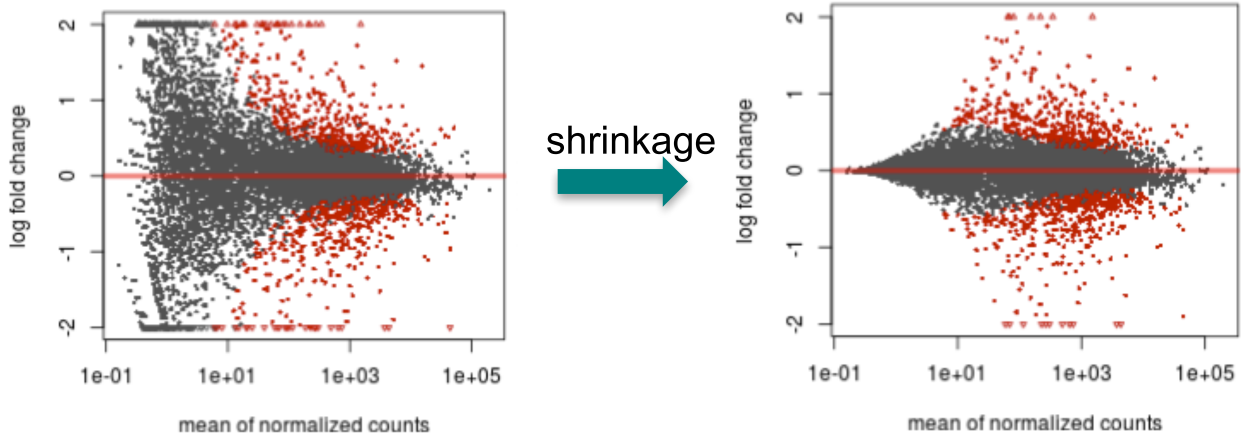


log2FoldChange

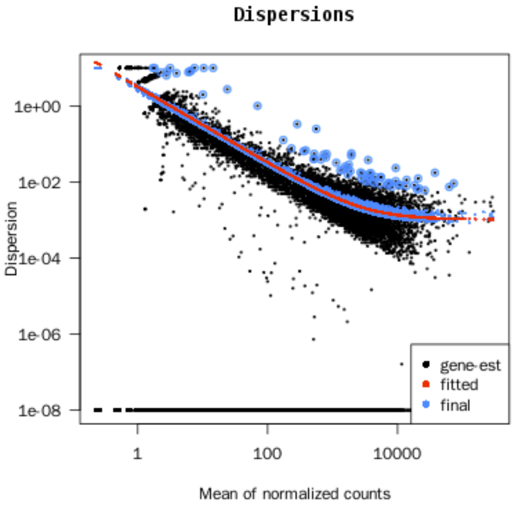
- log2FoldChange estimated by the generalized linear model
  - Reflects the differential expression between siMitf and siLuc
  - ~0 → similar gene expression in both conditions
  - >0 → over-expressed gene (siMitf > siLuc)
  - <0 → under-expressed gene (siMitf < siLuc)

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 (e.g. 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



pvalue	padj

- 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

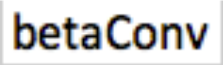
dispMAP

- 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 and circles 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	outlier

- 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 which have 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