

RNA-Seq Analysis Training



Introduction	2
Connecting to your Galaxy instance	2
Data analysis De novo - rsem	2
Import the data into a new history	2
Preprocessing phase with SARTools.	3
Analysis with Sartools	5
Analyze "Lobel" data with SARTools	8
Import data into new history	8
Analysis with SARTools	9
Data analysis "Stats Smash chr18" with SARTools	15
Import the data in a new history	15
Preprocessing phase with SARTools.	17
Analysis with Sartools	18
GSEA Analysis	21
Data Preparation	21
Analysis of data from the recount project	27
Retrieving data via the recount tool	27
Exercise: Perform the differential analysis between tum-IMMC and adj-IMMC conditions.	27
GSEA analysis on the msigdb site	28
Preparation of the data:	28
Exercise: Realizing the GSEA analysis	30

Introduction

Connecting to your Galaxy instance

- Open an internet browser and type in the IP address given by the trainers in the address bar.
- the following home page is displayed:



- Click on User -> login and enter the ID userX@bilille.fr and the password "bililleuser" (with X a number between 1 and 6 given by the trainer)
- The home page is displayed again. Click on User and "logged in as userX@bilille.fr" should appear. (with X a figure between 1 and 6 given by the trainer)

Data analysis De novo - rsem

Import the data into a new history

- Within Galaxy, click on Shared data -> Data library
- Click on De novo - rsem
- Select the 4 .txt files

Galaxy / Galaxy-RNA-seq Analyse de données Workflow Données partagées Visualisation Admin Aide Utilisateur Using 104.6 MB

DATA LIBRARIES 4 items shown (change) 4 total Include deleted [+ Create Folder](#) [+ Add](#) [To History](#) [Download](#) [Delete](#) [Details](#) [Help](#)

Libraries / De novo - rsem

<input checked="" type="checkbox"/>	name	description	data type	size	time updated (UTC)	state
<input checked="" type="checkbox"/>	rsem_sample.gene_abundances_C1		txt	53.4 KB	2018-08-14 02:30 PM	Manage
<input checked="" type="checkbox"/>	rsem_sample.gene_abundances_C2		txt	53.9 KB	2018-08-14 02:30 PM	Manage
<input checked="" type="checkbox"/>	rsem_sample.gene_abundances_T1		txt	54.0 KB	2018-08-14 02:30 PM	Manage
<input checked="" type="checkbox"/>	rsem_sample.gene_abundances_T2		txt	54.1 KB	2018-08-14 02:30 PM	Manage

4 items shown (change) 4 total

- Click on "To History" then "as Datasets"
- Give a name to the new history in the "or create new" field (eg De novo rsem).

Galaxy / Galaxy-RNA-seq Analyse de données Workflow Données partagées Visualisation Admin Aide Utilisateur Using 104.6 MB

DATA LIBRARIES 4 items shown (change) 4 total [Delete](#) [Details](#) [Help](#)

Libraries / De novo - rsem

Importer dans l'historique

Select history:

or create new:

[Import](#) [Close](#)

4 items shown (change) 4 total

- Your new history now appears in the "Analyze Data" section.

Galaxy / Galaxy-RNA-seq Analyse de données Workflow Données partagées Visualisation Admin Aide Utilisateur Using 104.6 MB

Tools [search tools](#)

[Get Data](#)
[Send Data](#)
[Collection Operations](#)
[Lift-Over](#)
[Text Manipulation](#)
[Filter and Sort](#)
[Join, Subtract and Group](#)
[Convert Formats](#)
[Extract Features](#)
[Fetch Sequences](#)
[Fetch Alignments](#)
[Statistics](#)
[Graph/Display Data](#)
[NGS: Differential Analysis](#)
[SAM Tools](#)
[BCFtools](#)
[NGS: Reads Manipulation](#)
[NGS: Mapping](#)
[NGS: Transcriptomics](#)

Bienvenue!

Vous êtes actuellement sur une instance Galaxy dédiée à l'analyse RNA-seq. Elle a été développée avec Docker, spécialement pour les formations de la plateforme BILILLE et est déployée sur le Cloud BILILLE.

[Guided Tour »](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors. The Galaxy Docker project is supported by the University of Freiburg, part of de NBI. The Galaxy Project is supported in part by MHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

History [Refresh](#) [Settings](#) [Close](#)

Rechercher des données

De novo - rsem
4 shown
215.4 KB [Check](#) [Refresh](#) [Close](#)

4: [View](#) [Edit](#) [Close](#)
rsem_sample.gene_abundances_T2

3: [View](#) [Edit](#) [Close](#)
rsem_sample.gene_abundances_T1

2: [View](#) [Edit](#) [Close](#)
rsem_sample.gene_abundances_C2

1: [View](#) [Edit](#) [Close](#)
rsem_sample.gene_abundances_C1

The data correspond to RNA-Seq count data from rsem for 4 replicates under 2 conditions: 2 replicates per condition, C (Control) and T (Treatment).

Preprocessing phase with SARTools.

Goal: This step creates datasets adapted to SARTools.

In the Tools panel, in the NGS: Differential Analysis section, click on the "preprocess files for SARTools" tool.

- Create 2 groups: Control and Treatment and add the 2 corresponding replicates to each of the 2 conditions
- Choose different replicate names for each replicate (for example repC_1, rep C_2 for the Control group and repT_1, rep T_2 for the Treatment group)

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

Add a blocking factor
 Yes No
Adjustment variable to use as a batch effect (default no).

Group

1: Group

Group name
Treatment

Raw counts

1: Raw counts

Replicate raw count
3: rsem_sample.gene_abundances_T1

Replicate label name
repT_1
You need to specify a unique label name for your replicates.

2: Raw counts

Replicate raw count
4: rsem_sample.gene_abundances_T2

Replicate label name
repT_2
You need to specify a unique label name for your replicates.

2: Group

Group name
Control

Raw counts

1: Raw counts

Replicate raw count
1: rsem_sample.gene_abundances_C1

Replicate label name
repC_1
You need to specify a unique label name for your replicates.

2: Raw counts

Replicate raw count
2: rsem_sample.gene_abundances_C2

Replicate label name
repC_2
You need to specify a unique label name for your replicates.

The tool returns 2 outputs

- a design file containing the conditions of the experiment in the .txt format.

1	2	3
label	files	group
repT_1	dataset_216.dat	Treatment
repT_2	dataset_217.dat	Treatment
repC_1	dataset_214.dat	Control
repC_2	dataset_215.dat	Control

- a .zip file containing all the count files.

Analysis with Sartools

Goal : Perform differential analysis of loaded data.

In the panel tool, click on the "SARTools DESeq2" tool

- Fill in the design / target file and the Zip file containing the raw counts.
- In the field "Reference biological condition" enter the value "Control" corresponding to the reference condition of the data.
- Leave the other fields unchanged.

SARTools DESeq2 Compare two or more biological conditions in a RNA-Seq framework with DESeq2 (Galaxy Version 1.3.2.0) Options

Name of the project used for the report

 (-P, --projectName) No space allowed.

Name of the report author

 (-A, --author) No space allowed.

Design / target file
 9: design file for SARTools (on data 2, data 1, and others) ▼
 (-t, --targetFile) See the help section below for details on the required format.

Zip file containing raw counts files
 10: counts files for SARTools (on data 2, data 1, and others) ▼
 (-r, --rawDir) See the help section below for details on the required format.

Names of the features to be removed

 (-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

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

Reference biological condition

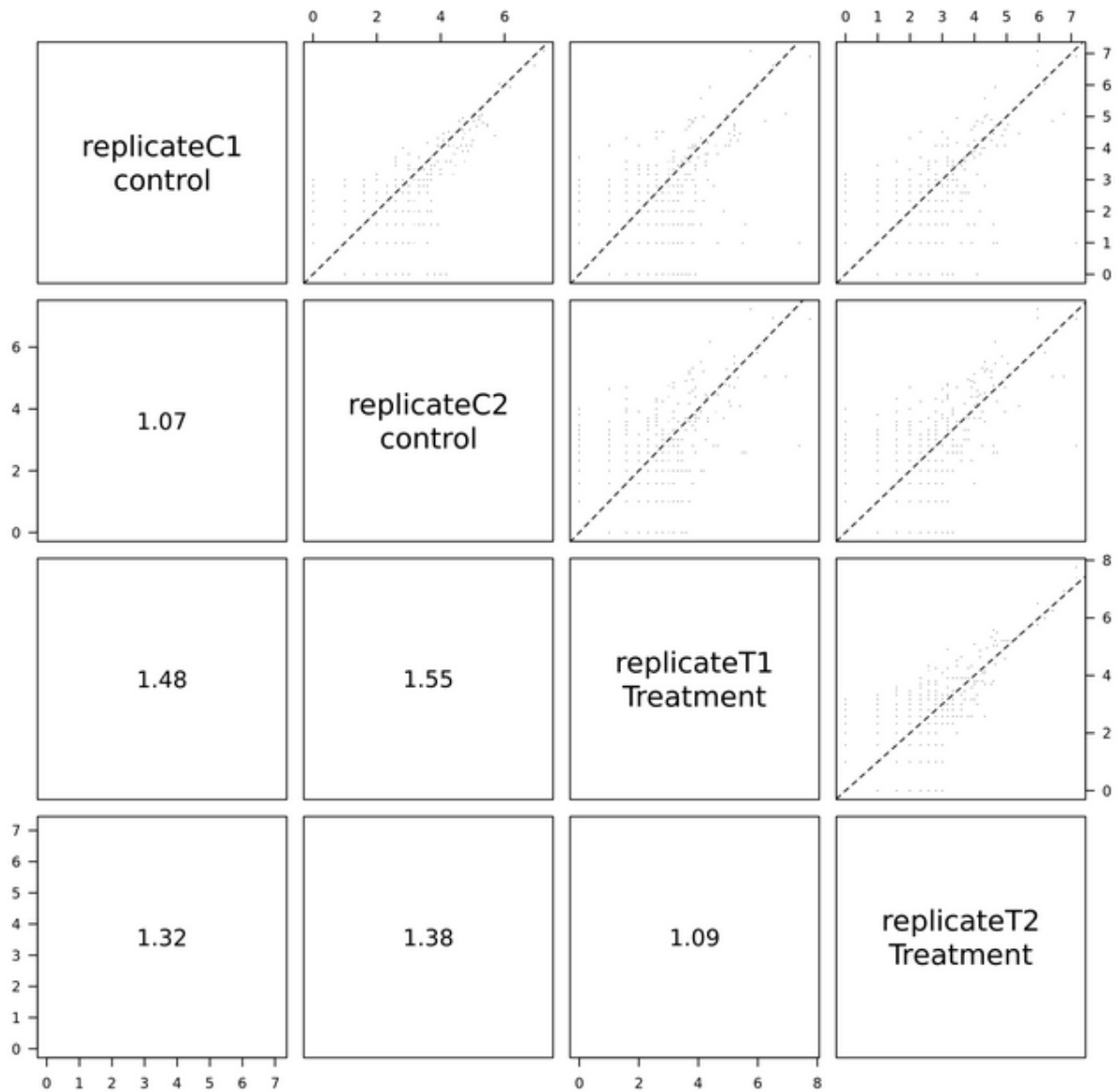
 (-c, --condRef) Reference biological condition used to compute fold-changes, must be one of the levels of 'Factor of interest'.

Advanced Parameters
 ▼

SARTools returns 5 datasets:

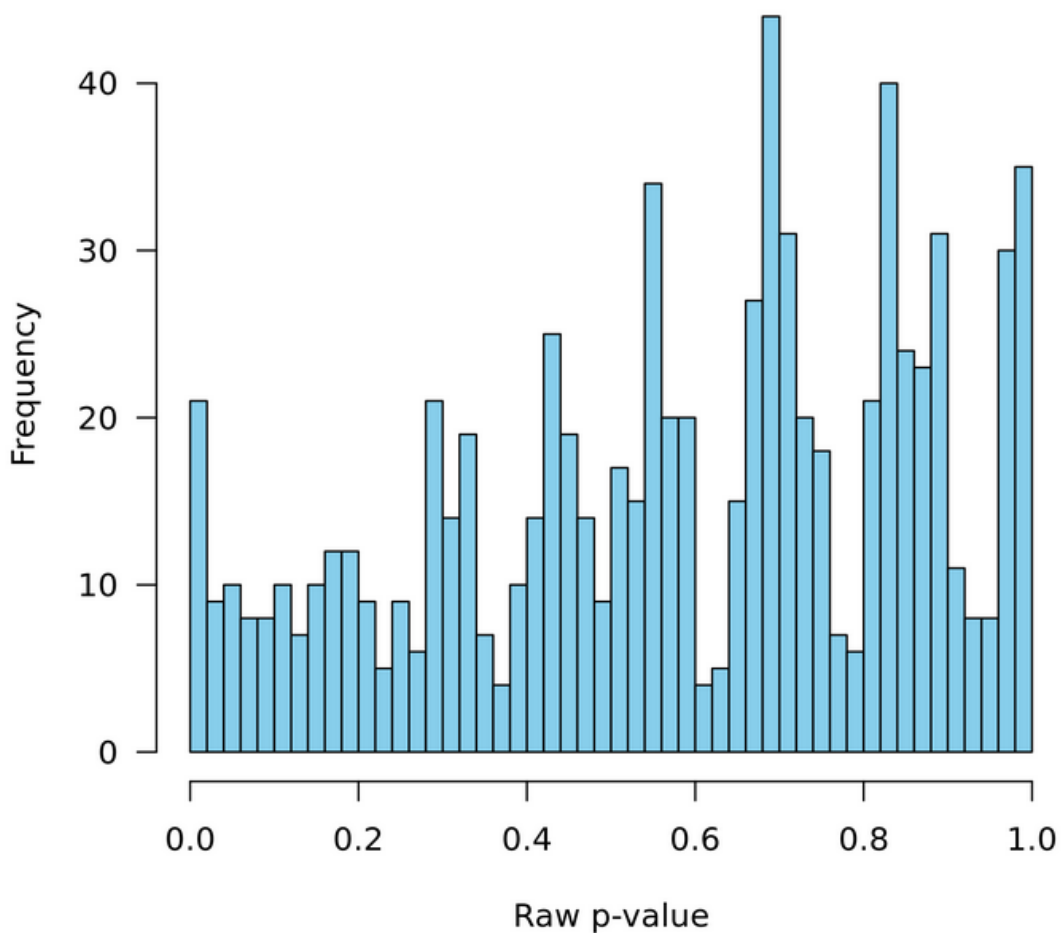
- A complete report in HTML to evaluate the quality of the data and the quality of differential analysis
- The complete list of the analyzed genes, as well as the under-expressed and over-expressed genes.
- The list of figures used in the HTML report
- A log check that all the stages of the analysis went well.
- An RData object that can be exported and used under R.

In the report, we can check that the SERE indicators are very close to 1, suggesting that the replicates of this dataset are technical and non-biological replicates.



A little later in the report, the histogram of raw p-values suggests that the statistical model used is not appropriate for this dataset.

Distribution of raw p-values - Treatment vs contro



Analyze "Lobel" data with SARTools

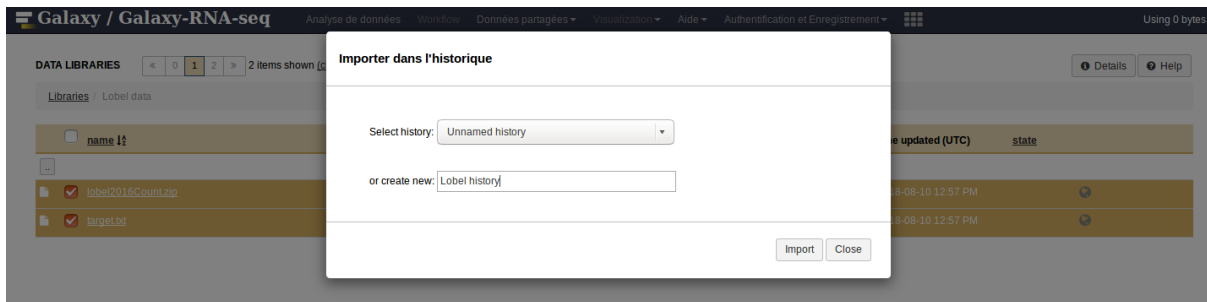
Import data into new history

- Under Galaxy click Shared Data -> Data Library
- Click Lobel Data.
- Select the 2 files "lobel2016Count.zip" and "target.txt".

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy / Galaxy-RNA-seq' and various utility links. Below the navigation bar, the 'DATA LIBRARIES' section is active, showing 'Lobel data' with 2 items shown. The table below lists the files:

<input type="checkbox"/>	name ↓	description	data type	size	time updated (UTC)	state
<input checked="" type="checkbox"/>	lobel2016Count.zip		zip	155.5 KB	2018-08-10 12:57 PM	📄
<input checked="" type="checkbox"/>	target.txt		txt	534 bytes	2018-08-10 12:57 PM	📄

- Click on "To History" then "as Datasets"
- Give a name to the new history in the "or create new" field (eg "Lobel history")



- Your new history now appears in the "Analyze Data" section.



The lobel.zip file contains counts from Lobel L, Herskovits AA (2016) Systems Level Analyzes Reveal Multiple Regulatory Activities of CodY Controlling Metabolism, Motility and Virulence in *Listeria monocytogenes*. PLoS Genet 12 (2): e1005870. doi: 10.1371 / journal.pgen.1005870.

The target.txt file contains the description of the conditions of the experiment for its analysis by SARTools: 11 replicates for 2 conditions (6 WT for 5 codY)

Analysis with SARTools

- Fill in the design / target file and the Zip file containing the raw counts.
- In the field "Factor of interest" enter the value "strain" corresponding to the 3rd column of the file target and containing the 2 conditions to be compared.
- In the "Reference biological condition" field enter the value "WT".
- Leave the other fields unchanged.

SARTools DESeq2 Compare two or more biological conditions in a RNA-Seq framework with DESeq2 (Galaxy Version 1.3.2.0) Options

Name of the project used for the report

 (-P, --projectName) No space allowed.

Name of the report author

 (-A, --author) No space allowed.

Design / target file

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

Zip file containing raw counts files

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

Names of the features to be removed

 (-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

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

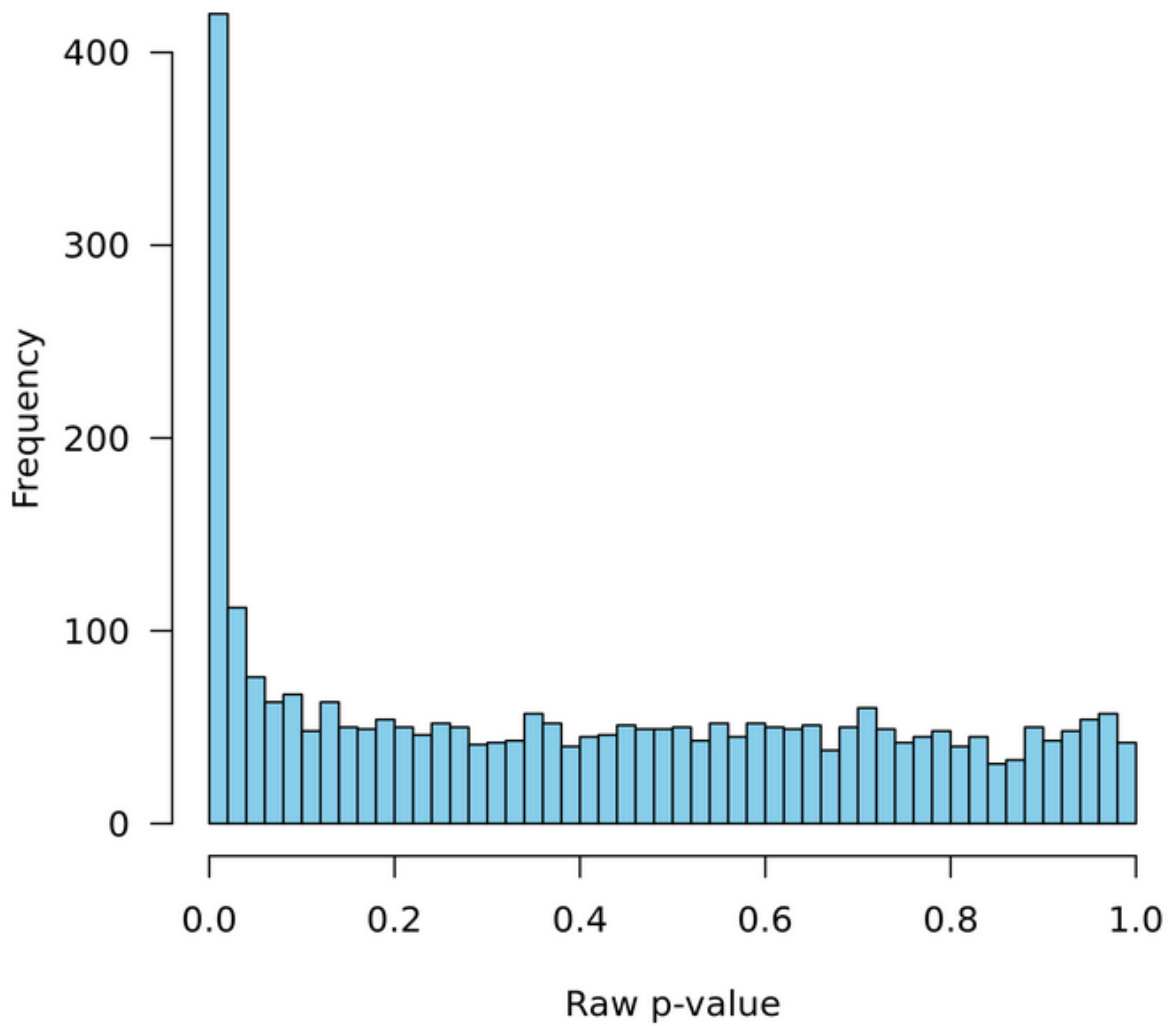
Reference biological condition

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

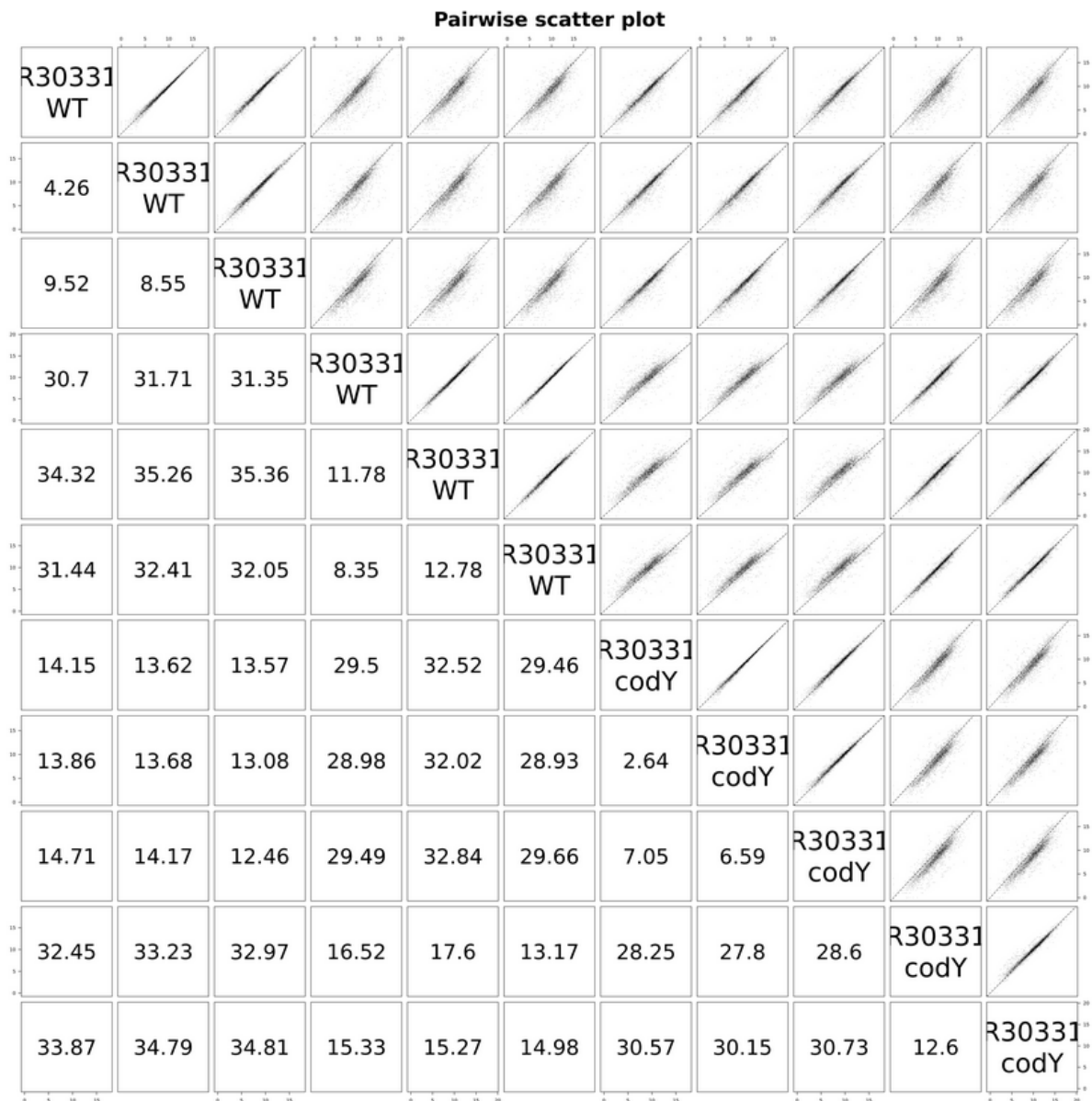
Advanced Parameters

In the generated report, we notice that the histogram of the raw p-values has an expected shape: a left peak corresponding to the differentially expressed genes and a uniform distribution elsewhere.

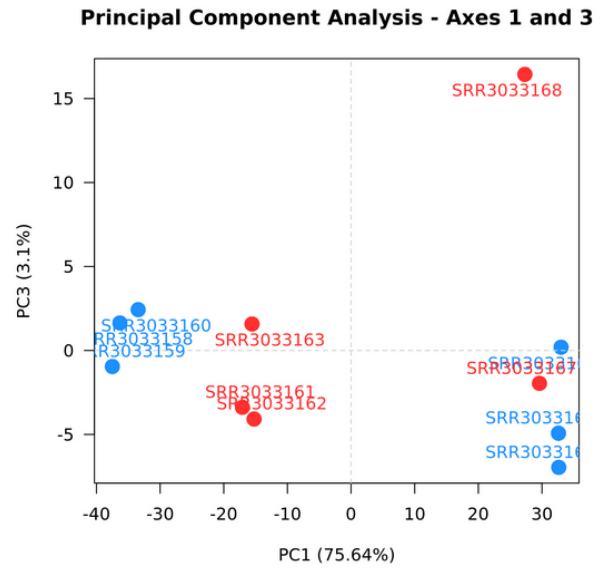
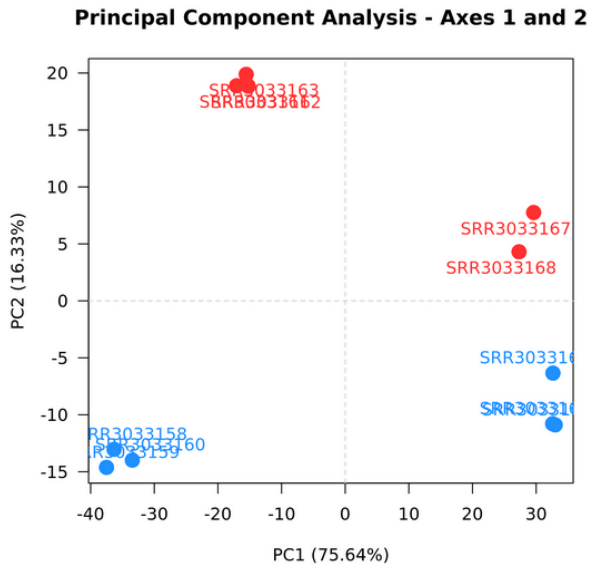
Distribution of raw p-values - codY vs WT



Let see the exploratory analysis:

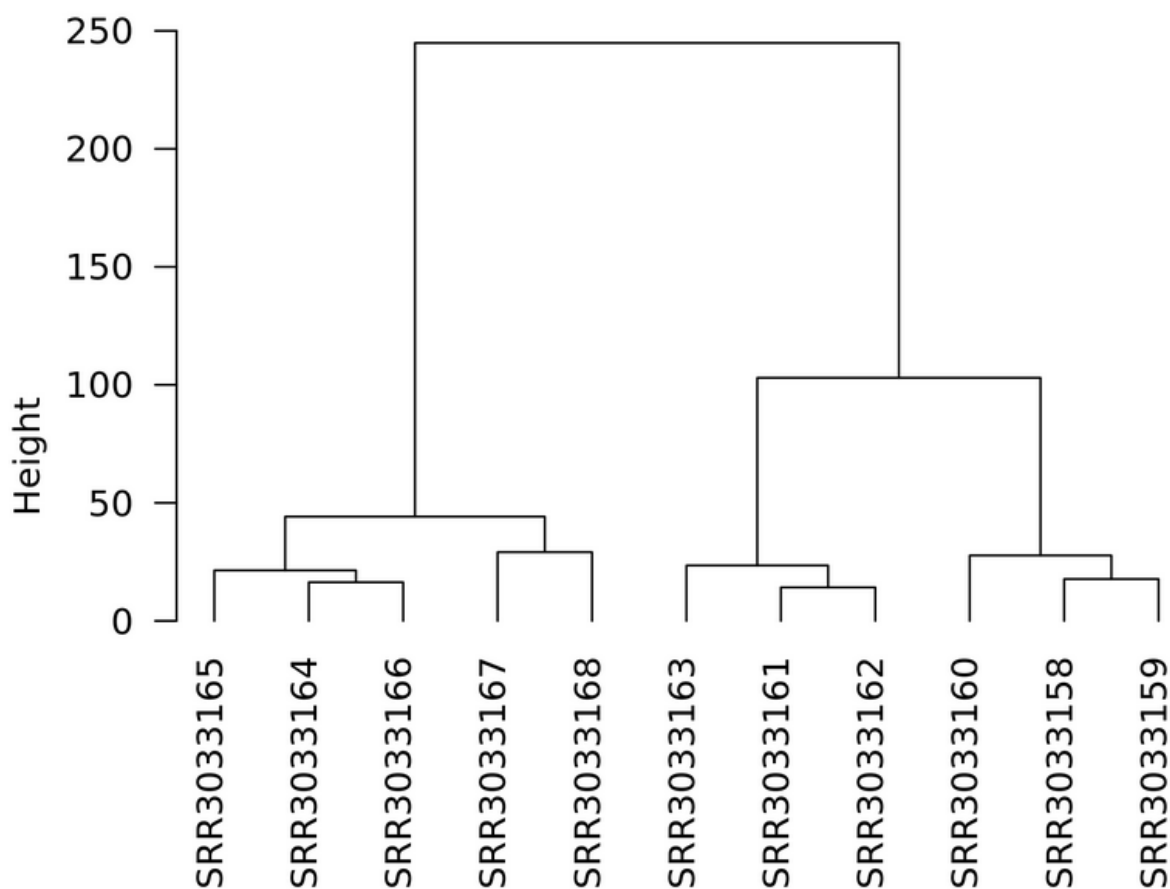


All SERE coefficients are much higher than 1 suggesting that there are only biological replicates here. Note that the coefficient between the 3rd WT and the 4th WT (31,35) is greater than the coefficient between the 3rd WT and the first CodY (13,57). This is explained very well a little further thanks to the ACP



The first axis that explains more than 75% of the variability separates the samples according to their culture environment ("medium" column in the target file).

Cluster dendrogram



The dendrogram shows that the BHI medium separates the WT from the CodY better than the LBMM medium.

In order to take into account the effect of the culture medium, the analysis is restarted by including this effect as a blocking factor.

To do this:

- Click on "show" at the end of the parameters
- Click on "YES" in the blocking factor field and indicate the value "medium".
- Relaunch the analysis

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

medium

Must be a column of the target file

Mean-variance relationship

parametric ▾

(-f, --fitType) Type of model for the mean-dispersion relationship. Parametric by default.

Perform the outliers detection

Yes No

(-o, --cooksCutoff) Checked by default.

Perform independent filtering

Yes No

(-i, --independentFiltering) Checked by default.

Threshold of statistical significance

0.05

(-a, --alpha) Significance threshold applied to the adjusted p-values to select the differentially expressed features. Default is 0.05. The comma is not allowed as decimal separator, use a point instead.

p-value adjustment method

BH ▾

(-p, --pAdjustMethod) p-value adjustment method for multiple testing. 'BH' by default, 'BY' or any value of p.adjust.methods.

Transformation for PCA/clustering

VST ▾

(-T, --typeTrans) Method of transformation of the counts for the clustering and the PCA: 'VST' (default) for Variance Stabilizing Transformation, or 'log' for Regularized Log Transformation.

Estimation of the size factors

median ▾

(-l --locfunc) 'median' (default) or 'shorth' from the genefilter package.

Colors of each biological condition on the plots: 'col1,col2,col3,col4'

dodgerblue,firebrick1,MediumVioletRed,SpringGreen,chartreuse,cyan,darkorchid,darkorange

(-C, --colors) Separate the colors with a comma, no space allowed. Default are 'dodgerblue,firebrick1,MediumVioletRed,SpringGreen,chartreuse,cyan,darkorchid,darkorange'.

This second analysis returns more genes differentially expressed than the previous one.

Data analysis "Stats Smash chr18" with SARTools

Import the data in a new history

- Click on Shared Data -> Data Library
- Click on Stats Smash chr18.
- Select the 6 .tsv files

Galaxy / Galaxy-RNA-seq

Analyse de données Workflow Données partagées Visualisation Aide Utilisateur

Using 0 bytes

DATA LIBRARIES 6 items shown (change) 6 total include deleted To History Download Delete Details Help

Libraries Stats Smash chr18

name	description	data type	size	time updated (UTC)	state
day_0_1.tsv		tabular	21.2 KB	2018-08-10 12:58 PM	
day_0_2.tsv		tabular	21.2 KB	2018-08-10 12:58 PM	
day_0_3.tsv		tabular	21.2 KB	2018-08-10 12:58 PM	
day_7_1.tsv		tabular	21.1 KB	2018-08-10 12:58 PM	
day_7_2.tsv		tabular	21.1 KB	2018-08-10 12:58 PM	
day_7_3.tsv		tabular	21.2 KB	2018-08-10 12:58 PM	

6 items shown (change) 6 total

- Click on "To History" then "as Datasets"
- Give a name to the new history in the "or create new" field (eg Stats smash chr18).

Galaxy / Galaxy-RNA-seq

Analyse de données Workflow Données partagées Visualisation Aide Utilisateur

Using 0 bytes

DATA LIBRARIES 6 items shown (change) 6 total include deleted To History Download Delete Details Help

Libraries Stats Smash chr18

Importer dans l'historique

Select history: Stats smash chr18

or create new: Stats smash chr18

Import Close

name	description	data type	size	time updated (UTC)	state
day_0_1.tsv		tabular	21.2 KB	2018-08-10 12:58 PM	
day_0_2.tsv		tabular	21.2 KB	2018-08-10 12:58 PM	
day_0_3.tsv		tabular	21.2 KB	2018-08-10 12:58 PM	
day_7_1.tsv		tabular	21.1 KB	2018-08-10 12:58 PM	
day_7_2.tsv		tabular	21.1 KB	2018-08-10 12:58 PM	
day_7_3.tsv		tabular	21.2 KB	2018-08-10 12:58 PM	

6 items shown (change) 6 total

- Your new history now appears in the "Data Analysis" section.

Galaxy / Galaxy-RNA-seq

Analyse de données Workflow Données partagées Visualisation Aide Utilisateur

Using 0 bytes

Tools

search tools

Get Data
Send Data
Collection Operations
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data
NGS: Differential Analysis
SAM Tools
BCFtools
NGS: Reads Manipulation
NGS: Mapping
NGS: Transcriptomics
NGS: RNA
NGS: Variant Analysis
RSEM (back-up)
Workflows
All workflows

Bienvenue!

Vous êtes actuellement sur une instance Galaxy dédiée à l'analyse RNA-seq. Elle a été développée avec Docker, spécialement pour les formations de la plateforme BILILLE et est déployée sur le Cloud BILILLE.

Guided Tour »

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors. The Galaxy Docker project is supported by the University of Freiburg, part of the NBI. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

History

Rechercher des données

Stats smash chr18
6 shown

(empty)

- 6: day_7_3.tsv
- 5: day_7_2.tsv
- 4: day_7_1.tsv
- 3: day_0_3.tsv
- 2: day_0_2.tsv
- 1: day_0_1.tsv

The data correspond to RNA-Seq count data for 6 replicates under 2 conditions (3 replicates per condition, day0 and day7).

Preprocessing phase with SARTools.

Goal: This step creates datasets adapted to SARTools.

In the Tools panel, click the "preprocess files for SARTools" tool.

- Create 2 groups: day0 and day7 and add the 3 corresponding replicates to each of the 2 conditions (click "Insert Raw counts" to add a replicate to the groups)
- Choose different replicate names for each replicate (for example rep0_1, rep 0_2 and rep0_3 for the group day0 and rep7_1, rep 7_2 and rep7_3 for the group day7)

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

Add a blocking factor
Yes No
Adjustment variable to use as a batch effect (default no).

Group

1: Group

Group name
day0

Raw counts

1: Raw counts

Replicate raw count
1: day_0_1.tsv

Replicate label name
rep0_1
You need to specify a unique label name for your replicates.

2: Raw counts

Replicate raw count
2: day_0_2.tsv

Replicate label name
rep0_2
You need to specify a unique label name for your replicates.

3: Raw counts

Replicate raw count
3: day_0_3.tsv

Replicate label name
rep0_3
You need to specify a unique label name for your replicates.

+ Insert Raw counts

2: Group

Group name
day7

Raw counts

1: Raw counts

Replicate raw count
4: day_7_1.tsv

Replicate label name
rep7_1
You need to specify a unique label name for your replicates.

2: Raw counts

Replicate raw count

5: day_7_2.tsv

Replicate label name

rep7_2

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

3: Raw counts

Replicate raw count

6: day_7_3.tsv

Replicate label name

rep7_3

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

+ Insert Raw counts

+ Insert Group

Execute

The tool returns 2 outputs

- a design file containing the conditions of the experiment in format txt

1	2	3
label	files	group
rep0_1	dataset_87.dat	day0
rep0_2	dataset_86.dat	day0
rep0_3	dataset_85.dat	day0
rep7_1	dataset_84.dat	day7
rep7_2	dataset_83.dat	day7
rep7_3	dataset_82.dat	day7

- a zip file containing all the count files.

Analysis with Sartools

Goal: Carry out the differential analysis.

In the panel tool, click on the "SARTools DESeq2" tool

- Fill in the design / target file and the Zip file containing the raw counts.
- In the field "Reference biological condition" enter the value "Day0" corresponding to the reference condition of the data.
- Leave the other fields unchanged.

SARTools DESeq2 Compare two or more biological conditions in a RNA-Seq framework with DESeq2 (Galaxy Version 1.3.2.0) Options

Name of the project used for the report

 (-P, --projectName) No space allowed.

Name of the report author

 (-A, --author) No space allowed.

Design / target file

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

Zip file containing raw counts files

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

Names of the features to be removed

 (-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

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

Reference biological condition

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

Advanced Parameters

The histogram of raw p-values has peaks in unexpected places. We restart the analysis with edgeR, keeping the same parameters as for DESeq2:

In the Tools panel, in the NGS: Differential Analysis part, click on the "SARTools edgeR" tool

- Fill in the design / target file and the Zip file containing the gross counts.
- In the field "Reference biological condition" enter the value "Day0" corresponding to the reference condition of the data.
- Leave the other fields unchanged.

SARTools edgeR Compare two or more biological conditions in a RNA-Seq framework with edgeR (Galaxy Version 1.3.2.0) Options

Name of the project used for the report

 (-P, --projectName) No space allowed.

Name of the report author

 (-A, --author) No space allowed.

Design / target file

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

Zip file containing raw counts files

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

Names of the features to be removed

 (-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

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

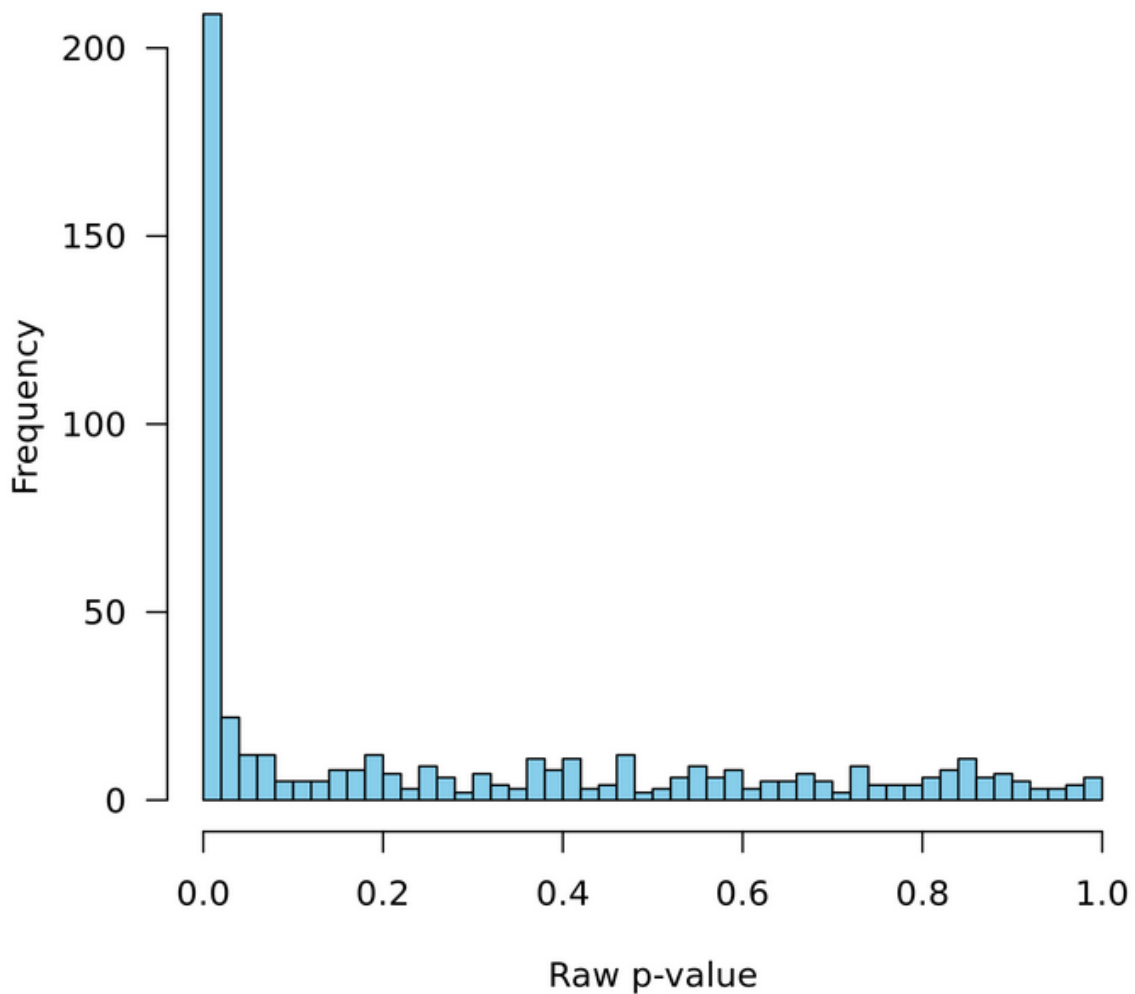
Reference biological condition

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

Advanced Parameters

It now becomes clear that EdgeR is more suited to this dataset than DESeq2.

Distribution of raw p-values - day7 vs day0



GSEA Analysis

Data Preparation

In order to perform the GSEA analysis of these data, we need to retrieve the identifiers of the differentially expressed genes. For this example we will focus on the overexpressed genes.

Reports generated by SARTools produce files that are not directly usable in Galaxy. So we will have to recover the file of interest and re-import it into Galaxy.

- Click on the "eye" icon of the "SARtools EdgeR tables" dataset, the following page appears.

Galaxy Tool SARTools_edgeR

Run at 17/08/2018 09:07:20

Tables available for downloading

Output File Name (click to view)	Size
day7vsday0.complete.txt	111.7 KB
day7vsday0.down.txt	14.2 KB
day7vsday0.up.txt	15.1 KB




With a right click of the mouse, we get the day7vsday0.up.txt file by clicking on "save link as".

Then upload this file on the current Galaxy history :




- Click Get Data -> Upload in the panel tools
- Choose type "tabular" to make sure of its good display under Galaxy.

Download from web or upload from disk

Regular Composite Collection

Name	Size	Type	Genome	Settings	Status
 day7vsday0.up.txt	15.1 KB	tabular	----- Additional Speci...		100% 

Type (set all): Auto-detect Genome (set all): ----- Additional Species Are B...

 Choose local file  Choose FTP file  Paste/Fetch data Pause Reset Start Close

Once the file has been uploaded, the first column containing the identifiers of the differentially expressed genes is retrieved:

- Click on Text Manipulation -> Cut in the panel tools
- In the Cut columns field, enter "c1" to keep only the first column.

Cut columns from a table (Galaxy Version 1.0.2) Options

Cut columns

c1

Delimited by

Tab

From

24: day7vsday0.up.txt

Execute

The list of ENSEMBL identifiers of the overexpressed genes is then obtained.

To carry out the GSEA analysis, go to the following link:

<http://software.broadinstitute.org/gsea/msigdb/annotate.jsp>

This analysis will make it possible to leave gene clusters whose genes are overrepresented among the list overexpressed genes.

Once identified on the site:

- Copy / paste the list of identifiers of our genes in the field on the left
- Select the gene sets of interest (in this example we select all the groups).
- Choose to display only the top 10 gene groups.

Investigate Gene Sets

Gain further insight into the biology behind a gene set by using the following tools:

- ▶ **compute overlaps** with other gene sets in MSigDB ([more...](#))
- ▶ **display the gene set expression profile** based on a selected compendium of expression data ([more...](#))
- ▶ **categorize** members of the gene set by gene families ([more...](#))

Gene Identifiers

ENSG00000170558
ENSG00000134769
ENSG00000173482
ENSG00000168461
ENSG00000154065
ENSG00000176014
ENSG00000132199
ENSG00000176890
ENSG00000166974
ENSG00000078142
ENSG00000196628
ENSG00000101665
ENSG00000150636
ENSG00000166479
ENSG00000074657
ENSG00000168234
ENSG00000166401
ENSG00000134508
ENSG00000134030
ENSG00000152223
ENSG00000206052
ENSG00000141447
ENSG00000141429
ENSG00000078043
ENSG00000154856
ENSG00000141646
ENSG00000154059
ENSG00000132205

Compute Overlaps

- H: hallmark gene sets [?](#)
- C1: positional gene sets [?](#)
- C2: curated gene sets [?](#)
- CGP: chemical and genetic perturbations [?](#)
- CP: Canonical pathways [?](#)
- CP:BIOCARTA: BioCarta gene sets [?](#)
- CP:KEGG: KEGG gene sets [?](#)
- CP:REACTOME: Reactome gene sets [?](#)
- C3: motif gene sets [?](#)
- MIR: microRNA targets [?](#)
- TFT: transcription factor targets [?](#)
- C4: computational gene sets [?](#)
- CGN: cancer gene neighborhoods [?](#)
- CM: cancer modules [?](#)
- C5: GO gene sets [?](#)
- BP: GO biological process [?](#)
- CC: GO cellular component [?](#)
- MF: GO molecular function [?](#)
- C6: oncogenic signatures [?](#)
- C7: immunologic signatures [?](#)

show genesets

with FDR q-value below

Compendia expression profiles

- Human tissue compendium (Novartis)
- NCI-60 cell lines (National Cancer Institute)

Gene families

Clicking on "compute overlaps" gives the following results:

- The list of gene sets that are overrepresented in the list of differentially expressed genes

Compute Overlaps for Selected Genes











Converted 110 submitted identifiers into 77 entrez genes. [click here for details](#).

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
C1, C2, C3, C4, C5, C6, C7, H	10	17810	77	45956

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates less significant FDR q-values (>= 0.05).

Save to: [Excel](#) | [GenomeSpace](#)

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value ?	FDR q-value ?
chr18p11 [149]	Genes in cytogenetic band chr18p11	16		7.51 e-25	1.34 e-20
chr18q21 [128]	Genes in cytogenetic band chr18q21	15		6.15 e-24	5.48 e-20
chr18q12 [70]	Genes in cytogenetic band chr18q12	13		1.24 e-23	7.36 e-20
chr18q22 [38]	Genes in cytogenetic band chr18q22	10		4.32 e-20	1.92 e-16
chr18q11 [55]	Genes in cytogenetic band chr18q11	9		3.83 e-16	1.36 e-12
chr18q23 [36]	Genes in cytogenetic band chr18q23	8		1.24 e-15	3.69 e-12
GO_SEQUENCE_SPECIFIC_DNA_BINDING [1037]	Interacting selectively and non-covalently with DNA of a specific nucleotide composition, e.g. GC-rich DNA binding, or with a specific sequence motif or type of DNA e.g. promotor binding or rDNA binding.	13		1.77 e-8	4.5 e-5
PILON_KLF1_TARGETS_DN [1972]	Genes down-regulated in erythroid progenitor cells from fetal livers of E13.5 embryos with KLF1 [GeneID=10661] knockout compared to those from the wild type embryos.	17		2.24 e-8	4.99 e-5
GO_TRANSCRIPTION_FROM_RNA_POLYMERASE_I_E_II_PROMOTER [724]	The synthesis of RNA from a DNA template by RNA polymerase II, originating at an RNA polymerase II promoter. Includes transcription of messenger RNA (mRNA) and certain small nuclear RNAs (snRNAs).	11		3.58 e-8	7.09 e-5
GSE11924_TH2_VS_TH17_CD4_TCELL_DN [200]	Genes down-regulated in comparison of Th2 cells versus Th17 cells.	7		4.95 e-8	8.81 e-5

- The overlay matrix between over-expressed genes and gene clusters.

Entrez Gene Id	Gene Symbol	Chr18p11	Chr18q21	Chr18q12	Chr18q22	Chr18q11	Chr18q23	GO_SEQUENCE_SPECIFIC_DNA_BINDING	PILON_KLF1_TARGETS_DN	GO_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	GSE11924_TH2_VS_TH17_CD4_TCELL_DN	Entrez	Source	Gene Description
8731	RNMT												S	RNA (guanine-7-) methyltransferase
11031	RAB31												S	RAB31, member RAS oncogene family
9989	PPP4R1												S	protein phosphatase 4, regulatory subunit 1
65258	MPPE1												S	metallophosphoesterase 1
9984	THOC1												S	THO complex 1
8774	NAPG												S	N-ethylmaleimide-sensitive factor attachment protein, gamma
84617	TUBB6												S	tubulin, beta 6 class V
84034	EMILIN2												S	elastin microfibril interfacier 2
23253	ANKRD12												S	ankyrin repeat domain 12
9229	DLGAP1												S	discs, large (Drosophila) homolog-associated protein 1
5797	PTPRM												S	protein tyrosine phosphatase, receptor type, M
55556	ENOSF1												S	enolase superfamily member 1
7298	TYMS												S	thymidylate synthetase
10939	AFG3L2												S	AFG3 ATPase family gene 3-like 2 (S. cerevisiae)
147495	APCDD1												S	adenomatous polyposis coli down-regulated 1
339290	LOC339290												S	uncharacterized LOC339290
4089	SMAD4												S	SMAD family member 4
4087	SMAD2												S	SMAD family member 2
4152	MBD1												S	methyl-CpG binding domain protein 1
6925	TCF4												S	transcription factor 4
55205	ZNF532												S	zinc finger protein 532
115701	ALPK2												S	alpha-kinase 2
4092	SMAD7												S	SMAD family member 7
2235	FECH												S	ferrochelatase
83473	KATNAL2												S	katanin p60 subunit A-like 2
57614	KIAA1468												S	KIAA1468
23335	WDR7												S	WD repeat domain 7
497661	C18orf32												S	chromosome 18 open reading frame 32
9811	CTIF												S	CBP80/20-dependent translation initiation factor
11201	POLI												S	polymerase (DNA directed) iota
5271	SERPINB8												S	serpin peptidase inhibitor, clade B (ovalbumin), member 8
30827	CXXC1												S	CXXC finger protein 1
7572	ZNF24												S	zinc finger protein 24
10982	MAPRE2												S	microtubule-associated protein, RP/EB family, member 2

Analysis of data from the recount project

Retrieving data via the recount tool

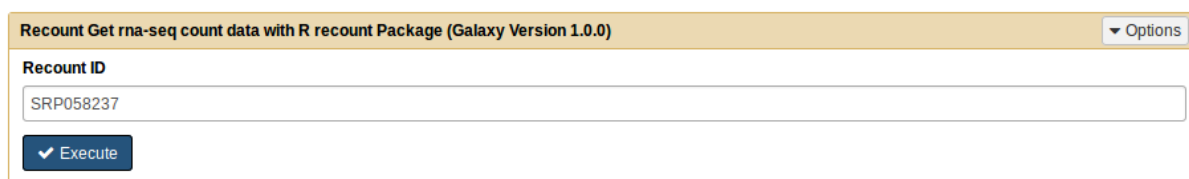
Goal: To retrieve count data via the recount2 tool (<https://jhubiostatistics.shinyapps.io/recount/>).

In this example we will treat the dataset SRP058237: This dataset contains 17 samples related to lung cancer.

- 2 conditions: Tumor (tumor cells) and adjacent (healthy cells taken next to the tumor)
- 3 types of cells (IMMCs, Neutrophil, Epithelial)

In the Tools panel, in the NGS: Differential Analysis part, click on the Recount tool .

- Fill in the "Recount ID" field using ID SRP058237.



The Recount tool returns 1 count file per sample, here 17 files, and 1 file summarizing the conditions of the samples.

1	2
"ENSG00000000003.14"	0
"ENSG00000000005.5"	0
"ENSG000000000419.12"	515
"ENSG000000000457.13"	91
"ENSG000000000460.16"	182
"ENSG000000000938.12"	13683
"ENSG000000000971.15"	136
"ENSG00000001036.13"	2538

Header of a count file generated by Recount.

Exercise: Run the differential analysis between tum-IMMC and adj-IMMC conditions.

GSEA analysis on the msigdb site

Preparation of the data:

In order to carry out the GSEA analysis, it is necessary to carry out some pre-treatments.


For this analysis we will retrieve the set of differentially expressed genes. First, it is necessary to retrieve the list of the differentially expressed genes generated by SARTools. We will proceed in the same way as in the previous chapter :

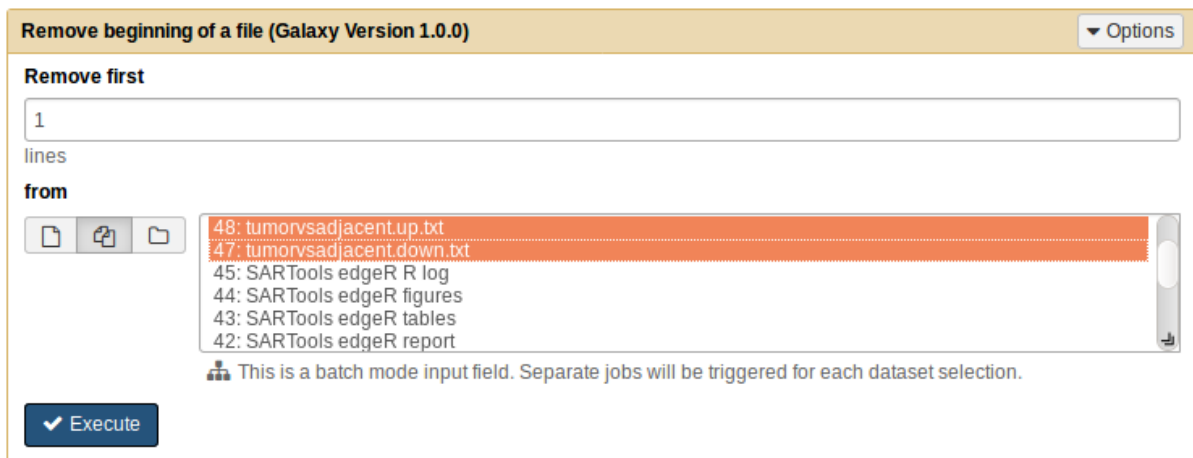
- Save the 2 files locally
- Reimport them under Galaxy using the upload tool.

Once reimported in Galaxy, it is necessary to concatenate and to modify the ENSEMBL identifiers because the broad institute website does not accept the suffixes of these identifiers.

First we will delete the first header line of the file.

In the "Text Manipulation" section click on the "Remove beginning of a file" tool

- Enter "1" in the "Remove first" field
- Click on the icon  to select the 2 files to be processed






Remove beginning of a file (Galaxy Version 1.0.0) Options

Remove first


1

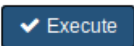
lines

from

- 48: tumorsadjacentLup.txt
- 47: tumorsadjacent.down.txt
- 45: SARTools edgeR R log
- 44: SARTools edgeR figures
- 43: SARTools edgeR tables
- 42: SARTools edgeR report

 This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

 Execute

To concatenate the files, in the section "Text Manipulation" click on the tool "Concatenate datasets tail-to-head"

- Choose the 2 files corresponding to the 2 resulting files of the previous step

Concatenate datasets tail-to-head (Galaxy Version 1.0.0) Options

Concatenate Dataset

61: Remove beginning on data 48

Dataset

1: Dataset

Select

60: Remove beginning on data 47

Then retrieve the first column of the resulting file with the tool "Cut" of the "Text Manipulation" section:

Cut columns from a table (Galaxy Version 1.0.2) Options

Cut columns

c1

Delimited by

Tab

From

62: Concatenate datasets on data 60 and data 61

We get the list of differentially expressed genes, but the identifiers still contain the suffixes. To delete them, use the "convert" tool in the "Text Manipulation" section and replace the points with tabs:

Convert delimiters to TAB (Galaxy Version 1.0.0) Options

Convert all

Dots

in Dataset

63: Cut on data 62

Strip leading and trailing whitespaces

Condense consecutive delimiters in one TAB

Finally, use the “cut” tool again to get the first column of the last resulting file and you should get the list of ENSEMBL identifiers of differentially expressed genes.

```
1  
ENSG00000116774  
ENSG00000091409  
ENSG00000131747  
ENSG00000133063  
ENSG00000134061  
ENSG00000114251  
ENSG00000262406  
ENSG00000088325  
ENSG00000166165  
ENSG00000143195  
ENSG00000143891  
ENSG00000117394
```

Exercise: Run the GSEA analysis