

TP RNA-Seq Biostats



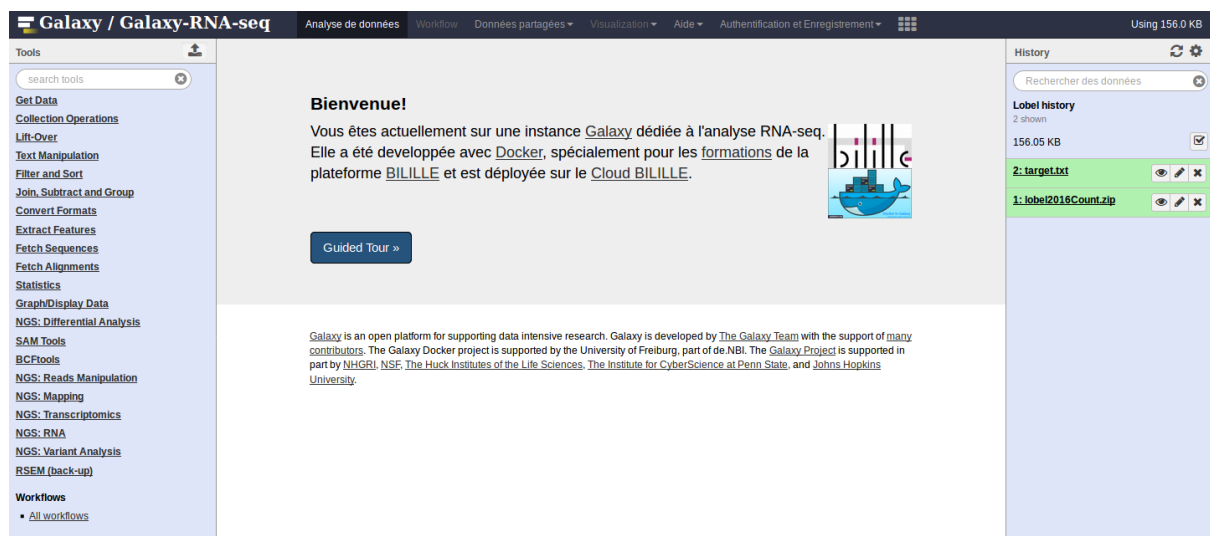
Pierre Pericard: pierre.pericard@univ-lille.fr
Samuel Blanck : samuel.blanck@univ-lille.fr
Guillemette Marot : guillemette.marot@univ-lille.fr

Analyse des données “Lobel” avec SARTools	2
Importer les données dans un nouvel historique	2
Analyse avec SARTools	3
Analyse d’enrichissement des données “Stats Smash chr18”	8
Importer les données dans un nouvel historique	9
Préparation des données	10
Analyse d’enrichissement	11
Analyse de données issues du projet recount	14
Prétraitement des données	14
Exercice :	16
Analyse d’enrichissement	16
Préparation des données	16
Exercice : Réaliser l’analyse d’enrichissement	18
Exercice bonus: Revigo	18

Analyse des données “Lobel” avec SARTools

Importer les données dans un nouvel historique

- Sous Galaxy cliquez sur Données partagées (Shared Data) -> Historiques (Histories)
- Importer l'historique “Bilille RNA-seq Biostats - Lobel”.
- Donner un nom au nouvel historique (par exemple “Lobel history”)



The screenshot shows the Galaxy web interface. The main content area displays a welcome message: "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." Below this is a "Guided Tour" button. The left sidebar contains a "Tools" menu with various categories like "Get Data", "Collection Operations", "Text Manipulation", etc. The right sidebar shows a "History" panel with a search bar and a list of data items: "Lobel history" (156.05 KB) and "1: lobel2016Count.zip".

Le fichier lobel.zip contient les comptages issus de la publication Lobel L, Herskovits AA (2016) Systems Level Analyses 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.

Le fichier target.txt contient la description des conditions de l'expérience en vue de son analyse par SARTools : 11 réplicats pour 2 conditions (6 WT pour 5 codY)

L'analyse avec “SARTools DESeq2” peut ensuite être lancée via le menu déroulant de gauche.

Analyse avec SARTools

- Renseignez le design/target file et le fichier Zip contenant les comptages bruts.
- Dans le champ “Factor of interest” entrez la valeur “strain” correspondant à la 3ème colonne du fichier target et contenant les 2 conditions à comparer.
- Dans le champ “Reference biological condition” entrez la valeur WT.
- Vous pouvez laisser les autres champs inchangés.

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

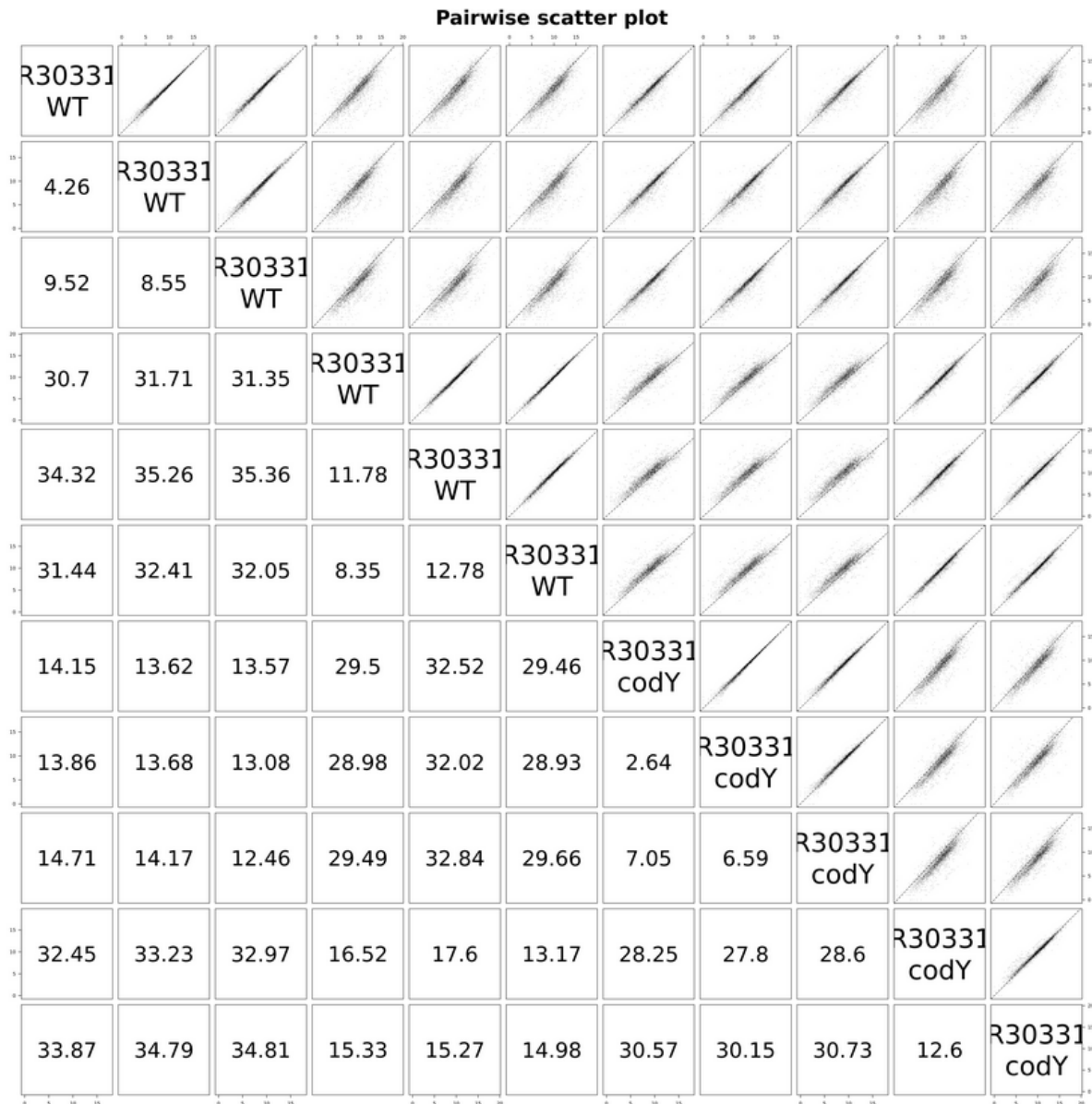
 (-v, --varInt) Biological condition in the target file. Default is 'group'.

Reference biological condition

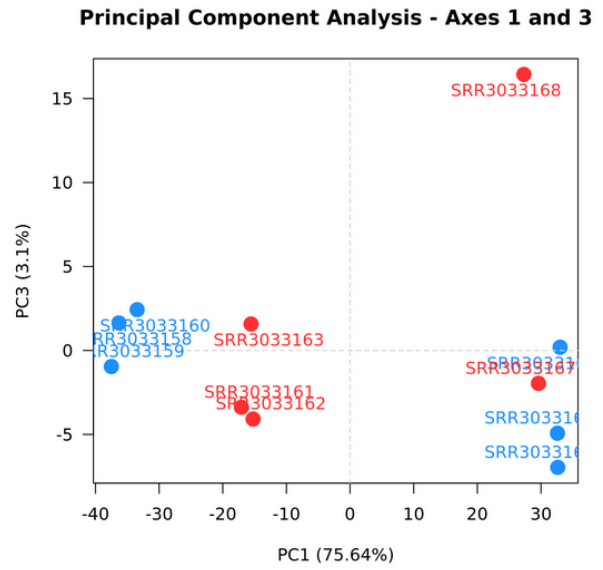
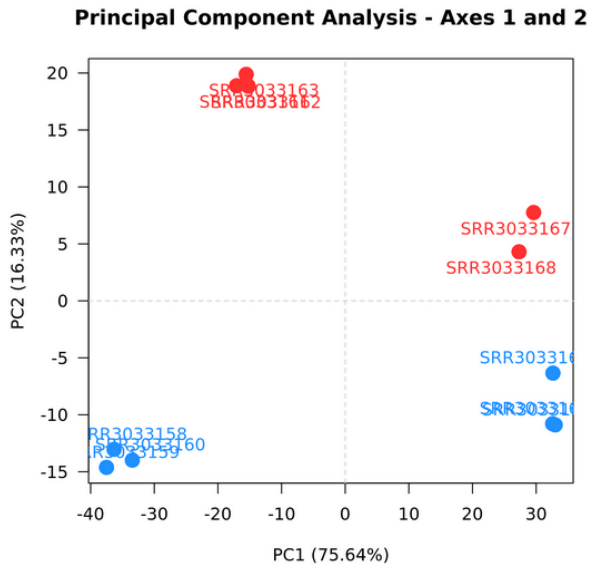
 (-c, --condRef) Reference biological condition used to compute fold-changes, must be one of the levels of 'Factor of interest'.

Advanced Parameters

Voici les résultats présentés dans le rapport d'analyse :

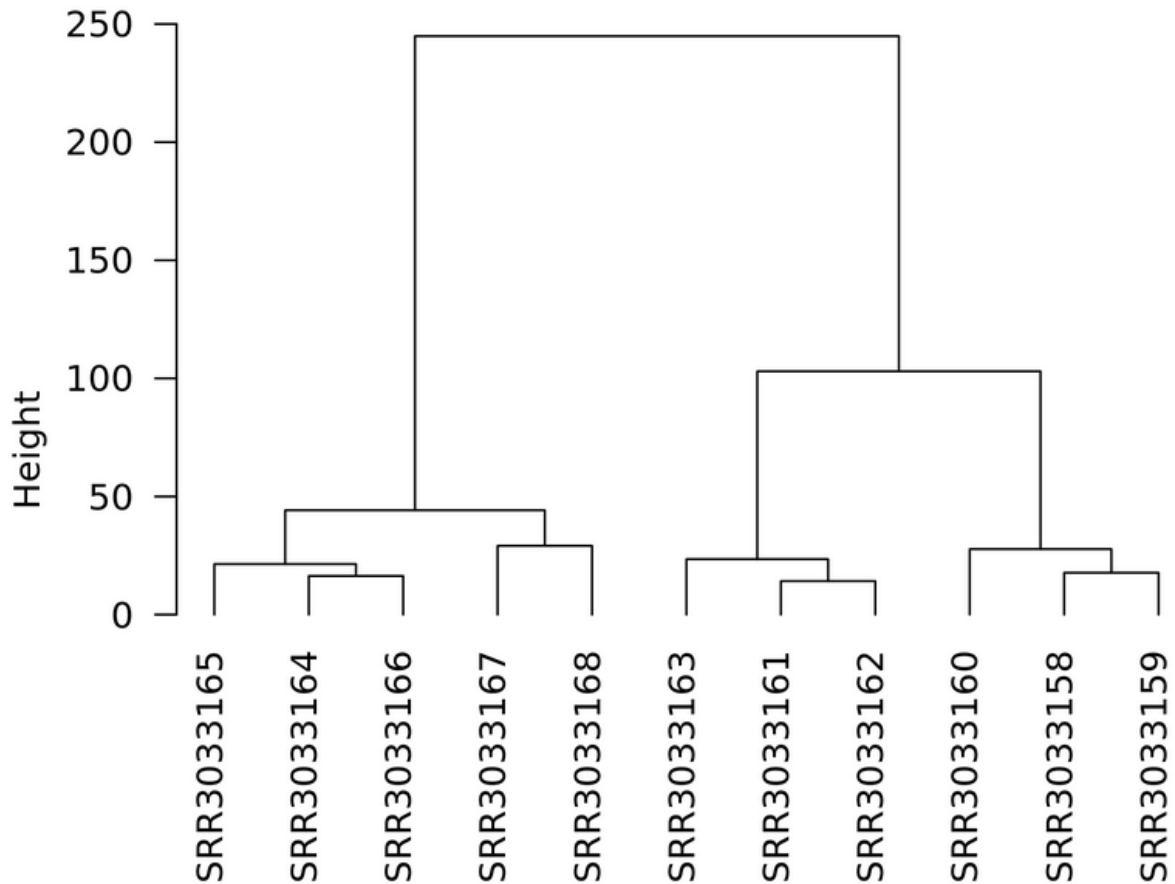


Tous les coefficients SERE sont nettement supérieurs à 1 laissant penser qu'il n'y a ici que des répliqués biologiques. On remarque que le coefficient entre le 3ème WT et le 4ème WT (31,35) est supérieur au coefficient entre le 3ème WT et le premier CodY (13,57). Ceci s'explique très bien un peu plus loin grâce à l'ACP



Le premier axe qui explique plus de 75% de la variabilité sépare les échantillons suivant leur environnement de culture (colonne "medium" dans le fichier target).

Cluster dendrogram



Le dendrogramme permet de voir que le milieu BHI sépare mieux les WT des CodY que le milieu LBMM.

Afin de prendre en compte l'effet du milieu de culture, on relance l'analyse en incluant cet effet comme blocking factor.

Pour cela :

- Cliquez sur "show" à la fin des paramètres
- Cliquez sur "YES" dans le champs blocking factor et indiquer la valeur "medium".
- Relancez l'analyse

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

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

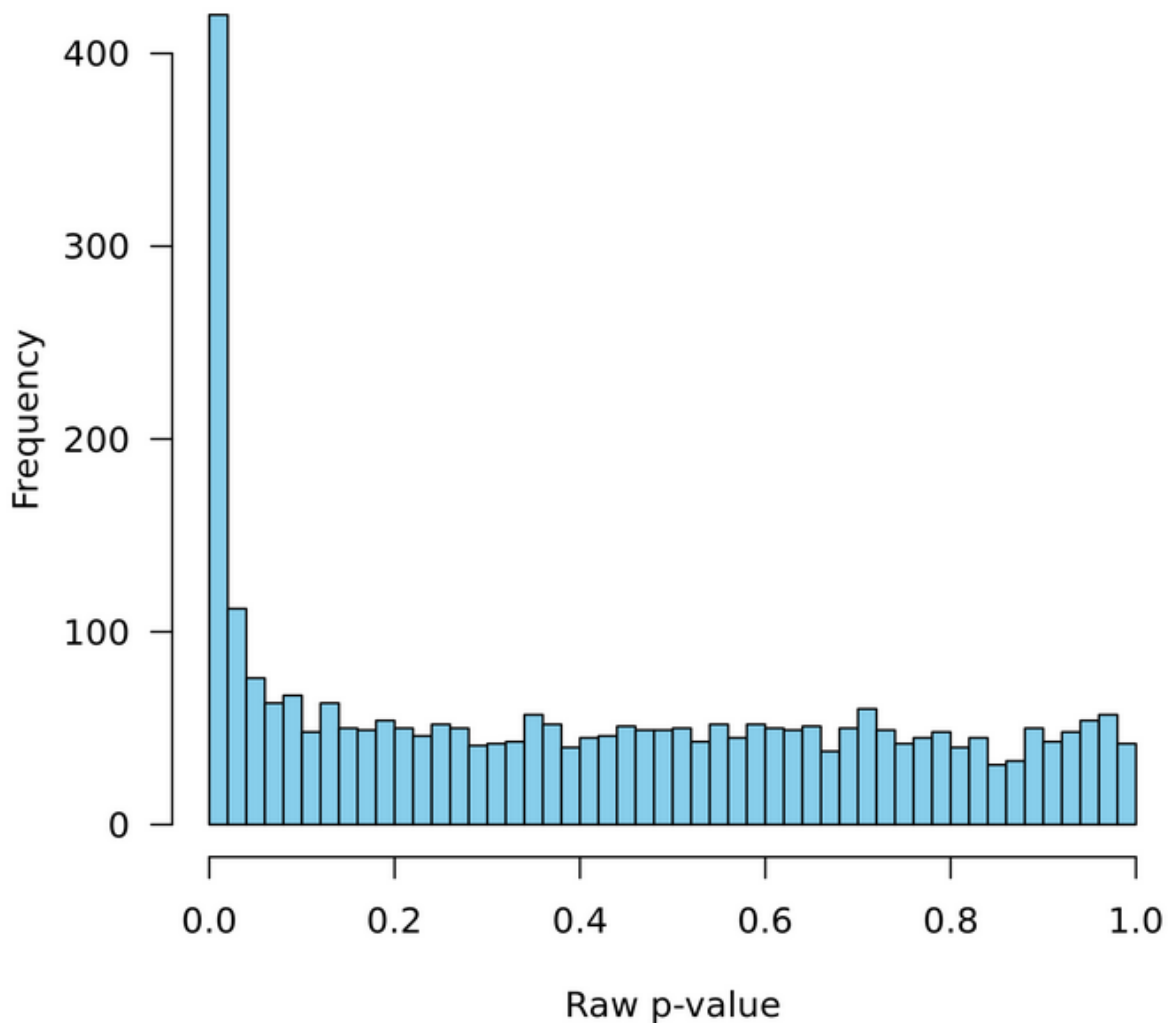
dodgerblue,firebrick1,MediumVioletRed,SpringGreen,chartreuse,cyan,darkorchid,darkorange

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

Cette seconde analyse permet de ressortir plus de gènes différentiellement exprimés que la précédente.

Dans le rapport généré, on remarque que l’histogramme des p-values brutes présente une forme attendue : un pic à gauche correspondant aux gènes différentiellement exprimés et une distribution uniforme par ailleurs.

Distribution of raw p-values - codY vs WT



Analyse d’enrichissement des données “Stats Smash chr18”

Les données correspondent à des comptages RNA-Seq humains pour 6 réplicats dans 2 conditions (3 réplicats pour 2 conditions, day0 et day7). Ces données ont été analysées avec DESeq2 et ce sont ces résultats d’analyses qui serviront de base pour l’analyse d’enrichissement

Importer les données dans un nouvel historique

- Sous Galaxy cliquez sur Données partagées (Shared Data) -> Histories

Published Histories

search name, annotation, owner 🔍

Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
Recount		admin	★★★★★		Mar 09, 2022
rsem		admin	★★★★★		Mar 09, 2022
Stats_smash_chr18		admin	★★★★★		Mar 09, 2022
Lobel		admin	★★★★★		Mar 09, 2022

- Cliquez sur “Bilille RNA-seq Biostats - Stats_smash_chr18”.
- Cliquez sur Import history et choisissez un nom d'historique

Published Histories | admin | Stats_smash_chr18 Import history

Stats_smash_chr18

3.99 MB

Rechercher des données 🔍

Jeu de données	Annotation
128: SARTools edgeR R objects (.RData)	
127: SARTools edgeR R log	
126: SARTools edgeR figures	
125: SARTools edgeR tables	
124: SARTools edgeR report	
123: SARTools DESeq2 R objects (.RData)	
122: SARTools DESeq2 R log	
121: SARTools DESeq2 figures	
120: SARTools DESeq2 tables	
119: SARTools DESeq2 report	

- Votre nouvel historique apparaît maintenant dans la partie “Analyses de données”.

The screenshot shows the Galaxy web interface. On the left is a 'Tools' sidebar with a search bar and a list of tool categories. The main area displays a 'Bienvenue!' (Welcome!) message with a 'Guided Tour' button and a description of the Galaxy instance. On the right, the 'History' panel is open, showing the 'Stats_smash_chr18' dataset with a list of 12 sub-histories, each with a search icon, an edit icon, and a delete icon.

Préparation des données

Afin de réaliser l'analyse d'enrichissement de ces données, nous devons récupérer les identifiants des gènes différentiellement exprimés. Pour cet exemple nous allons nous concentrer sur les gènes surexprimés.

Les rapports générés par SARTools produisent des fichiers qui ne sont pas directement exploitables dans Galaxy. Nous allons donc devoir récupérer le fichier qui nous intéresse et le réimporter dans Galaxy.

En cliquant sur l'icône "oeil" du dataset "SARtools DESeq2 tables", la page suivante apparaît

Galaxy Tool SARTools_DESeq2

Run at 03/10/2023 22:47:09

Tables available for downloading




Output File Name (click to view)	Size
day7vsday0.complete.txt	134.6 KB
day7vsday0.down.txt	17.3 KB
day7vsday0.up.txt	14.4 KB

A l'aide du clic droit de la souris on récupère le fichier day7vsday0.up.txt en cliquant sur "enregistrer la cible du lien sous" et on l'enregistre sur la machine locale.




Puis en cliquant sur Get Data -> Upload dans le panel tools, on l'upload sous Galaxy :
De plus, en choisissant le type "tabular" on s'assure de son bon affichage sous 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

Une fois le fichier uploadé, on récupère la première colonne contenant les identifiants des gènes différentiellement exprimés :




Cliquez sur Text Manipulation -> Cut dans le panel tools :

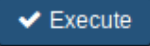
- Dans le champ Cut columns, saisissez "c1" pour ne garder que la première colonne.

Cut columns from a table (Galaxy Version 1.0.2) Options

Cut columns
c1

Delimited by
Tab

From
   24: day7vsday0.up.txt



On obtient alors la liste des identifiants ENSEMBL des gènes surexprimés.

Analyse d'enrichissement

Pour réaliser l'analyse d'enrichissement, il faut aller sur le lien suivant :

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

Cette analyse va permettre de sortir des groupes de gènes dont les gènes sont surreprésentés parmi la liste des gènes surexprimés.

Une fois identifié sur le site :

- Copiez/collez la liste des identifiants de nos gènes dans le champ à gauche
- Sélectionnez les groupes de gènes qui nous intéressent (dans l'exemple nous sélectionnons tous les groupes).
- Choisissez de n'afficher que le top 10 des groupes de gènes.

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

En cliquant sur "compute overlaps" on obtient les résultats suivants :

- La liste des groupes de gènes qui sont surreprésentés dans la liste des gènes différentiellement exprimés :

Converted 109 submitted identifiers into 85 NCBI (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, C8, H	10	32880	85	40786

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: [Text](#) (as Tab separated values; *.tsv)

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
chr18q21 [195]	Ensembl 103 genes in cytogenetic band chr18q21	20		1.54 e-28	4.73 e-24
chr18p11 [201]	Ensembl 103 genes in cytogenetic band chr18p11	20		2.88 e-28	4.73 e-24
chr18q12 [98]	Ensembl 103 genes in cytogenetic band chr18q12	14		2.68 e-22	2.94 e-18
chr18q11 [81]	Ensembl 103 genes in cytogenetic band chr18q11	12		1.88 e-19	1.54 e-15
chr18q22 [61]	Ensembl 103 genes in cytogenetic band chr18q22	10		7.39 e-17	4.86 e-13
chr18q23 [39]	Ensembl 103 genes in cytogenetic band chr18q23	9		9.65 e-17	5.29 e-13
GU_PDEF_TARGETS_UP [71]	Integrin, VEGF, Wnt and TGFbeta signaling pathway genes up-regulated in PC-3 cells (prostate cancer) after knockdown of PDEF [GeneID=25803] by RNAi.	5		4.08 e-7	1.92 e-3

- La matrice de superposition entre les gènes surexprimés et les groupes de gènes.

Entrez Gene Id	Gene Symbol	chr18q21	chr18p11	chr18q12	chr18q11	chr18q22	chr18q23	GU_PDEF_TARGETS_UP	MIR3942_3P	GOCC_SMAD_PROTEIN_COMPLEX	MIR348AN	Entrez	Ensembl	Gene Description	
4087	SMAD2														SMAD family member 2 [Source:HGNC Symbol;Acc:HGNC:6768]
4089	SMAD4														SMAD family member 4 [Source:HGNC Symbol;Acc:HGNC:6770]
6925	TCF4														transcription factor 4 [Source:HGNC Symbol;Acc:HGNC:11634]
57614	RELCH														"RAB11 binding and LisH domain, coiled-coil and HEAT repeat containing [Source:HGNC Symb
2235	FECH														ferrochelatase [Source:HGNC Symbol;Acc:HGNC:3647]
4092	SMAD7														SMAD family member 7 [Source:HGNC Symbol;Acc:HGNC:6773]
55205	ZNF532														zinc finger protein 532 [Source:HGNC Symbol;Acc:HGNC:30940]
28316	CDH20														cadherin 20 [Source:HGNC Symbol;Acc:HGNC:1760]
23335	WDR7														WD repeat domain 7 [Source:HGNC Symbol;Acc:HGNC:13490]
9063	PIAS2														protein inhibitor of activated STAT 2 [Source:HGNC Symbol;Acc:HGNC:17311]

Analyse de données issues du projet recount

Prétraitement des données

Dans cet exemple nous traiterons le jeu de données SRP058237 : Ce jeu de données contient 17 échantillons liés au cancer du poumon.

- 2 conditions : Tumeur pour les cellules tumorales et adjacent pour les cellules saines prélevées à côté de la tumeur
- 3 types de cellules (IMMCs, Neutrophile, Épithéliales)

Nous avons récupéré les données de comptage via le projet recount3 (<https://rna.recount.bio/>)

Importer l'historique partagé: "Bilille RNA-seq - SRP058237 data"

Dans le panel Tools, cliquez sur l'outil "preprocess files for SARTools".

- Créez 2 groupes : Tumeur (TumIMMC) et Adjacent (AdjIMMC) et ajoutez-y les 3 réplicats correspondants à chacun des 2 conditions
- Choisissez des noms de réplicats différents pour chaque réplicat (par exemple Tum1, Tum2, Tum3, pour le groupe Tumeur et Adj1, Adj2, Adj3 Adjacent)

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

Group

1: Group

Group name

Raw counts

1: Raw counts

Replicate raw count

Replicate label name

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

2: Raw counts

Replicate raw count

Replicate label name

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

3: Raw counts

Replicate raw count

Replicate label name

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

Insert Raw counts

2: Group

Group name
AdjIMMC

Raw counts

1: Raw counts

Replicate raw count
9: Recount (SRR2016911_Adj-IMMC01)

Replicate label name
Adj1
You need to specify an unique label name for your replicates.

2: Raw counts

Replicate raw count
10: Recount (SRR2016912_Adj-IMMC02)

Replicate label name
Adj2
You need to specify an unique label name for your replicates.

3: Raw counts

Replicate raw count
11: Recount (SRR2016913_Adj-IMMC03)

Replicate label name
Adj3
You need to specify an unique label name for your replicates.

+ Insert Raw counts

+ Insert Group

✓ Execute

L'outil renvoie 2 sorties

- un fichier design reprenant les conditions de l'expérience au format txt

1	2	3
label	files	group
Tum1	dataset_2.dat	TumIMMC
Tum2	dataset_3.dat	TumIMMC
Tum3	dataset_4.dat	TumIMMC
Adj1	dataset_9.dat	AdjIMMC
Adj2	dataset_10.dat	AdjIMMC
Adj3	dataset_11.dat	AdjIMMC

- un fichier zip contenant l'ensemble des fichiers de comptages.

Exercice :

Réaliser l'analyse différentielle entre les conditions TumIMMC et AdjIMMC.

Analyse d'enrichissement


Préparation des données

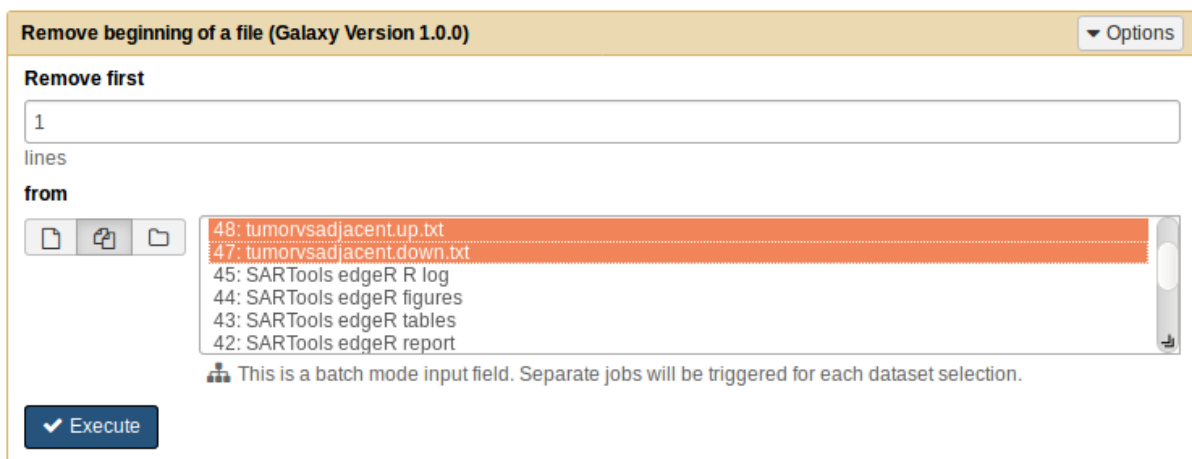
Afin de réaliser l'analyse d'enrichissement, il est nécessaire de procéder à quelques prétraitements.

Pour cette analyse nous allons récupérer l'ensemble des gènes différentiellement exprimés. Pour cela, il faut d'abord récupérer la liste des gènes sur et sous exprimés générée par SARTools. On procédera de la même façon que dans le chapitre précédent, à savoir enregistrer les 2 fichiers en local et les réimporter sous Galaxy à l'aide de l'outil upload.

Une fois ces 2 listes de gènes réimportés dans Galaxy, il faut les concaténer puis retravailler les identifiants ENSEMBL car le site du broad institute n'accepte pas les suffixes pour réaliser l'analyse.

Tout d'abord nous allons supprimer la première ligne du fichier qui sert d'en-tête. Dans la section "Text Manipulation" cliquez sur l'outil "Remove beginning of a file"

- Supprimez juste la première ligne en indiquant "1" dans le champ "Remove first"
- Cliquez sur l'icône  pour sélectionner les 2 fichiers à traiter



Pour concaténer les fichiers dans la section "Text Manipulation" cliquez sur l'outil "Concatenate datasets tail-to-head"

- Choisissez les 2 fichiers correspondants aux 2 fichiers résultats de l'étape précédente

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

Puis récupérez la première colonne du fichier ainsi obtenu avec l'outil "cut" de la section "Text Manipulation" :

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

On obtient bien la liste des gènes différentiellement exprimés, mais les identifiants contiennent encore les suffixes. Pour les supprimer utilisez l'outil "convert dans la section "Text Manipulation" et remplacez les points par des tabulations :

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

Enfin, utilisez une nouvelle fois l'outil cut pour récupérer la première colonne du dernier fichier résultat et vous devriez obtenir la liste des identifiants ENSEMBL des gènes différentiellement exprimés.

1

ENSG00000116774

ENSG00000091409

ENSG00000131747

ENSG00000133063

ENSG00000134061

ENSG00000114251

ENSG00000262406

ENSG00000088325

ENSG00000166165

ENSG00000143195

ENSG00000143891

ENSG00000117394

Exercice : Réaliser l'analyse d'enrichissement

1. En utilisant Metascape (<https://metascape.org>)
2. En utilisant g:Profiler (<https://biit.cs.ut.ee/gprofiler/gost>)

Exercice bonus: Revigo

Extrayez les GO:ID enrichis avec leurs p-values et réduisez l'information en utilisant Revigo (<http://revigo.irb.hr/>)