

Roscoff-RNASeq

A new pipeline for Differential Expression

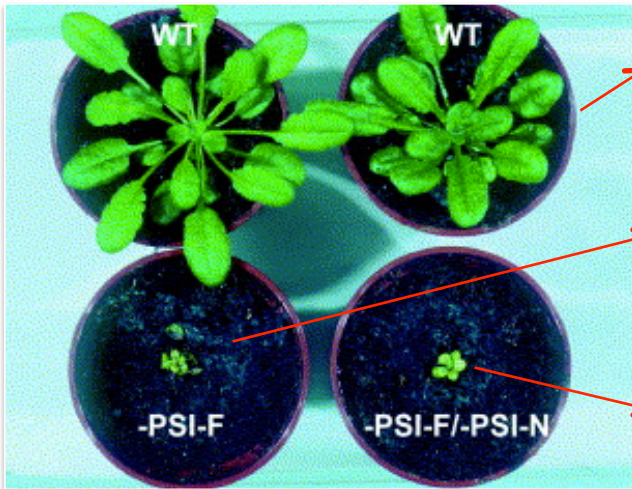
15-01-2013

Yufei Luo, Marie-Agnès Dillies, Matthias Zytnicki, Delphine Steinbach



Differential Expression

arabidopsis

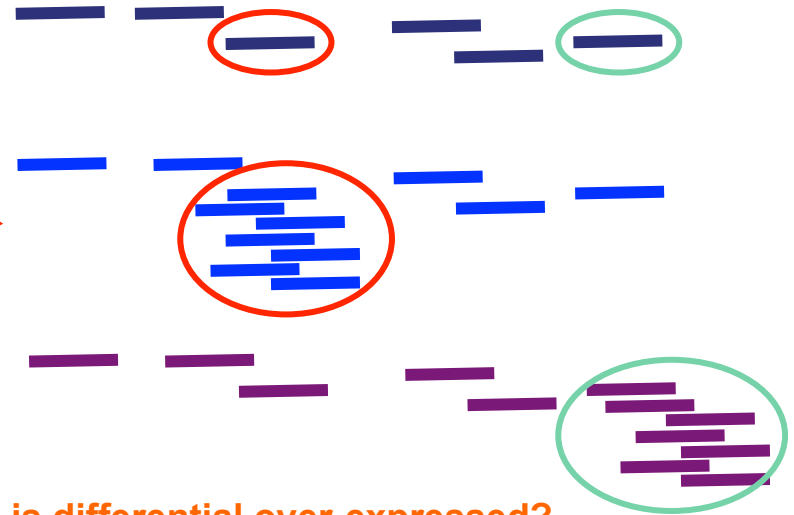


genome

Condition 1
(At least 2 replicates)

Condition 2
(At least 2 replicates)

Condition 3
(At least 2 replicates)



Which gene is differential over-expressed?

How do we analyze the data for differential expression?

How to do a differential expression analysis

- I. Mapping (**Tophat**, BWA, Bowtie...)
- II. Count (HTSeq, **CompareOverlapping (S-MART)**, Matlab ...)
- III. Statistical analysis (**DESeq**, EdgeR, Goseq...)

Why DESeq?

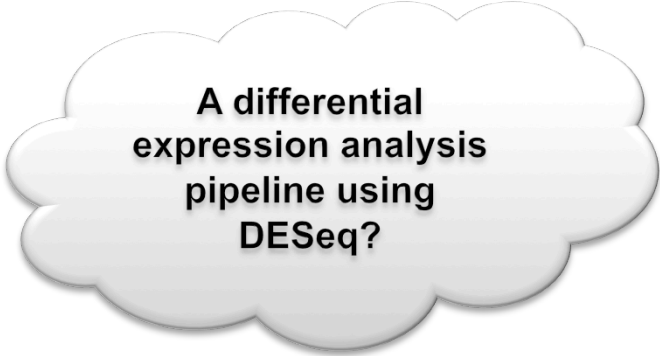
--negative binomial distribution, with variance and mean linked by local regression for 2 conditions

- **Infer differential signal correctly**
- **Good statistical power**
- **Estimation of data variability throughout the dynamic range**
- **Suitable error model**

DESeq Implementation

*Marie-Agnès Dillies, Platform Transcriptome and Epigenome,
Génoptole, Pasteur Institute*

- Total read count per sample
- Non-expressed gene per sample
- Density of counts distribution
- Proportion of reads associated to the most expressed gene
- Scatter plot of samples
- Diagnostic of samples
- Estimated dispersion distribution
- Raw p-value histogram
- MAplot
- All over-expressed gene



A differential
expression analysis
pipeline using
DESeq?

Differential expression analysis pipeline

--steps

- ① **Provide** replicate files per condition, reference genome, gene annotations
- ② **Fastq groomer** to convert an identical fastq format
- ③ **Tophat** to align the reads against reference genome
- ④ **SamTools** to convert bam format to sam format
- ⑤ **CompareOverlapping** to calculate the number of overlapping reads
- ⑥ **CountNumber** to convert the previous output to a tabular file
- ⑦ **DESeq** to analyze differential expression

Import the workflow to your galaxy work sapce

Published Workflows

search name, annotation, owner, and tags

[Advanced Search](#)

Name	Annotation	Owner
Differential_expression_DESeq		yluo
RNAseq de-novo ABiMS		ppericard
MAPHiTS Pipeline BWA		nchoisne
MAPHiTS Pipeline BOWTIE		nchoisne
MAPHiTS Pipeline		oinizan

Galaxy / ABiMS

Analyze Data Workflow Shared Data Visualization Help User

Using 5.3 GB

Published Workflows | yluo | Differential_expression_DESeq

Galaxy Workflow 'Differential_expression_DESeq'

Step

Annotation

Step 1: Input dataset

[+ Import workflow](#) [About this workflow](#)

Import workflow

yluo

Related Workflows

[All published workflows](#)

[Published workflows by yluo](#)

Differential expression workflow

Your workflows

Name

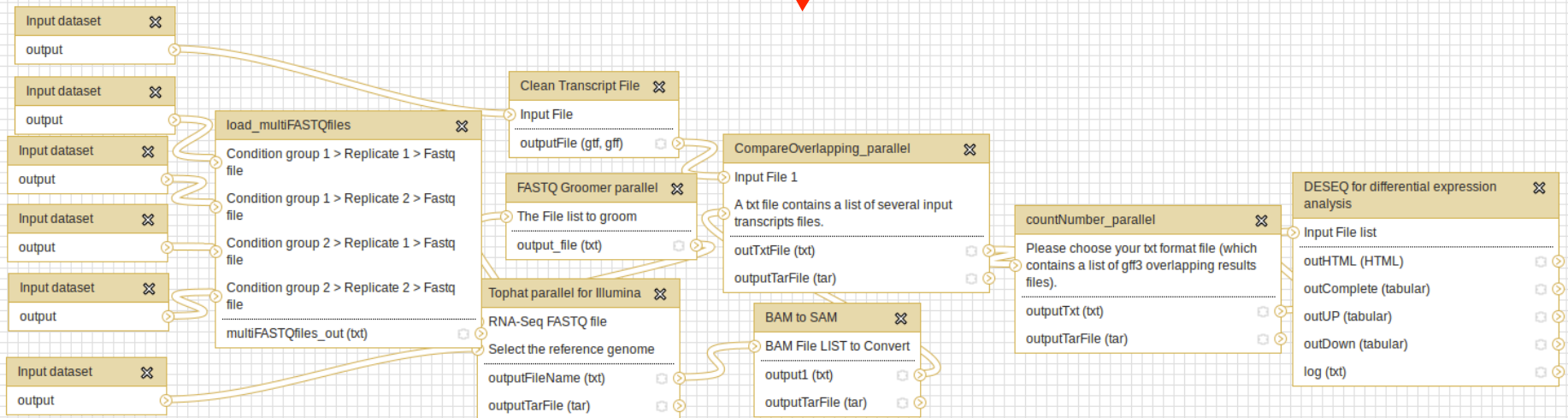
Differential_expression_DESeq

Workflows shared with you by others

No workflows have been shared with you.

Other options

Configure your workflow menu



Differential expression workflow usage

Tools

DATA MANAGEMENT

[Get Data](#)

[Send Data](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

ECOLE GGB GROUPE 1

[Mardi 15 RNA-seq](#)

[Mercredi 16 RNA-seq de novo ABiMS](#)

[Mercredi 16 ChIP-seq J. van Helden](#)

[Jeudi 17 ChIP-seq J. van Helden](#)

[Jeudi 17 - miRNAs Qualite / Nettoyage / Mirdeep2 / Annotation](#)

ECOLE GGB GROUPE 2

[Mardi 15 URGI: MAPHITS - PreProcess Tools](#)

[Mardi 15 URGI: MAPHITS - Tools](#)

[Mardi 15 URGI: MAPHITS - PostProcess Tools](#)

[Mardi 15 URGI: MAPHITS - SNPs Chip Tools](#)

[Mardi 15 URGI: S-MART - Tools](#)

[Mardi 15 URGI: S-MART - Differential Expression Pipeline Tools](#)

[Mercredi 16 - Annotation fonctionnelle des SNPs](#)

[Jeudi 17 - SVDetect](#)

ABIMS TOOLS

[Tools](#)

[Metabolomic](#)

COMPARISON TOOLS

[Evolution](#)

[Multiple Alignments](#)

[FASTA manipulation](#)

NGS TOOLS

[NGS QC and manipulation](#)

Running workflow "Differential_expression_DESeq" Expand All Collapse

- Step 1: [Input dataset](#)
- Step 2: [Input dataset](#)
- Step 3: [Input dataset](#)
- Step 4: [Input dataset](#)
- Step 5: [Input dataset](#)
- Step 6: [Input dataset](#)
- Step 7: [Input dataset](#)
- Step 8: [Clean Transcript File \(version 1.0.0\)](#)
- Step 9: [load multiFASTQfiles \(version 1.0.0\)](#)
- Step 10: [FASTQ Groomer parallel \(version 1.0.0\)](#)
- Step 11: [Tophat parallel for illumina \(version 1.0.0\)](#)
- Step 12: [BAM to SAM \(version 1.0.0\)](#)
- Step 13: [CompareOverlapping_parallel \(version 1.0.0\)](#)
- Step 14: [countNumber_parallel \(version 1.0.0\)](#)
- Step 15: [DESEQ for differential expression analysis \(version 1.0.0\)](#)

Send results to a new history

[Run workflow](#)

History

Differential_expression_DESe 776.9 MB

- 19: [\[DESEQ\] Output log File](#)
- 18: [\[DESEQ\] Output down File](#)
- 17: [\[DESEQ\] Output up File](#)
- 16: [\[DESEQ\] Output complete File](#)
- 15: [\[DESEQ\] Output HTML File](#)
- 14: [countNumber Output](#)
- 13: [overlapping output files](#)
- 12: [converted SAM LIST files](#)
- 11: [\[tophat_parallel\] txt File](#)
- 10: [FASTQ Groomer parallel on data 9](#)
- 9: [loadMultiFASTQFiles result](#)
- 8: [Clean Transcript File on data 2](#)
- 7: [genome.fa](#)
- 6: [G12S4.fastq](#)
- 5: [G12S2.fastq](#)
- 4: [G12S1.fastq](#)
- 3: [G13S5.fastq](#)
- 2: [G12S3.fastq](#)
- 1: [genes.gtf](#)

Where to find tutorial page?

Shared Data ▾ Visualizat

- Data Libraries
- Published Histories
- Published Workflows
- Published Visualizations
- Published Pages**

Galaxy / ABiMS Analyze Data Workflow Shared Da

Published Pages

search title, annotation, owner, and tags

[Advanced Search](#)

Title	Annotation
Differential expression pipeline Documentation	This pipeline performs differential expression analysis with two different conditions.
MAPHiTS Tutorial	

Galaxy / ABiMS Analyze Data Workflow Shared Data ▾ Visualization ▾ Help ▾ User ▾ Using 5.3 GB

Published Pages | yluo | Differential expression pipeline Documentation

Welcome to Differential Expression Analysis tutorial page.

This pipeline performs differential expression analysis with two different conditions. One reference genome (fasta format), one annotation (gtf format) and several RNA-seq samples are required.

Step 0: Upload RNA-seq samples.
 Step 1: Clean the annotation file.
 Step 2: Clean the RNA-seq files.
 Step 3: Map the RNA-seq samples to the genome reference, using Tophat.
 Step 4: Convert the bam files (given by Tophat) to sam files.
 Step 5: Count the reads number per annotation using S-MART (CompareOverlapping).
 Step 6: Build the input files for DESeq.
 Step 7: Differential expression analysis using DESeq, and output graphical results.

Import data in your current history.

Data Pre-processing:

Step0:
 Upload RNA-seq samples using 'load_MultiFASTQFiles' tool. You can choose 'single-end' or 'paired-end' as data style (only 'single-end' option for this pipeline version).

load_multiFASTQFiles (version 1.0.0)

The uploading fastq files for single-end or paire-end mapping mode:

Condition groups

Condition group 1

Replicates

Condition group 2

Replicates

About this Page

Author
yluo

Related Pages
[All published pages](#)
[Published pages by yluo](#)

Rating
 Community (0 ratings, 0.0 average) ★★★★★
 Yours ★★★★★

Tags
 Community: none
 Yours:

Example

input data (GEO)

Tools

search tools

DATA MANAGEMENT

[Get Data](#)

[Send Data](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

ECOLE GGB GROUPE 1

[Mardi 15 RNA-seq](#)

[Mercredi 16 RNA-seq de novo ABIMS](#)

[Mercredi 16 ChIP-seq J. van Helden](#)

[Jeudi 17 ChIP-seq J. van Helden](#)

[Jeudi 17 - miRNAs Qualite / Nettoyage / Mirdeep2 / Annotation](#)

ECOLE GGB GROUPE 2

[Mardi 15 URGI: MAPHITS - PreProcess Tools](#)

[Mardi 15 URGI: MAPHITS - Tools](#)

[Mardi 15 URGI: MAPHITS - PostProcess Tools](#)

[Mardi 15 URGI: MAPHITS - SNPs Chip Tools](#)

[Mardi 15 URGI: S-MART - Tools](#)

[Mardi 15 URGI: S-MART - Differential Expression Pipeline Tools](#)

[Mercredi 16 - Annotation fonctionnelle des SNPs](#)

[Jeudi 17 - SVDetect](#)

ABIMS TOOLS

[Tools](#)

[Metabolic](#)

COMPARISON TOOLS

[Evolution](#)

[Multiple Alignments](#)

[FASTA manipulation](#)

NGS TOOLS

[NGS: QC and manipulation](#)

[NGS: Picard \(beta\)](#)

Expand All Collapse

Running workflow "Differential_expression_DESeq"

Step 1: Input dataset

annotation.gtf

8: Clean Transcript...e on data 2

type to filter

Step 2: Input dataset

Input replicat 1 for condition 1, fastq format

6: G12S4.fastq

type to filter

Step 3: Input dataset

Input replicat 2 for condition 1, fastq format

6: G12S4.fastq

type to filter

Step 4: Input dataset

Input replicat 1 for condition 2, fastq format

6: G12S4.fastq

type to filter

Step 5: Input dataset

Input replicat 2 for condition 2, fastq format

6: G12S4.fastq

type to filter

Step 6: Input dataset

reference genome .fasta

7: genome.fa

type to filter

History

Differential_expression_DESe 776.9 MB

19: [DESEQ] Output log File

18: [DESEQ] Output down File

17: [DESEQ] Output up File

16: [DESEQ] Output complete File

15: [DESEQ] Output HTML File

14: countNumber Output

13: overlapping output files

12: converted SAM LIST files

11: [tophat_parallel] txt File

10: FASTQ Groomer parallel on data 9

9: loadMultiFASTQFiles result

8: Clean Transcript File on data 2

7: genome.fa

6: G12S4.fastq

5: G12S2.fastq

4: G12S1.fastq

3: G13S5.fastq

2: G12S3.fastq

1: genes.gtf

→

Annotation

↳

Replicates inputs per condition

→

Reference genome

Example

Step 9: FASTQ Groomer parallel (version 1.0.0)

The File list to groom

Output dataset 'multiFASTQfiles_out' from step 8

Input FASTQ quality scores type
illumina 1.3-1.7



Option: illumina by default.

Advanced Options

Hide Advanced Options

Step 10: Tophat parallel for Illumina (version 1.0.0)

RNA-Seq FASTQ file

Output dataset 'output_file' from step 9

Will you select a reference genome from your history or use a built-in index?

Use one from the history

Select the reference genome

Output dataset 'output' from step 6

Is this library mate-paired?

Single-end

TopHat settings to use

Use Defaults

tar option



Option: for each step, you can choose to output a tar file to visualize the results.

Step 11: BAM to SAM (version 1.0.0)

BAM File LIST to Convert

Output dataset 'outputFileName' from step 10

Include header in output

tar option

Step 12: CompareOverlapping_parallel (version 1.0.0)

Example

search tools

DATA MANAGEMENT

[Get Data](#)

[Send Data](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

ECOLE GGB GROUPE 1

[Mardi 15 RNA-seq](#)

[Mercredi 16 RNA-seq de novo ABIMS](#)

[Mercredi 16 ChIP-seq J. van Helden](#)

[Jeudi 17 ChIP-seq J. van Helden](#)

[Jeudi 17 - miRNAs Qualite / Nettoyage / Mirdeep2 / Annotation](#)

ECOLE GGB GROUPE 2

[Mardi 15 URGI: MAPHITS - PreProcess Tools](#)

[Mardi 15 URGI: MAPHITS - Tools](#)

[Mardi 15 URGI: MAPHITS - PostProcess Tools](#)

[Mardi 15 URGI: MAPHITS - SNPs Chip Tools](#)

[Mardi 15 URGI: S-MART - Tools](#)

[Mardi 15 URGI: S-MART - Differential Expression Pipeline Tools](#)

[Mercredi 16 - Annotation fonctionnelle des SNPs](#)

[Jeudi 17 - SVDetect](#)

ABIMS TOOLS

[Tools](#)

[Metabolomic](#)

COMPARISON TOOLS

[Evolution](#)

[Multiple Alignments](#)

[FASTA manipulation](#)

NGS TOOLS

[NGS: QC and manipulation](#)

[NGS: Picard \(beta\)](#)

Extension towards 5 for file 2
No

Extension towards 3 for file 1
No

Extension towards 3 for file 2
No

Colinear or anti-sens
NONE

Maximum Distance between two reads
No

Minimum number of overlapping between two reads
No

Invert match
False

Report intron
False

When there is no overlapping, the number of Overlapping will be set to 0 by default.
True

tar option
False

Step 13: countNumber_parallel (version 1.0.0)

Please choose your txt format file (which contains a list of gtf3 overlapping results files).
Output dataset 'outTxtFile' from step 12

tar option

Step 14: DESEQ for differential expression analysis (version 1.0.0)

Input File list
Output dataset 'outputTxt' from step 13

If there is a header for your count files, please choose this case.
True

If your data has not replicates, please choose this case.

Send results to a new history named:

Differential_expression_DESe 776.9 MB

19: [DESEQ] Output log File

18: [DESEQ] Output down File

17: [DESEQ] Output up File

16: [DESEQ] Output complete File

15: [DESEQ] Output HTML File

14: countNumber Output

13: overlapping output files

12: converted SAM LIST files

11: [tophat_parallel] txt File

10: FASTQ Groomer parallel on data 9

9: loadMultiFASTQFiles result

8: Clean Transcript File on data 2

7: genome.fa

6: G12S4.fastq

5: G12S2.fastq

4: G12S1.fastq

3: G13S5.fastq

1: genes.gtf

Option: check the box if your data has a header.

Option: check the box if you do not have replicates data.

Choose to send the analysis results to a new history.

Example

Galaxy / ABiMS Analyze Data Workflow Shared Data Visualization Help User Using 5.3 GB

Tools

DATA MANAGEMENT

[Get Data](#)

[Send Data](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

ECOLE GGB GROUPE 1

[Mardi 15 RNA-seq](#)

[Mercredi 16 RNA-seq de novo ABiMS](#)

[Mercredi 16 ChIP-seq J. van Helden](#)

[Jeudi 17 ChIP-seq J. van Helden](#)

[Jeudi 17 - miRNAs Qualite / Nettoyage / Mirdeep2 / Annotation](#)

ECOLE GGB GROUPE 2

[Mardi 15 URGI: MAPHITS - PreProcess Tools](#)

[Mardi 15 URGI: MAPHITS - Tools](#)

[Mardi 15 URGI: MAPHITS - PostProcess Tools](#)

[Mardi 15 URGI: MAPHITS - SNPs Chip Tools](#)

[Mardi 15 URGI: S-MART - Tools](#)

[Mardi 15 URGI: S-MART - Differential Expression Pipeline Tools](#)

[Mercredi 16 - Annotation fonctionnelle des SNPs](#)

[Jeudi 17 - SVDetect](#)

ABiMS TOOLS

[Tools](#)

[Primer](#)

[RNASeq](#)

[Metabolomic](#)

[Debug](#)

Saved Histories

Advanced Search

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status	
<input type="checkbox"/> new_small_dataset	11	1	6	0 Tags	Accessible	829.3 MB	Dec 07, 2012 ~ 1 hour ago	current history
<input type="checkbox"/> Unnamed history				0 Tags		0 bytes	2 days ago	2 days ago
<input type="checkbox"/> new_data_test	6			0 Tags	Accessible	3.7 GB	Dec 06, 2012	Dec 07, 2012
<input type="checkbox"/> Differential_expression_DESeq	19			0 Tags		776.9 MB	Dec 03, 2012	Dec 03, 2012

For 0 selected histories:

Histories that have been deleted for more than a time period specified by the Galaxy administrator(s) may be permanently deleted.

History

- [new_small_dataset](#) 829.3 MB
- [18: \[DESEQ\] Output log File](#)
- [17: \[DESEQ\] Output down File](#)
- [16: \[DESEQ\] Output up File](#)
- [15: \[DESEQ\] Output complete File](#)
- [14: \[DESEQ\] Output HTML File](#)
- [13: countNumber Output](#)
- [12: overlapping output files](#)
- [11: converted SAM LIST files](#)
- [10: \[tophat_parallel\] txt File](#)
- [9: FASTQ Groomer parallel on data 8](#)
- [8: loadMultiFASTQFiles result](#)
- [7: Clean Transcript File on data 5](#)
- [6: arabidopsis_thaliana.fa](#)
- [5: arabidopsis_thaliana.gtf](#)
- [4: Group1_1.fastq](#)
- [3: Group1_2.fastq](#)
- [2: Group2_1.fastq](#)
- [1: Group2_2.fastq](#)

Results of Example