

Transcript quantification using Galaxy

Prepare input data for 3D RNA-seq App

Wenbin Guo

28 May 2019

Information & Computational Sciences, James Hutton Institute, Dundee DD2 5DA, UK

- Introduction
- Example data
- Registration of Galaxy Europe
- Galaxy user manual
- Structure of Galaxy interface
- Upload data to Galaxy
- Salmon quantification
- Kallisto quantification
- Prepare input files for 3D RNA-seq App
- References

Introduction

The 3D RNA-seq App takes transcript quantifications from Salmon (Patro et al., 2017) or Kallisto (Bray et al., 2016) as input for 3D analysis. The user manuals of command lines to generate quantifications can be found in:

- Salmon: <https://combine-lab.github.io/salmon/> (<https://combine-lab.github.io/salmon/>)
- Kallisto: <https://pachterlab.github.io/kallisto/about> (<https://pachterlab.github.io/kallisto/about>)

For biologists, we recommend to use the Salmon/Kallisto tool in web-based graphical user interface **Galaxy Europe** (<https://usegalaxy.eu/> (<https://usegalaxy.eu/>)) or other Galaxy resources (<https://galaxyproject.org/use/> (<https://galaxyproject.org/use/>)), in which users can perform transcript quantification and download results by “clicking mouse”.

Example data

Download link:

https://www.dropbox.com/s/k42kvxw9adrrcgp/Galaxy_example_data.zip?dl=0
(https://www.dropbox.com/s/k42kvxw9adrrcgp/Galaxy_example_data.zip?dl=0)

Transcriptome:

- A subset of AtRTD2 Arabidopsis transcriptome (Zhang et al, 2017) with 4679 transcripts from 1000 genes.

RNA-seq reads:

- Two conditions, 20°C vs 4°C.
- Each has 3 biological replicates.
- 150 bp paired-end reads.

Samples	Temperature	Bio-reps	Read1	Read2
Sample1	20	Brep1	sample_01_1	sample_01_2

Samples	Temperature	Bio-reps	Read1	Read2
Sample2	20	Brep2	sample_02_1	sample_02_2
Sample3	20	Brep3	sample_03_1	sample_03_2
Sample4	4	Brep1	sample_04_1	sample_04_2
Sample5	4	Brep2	sample_05_1	sample_05_2
Sample6	4	Brep3	sample_06_1	sample_06_2

Registration of Galaxy Europe

<https://usegalaxy.eu/> (<https://usegalaxy.eu/>)

The screenshot shows the Galaxy Europe web interface. The top navigation bar contains the following items: Galaxy / Europe, Analyze Data, Workflow, Visualize, Shared Data, Help, and Login or Register (highlighted with a red box). The 'Using 0 bytes' status is visible on the right. The main content area is divided into three sections: Tools (with a search bar and categories like Get Data, Send Data, etc.), News (with a quote from Prof. Stephen Hawking and a list of tool updates), and History (with a search bar and a message indicating it is empty).

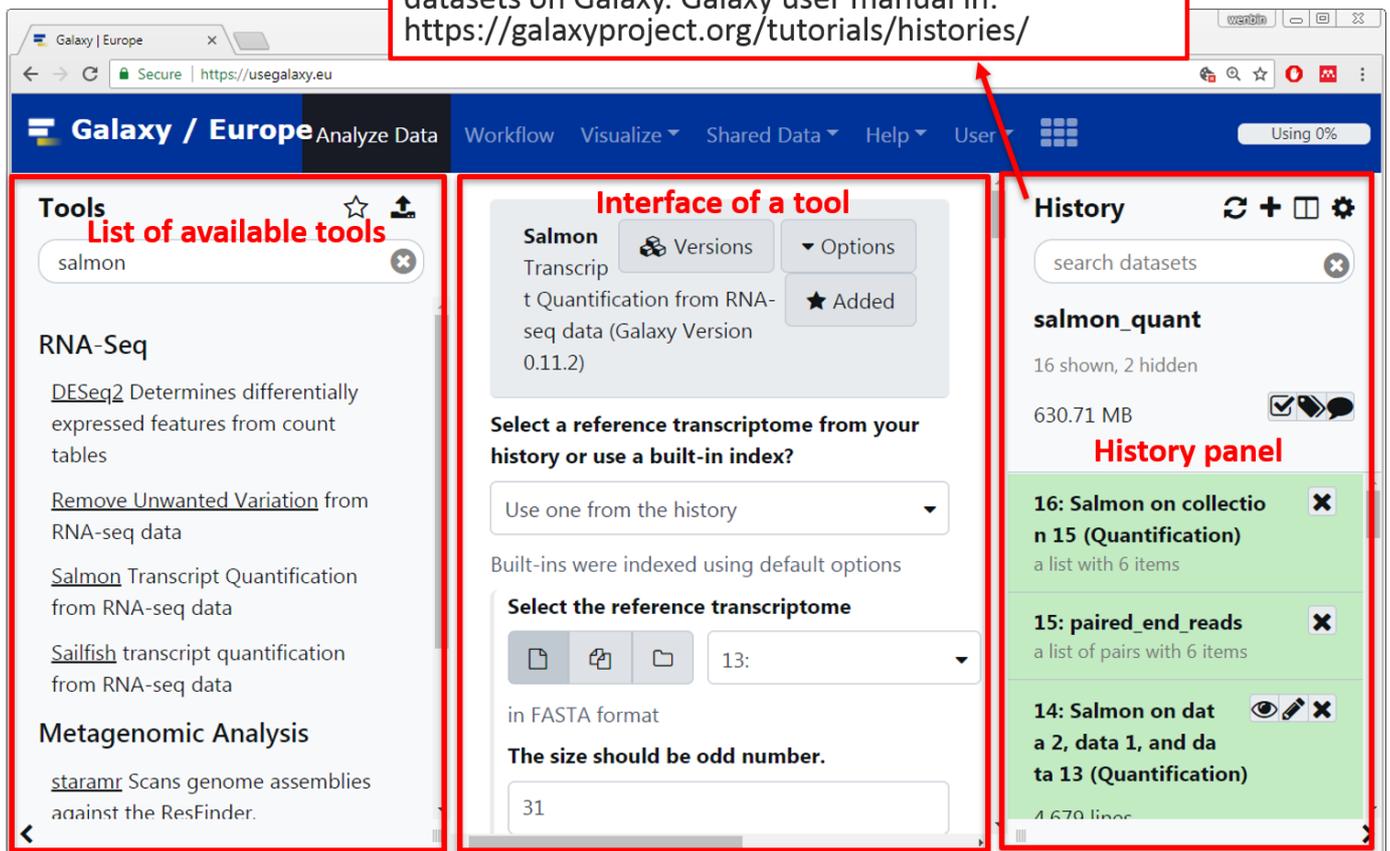
Note: The account needs email activation.

Galaxy user manual

Galaxy user manual can be found in: <https://galaxyproject.org/learn/> (<https://galaxyproject.org/learn/>)

Structure of Galaxy interface

History refers to the user's performance/stored datasets on Galaxy. Galaxy user manual in: <https://galaxyproject.org/tutorials/histories/>



Upload data to Galaxy

Galaxy user manual: <https://galaxyproject.org/tutorials/upload/> (<https://galaxyproject.org/tutorials/upload/>)

Two types of input files are required for transcript quantification using Salmon:

- RNA-seq reads in fasta/fastq format.
- Transcript sequence file in fasta (.fa) format.

Galaxy | Europe

Secure | https://usegalaxy.eu

Galaxy / Europe Analyze Data Workflow Visualize Shared Data Help User Using 0%

Tools 

search tools **Click to open upload data panel**

Get Data
Send Data
Collection Operations

GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group

GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
FASTQ Quality Control
CAM/DRAM

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

News

- Aug 3, 2019 **UseGalaxy.eu Tool Updates for 2019-08-03**
- Aug 2, 2019 **MaxQuant goes Galaxy**
- Aug 1, 2019 **Galactic News**
- Jul 27, 2019 **UseGalaxy.eu Tool Updates for 2019-07-27**
- Jul 20, 2019 **7000 users and 10.000.000 datasets**
- Jul 19, 2019 **UseGalaxy.eu Tool Updates for 2019-07-19**

History

search datasets

Unnamed history (empty)

This history has been purged and deleted

This history is empty. You can load your own data or get data from an external source

OPEN CHAT

Galaxy | Europe

Secure | https://usegalaxy.eu

Galaxy / Europe Analyze Data Workflow Visualize Shared Data Help User Using 0%

Tools

search tools

Get Data
Send Data
Collection Op
GENERAL TEXT
Text Manipul
Filter and So
Join, Subtrac
GENOMIC FILE I
Convert Form
FASTA/FASTQ
FASTQ Quali
CAM/DRAM

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 12 file(s) to the queue. Add more files or click 'Start' to proceed.

sample 05 1.fasta.	17.4 MB	fasta	Arabidopsis th...	0%
sample 05 2.fasta.	17.4 MB	fasta	Arabidopsis th...	0%
sample 06 1.fasta.	19.6 MB	fasta		%
sample 06 2.fasta.	19.6 MB	fasta		%

1. Upload RNA-seq read files of samples
2. Upload transcriptome sequence fasta file

Note: Salmon can take RNA-seq reads in either fasta.gz or fastq.gz format; Kallisto can only take fastq.gz

Select file type for all files or let auto-detect

Choose files to upload

Type (set all): fasta Genome (set all): Arabidopsis th...

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

OPEN CHAT

The screenshot shows the Galaxy Europe interface. The top navigation bar includes 'Galaxy / Europe', 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', and 'User'. The left sidebar contains a 'Tools' section with categories like 'Get Data', 'Send Data', 'Collection Operations', 'GENERAL TEXT TOOLS', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'GENOMIC FILE MANIPULATION', 'Convert Formats', 'FASTA/FASTQ', 'FASTQ Quality Control', and 'CAM/RAM'. The main content area features a quote by Stephen Hawking, a 'News' section with updates from August and July 2019, and an 'OPEN CHAT' button. On the right, the 'History' panel shows a search bar, a '+', a refresh icon, and a settings icon. Below this, it displays 'Unnamed history', '12 shown', and '1.86 GB'. A warning message states 'This history has been purged and deleted'. A list of datasets is shown, including '12: sample_06_2.fasta.gz', '11: sample_06_1.fasta.gz', '10: sample_05_2.fasta.gz', '9: sample_05_1.fasta.gz', '8: sample_04_2.fasta.gz', and '7: sample_04_1.fasta.gz'. Each dataset entry has an eye icon, a pencil icon, and an 'X' icon. Red annotations highlight the '+' icon, the 'Unnamed history' title, and the dataset list. Other annotations point to the quote and the 'MaxQuant goes Galaxy' news item.

Salmon quantification

Input files:

- Transcriptome sequence fasta file with extension “.fa”.
- RNA-seq reads files of samples with extension “.fastq.gz” or “.fasta.gz”

Salmon documentation:

- Salmon: <https://combine-lab.github.io/salmon/> (<https://combine-lab.github.io/salmon/>)

In most cases, RNA-seq data includes many samples and replicates. Galaxy allows users to build a list of data pairs from RNA-seq data. The user manual can be found in: Galaxy user manual can be found in: <https://galaxyproject.org/tutorials/collections/> (<https://galaxyproject.org/tutorials/collections/>)

History

search datasets

salmon_quant

14 shown

630.71 MB

Operations on multiple datasets

(1) In the history panel, click the tick sign to manipulate multiple datasets

14: Salmon on data 2, data 1, and data 13 (Quantification)

13: AtRTD2_1000genes.fa

12: sample_06_2.fasta.gz

11: sample_06_1.fasta.gz

10: sample_05_2.fasta.gz

9: sample_05_1.fasta.gz

8: sample_04_2.fasta.gz

Tick boxes are appeared

History

search datasets

salmon_quant

14 shown

630.71 MB

All None For all selected...

(4) In the select input list, select "Build List of Data pairs"

(3) Click here

(2) Select the paired-end RNA-seq read files

14: Salmon on data 2, data 1, and data 13 (Quantification)

13: AtRTD2_1000genes.fa

12: sample_06_2.fasta.gz

11: sample_06_1.fasta.gz

10: sample_05_2.fasta.gz

9: sample_05_1.fasta.gz

Create a collection of paired datasets

(5) Pop-up panel

6 pairs created: all datasets have been successfully paired

0 unpaired forward - (0 filtered out)

Choose filters Clear filters

0 unpaired reverse - (0 filtered out)

sample_01_1.fasta.gz → sample_01.fasta ← sample_01_2.fasta.gz

sample_02_1.fasta.gz → sample_02.fasta ← sample_02_2.fasta.gz

sample_03_1.fasta.gz → sample_03.fasta ← sample_03_2.fasta.gz

sample_04_1.fasta.gz → sample_04.fasta ← sample_04_2.fasta.gz

sample_05_1.fasta.gz → sample_05.fasta ← sample_05_2.fasta.gz

sample_06_1.fasta.gz → sample_06.fasta ← sample_06_2.fasta.gz

6 paired Unpair all

Remove file extensions from pair names? Hide original elements?

Name: paired_end_reads (7) Give a name

(8) Click to generate the collection Create list

(6) Pairs are automatically detected. If incorrect, modify them

(9) A new history of paired data collection is added in the history panel

History

search datasets

salmon_quant

14 shown, 2 hidden

630.71 MB

All None For all selected...

15: paired_end_reads

a list of pairs with 6 items

(10) Back to the salmon quantification panel. If lost, search "salmon" again in the "Tools"

Basic settings

Salmon Transcript Quantification from RNA-seq data (Galaxy Version 0.11.2)

Select a reference transcriptome from your history or use a built-in index?

Use one from the history

Built-ins were indexed using default options

Select the reference transcriptome

13: AtRTD2_1000genes.fa (11) Select transcriptome sequence fasta file

in FASTA format

The size should be odd number.

31 (12) Set k-mer, default is 31

(kmerLen)

Is this library mate-paired?

Paired-end Dataset Collection (13) Select "Paired-end Data Collection"

FASTQ Paired Dataset

15: paired_end_reads (with implicit datatype conversion) (14) Select the paired-end reads collection generated in previous step

Must be of datatype "fastqsanger" or "fasta"

Relative orientation of reads within a pair

Mates are oriented toward each other (I = inward)

(15) Other settings according to data details

(16) Execute quantification

Execute

Galaxy / Europe

Tools

salmon

RNA-Seq

DESeq2 Determines differentially expressed features from count tables

Remove Unwanted Variation from RNA-seq data

Salmon Transcript Quantification from RNA-seq data

Sailfish transcript quantification from RNA-seq data

Metagenomic Analysis

staramr Scans genome assemblies against the ResFinder.

Executed **Salmon** and successfully added 6 jobs to the queue.

The tool uses 2 inputs:

15: paired_end_reads (with implicit datatype conversion)

13: AtRTD2_1000genes.fa

It produces 6 outputs:

22: Salmon on data 12, data 11, and data 13 (Quantification)

21: Salmon on data 10, data 9 and data 13

History

search datasets

salmon_quant

16 shown, 2 hidden

630.71 MB

All None

For all selected...

16: Salmon on collect on 15 (Quantification)

6 jobs generating a list

15: paired_end_reads

a list of pairs with 6 items

14: Salmon on data 2, dat

(17) A new item in history, 6 job is running for 6 samples

A few minutes later ...

History ↻ + ▢ ⚙

search datasets ✕

salmon_quant

16 shown, 2 hidden

630.71 MB ☑️ 🗑️ 💬

(18) Quantification is done

16: Salmon on collection 15 (Quantification) ✕

a list with 6 items

15: paired_end_reads ✕

a list of pairs with 6 items

Click this history

History ↻ + ▢ ⚙

< Back to salmon_quant

Salmon on collection 15 (Quantification)

a list with 6 items (20) Click to download to local computer 📄

sample_01.fasta 👁️ 🗑️

sample_02.fasta 👁️ 🗑️

sample_03.fasta 👁️ 🗑️

sample_04.fasta 👁️ 🗑️

sample_05.fasta 👁️ 🗑️

sample_06.fasta 👁️ 🗑️

sample_01.fasta.tabular

sample_02.fasta.tabular

sample_03.fasta.tabular

sample_04.fasta.tabular

sample_05.fasta.tabular

sample_06.fasta.tabular

(21) Unzip the download, a list of quantification in .tabular format, can be visualised by "excel"

(19) A list, each element is the quantification for a sample

Kallisto quantification

Input files:

- Transcriptome sequence fasta file with estension ".fa".
- RNA-seq reads files of samples with estension ".fastq.gz" (Kallisto in Galaxy does not take ".fasta.gz" format).

Kallisto documentation:

- Kallisto: <https://pachterlab.github.io/kallisto/about> (<https://pachterlab.github.io/kallisto/about>)

Basic settings

Kallisto quant - quantify abundances of RNA-Seq transcripts (Galaxy Version 0.43.1.4)

☆ Favorite 🔄 Versions ▾ Options

Reference transcriptome for quantification

Use a transcriptome from history (3) Select transcriptome from history

FASTA reference transcriptome

13: AtRTD2_1000genes.fa (4) Select the transcriptome fasta file

Single-end or paired reads

Paired (5) Select "paired" for this example

Collection or individual datasets

Pair or list of pairs (6) Select "Pair or list of pairs"

Collection of reads

15: paired_end_reads (7) Select the list of data collection generated in previous step

Perform sequence based bias correction

Yes No (8) Other settings according to data details

(9) Execute quantification

Execute

History

search datasets

Transcript_quant

17 shown

283.54 MB

(10) Job is added to history

17: Kallisto quant on collection 15: Abundances (tabular) 6 jobs generating a list

16: Kallisto quant on collection 15: Abundances (HDF5) 6 jobs generating a list

15: paired_end_reads a list of pairs with 6 items

.h5 format

sample_01.fastq.h5
sample_02.fastq.h5
sample_03.fastq.h5
sample_04.fastq.h5
sample_05.fastq.h5
sample_06.fastq.h5

.tabular format

sample_01.fastq.tabular
sample_02.fastq.tabular
sample_03.fastq.tabular
sample_04.fastq.tabular
sample_05.fastq.tabular
sample_06.fastq.tabular

Unzipped files in local folder

History

search datasets

Transcript_quant

17 shown

283.54 MB

(11) Job is done

17: Kallisto quant on collection 15: Abundances (tabular) a list with 6 items

16: Kallisto quant on collection 15: Abundances (HDF5) a list with 6 items

15: paired_end_reads a list of pairs with 6 items

sample_01.fastq

sample_02.fastq

(12) Click history to access the data and download to local folder

A few minutes later ...

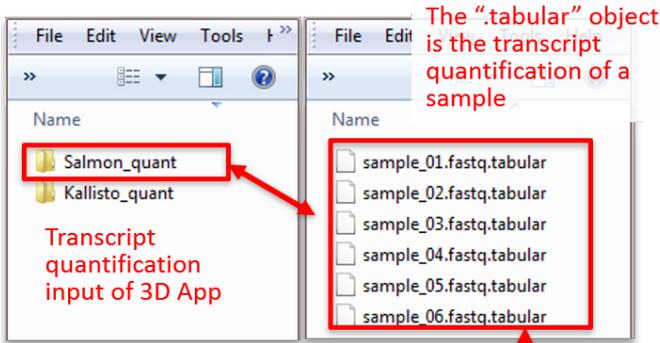
Output in tabular format

Output in .h5 format

Prepare input files for 3D RNA-seq App

The 3D RNA-seq App reads transcript quantifications in ".tabular" files from Galaxy outputs.

Salmon

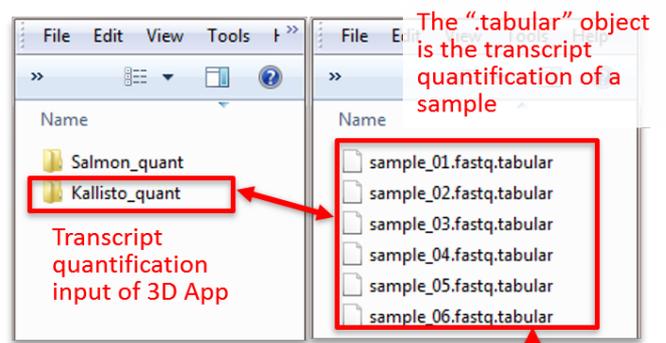


Put the file names to the meta-table input of 3D App

Samples	Condition	Bio-reps	Quant_files
Sample1	20°C	Brep1	sample_01.fastq.tabular
Sample2	20°C	Brep2	sample_02.fastq.tabular
Sample3	20°C	Brep3	sample_03.fastq.tabular
Sample4	4°C	Brep1	sample_04.fastq.tabular
Sample5	4°C	Brep2	sample_05.fastq.tabular
Sample6	4°C	Brep3	sample_06.fastq.tabular

Meta-data table in csv

Kallisto



Put the file names to the meta-table input of 3D App

Samples	Condition	Bio-reps	Quant_files
Sample1	20°C	Brep1	sample_01.fastq.tabular
Sample2	20°C	Brep2	sample_02.fastq.tabular
Sample3	20°C	Brep3	sample_03.fastq.tabular
Sample4	4°C	Brep1	sample_04.fastq.tabular
Sample5	4°C	Brep2	sample_05.fastq.tabular
Sample6	4°C	Brep3	sample_06.fastq.tabular

Meta-data table in csv

Note: Galaxy can generate ".tabular" files for both Salmon and Kallisto

References

Bray,N.L., Pimentel,H., Melsted,P., and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34, 525–527.

Calixto,C.P.G., Guo,W., James,A.B., Tzioutziou,N.A., Entizne,J.C., Panter,P.E., Knight,H., Nimmo,H.G., Zhang,R., and Brown,J.W.S. (2018) Rapid and Dynamic Alternative Splicing Impacts the Arabidopsis Cold Response Transcriptome. *Plant Cell*, 30, 1424–1444.

Guo,W., Tzioutziou,N., Stephen,G., Milne,I., Calixto,C., Waugh,R., Brown,J.W., and Zhang,R. (2019) 3D RNA-seq - a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists. *bioRxiv*, 656686. doi: <https://doi.org/10.1101/656686> (<https://doi.org/10.1101/656686>).

Patro,R., Duggal,G., Love,M.I., Irizarry,R.A., and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14, 417–419.