# 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**:

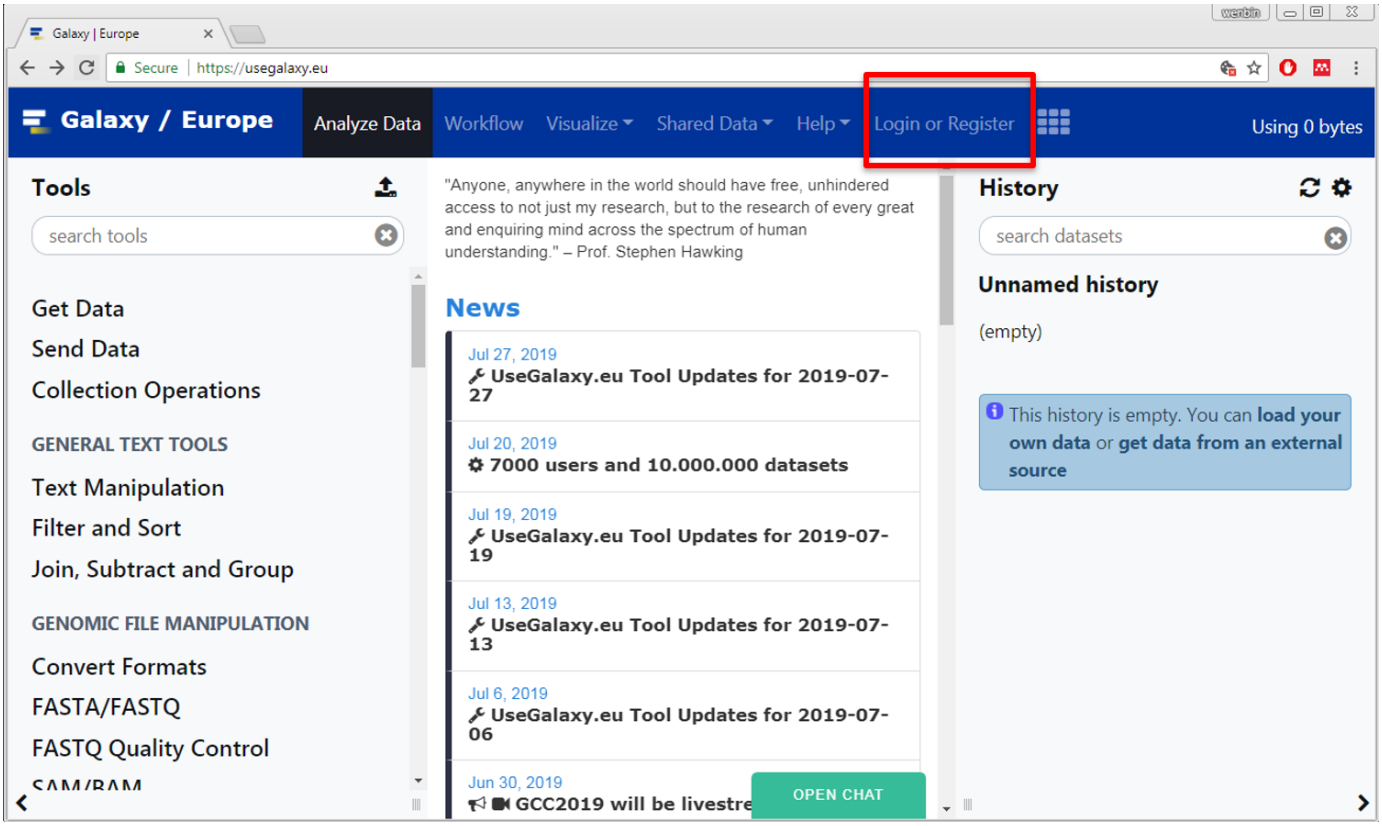
- Two conditions, $20^oC$ vs $4^oC$.
- Each has 3 biological replicates.
- 150 bp paired-end reads.

| 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/)

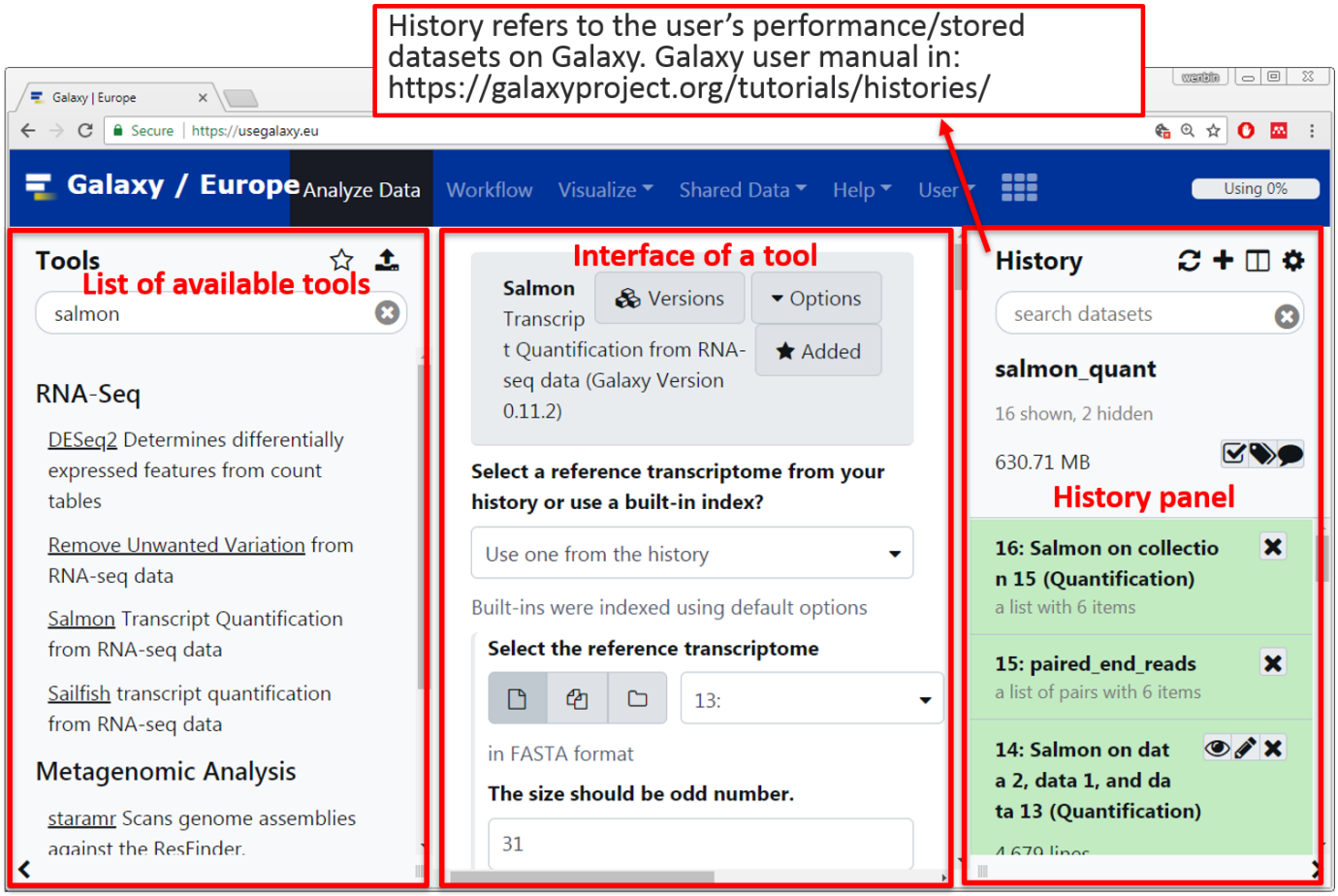



**Note**: The account needs email activation.

# Galaxy user manual

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

# Structure of Galaxy interface

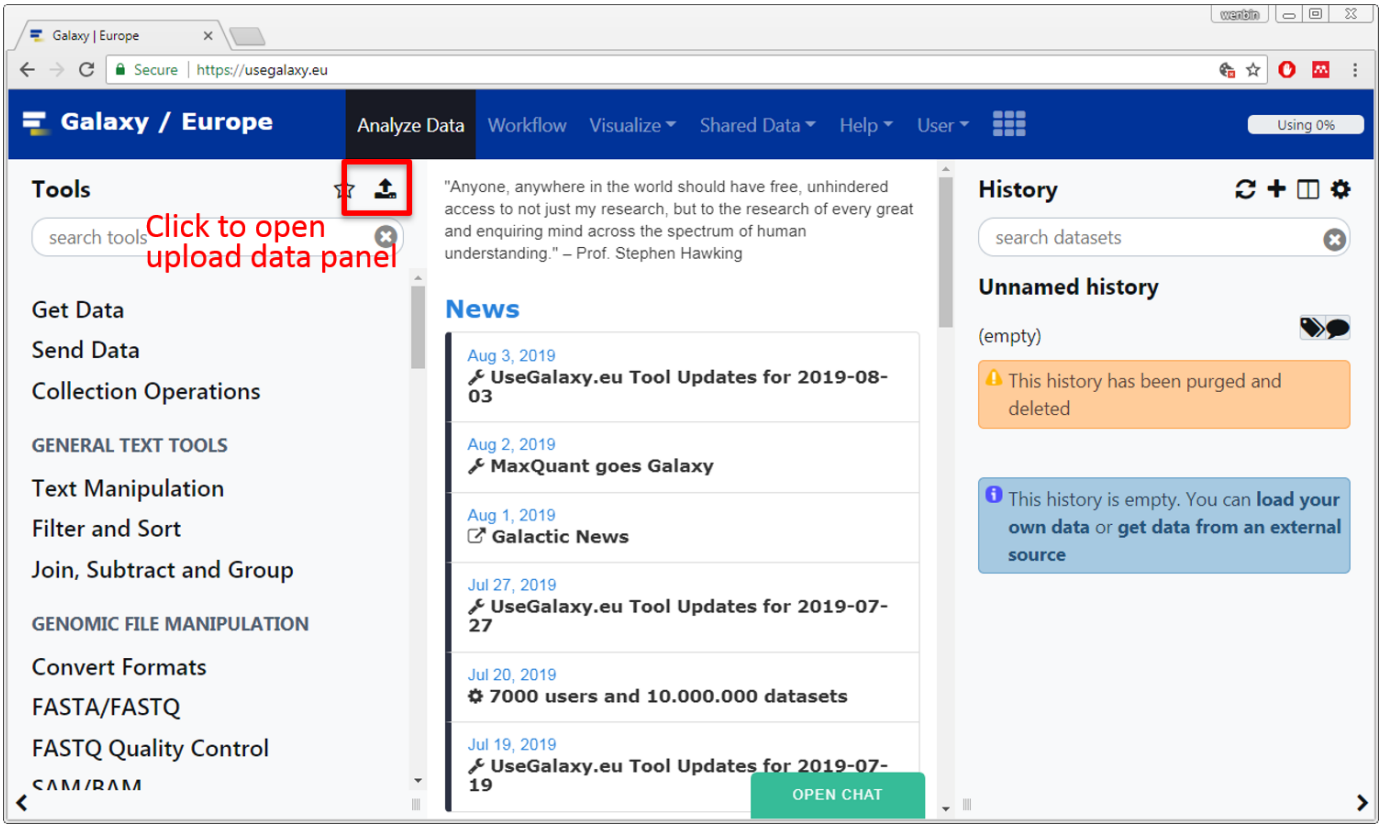


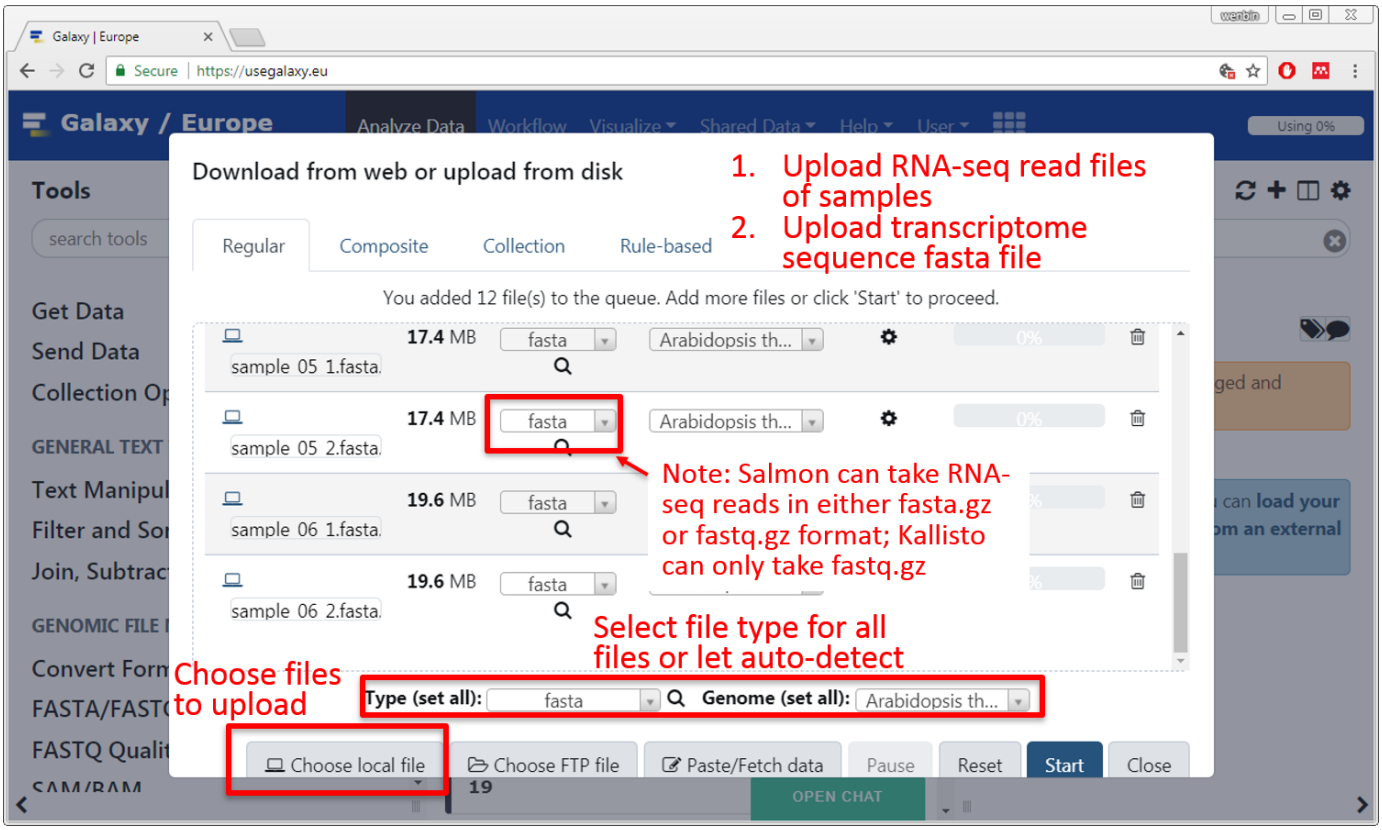

# Upload data to Galaxy

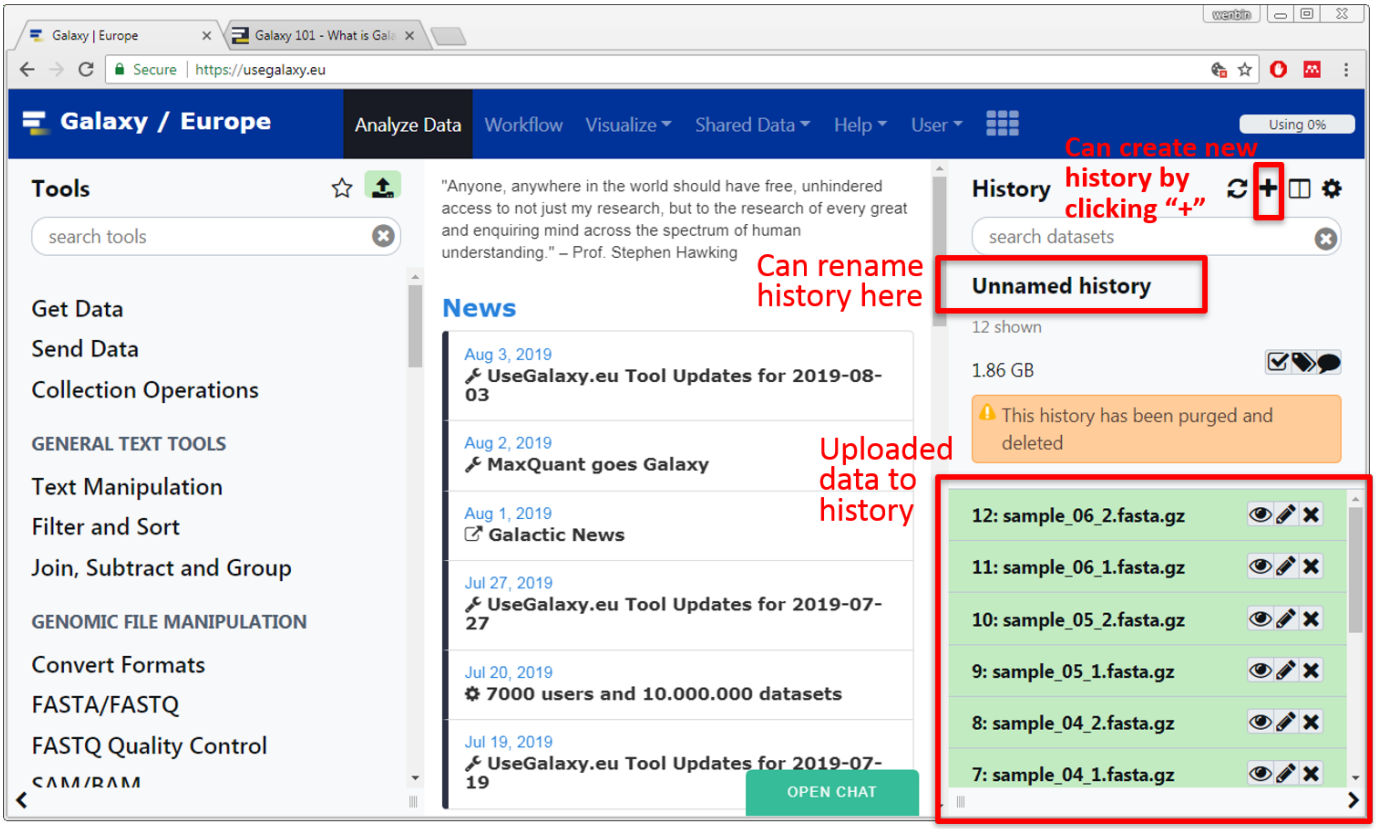
Galaxy user manual: https://galaxyproject.org/tutorials/upload/ (https://galaxyproject.org/tutorials/upload/)

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
"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
OPEN CHAT
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


Galaxy | Europe
Secure | https://usegalaxy.eu
Galaxy / Europe
Download from web or upload from disk
1. Upload RNA-seq read files of samples
2. Upload transcriptome sequence fasta file
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...
sample_05_2.fasta
17.4 MB
fasta
Arabidopsis th...
sample_06_1.fasta
19.6 MB
fasta
sample_06_2.fasta
19.6 MB
fasta
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




# Salmon quantification

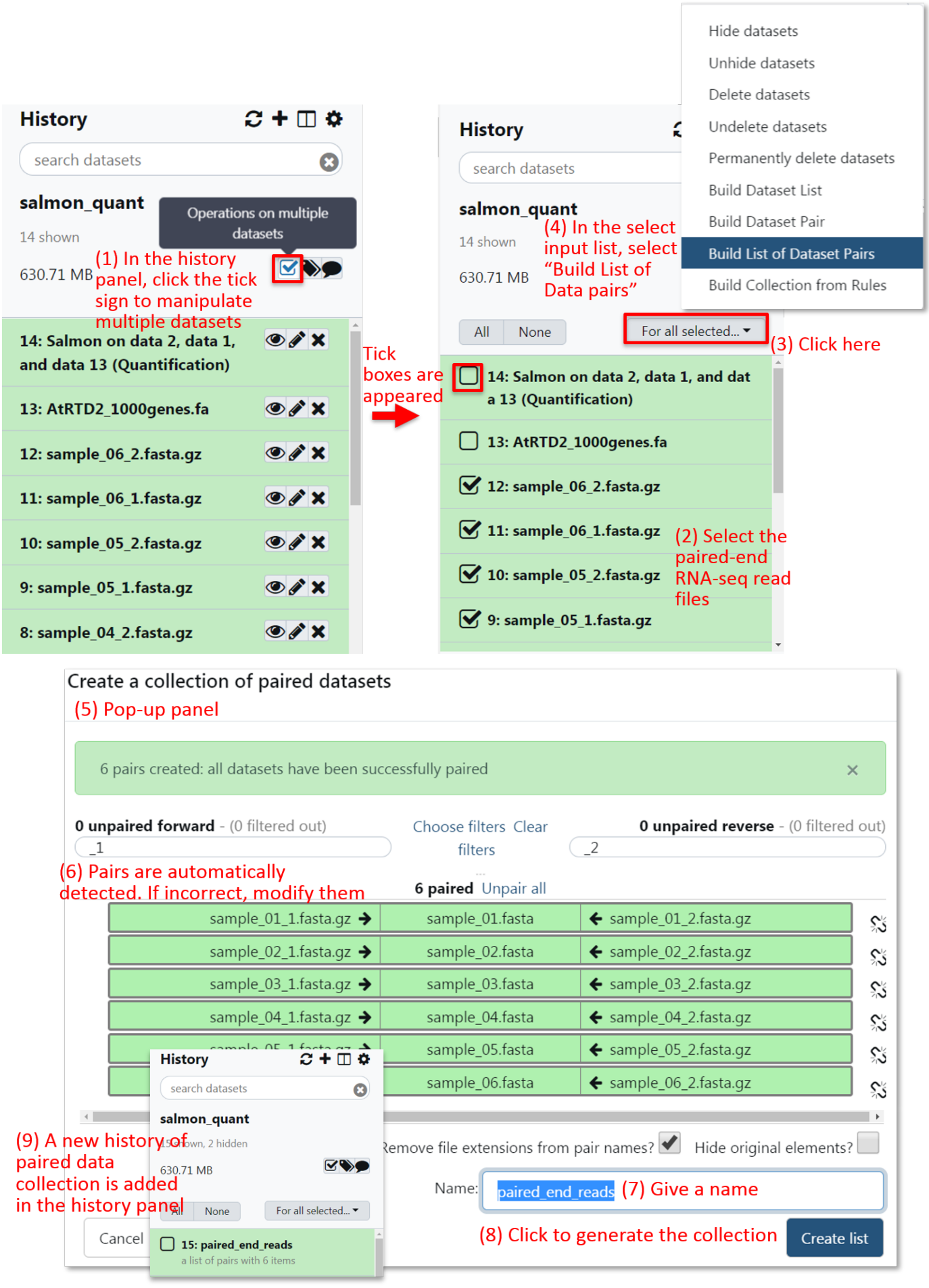
Input files:

- Transcriptome sequence fasta file with estension ".fa".
- RNA-seq reads files of samples with estension ".fastq.gz" or ".fasta.gz"

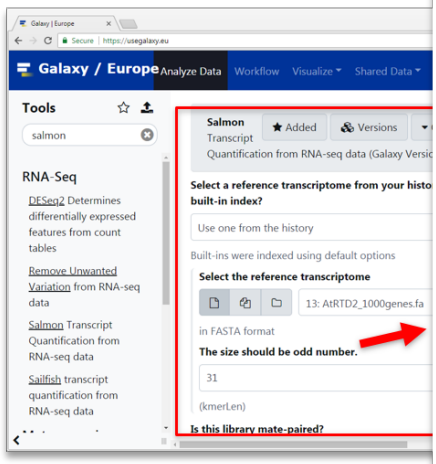
Salmon documentation:

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

Im most cases, RNA-sesq 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
Hide datasets
Unhide datasets
Delete datasets
Undelete datasets
Permanently delete datasets
Build Dataset List
Build Dataset Pair
Build List of Dataset Pairs
Build Collection from Rules
(4) In the select input list, select "Build List of Data pairs"
All
None
For all selected...
(3) Click here
(2) Select the paired-end RNA-seq read files
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)
_1
Choose filters Clear filters
0 unpaired reverse - (0 filtered out)
_2
(6) Pairs are automatically detected. If incorrect, modify them
6 paired Unpair all
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.fasta
sample_05_2.fasta.gz
sample_06.fasta
sample_06_2.fasta.gz
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
Cancel
(9) A new history of paired data collection is added in the history panel
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) ★ Added | Versions | Options

**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"

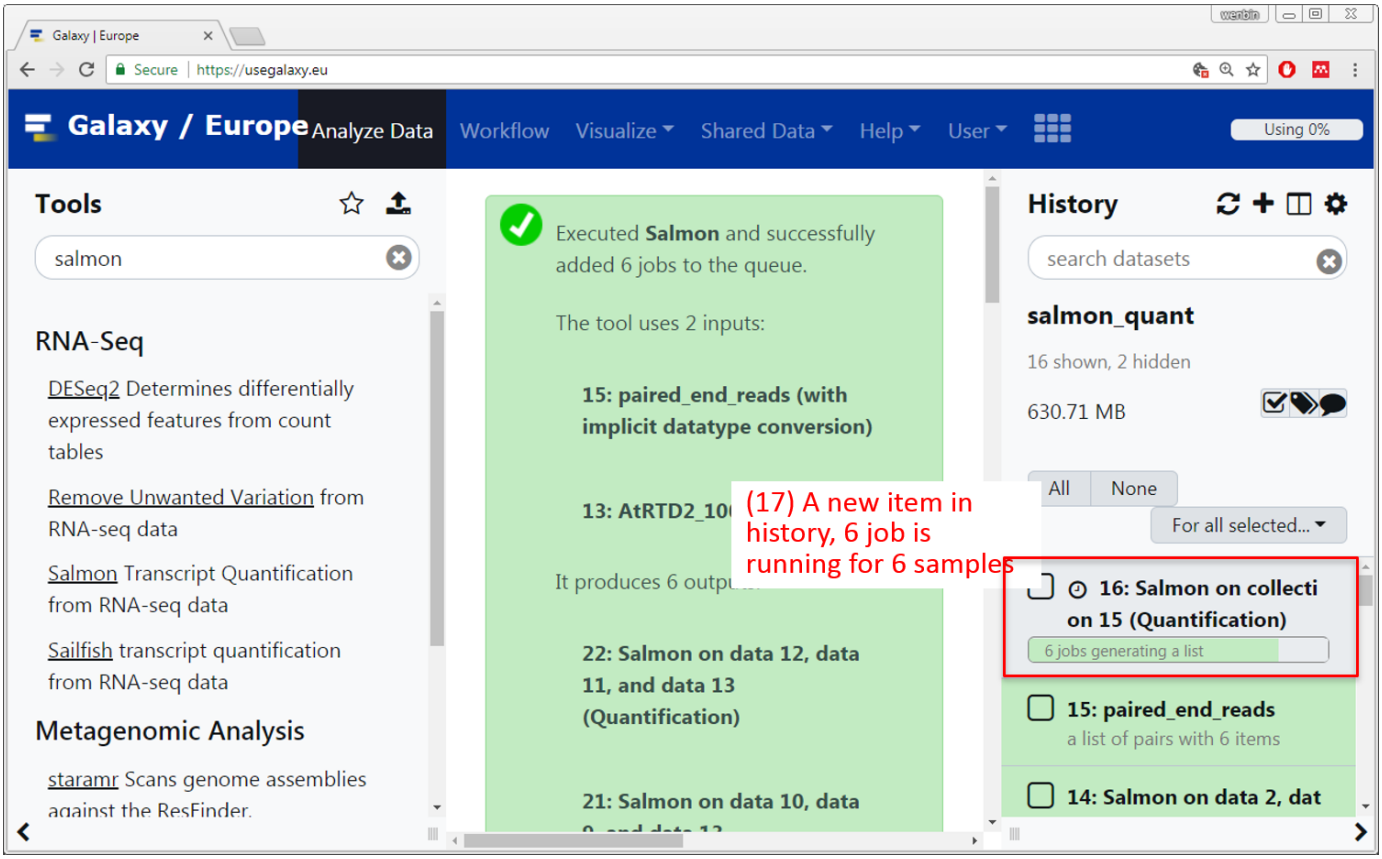
**Relative orientation of reads within a pair**

Mates are oriented toward each other (I = inward)

(15) Other settings according to data details

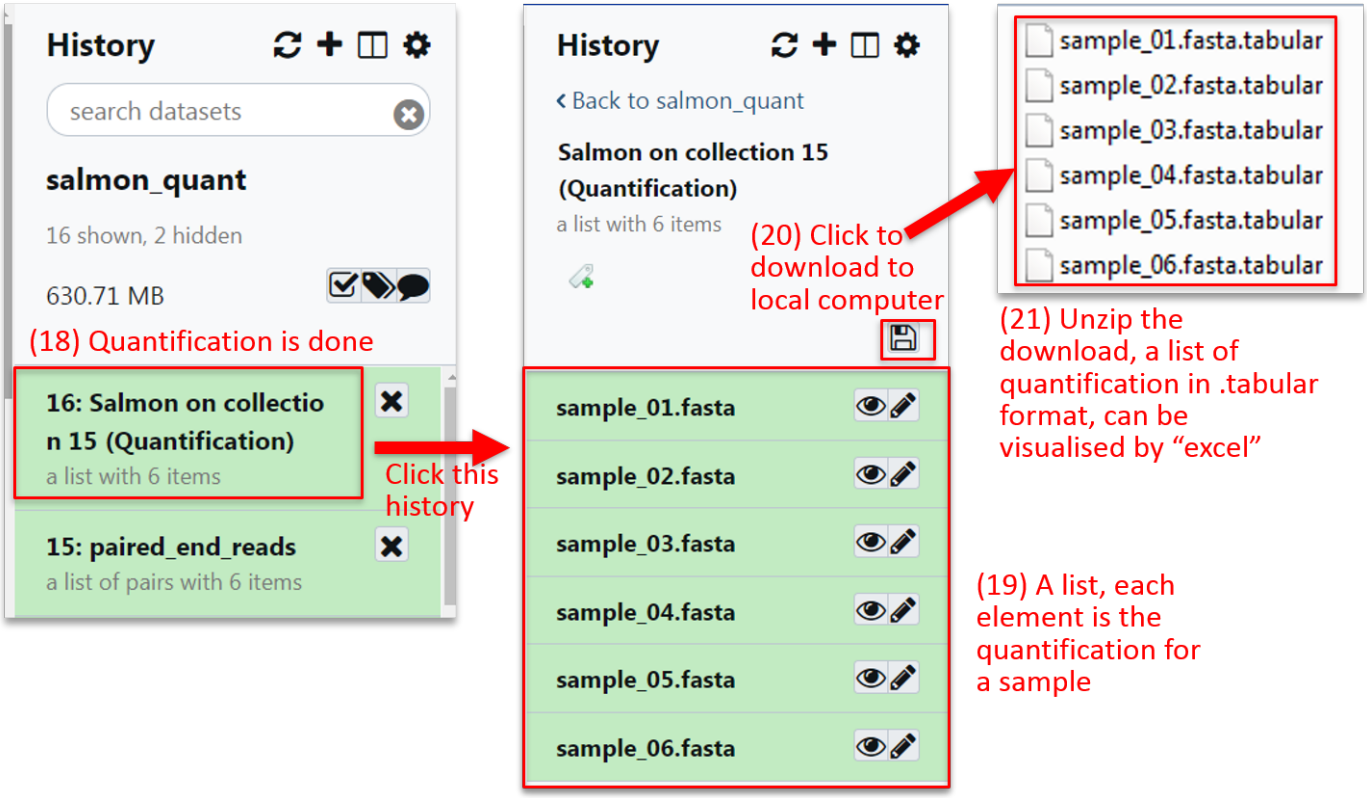
(16) Execute quantification

✔ Execute




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

# A few minutes later ...




# Kallisto quantification

Input files:

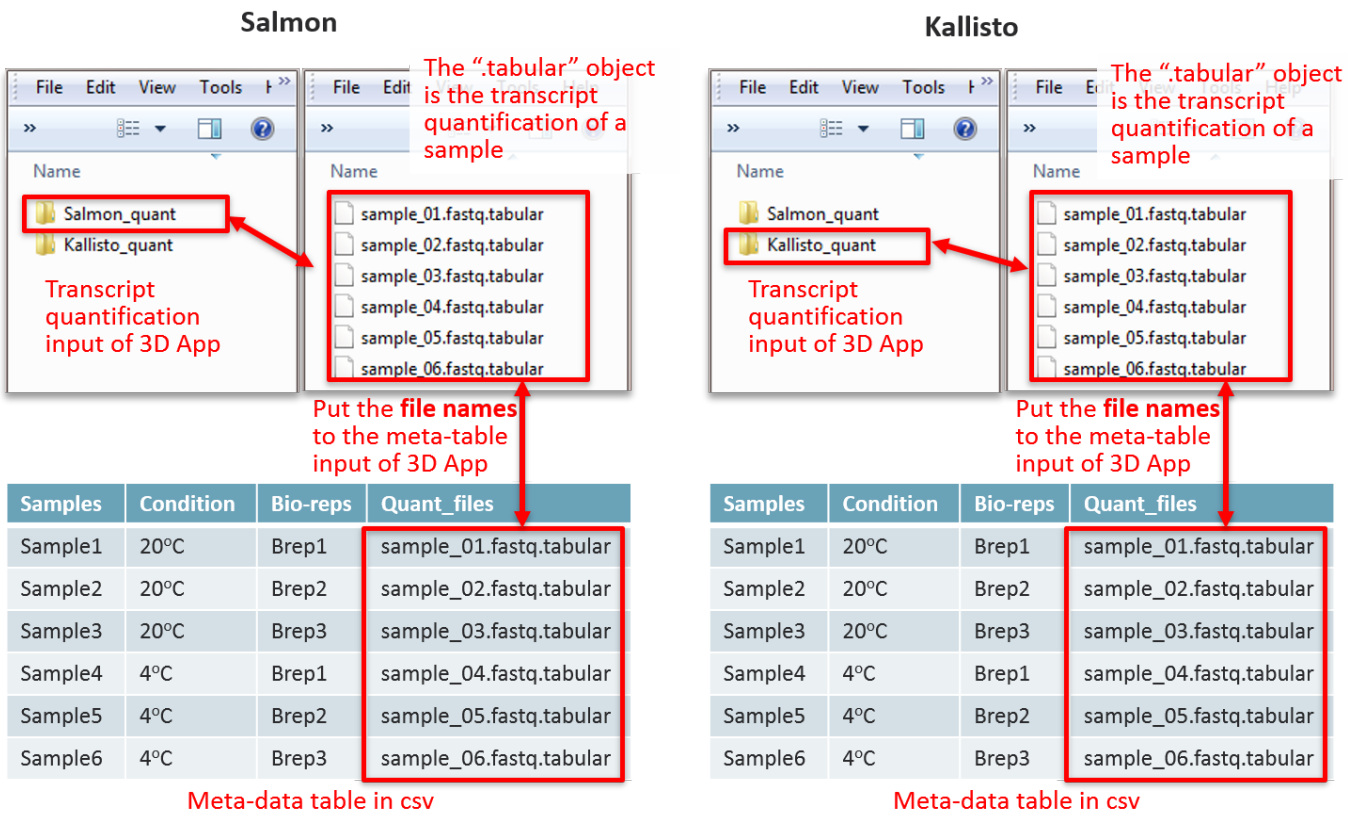
- Transcriptome sequence fasta file with estension “.fa”.
- RNA-seq reads files of samples with estension “.fastq.gz” (Kallisto in Galaxy does not take “.fasta.gz” format).

Kallisto documentation:

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


# Prepare input files for 3D RNA-seq App

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




**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.