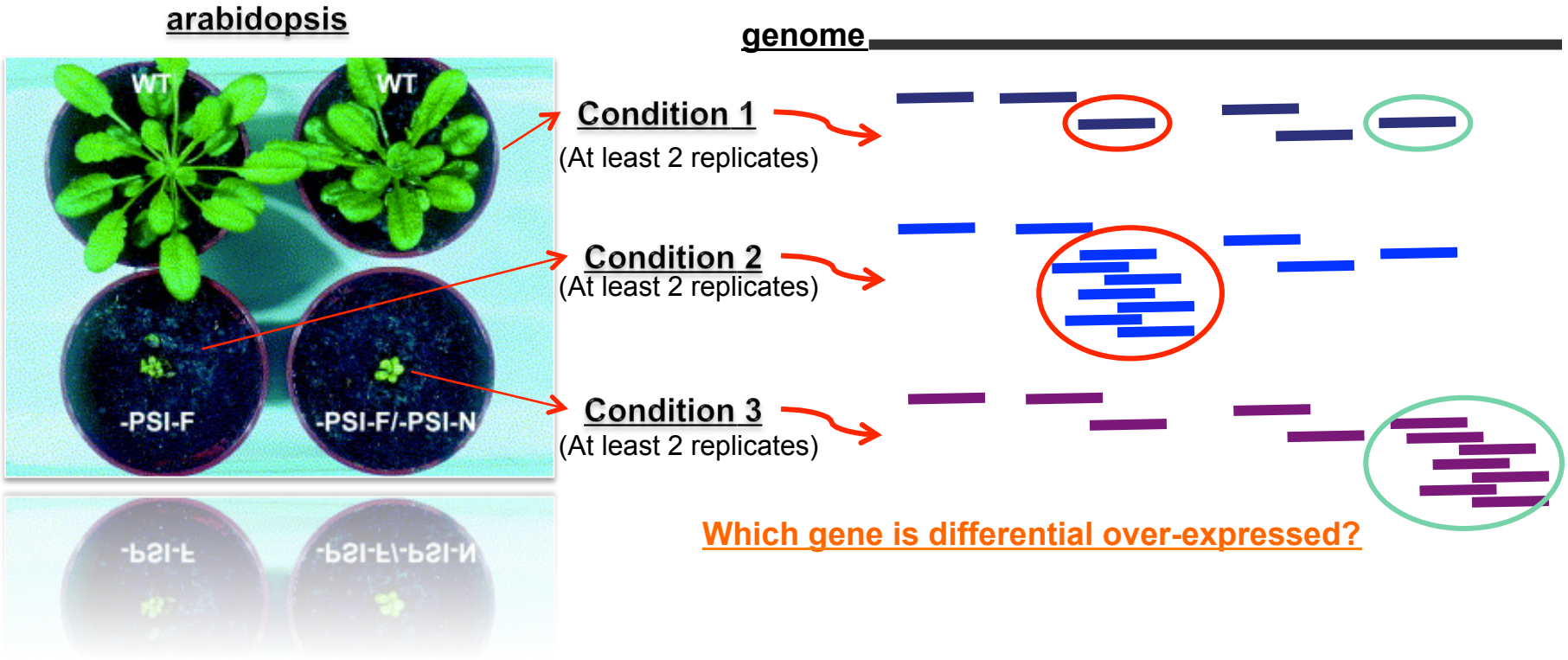
Roscoff-RNASeq

A new pipeline for Differential Expression

15-01-2013

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

Differential Expression



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, Génopole, 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

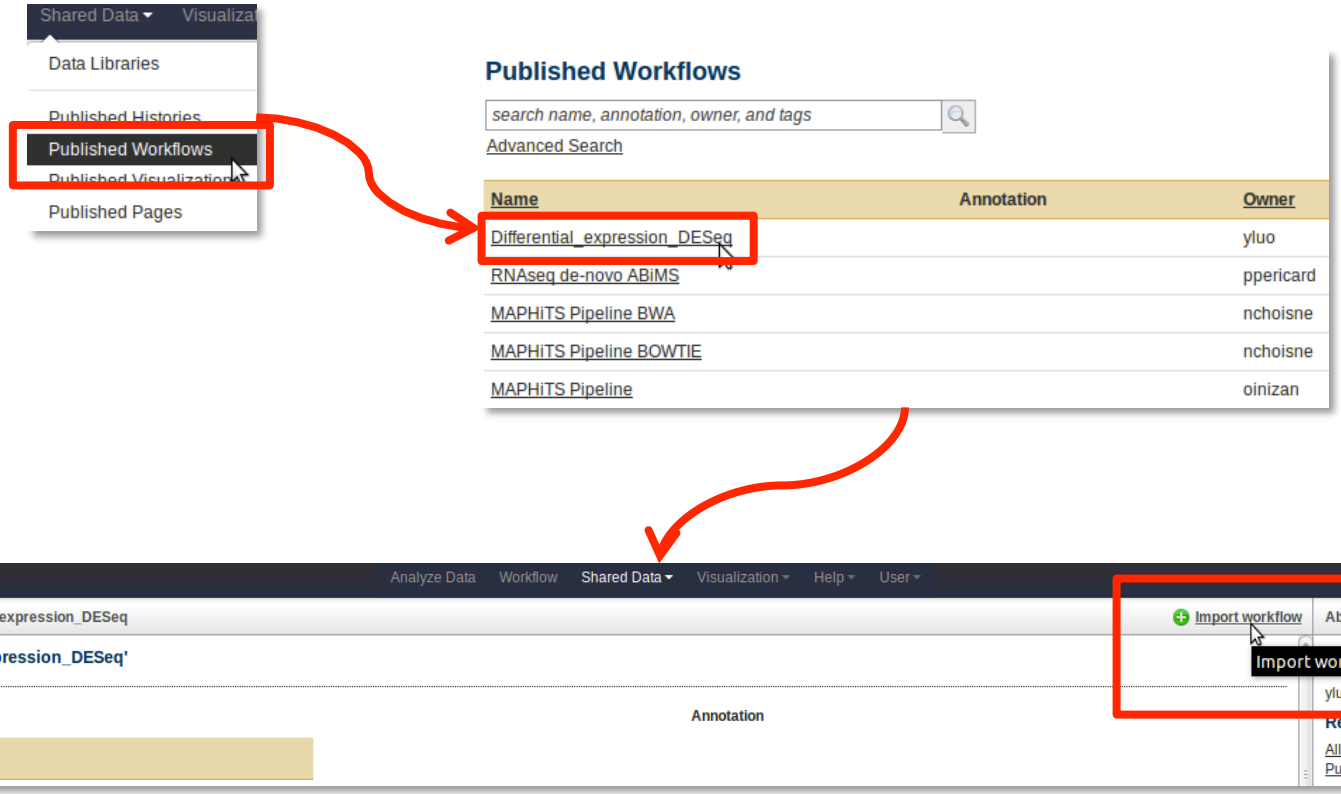
④ **SamTools** to convert bam format to sam format

⑤ **CompareOverlapping** to calculate the number of overlapping reads

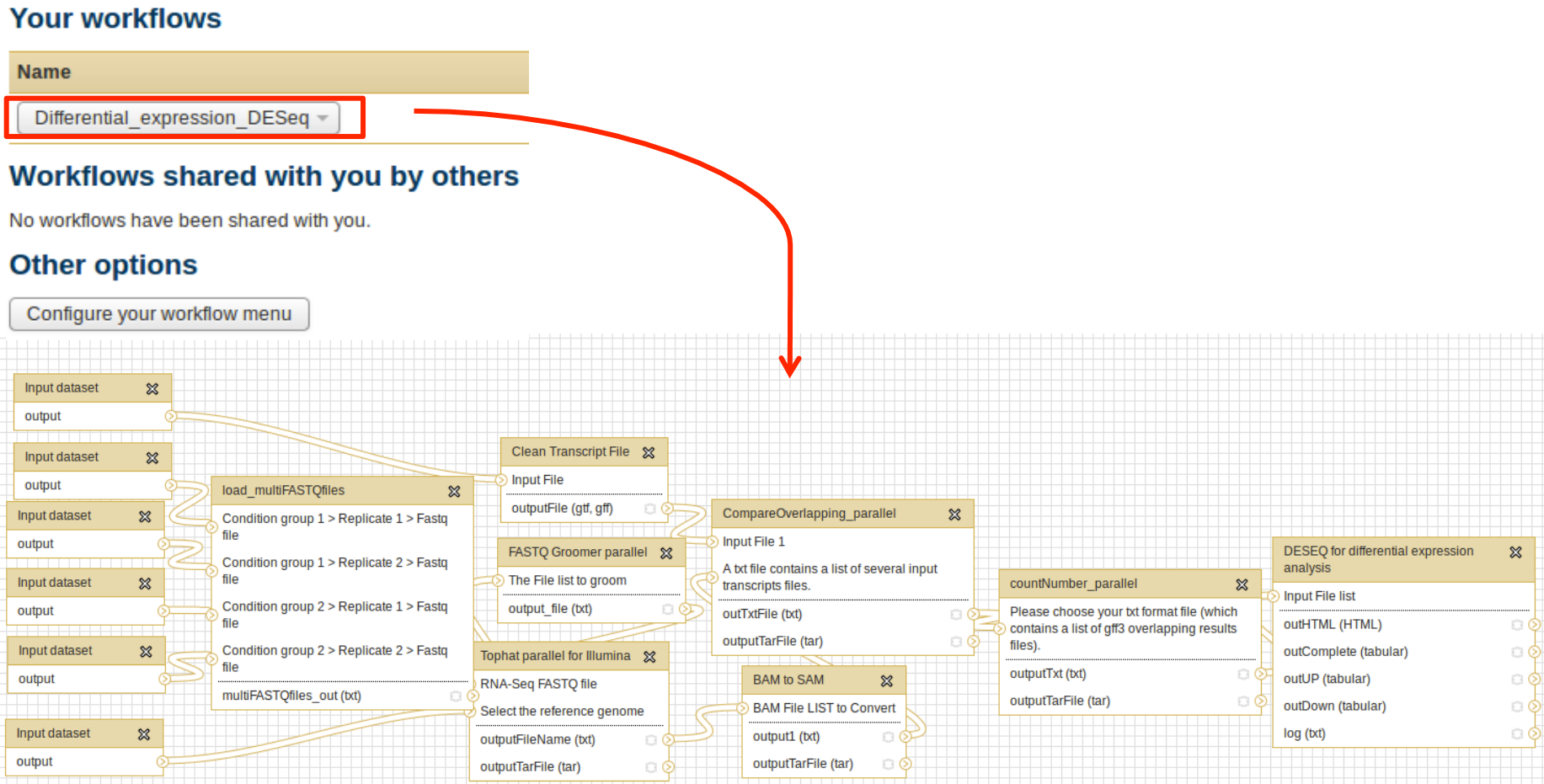
⑥ **CountNumber** to convert the previous output to a tabular file

⑦ **DESeq** to analyze differential expression

Import the workflow to your galaxy work sapce

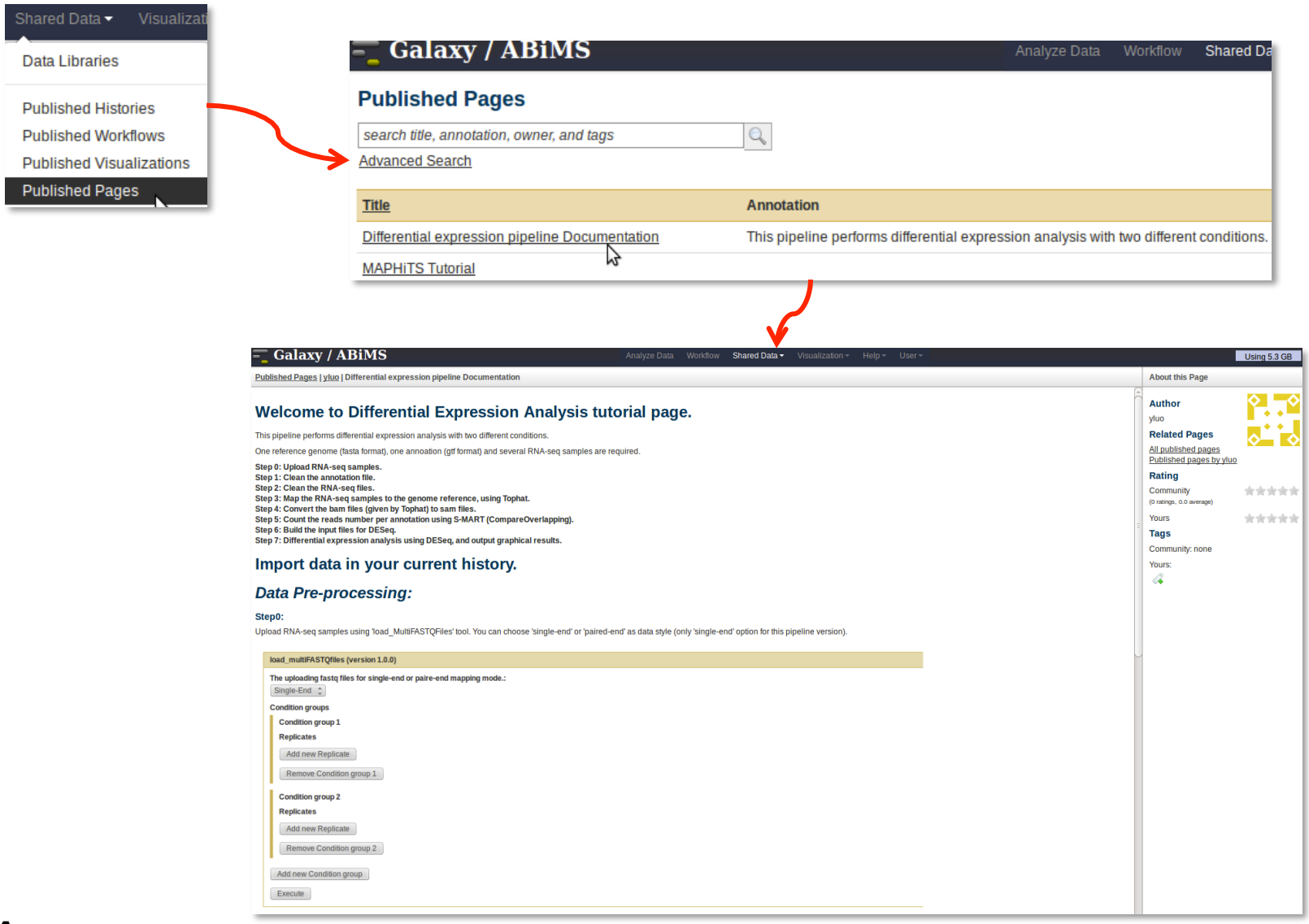


Differential expression workflow



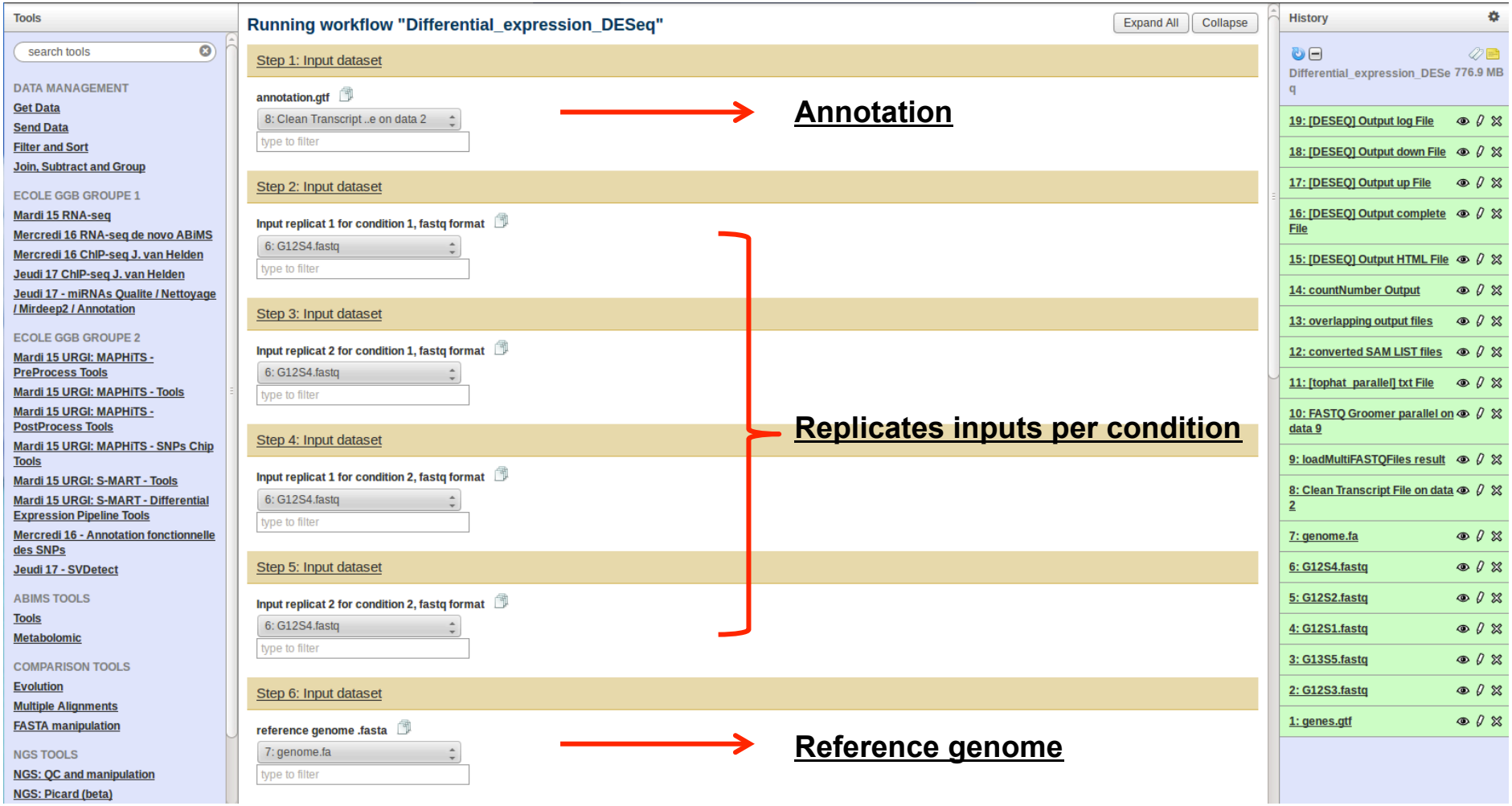
Differential expression workflow usage

Where to find tutorial page?



Example

input data (GEO)



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

Advanced Options
Hide Advanced Options

Option: illumina by default.

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)

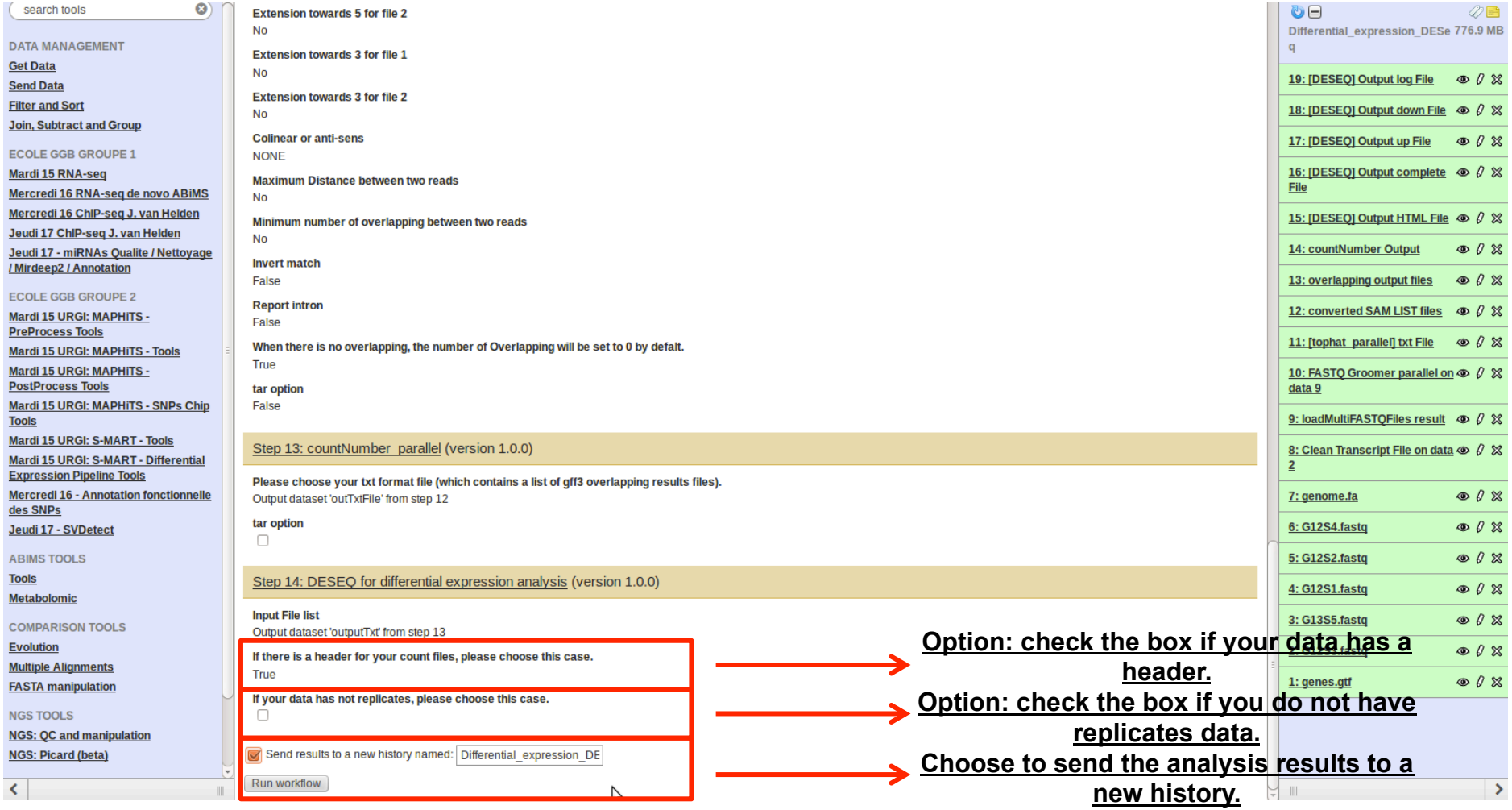
BAM File LIST to Convert
Output dataset 'outputFileName' from step 10

Include header in output
☐

tar option
☐

Step 12: CompareOverlapping_parallel (version 1.0.0)

Example

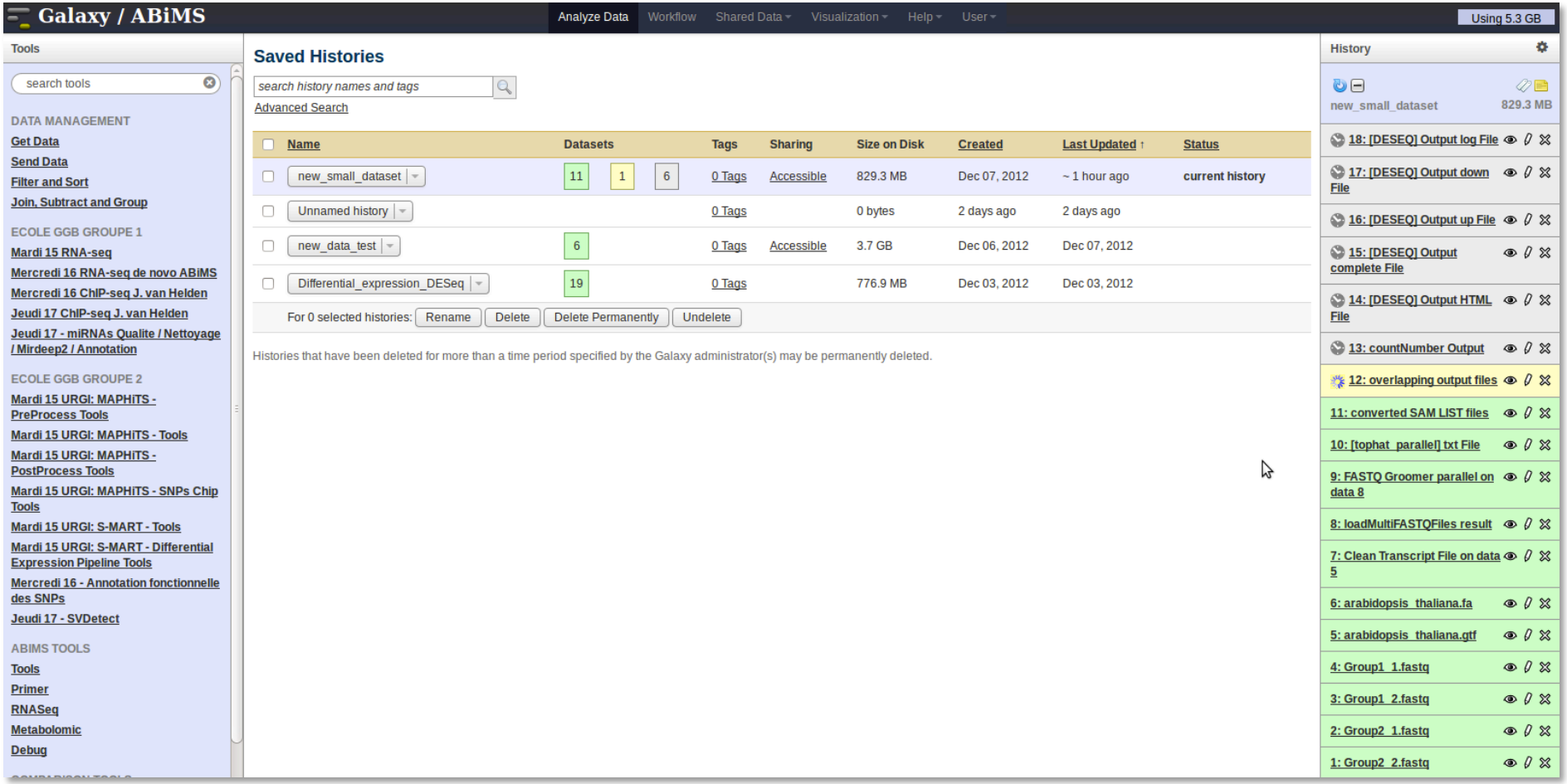


Option: check the box if your data has a header.

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
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
Primer
RNASeq
Metabolomic
Debug
Saved Histories
search history names and tags
Advanced Search
Name
Datasets
Tags
Sharing
Size on Disk
Created
Last Updated
Status
new_small_dataset
11
1
6
0 Tags
Accessible
829.3 MB
Dec 07, 2012
~ 1 hour ago
current history
Unnamed history
0 Tags
0 bytes
2 days ago
2 days ago
new_data_test
6
0 Tags
Accessible
3.7 GB
Dec 06, 2012
Dec 07, 2012
Differential_expression_DESeq
19
0 Tags
776.9 MB
Dec 03, 2012
Dec 03, 2012
For 0 selected histories:
Rename
Delete
Delete Permanently
Undelete
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