

Overall aim

This hands-on should familiarize the participants with Galaxy workflows for processing multiple files in parallel through multiple steps. We will be using RNA-seq workflows developed by Elixir-No, on mouse cell line data scaled down to chromosome 19 only (data kindly contributed by NGC/S. Lorenz).

In particular we'll look at how to retrieve datasets from NeLS storage (alternatively uploading them manually), define a set of files into a dataset called a collection in Galaxy terminology, retrieve and modify a copy of a standard NeLS RNA-seq differential expression workflow to other needs, initiate a run and inspect the resulting intermediate and final results.

We would value your input both on the selection of tools and parameters of the tools that should be emphasized. The value of the curated workflows lies in the right tools being made available with the right input parameters highlighted and the interface properly guiding the user to pay attention to these.

We'll first run a couple of the tools manually first to get to know some of the parameters, and then run the full pipeline.

Load data

- Log in to galaxy-ntnu.bioinfo.no
- Make sure “Analyze Data” tab is selected in the Galaxt top menubar
- Next,
 - If you are a FEIDE user, retrieve the four *.chr19.fastq files from NeLS storage through the “Get Data | Get Data from NeLS storage” menu choice in the leftmost menu.
 - If you are a non-FEIDE user you should make sure someone with a FEIDE user shares their history with the loadad data with you like this:
 - The FEIDE user prepares a history with the 4 files loaded, then
 - selects the “Share or publish” option in the history menu (gear icon top right in the history panel on the right side of the screen).
 - Clicks the “Share with user” button under the “Share with Individual users” heading.
 - Enter the e-mail address of the non-FEIDE user used as Galaxy login.
 - The non-FEIDE user selects the “Histories shared with me” option from the history menu (gear icon top right), then
 - Select the history in the middle panel and press the “Copy” button

- Inspect your files/data items in the history (click the name of the dataset and the eye icon)
- Annotate the data properly with “mm9” as database and “fastqsanger” as datatype/format. (tip: pencil icon top right of the data item in your history, select correct option and save for each tab, see that the values have been updated in the history afterwards).
- (FastQC and Fastq Summary Statistics tools are available under “NGS: QC and manipulation” submenu).

Run mapping step manually

We will now run a couple of the tools manually to get familiarized with the outputs of these steps before performing automatic execution in a workflow.

- Open the “NGS: RNA Analysis” submenu in the tool menu (at the left) by clicking the name.
- Click the “Tophat2” tool, and it will appear in the middle panel.
- Select one of your fastq files as input from the second dropdown menu. (If it appears empty, double check that you annotated it correctly above as “fastqsanger” formatted data).
- Select the right genome for your input files (chr19_mm9)
 - Note that you could provide a simple fasta file that you as a user upload yourself by selecting.....
- We would like to use a preindexed transcriptome rather than predicting a new gene model:
 - Select “Yes” to preindexed transcriptome
 - Select chr19_mm9 from the dropdown list
- Leave the defaults of other options (“single-end” etc)
- If you like you could look at the “Full parameter list” of Tophat2
- Hit the “Execute” button at the bottom

Your expected result data items from the analysis tool will then appear in your history, and should first be labeled as queued (grey color background), then labeled as running (yellow color background) and finally as completed successfully (green color background). In case of processing failure, they will appear as red.

You should get 5 result data items. Inspect one of them (i.e. accepted_hits) to learn more about the details of the job just run by:

- clicking the name
- press the “i” information icon to see parameters etc for the job that produced this item
- click the stdout link in the middle panel
- click the “i” again and the stderr link in the middle panel

Run read counting manually

We have opted to implement read counting using the efficient subread commandline tool. The user needs to specify a set of features to be counted, and the tool will produce a tab separated file of read

counts for all bam files provided.

- For now run the tool on a single file, selecting the chr19_mm9 gene annotations as feature set
- Options to allow counting of reads with multiple hits, and reads spanning multiple features are available under “Extended settings”. The default is to only count reads with unique hits and not allow contribution to multiple features.

Define collections

To prepare for a workflow run over all 4 files, we will define 2 collections with two files in each:

- Click the “Operations on multiple datasets” icon (checkbox icon) top right in the history
- Check 1.chr19.fastq and 2.chr19.fastq data items, and “For all selected ... | Build Dataset List”
- Uncheck the two previous and repeat for 3.chr19.fastq and 4.chr19.fastq
- Toggle off the “Operation on multiple datasets” by clicking the icon again.
- Renaming dataset collections is not yet possible, but expected as a standard feature in Galaxy very soon.
- Click the name of a dataset collection to see the contained dataset names. Click “Back to ... history” to return.

Prepare the DESeq workflow

Workflows can be shared to you either directly or as a published workflow accessible to all users of the Galaxy server. We will here use a published workflow that we import to our workspace. Many parameters such as reference genome etc can easily be modified at run time when you execute any workflow. Sometimes you know however you will set the same parameters over and over again to a non-default value, then it makes sense to make your own personal version of the workflow that saves you some time later:

- Find available public workflows in a Galaxy instance through “Shared Data | Public Workflows” in the top menu
- Click the “RNA seq pipeline for differential expression analysis (single-end, pooled samples)” workflow button/dropdown menu, and select “Import”. This WF will now be available in your private list of workflows, available through the “Workflow” top menu item.
- (If not already there click “Workflow” in top menu)
- Rename workflow to “Mouse chr19 mm9 RNA-seq...” instead of “Imported: RNA-seq...”

- To get a graphical view of the workflow, click the button/dropdown menu with the name of the workflow, and select “Edit”.
- Pan around the canvas by using the lower right overview window
- By clicking a tool, you can change the parameter settings of the workflow to suit your needs. For instance change the default choice of the reference genome: click on one of Tophat2 tool icons, and change both the reference genome and pre-indexed reference transcriptome to mm9_chr19.
- Do the same for the second Tophat2 tool icon on the workflow.
- Save your edits: the gear icon top right of the middle panel gives you a menu of options including save.
- Also change the gene annotation features in featureCount to mm9_chr19 and save.

Run the DESeq workflow

Now we are ready to run the workflow:

- Go to the workflow page (“Workflows” choice top menu).
- Click the name of your workflow, and select “Run” from the dropdown menu.
- In the middle pane you will now see all steps in the WF listet from the start to the end
 - Select the two different collections you created earlier as the inputs in step 1 (the two different “New dataset list” items).
 - Inspect the other steps (open and close by clicking on the header of each step), double check that the genome and transcriptome references are all set correctly to chr19_mm9.
 - When you're done, click the “Run workflow” button at the bottom below alt the listed WF steps.

As when executing a single job, multiple new history elements are created, and should progress through the cycle of grey (idle/waiting to run), yellow (running), green (successfully completed). If you get a red history item, this step in the WF failed to complete successfully.

Inspect final results

The DESeq tool produce 4 result items:

1. List of differentially expressed genes (not sorted on score but by ID) with statistics
2. MA plot of features to provide a global display of differentially expressed features

3. Hierarchical clustering of the 30 most differentially expressed genes
 4. PCA plot of samples (over all genes) to display global overall relationship between the samples.
- Inspect the 4 results
 - Also look into the stdout and stderr of the the DESeq step
 - Try the “Sort” tool under “Filter and sort” tool sub-menu, to sort the DESeq results in ascending order on pAdj column (column 7).

Inspect intermediate results

Although not all intermediate steps of a WF is shown automatically in the history, they are all present in a hidden state. At the top of the history panel to the right, you will see an overview of how many datasets in total, how many deleted and how many hidden datasets your history have at this point. Clicking these links will toggle the display of hidden and deleted datasets in your history.

- Toggle the hidden datasets in your history
- Try to identify the bam file and the other output datasets corresponding to 2.chr19.fastq input file, inspect the stdout and stderr of this job. Find out how many reads mapped.
- Look into the corresponding featureCount results of this first collection (1 and 2). How many reads were counted towards a feature for the bam file corresponding to 2.chr19.fastq?

A second run of DESeq (optional)

We'll now adjust a couple of parameters for the featureCount step and run the WF again.

- Clean up if needed, rename your history as “DESeq WF run 1” (click the name of your history at the top)
- Copy history (from the history menu, gear icon top right)
- Delete all items except the input data and the two collections
- From the history menu, select “Delete permanently all deleted items”
- Toggle deleted items on and off, the items are still there as empty placeholders, but the associated data has been removed (and disk space freed).
- Rename this history as “DESeq WF run 2”
- Go to your workflow and select to edit it
- For the two featureCount steps in the WF :
 - Select the box

- In the right panel, select “Extended settings” for “featureCounts parameters.”
- Allow reads with multiple matches in the reference to be counted as well (default only unique hits are counted).
- Allow reads to be counted in multiple features when they hit more than one feature.
- Save your workflow changes after you've changed both featureCount steps in the workflows.
- Set inputs and check parameters before you start a new run (“Workflow” top menu, click name of workflow and select run, set/check input/parameters in middle pane).
- Hit “Run Workflow” button at the bottom.
- Inspect both the DESeq2 end results and featureCount results again (for mapped and counted reads)

Running an alternative pipeline: the cuffdiff WF

The cuffdiff tool from the CuffLinks package uses a different statistic for differential expression. We generally advise to use the DESeq2 tool for differential expression, but sometime one would like to investigate using a second approach or for comparison with existing results using cuffdiff.

We could run the whole WF from start, but for real size data, starting after the TopHat2 mapping step is probably desirable.

- Copy your history into a new one and name it “Cuffdiff run”
- Delete the unnecessary DESeq and featureCount items in your history
- Show hidden data in your history
- Select to unhide the two outputted collections of bam files (accepted_hits) from the two TopHat2 mapping steps.
- Use these 2 as inputs to a manual run of the Cuffdiff tool from the left panel (under “NGS: RNA Analysis” sub-menu).
- Remember to specify the correct reference annotation (mm9, chr19) parameter setting.
- Hit the “Execute” button.

Alternatively:

- Go to “Shared data | Published workflows”
- Select and import the Cuff diff workflow
- Start this workflow from your fastq collection items, setting all the reference parameters correctly
- Run it

Navigate around to compare briefly the output of the Cuffdiff steps, and the DESeq results in your

other history (you can swap between histories using “User | Saved histories” in the top menu (or alternatively from the history menu if you have your history visible)).