HANDS-ON BIOINFORMATICS WORKSHOP

ANALYSIS OF RNA-SEQ DATA VIA CUSTOMIZABLE GENOMICS PIPELINES (GALAXY)

ELIXIR, NORWAY

NELS STORAGE HANDS-ON

In this hands-on exercise, you will upload a couple of files to the NeLS storage, both via the web portal and Secure copy (SCP). You will also do some manipulation of the directory structure and contents, both using web portal and SSH login.

IMPORTANT NOTICE: To carry out this hands-on you currently need a Feide Identity or NeLS Identity. If you do not have a Feide Identity or NeLS Identity, please contact one of the organizers of the workshop.

STEP 1: LOG IN TO THE NELS PORTAL AND MANIPULATE THE DIRECTORY STRUCTURE

- In a browser, open http://nels.bioinfo.no
- Click "Login with FEIDE Identity" (or "Login with NeLS Identity")
- Select your university from the list (if asked)
- Type in you username and password

If this is the first time you log in, you should get a notice from Feide about what kind of information about your identity that will be transferred to the NeLS Portal. Accept this to be fully logged in.

You should now see an overview of your Personal folder (which should be empty, if you are a first-time user). Also notice that you should have received a NeLS ID number (top right corner). This ID is an identifier in the NeLS user database which is uniquely matched to you Feide user identity.

- Click "My Projects" to see any projects you might be a part of. You should have access to “Elixir_workshops”.
- Click "My Data" to go back to your personal directory.
- To upload the file, click "New File"
- Click "Browse" and select any file on your machine.
- Click "Upload"
- Close the dialog box by clicking the "X" in the upper right corner The Word file should then be shown in your Personal directory.
- If you click on the file, it will be downloaded back to you local computer.

You should now try to move this file into a new directory:

- Click "New Folder", type "NeLS_workshop", click "Save" and close the box
- Click the check box to the left of your uploaded file to select it
- Click the "Cut" button
- Click the new folder to enter it
- Click the "Paste" button to finish moving the file
- Click "Personal" in the path "Personal/NeLS_workshop" above the file list to go up a level
- Also note that you can easily delete and rename files with separate buttons

STEP 2: TRANSFER A FILE USING SCP

Using the web portal to upload and download files is practical **for only small files.** If you have larger files, the best way to upload them are by using secure copy (SCP). Now, as an example we will use small file for scp command.

- On your Desktop create directory with name "NELS".
- Download file "sequence.fasta.gz" by this link https://goo.gl/JcTrKy and place it to the NELS directory.

In the NeLS Portal, click "Connection Details".

- Press "Download key"
- Rename downloaded file to key.txt
- Move key.txt to the "NELS" directory, which you have just created.
- Move a file which you would like to upload to this directory as well.
- In the NeLS web-page note the **Host** (which is the address of the NeLS storage login server), and the **Username** fields.

FOR WINDOWS USERS:

Download Git for Windows https://git-for-windows.github.io/ and install it (default options should be fine). It will install a bash emulator on your machine. Once it is installed, open the Git Bash (it will be referred as a Terminal) and follow the instructions for Mac and Linux users starting from the second point.

FOR MAC AND LINUX USERS:

- Open the Terminal application (under Application/Utilities or Programmer/Verktøy).
- Move into the "NELS" directory on your desktop. Use ***cd*** command for navigation, e.g "***cd Desktop/NELS***". To check your current location type ***pwd***
- Make sure that "key.txt" and "sequence.fasta.gz" are in the folder.
- Use **ls** command to display content of the folder.

- Type: ***chmod 600 key.txt***
  this command is needed to set correct permissions of the private key file.
- In the terminal type following command:
  scp -i key.txt **sequence.fasta.gz** ***USERNAME@HOST:Personal/***
  where sequence.fasta.gz - is the file your are uploading,
  USERNAME and **HOST** are the values found under "Connection details" in the NeLS Portal

Now check that the file was successfully uploaded in the NeLS Portal. The file should be located in the "Personal" directory.

STEP 3: MANIPULATE THE DIRECTORY STRUCTURE AND FILE CONTENTS USING SSH

The NeLS storage also supports command-line access to file and directory manipulation, using SSH. In this part, you first edit the directory structure, as you did above with the NeLS portal. Also, the file you have uploaded is a compressed file to reduce transfer time. As Galaxy needs the files to be uncompressed, you will need to do this before carrying out the RNA-seq analysis in the next session. This can then be done in SSH (after the upload is complete)

- Open the Terminal application (Git Bash for Windows users).
- Move into the directory with the files to upload.
- As with SCP, you need to type "chmod 600 key.txt" to set the permissions of the key file. However, as this is already done above, you do not need to do this again.
- Type "ssh -i key.txt USERNAME@HOST", where USERNAME and HOST are the values found under "Connection details" in the NeLS Portal
- Now you navigate to Personal directory "***cd Personal***"
- Check content of the folder by ***ls*** command, you should see your **sequence.fasta.gz** file here.
  sequence.fasta.gz is compressed, to decompress it type:
  gzip –d sequence.fasta.gz
  it will create "sequence.fasta", which can be used for the analysis.

GALAXY WORKFLOW

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-Norway, on mouse RNA-seq data.

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 nels.bioinfo.no
- Click on "Available Pipelines"
- Click on "RNA-seq"
- Click on Eukaryotic RNA-Seq pipelines. Then click on RNA-seq pipeline for differential gene expression analysis (single-end, pooled samples)
- Click on the green + symbol to import the pipeline to your galaxy.
- Make sure "Analyze Data" tab is selected in the Galaxy top menu bar
- Next, retrieve the four fastq files from NeLS storage through the "Get Data | Get Data from NeLS storage" menu choice in the leftmost menu (/Projects/Elixir_workshops/mouse-chr19).

- Select all the files under the above mentioned directory and click on send to galaxy
- 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 FASTQC MANNUALY

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: QC and manipulation" submenu in the tool menu (at the left) by clicking the name.
- Click the "FastQC:Read QC" 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).
- Leave the defaults of Contaminant list and Submodule and Limit specifying file (Nothing selected)
- 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 2 result data items. Inspect html document (i.e. accepted_hits)

RUN TRIM GALORE MANNUALY

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: QC and manipulation" submenu in the tool menu (at the left) by clicking the name.
- Click the "Trim Galore" 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).
- Go to Advanced settings and chose "Full parameter list" then go to "Generate a report file" and chose "Yes"
- Hit the "Execute" button at the bottom

Your expected result data items from the analysis tool will then appear in your history. You should get 2 result data items - the report file and the trimmed file. Inspect the report file.

RUN MAPPING STEP MANUALLY

- 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 trimmed fastq files (output from "TrimGalore!") as input from the second dropdown menu.
- Select the right genome for your input files (chr19_mm9)
- We would like to use a predindexed 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)
- 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

RUN READ COUNTING MANUALLY

We have opted to implement read counting using the efficient subread command line 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.

- On the left menu go to "NGS: RNA Analysis | featureCounts"
- In the alignment file leave you file created in previous step (e.g. Tophat2 on data 1:accepted hits)
- Check that Reference Gene Sets used during alignment is set to mm9, chr19.
- Leave default settings for the rest of the options.

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:

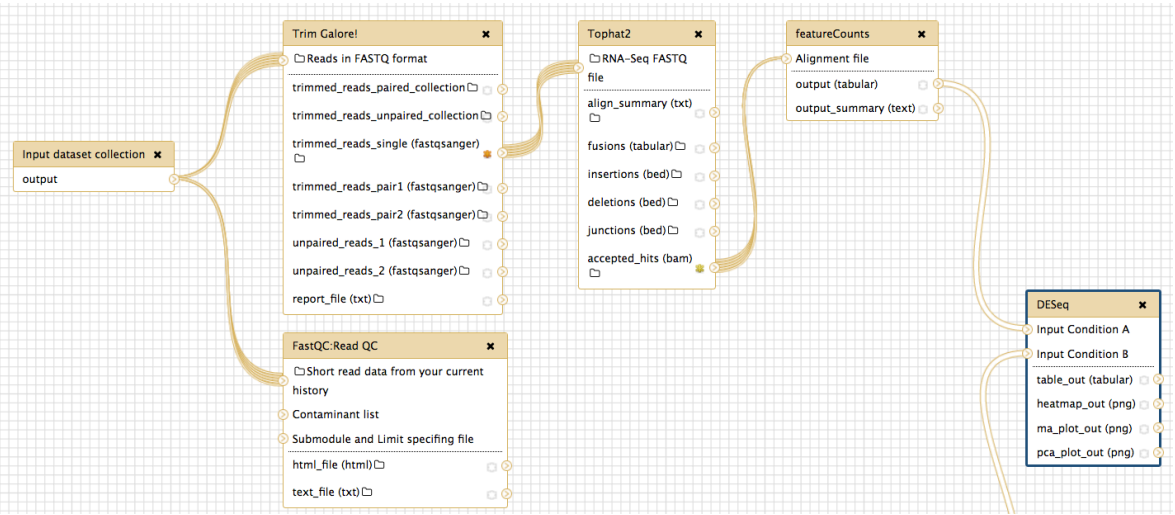
- Click the "Operations on multiple datasets" icon (checkbox icon) top right in the history

- 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 COMPLETE RNASEQ 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".
- Add "FastQC" and "TrimGalore" tools for each data list input (2 blocks of FastQC and two blocks of "TrimGalore")
- Adjust inputs and outputs according to the figure for each Input dataset collection:



- 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.
    - Change the default in both TrimGalore blocks: Go to Advanced settings and chose " Full parameter list" then go to "Generate a report file" and chose "Yes".
    - Change the default choice of the reference genome in both Tophat2 blocks: click on one of Tophat2 tool icons, and change **both** the **reference genome** and **pre-indexed reference transcriptome** to mm9_chr19.
- 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 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).
- 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 workflow failed to complete successfully.

INSPECT FINAL RESULTS

The DESeq tool produce four 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 four 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 corrsesponding to the 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 accociated data has been removed (and diskspace 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 advice 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 two as inputs to a manual run of the Cuffdiff tool from the left panel (under “NGS: RNA Analysis” sub-menu).
- Hit the “Execute” button.

Alternatively:

- Go to “Shared data | Publised 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)).