NeLS RNASeq Galaxy exercise

Overall aim

This hands-on should familiarize the participants with Galaxy workflows in general for processing multiple files in parallel through multiple steps. In addition, we particularly introduce the RNA-seq workflow developed by Elixir Norway. We will use the workflow to analyze six human prostate cell-line samples (LNCaP and RWPE) from paired-end sequencing. The data have been reduced in size to between 1M and 4M reads pr. sample. The sequence data are publicly available at NCBI, GEO with accession GSE75035. The end goal is to find genes differentially expressed between prostate cancer cell lines (LNCaP) and cell lines representing normal prostate epithelium (RWPE).

We will look at how to retrieve datasets from NeLS storage, or alternatively uploading them manually from an URL or your own computer. We will then run the tools manually step-by-step to get to know the individual tools and their parameters, before we run the full mRNA workflow, starting with fastq sequencing files and ending up with differentially expressed mRNAs.

Note1! For the Elixir WS today, you are expected to run only instructions written with ordinary black font. Some instructions are written with grey font. They are included as they may be relevant if you later want to run the analysis on your own data with (slightly) different setup (like having single-end data, uploading data from your own computer, having different experimental design etc).

Note2! Galaxy does not work equally well with all browser types. We recommend Firefox, Chrome (primary choice), and Safari, Edge (secondary choice).

We strongly encourage you to work in pairs, as this will reduce the burden on the (NTNU) Galaxy server. (Advice: Show the exercise on one of the computers and the Galaxy on the other computer.)

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.

Load data into Galaxy

Before you start the analyses, please make sure you have NeLS user, which will automatically be crated the first time you log in to nels.bioinfo.no. If you have a Feide user (UiO, NMBU etc) you can log in with Feide Identity. If you don’t have a Feide user, please contact one of the teachers and we will create a NeLS user for you. After you have gotten a NeLS user, one of the teachers will add you to the NeLS project Elixir_workshops, where the data files are located.

Open https://galaxy-ntnu.bioinfo.no in a web-browser and log in with your FEIDE or NeLS user and password.

Make sure “Analyze Data” tab is selected in the Galaxy top menu bar

Upload data from NeLS:

- Select “Get Data | Get Files from NeLS storage” in the Tools menu (the leftmost menu)
- The data are placed in /Projects/Elixir_workshops/ws_June3rd2019
  Note that there are 6 samples, but 12 fastq files. This is because these are paired-end reads. Underscore 1 (_1) indicates samples where the reads are read from the left end (forward reads), and underscore 2 (_2) indicates right-end (reverse) reads.
- Select all the 12 fastq files, then click “Send to Galaxy”
- The files should now appear in your History menu (the rightmost menu)
- The files will turn green when they are uploaded.

Upload data from your own computer:

- Select “Get Data | Upload File from your computer” in the Tools menu (the leftmost menu)
- Select “Choose local file” in the upload menu
- Select “Start”. Wait for the files to upload (100% in progress bar).
- The files should now appear in your History menu (the rightmost menu)
- Select “Close” in the upload menu
- The files will turn green when they are uploaded. This will take a couple of minutes.

You shall now select **one pair of the** .fastq sequence files (for example “LNCaP_rep1_1.fastq” and “LNCaP_rep1_2.fastq”) in your history to perform a step-by-step run-through of two of the tools in the workflow using these files only. We will perform analysis on collections of multiple files later.

Inspect and edit information of your data

You can look at your data file by clicking the “View data” (the button with an eye) for each data file in the History menu. You can click the “Edit Attributes” button (the button with a pencil) for each datafile to inspect, edit or add attributes to your data-file. Remember to save after you edit or add an attribute. As long as your datatype is “fastq” you would not need to change the datatype when you run the entire workflow. However, when we run each step manually, some of the steps will require the sequence file to be of datatype “fastqsanger”.

- View your selected input data files.
- Select “Edit Attributes”, then under “Datatypes” in the menu, select “fastqsanger” (NB! not fastqcsanger, which may com up first).

- Save the changes
- Go back to “Attributes” in the menu
- Under “Database/Build”, select “Human Dec. 2013 (GRCh38/hg38)(hg38)”
- Save the changes

Note1: You only have to do the steps above to the pair of files you have chosen . (When we later will run the whole workflow on all files, these steps are not necessary).

Note2: If you upload files from your computer or import from URLs, you can set both the data type (fastqsanger) and genome build (hg38) for all the files together or for individual files when you import.

Run alignment and mapping 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.

HISAT2 – Genome and transcriptome alignment for RNA-Seq data

HISAT2 is a sequence aligner, meaning that its purpose is to align sequences to a genome. It uses your library of sequences as input and a reference genome of the organism the samples was sequenced from, and finds where in the reference genome the sequences belong. HISAT2 can align both single-end and paired-end RNA-Seq reads, however the settings are slightly different for the two options.

To run HISAT2:

- In the tools menu, select: “NGS: RNA Analysis | HISAT2”
- Check that the “Input data format” is “FASTQ”.

- **For single-end reads:** Under “single end or paired reads?” drop-down menu, select “Individual unpaired reads”.
- Select a single-end file (for example LNCaP_rep1_1.fastq) in the “Reads” drop down menu
- You can also run several files, or a collection of single-end files by using the buttons next to the “Reads” drop-down menu.
- **For paired-end reads:** Under “single end or paired reads?” drop-down menu, select “Individual paired reads”.
- Select two paired files (for example LNCaP_rep1_1.fastq and LNCaP_rep1_2.fastq) in the “Forward Reads” and “Reverse Reads” drop down menus
- **For collections of paired reads:** Under “single end or paired reads?” drop-down menu, select “Collection of paired reads”. Your collection of paired reads should now appear automatically in the “Paired reads” drop-down menu
- Under “Source of reference genome to align against”. Select “Use a built-in genome” in the first drop-down menu, and then select “Human (GRCh38/hg38) with Ensembl transcripts” in the “Select a reference genome” drop-down menu.
- Leave all other parameters unchanged, and select “Execute”.

The main output from HISAT2 is a **.bam** file with aligned sequences, that is, the file contains information on where each of your sequences matches to the genome. The alignment has been made in compressed .bam format to save space. This is a binary format which is not readable by humans, so you will not be able to look at the actual alignments. If you click on the HISAT2 output file in your History-menu, it will display an alignment summary. You can also view additional details from the alignment by clicking the “View details” button (an “i”-inside a circle) in the History file-display.

Note 1: Here we align all the reads to the latest human version of the reference genome (hg38) with ensembl transcripts included, which have been specifically indexed for the HISAT2 aligner and made available in the NeLS Galaxy framework by us. If you are analyzing another organism, and cannot find your indexed reference genome in the menu, contact the ELIXIR helpdesk (contact@bioinfo.no) and we can index it for you.

Note 2: It is also possible to increase the number of computational cores to use to speed up the alignment process. For this exercise using only one core should be sufficient.

Visualize mapped reads in Trackster

The contents of the BAM file cannot be viewed directly by clicking the “View data” button (the eye), but Galaxy has an internal sequence browser called Trackster which is capable of visualizing the mapped reads graphically.

- Click on the name of the BAM dataset. This will expand the dataset in the history panel and display some more information as well as a few new buttons.
- Click the “Visualize” button, which is the one that looks like a bar chart (the fourth button in the row of five buttons), and select “Trackster” from the menu.
- In the popup dialog, click “View in new visualization”.
- Give the visualization a name in the “Browser name” box and make sure “hg38” is selected as the reference genome build before clicking the “Create” button.
- It may take some time for Trackster to set up the visualization, but after a while the mapped reads should appear in the track.
- You can select a region to zoom in on by clicking and dragging within the ruler area between the gray tool bar and the reads track.
- If you zoom in close enough you will be able to see the differences between the mapped reads and the reference genome indicated with colors.
- To exit Trackster, click the “Close” button at the right end of the gray tool bar or the “Analyze Data” link in the blue menu bar.
- If you saved the visualization before exiting Trackster, you can revisit it by selecting “Saved Visualizations” from the “Visualize” menu in the top menu bar.

featureCounts – Measure gene expression in RNA-Seq experiments from SAM or BAM files

Now that we have aligned our sequences to the genome, we want to associate the sequences with annotated features from the genome. In this case we want to count the number of sequences which overlap known genes in the genome. For this, we need to use a genome annotation file (usually in .gff or .gtf format) for genes and transcripts. To associate the aligned sequences with an annotation we use a tool called featureCounts.

To setup and run featureCounts for gene analysis:

- In the tools menu, select: “NGS: RNA Analysis | featureCounts”
- Select your output .bam files from the HISAT2 alignment under “Alignment file”.
- Select any of the “Use a built-in index” options under “GFF/GTF Source” and then make sure that “Ensembl genes (hg38)” is selected as the annotation file.
- Make sure the “Output format” is set to “Gene-name “/t” gene-count (tab-delimited)”
- Select “Execute”

The main output file from featureCounts is called “featureCounts on [filename on .bam file from HISAT2]” (the one without any suffix, that is, not the file with suffix ““summary”). If you view this file (the “eye” button), you see that it is a table where all the genes from the annotation are listed with a count of the number of sequences that is assigned to each gene (from now on referred to as the “count-table”)

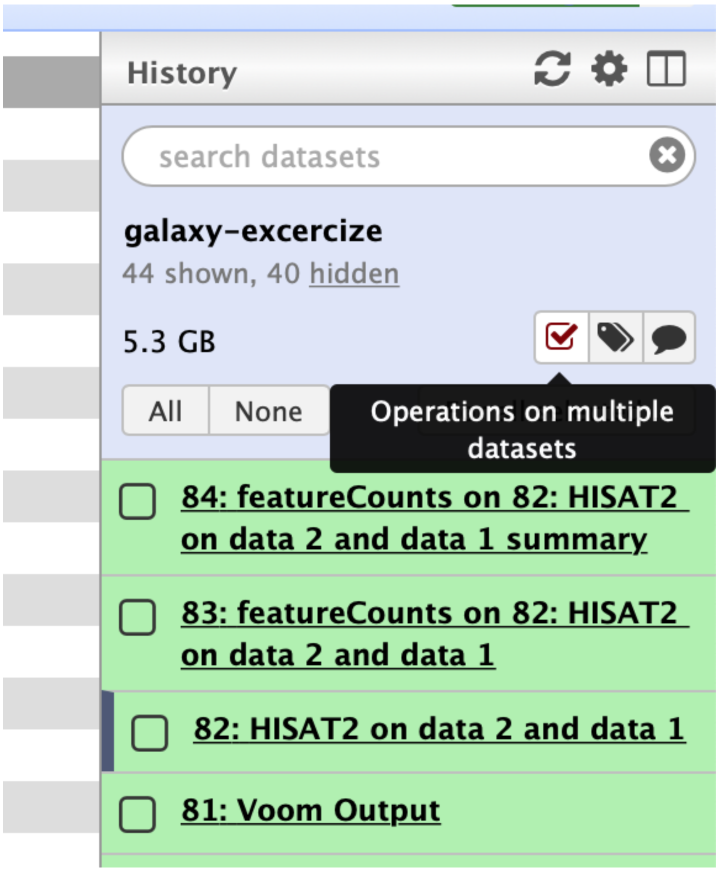
It is the count-table which is eventually the input you need to generate differential gene expression results. However, to have a sensible setup for differential gene expression analysis, you need an experimental design with at least two conditions to compare, and a number of replicate samples for each condition. Thus the count-table with only one sample created above cannot be used for differential expression; you will need a count-table with multiple samples. This basically means that you have to do the steps above (sequence-processing, alignment and assignment) for each of your sequence samples, or on a collection of samples. For differential expression analysis we use the tool Voom, which takes the count-table as input and returns a list of differentially expressed mRNAs with fold-changes, FDR adjusted p-values as well as other statistical measures.

Create dataset collections with single- and paired-end samples

To speed up analysis It is convenient to be able to process all samples at once. To achieve this we create dataset collections of all the samples in the experiment. However the options are somewhat different for single-end and paired-end sequences. In this exercise the total number of samples is six in both cases.

To create a dataset collection:

- On the top of the History menu there is an icon shaped as a “ticked box”. If you move your mouse pointer across it you will see it is named “Operations on multiple datasets”. When you select this box, checkboxes will appear before all the datasets in the history.


History
search datasets
galaxy-excercize
44 shown, 40 hidden
5.3 GB
All
None
Operations on multiple datasets
84: featureCounts on 82: HISAT2 on data 2 and data 1 summary
83: featureCounts on 82: HISAT2 on data 2 and data 1
82: HISAT2 on data 2 and data 1
81: Voom Output

For paired-end reads

- Select all 12 ".fastq" files (3 LNCaP and 3 RWPE replicates with paired-end suffixes "_1.fastq" and "_2.fastq"). To make selection easier you can use the "All" button on top of all the files.
- Click the "For all selected.." and select "Build List of Dataset Pairs" in the appearing menu.
- A window will now appear showing all the files you selected. If you have uploaded your paired-end files with correct suffix-annotation, Galaxy will automatically suggest which pairs that belong together, and suggest a sample name for the paired files (see screenshot below). Check that files are listed and combined correctly. Create a name for the file selection (for example "paired-end_sample_collection") and then click "Create list".
- You have now created a collection of paired files, which should appear as a new dataset in your history menu.

For single-end reads

- Select only the .fastq files having the suffix "_1.fastq".
- Click the "For all selected.." and select "Build Dataset List" in the appearing menu.
- A window will now appear showing all the files you selected. Check that the correct files are listed. You can reorder the list by dragging up and down. Create a name for the file selection (for example "single-end_sample_collection") and then click "Create list".
- You have now created a collection of single files, which should appear as a new dataset in your history menu.

The HISAT2-featureCounts-Voom RNA-Seq analysis workflow

Now that you have become familiar with each of the tools HISAT2, featureCounts and Voom, it is time to test the entire workflow running all steps at once. This is why we have created a workflow for the entire mRNA pipeline to handle this for you.

Clean and reload data files

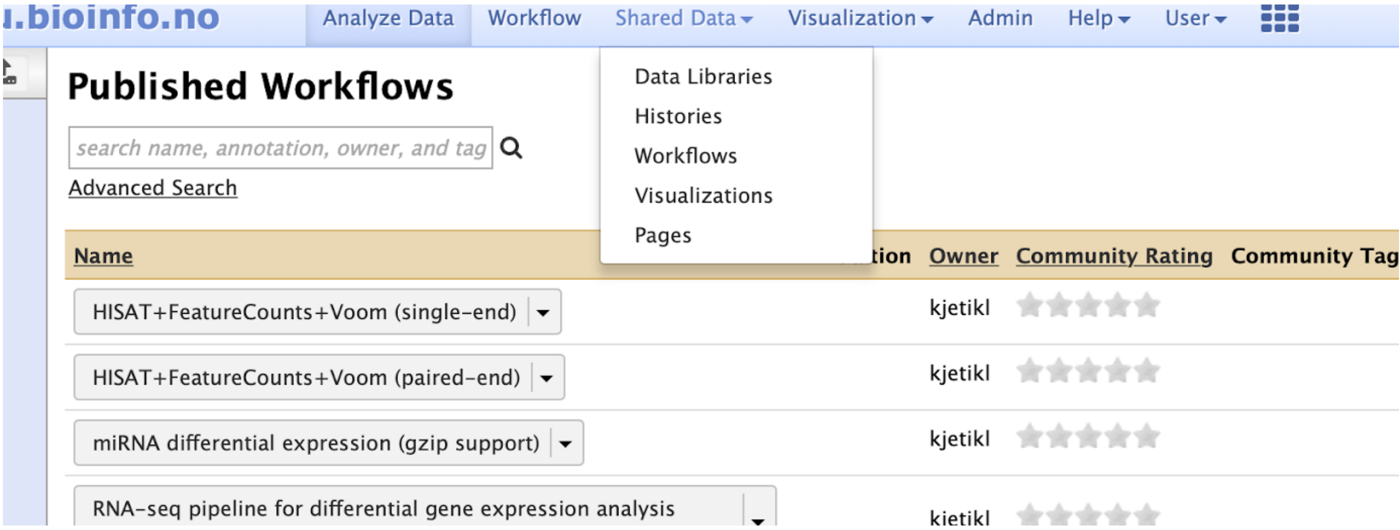
Before you run the workflow, it could be a good idea to clean up your history, so you only have the original, fastq-files as well as the data-collection of .fastq files to use as input to your workflow.

Import and inspect RNA-Seq workflows

We have created separate workflows for single-end and paired-end reads. These workflows are named:

- **HISAT+FeatureCounts+Voom** for single-end reads
- **HISAT+FeatureCounts+Voom (paired-end)** for paired-end reads

Both workflows are available under "Shared Data => Workflows" at the NTNU Galaxy Server.



Published Workflows

Name	tion	Owner	Community Rating	Community Tag
HISAT+FeatureCounts+Voom (single-end) ▾		kjetikl	☆☆☆☆☆	
HISAT+FeatureCounts+Voom (paired-end) ▾		kjetikl	☆☆☆☆☆	
miRNA differential expression (gzip support) ▾		kjetikl	☆☆☆☆☆	
RNA-seq pipeline for differential gene expression analysis ▾		kjetikl	☆☆☆☆☆	

(*Note: You can also get to this page by pressing the "Published Workflows" button on the welcome page.*) Press the arrow at the right edge of the workflow, and select "Import" from the drop-down menu to import the workflow to your own Galaxy session. After you have imported the workflow you will find it again by clicking on "Workflow" in the top menu bar. Click on the imported workflow and select "run" from the menu.

Note 1: The workflow's documentation page will be made available under "Shared Data => Pages". It will be possible to import the workflow directly from this page by clicking the green "+" button near the right edge of the orange workflow box.

Note 2: If you press the "Analysis Pipelines" button under "Tools and Workflows" on the welcome page, you will see an overview of all the NeLS workflows. Clicking on a workflow in this overview will take you to the workflow's documentation page.

Assign samples to groups, and define which samples to compare

Since we can now analyze more samples at once, we can also define sample groups which should be compared in the differential expression analysis by Voom at the end of the workflow. Actually, this is the first thing you are required to define in the workflow:

- Under “Workflow parameters”, define the group assignment for each of the samples in your collection under “Factor Values”. Be careful that the order of sample-assignments are the same as the order of files displayed in your dataset collection. For example the assignment “L,L,L,R,R,R” would indicate that the collection consist of two sample groups (LNCaP-L and RWPE-R) with three samples in each group, and where the three L-samples are listed first, followed by three R-samples.
- Under “Contrast of Interest” you define which groups you want to compare. Write “L-R”, which means that you want to compare LNCaP to RWPE samples (you can define more contrasts as well, but for now we will use only one contrast).

Run entire workflow

Now that the groups are defined, you can run the entire workflow.

- Select your dataset collection under “Step 1: Input dataset collection”. (Check that you select the right collection corresponding to the single-end or paired-end workflow)
- Make sure that the correct reference genome (Human (GRCh38/hg38) with Ensembl transcripts) and annotation file (Ensembl genes (hg38)) are selected in HISAT2 and featureCounts, respectively. (You have to mouse click on 2: HISAT2 (Galaxy version …) and 3: featureCounts (Galaxy version …))
- Otherwise you should run the entire workflow with the present parameters (however, it is generally a good idea to check that they are set correctly).
- Select “Run workflow” at the top of the workflow menu to run the entire workflow.

Inspect the result files. You should particularly look at the count-table produced by featureCounts and the differentially expressed genes from the Voom analysis. To view the differentially expressed genes you should download the expression table (“Top differential expressions (L-R).tsv”) which is found a little bit down on the Voom result page (Go to “Voom output” and “View data” (the eye icon).

And you are almost done!
Next time you log in to the Galaxy your history will be there. Before you leave you might want to share the history with your colleague, which can be done by going to

“history options” (upper right corner), click on “share or publish”, and chose “Make History Accessible via Link”. Then email the link to your colleague. We also kindly ask you to complete two short surveys about NeLS/Galaxy and this course (see below).

Run workflow with more than one contrast

In the workflow example above we ran the workflow using only one contrast L-R. For example, the LNCaP and RWPE replicates 2 and 3 used siRNA knockdown of the gene *ECI2*, while replicate 1 was used as control. In this setting you could define another contrast of siRNA-knockdown (KD) vs control (CR), that is, your Factor Values are now “CR,KD,KD,CR,KD,KD” and your contrast is “KD-CR”.

Run Voom on count-table with new contrast

It is possible to select another contrast without running the whole workflow again. Here we describe two ways of doing this.

Rerun the differential expression analysis with Voom from History menu

- In the History menu, select the data named “Voom Output”. Click on the name to expand it.
- Click the button with an icon which looks like two arrows forming a circle, named “run this job again”. A setup of Voom with the same parameters will appear in the central window.
- To define another contrast, type “C-B” under “Contrast of interest”.
- Select “Execute”

Rerun the differential expression analysis by selecting Voom from tools menu

- Find the dataset ID (number) for the count-matrix from featureCounts.
- In the tools menu, select: “NGS: RNA Analysis | Voom Rnaseq”.
- Select the featureCounts matrix under “Counts data”.
- Re-type your “Factor Values” exactly as you did for the workflow (Here: “M,M,M,C,C,C,B,B,B”)
- Select your “Contrasts of interest“, for example “M-B”.
- Change “Minimum CPM” from 0.5 to 10 (the CPM value used in the workflow)
- Leave other parameters as default.
- Select “Execute”

This rerun with a new contrast is most relevant if you have a collection of large files which takes a long time to process.

What do you think about NeLS and Galaxy?

We highly value your opinion on the NeLS and Galaxy system to continuously improve the service, and would highly appreciate if you spent a few minutes to take our online feedback survey at:

https://nettskjema.uio.no/answer/72165.html

What do you think about this course?

We would also highly appreciate if you gave feedback about the course by taking the short survey here:

https://nettskjema.uio.no/a/119484

Thank you for your attendance! :-D