INF-BIO5121/9121 - Oct 7, 2014

Analyzing miRNA data using Lifeportal

PRACTICALS

In this experiment we have miRNA data from the livers of baboons (*Papio Hamadryas*) before and after they are given a high cholesterol high fat diet. We have three samples from baboons on the baseline diet (Chow), and three samples from the baboons on the high cholesterol high fat diet (HCHF). The goal of the analysis is to find miRNAs that are differentially expressed in these two groups. We will start out with the fastq read files and follow these steps to get the results:

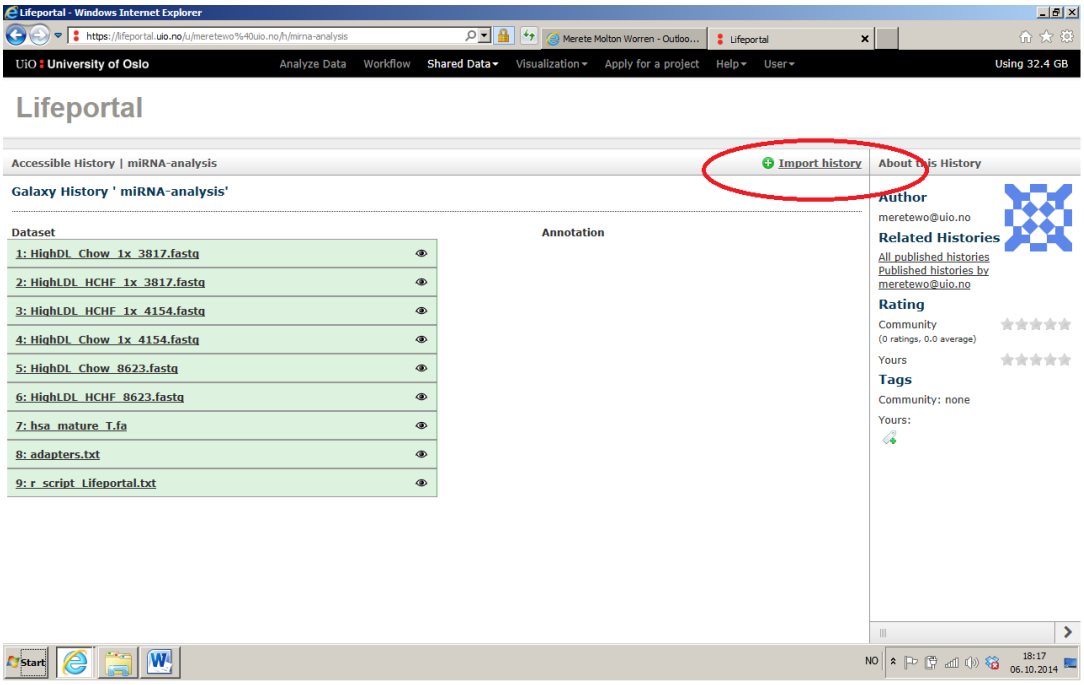
1. Logging in to lifeportal and get the read files from a shared history
2. Removing adapters
3. Looking at read qualities with FastQC
4. Align the reads to human miRNAs
5. Count how many reads map to each miRNA
6. Make a workflow, and run it
7. Use join and cut to get the appropriate file format
8. Run a differential expression analysis with EdgeR
9. Make a Galaxy page

1. Getting the read files

Start by using this link to go to the history page:

https://lifeportal.uio.no/u/sveinugu%40uio.no/h/infbiox121-mirna-demo

You should then see a page like this:



S

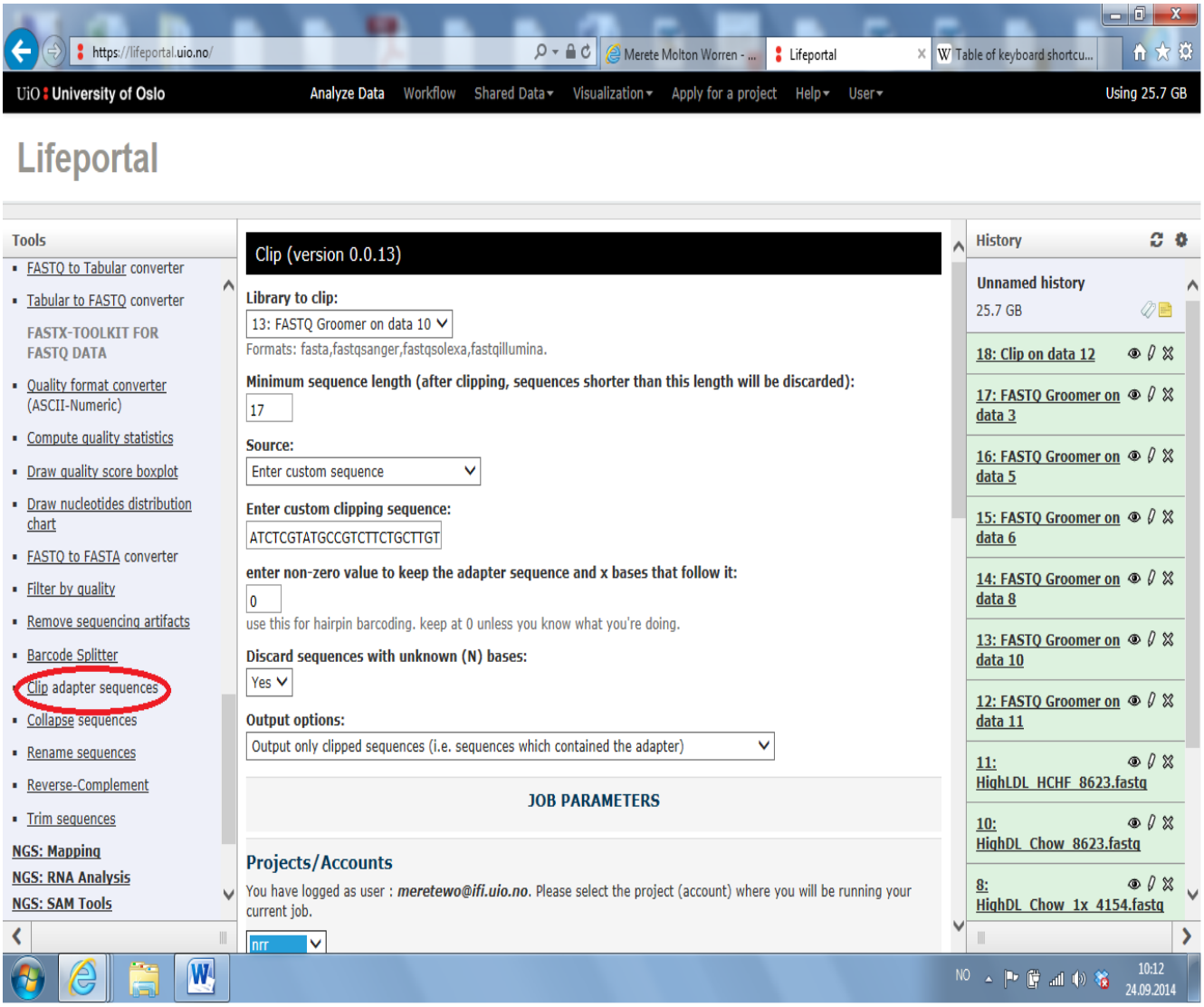
Click **import history**, and on the next page click **start using this history.**

This will create a new history for you to use, with all the needed files in it.

In the right pane, you will have your history items. In each of them you will find a little symbol of an eye. Click on this to view the data, for at least one of the files. Expand one of the history items by clicking on the name, and see what kind of info you get there.

2. Removing adapters

Since miRNAs are shorter (~22 bp) than the read length (36 bp), the reads will contain a stretch of the 3' adapter. These should be removed before further analysis is done. In Lifeportal, under **NGS:QC and manipulation** we can find the **fastx-toolkit** with an option to **Clip adapter sequences** that can be used for this. Below is a picture of how it looks after clicking the **Clip**. Take a look at it before starting.



There are several options that needs to be set before running this program.

First, choose a file for the **Library to clip** option. A list will come up where you can choose from the ones Lifeportal can see has the right format. Choose the first of the provided HCHF fastq read files (("1: HighLDL_HCHF_1x_3817").

Next, set the **Minimum sequence length** to 17.

Then it is time to put in the adapter sequence that is to be removed. Under **Source** we choose **Enter custom sequence**. The 3' adapter sequence is ATCTCGTATGCCGTCTTCTGCTTGT, copy and paste this sequence into the **Enter custom clipping sequence** field.

In the next field, leave the 0. We do not want to keep the adapter sequence.

For **Discard sequences with unknown (N) bases**, choose **Yes**

For **Output options** choose **Output only clipped sequences**

Now, all you need to do is to choose a project under **Projects/Accounts** and click the blue **Execute** button.

While your job is running, click the **back arrow** on your browser and check the description of what it does at the bottom of the middle pane.

When your job is finished, look at the file using the little eye button in your history item. You should see right away that the reads are shorter than they were before the clipping. Click on the history item. After the description field you will find a **rerun** button . Click on it to get the options for this run back in the middle pane. That way, for the next **Clip adaper** run there will be less work.

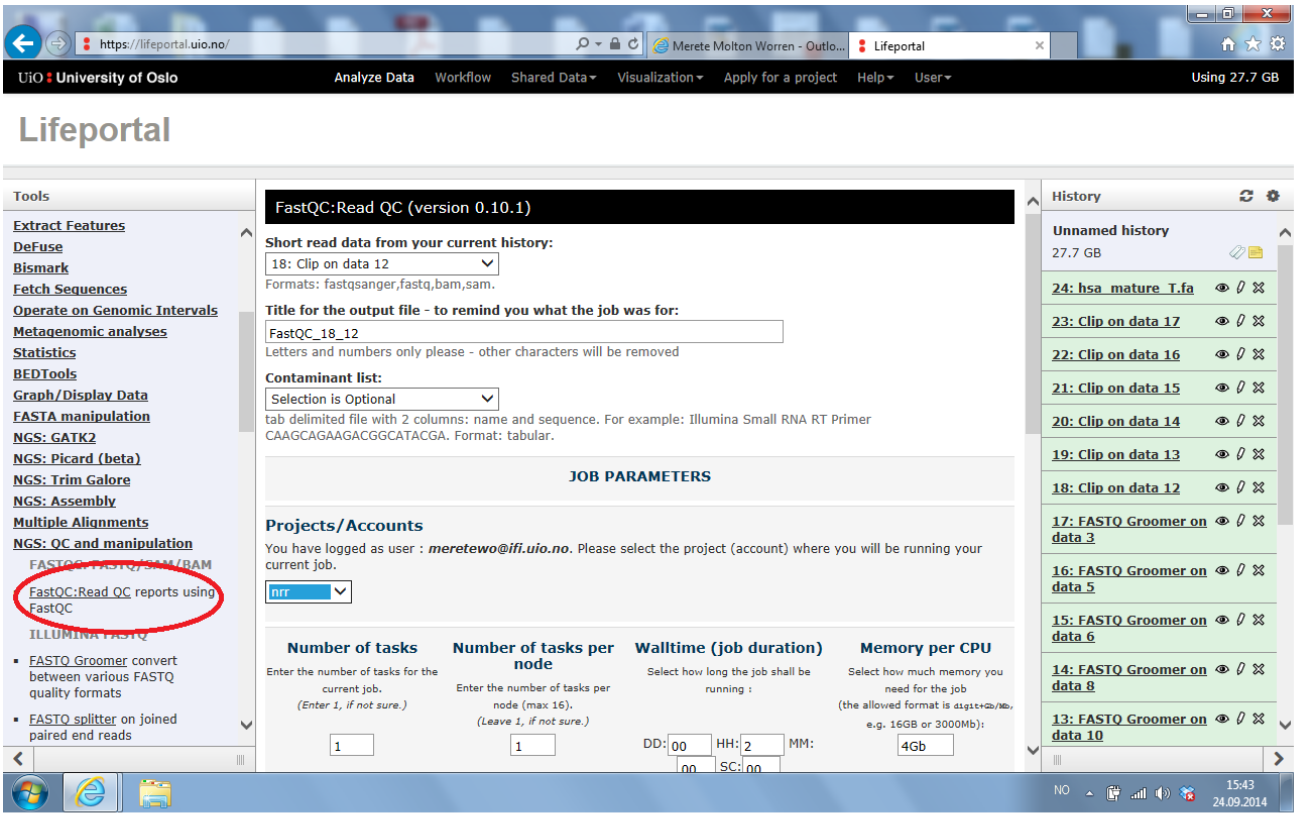
For now, we will only work with this file, so do not rerun the job.

3. Looking at read qualities with FastQC

Now that the adapter is removed, it is time to look at the reads with FastQC, to make sure they look ok.

Choose **FastQC:Read QC report using FastQC** under **NGS:QC and manipulation**.

Choose the Clip history item you just made. Make a title for the output. We do not use a contamination list in this instance. Choose the project and click execute.



4. Mapping

Now that the reads are cleaned up, it is time to map them to the reference. The reference we use are the human mature miRNAs found in miRBase. A file with the sequences are found in your imported history ("7: hsa_mature_T.fa").

Start by finding **Bowtie2** under **NGS:Mapping** and click on it.

Choose **Single-end** under **Is this library mate paired**?.

Choose the result of the previous Clip job as the Fastq file.

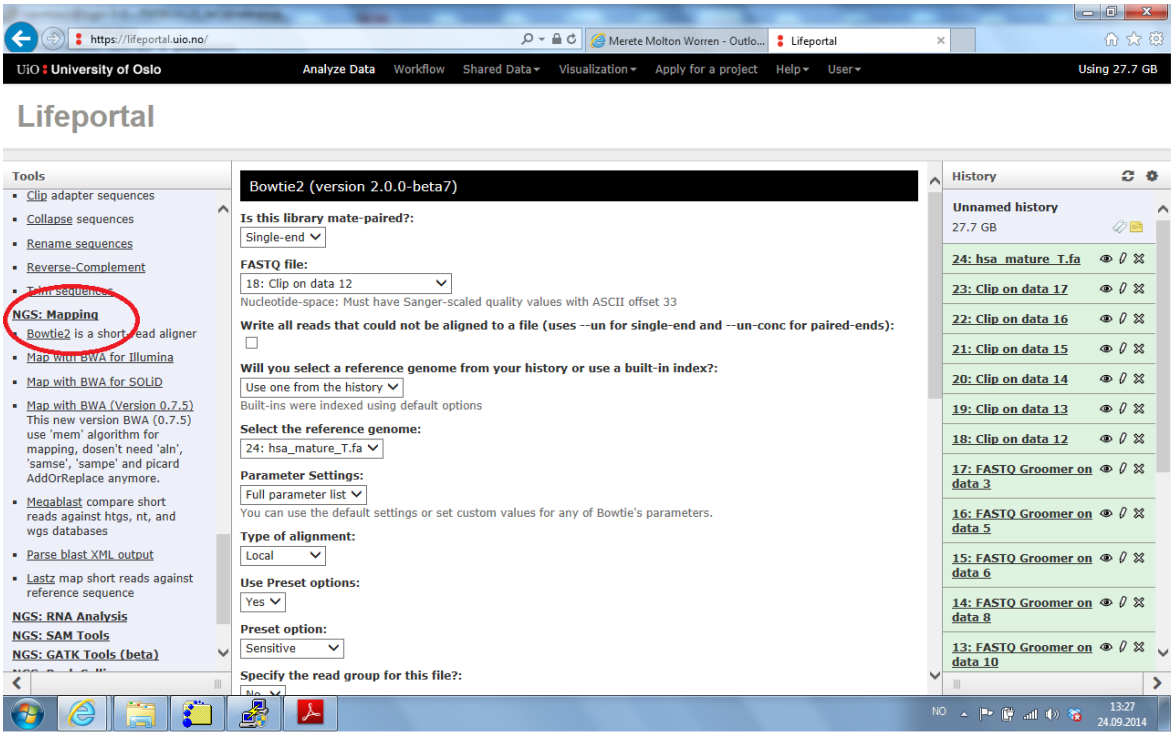
Do not check the box for writing unaligned reads to a file.

Choose **Use one from the history** under the question **Will you select a reference genome from your history or use a built in index?**.

Select the history item called **hsa_mature_T.fa** as your reference.

Choose to have the **full parameter list**. Change the **Type of alignment** to **Local**, and leave the others as they are.

Choose your project and click execute.

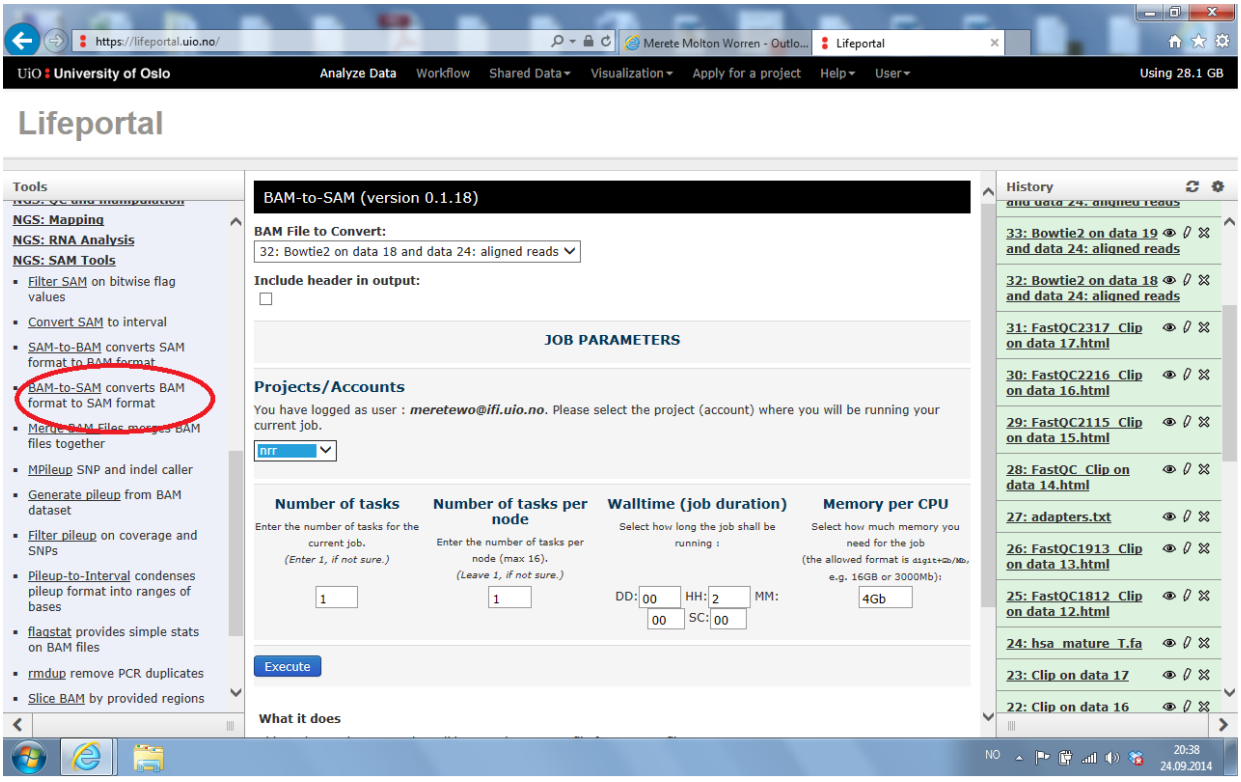


When the run is finished, the result is a .bam file, which is a binary file. Since we want to count the occurrences of the different miRNAs using textual manipulation, we need to transform it to a .sam file, which is a tabular file (i.e. a tab-delimited text file).

Under **NGS:SAM tools** choose **BAM-to-SAM**. Choose the Bowtie2 result file.

Do not check **Include header in output**.

Choose your project and run.



5. Counting

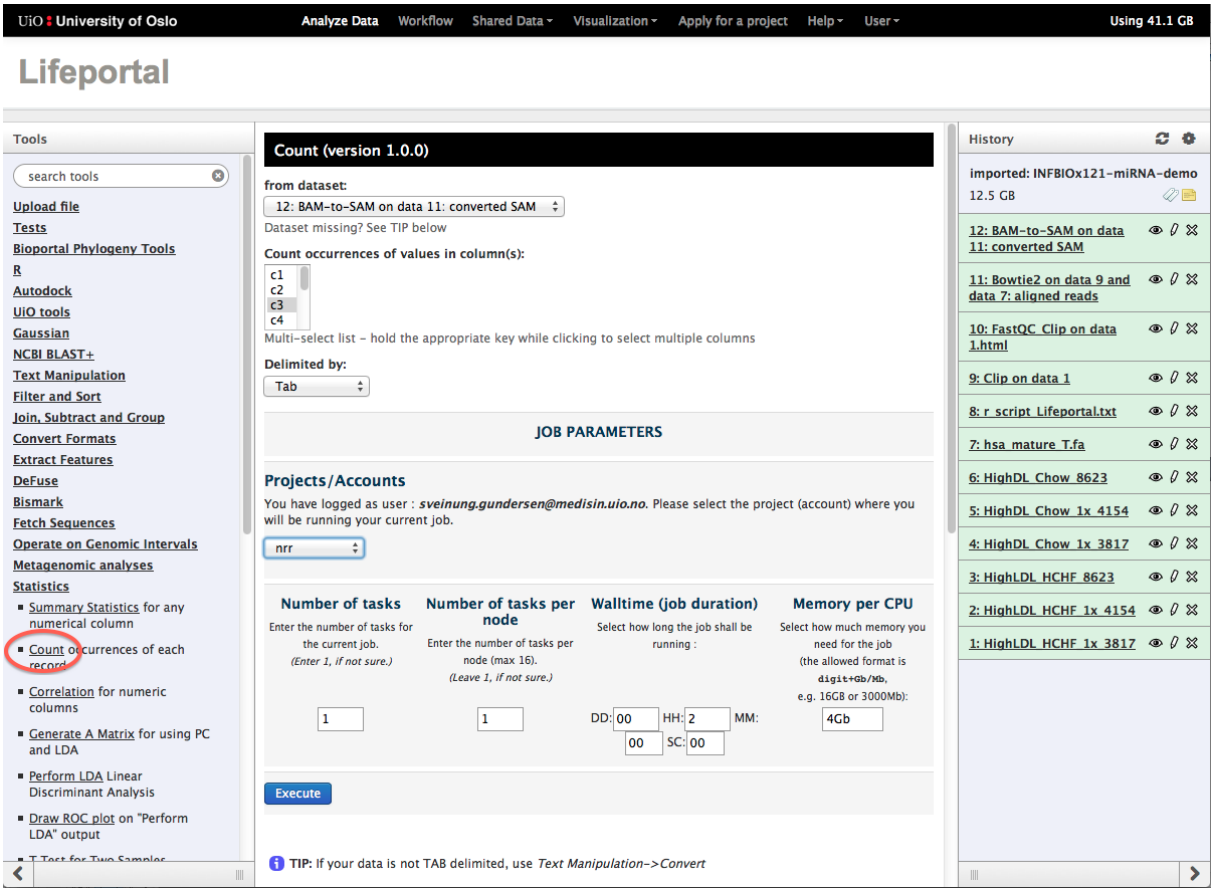
The next thing we need to do is to count the reads for each individual miRNA.

Under **Statistics**, choose **Count occurrences of each record**.

Under **from dataset** choose your **BAM-to-SAM** history item.

Count occurrences of values in column(s) should be c3.

Keep the **delimited by Tab** and run.



Click the **eye** icon of the resulting history item. Note that the first column in the file contains the counts and the second column contains the name of the miRNA that is counted. In order to better manipulate the data, we need to reverse the order of the two columns. To do this, we will use the **cut** tool.

Under **Text Manipulation**, select the **cut** tool. This is typically used for extracting some of the columns, in any order. In our case, we will use the tool to extract *both* columns, but in the reverse order, i.e. column 2 before column 1.

Write **c2,c1** in the **cut columns** box.

Select the result of the **Count** job after **From**. Select a project and execute the job.

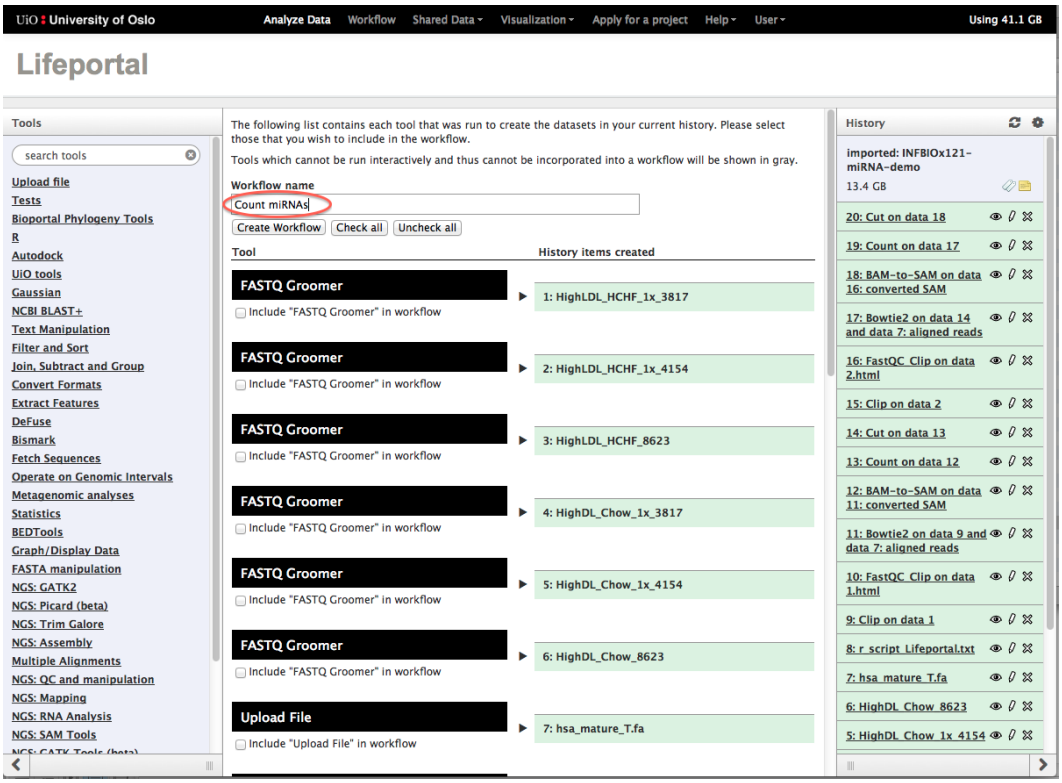
2-5. NOW, REDO ALL THE STEPS FOR THE NEXT HCHF FASTQ FILE (2: HighLDL_HCHF_1x_4154).

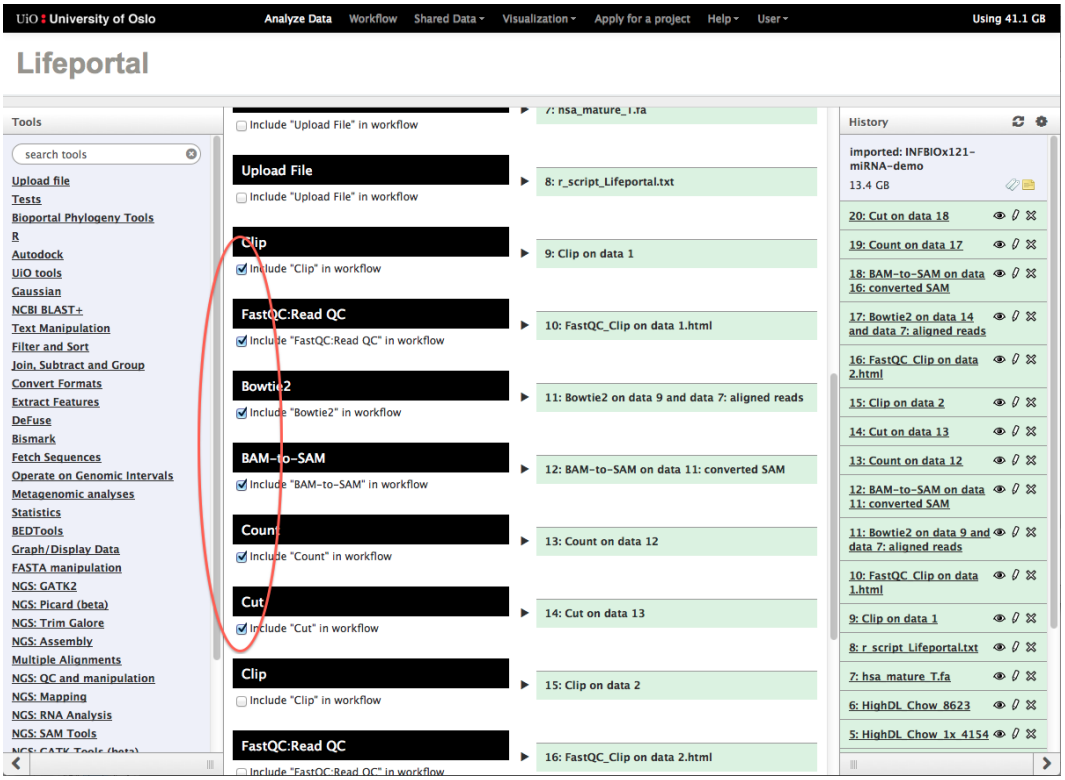
You should redo all six steps, from **Clip** to **Cut**. To make this easier, you can expand each of the previous history items and click the **rerun** button . Then change the input dataset accordingly, but keep the other options.

Note that you do not need to wait for each step to finish before you rerun the next one. Each step can be set up in advance, and will run when the previous step is finished. Note also that if you rerun the steps in this fashion (without waiting for each step to finish), you need to manually write "c3" in the **Count** step instead of selecting c3 from a list.

6. Create a workflow and rerun the steps for the last HCHF file

Click the **cogwheel** icon on the top left of the history pane, and select **Extract workflow**. In the middle pane, you will now be able to select which previous jobs to base the new workflow on.

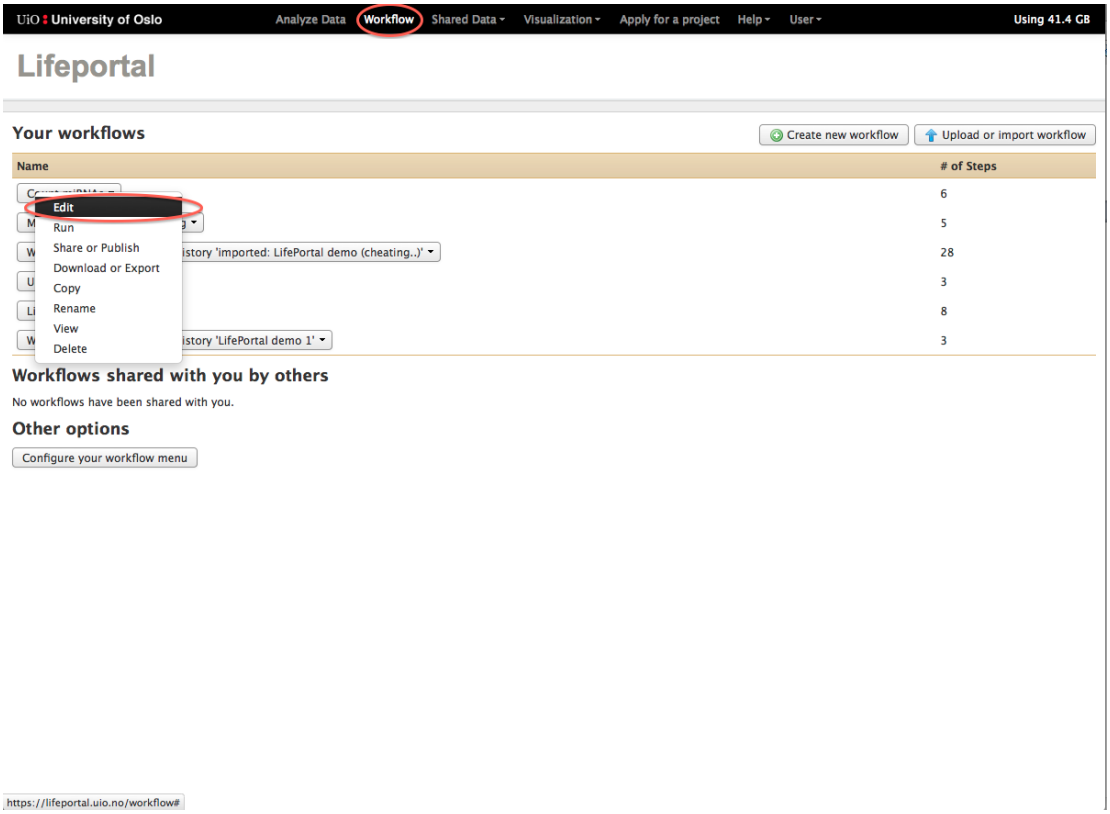




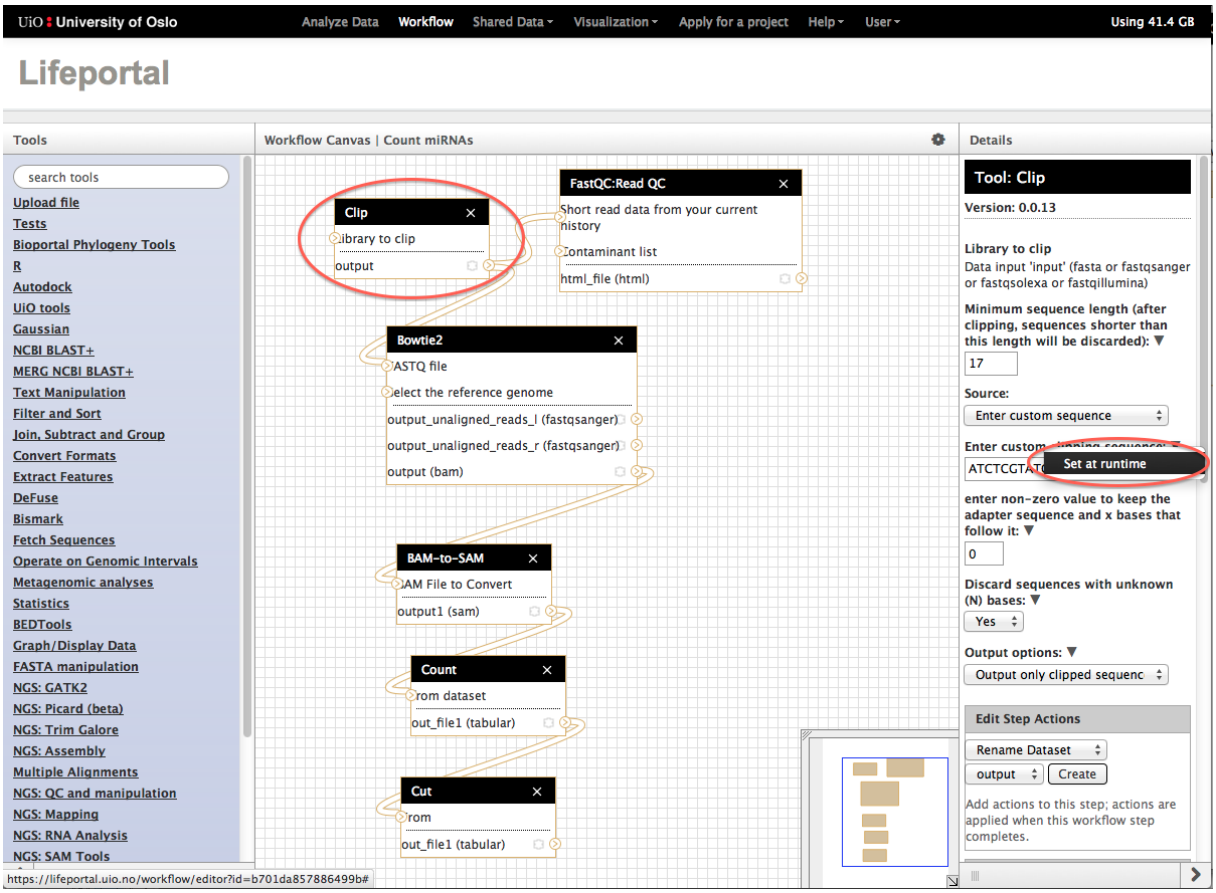
Name the workflow “Count miRNAs”, or something similar.

Select only the six steps from **Clip** to **Cut**. Click **Create Workflow** on the top of the page.

Next, click **Workflow** on the top menu of the LifePortal screen. Then, click on the “Count miRNAs” workflow and select **Edit.**



You will now see a graphical representation of the workflow:



Drag and rearrange the boxes in a way that looks nice. Try to understand what the boxes and lines represent, as well as the input and output arrows.

A workflow is not connected to specific datasets, but may be reused for any dataset. As the 3' adapter sequence will differ between datasets, the choice of adapter sequence should be up to the user. To configure this:

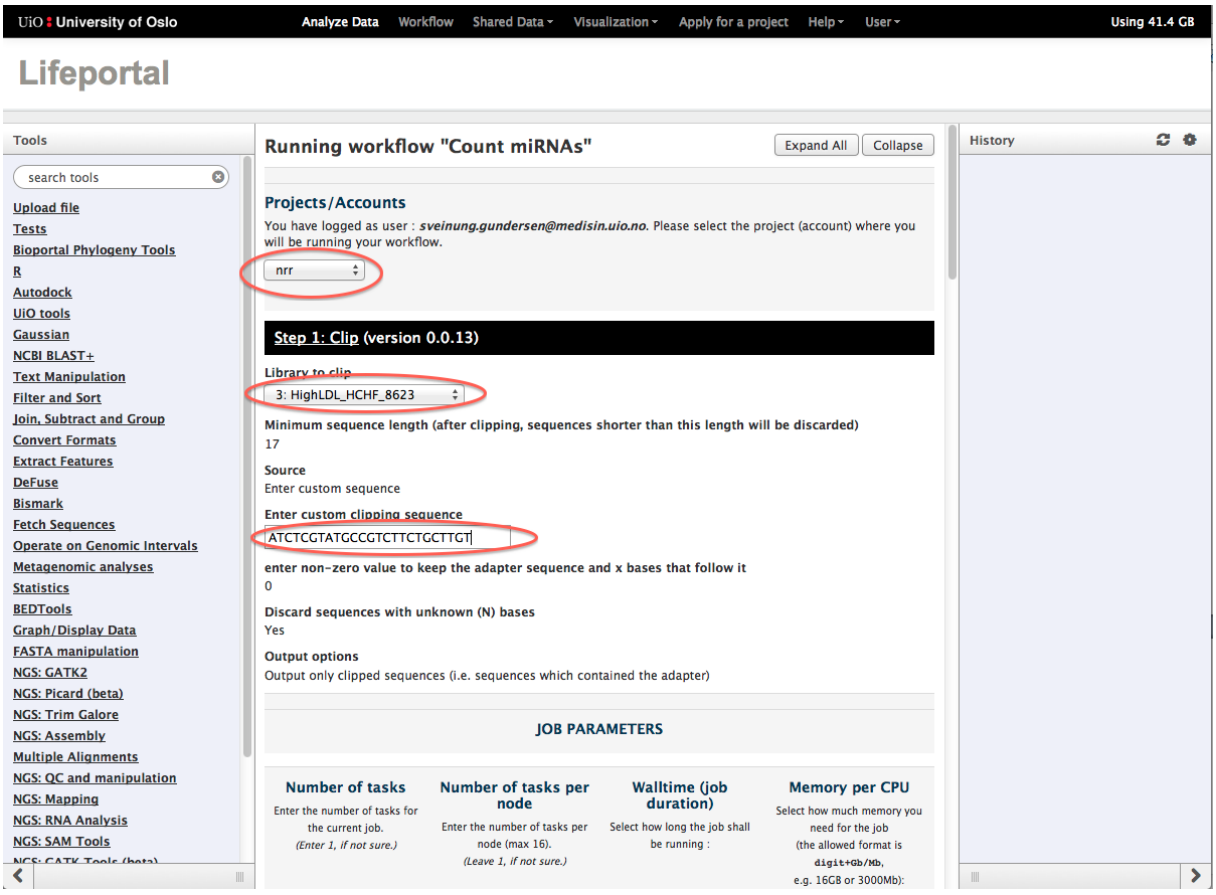
Click the box for the **Clip** job.

On the right-hand side, click the arrow above **Enter custom clipping sequence** and select **Set at runtime.**

To save the workflow, click the **cogwheel** icon and select **Save**.

Now, we want to run the workflow on the last HCHF dataset:

Click the **cogwheel** icon and select **Run**.



Select a project.

Under the **Clip** job, select the next HCHF input file ("3: HighLDL_HCHF_8623"). Enter the custom clipping sequence: ATCTCGTATGCCGTCTTCTGCTTGT

Under the **Bowtie2** step, make sure that the correct reference genome is selected.

Click **Run workflow** at the bottom of the page. Now, it is time to take a break, as it will take some minutes to run the workflow.

7. Create a full workflow that also joins the counts for all three input datasets

At this point we have the counts for the different experiments in different files. To simplify subsequent analysis, it is better to have only two files. One with the counts for HCHF and one with the counts for Chow. The first step is to join the HCHF count files.

You will add a new step for this, but in this exercise, you will add the step via the workflow editor. Also, you should now be quite skillful in using Galaxy, so you will get a little less step-by-step input at this stage.

First, extract a new workflow "Count 3 miRNAs datasets - joint output" (or similar), based upon all six steps, for all three datasets, in total eighteen steps.

In the workflow editor, align the boxes in an orderly fashion, so that the output files (the **Cut** boxes) are beside (perhaps above?) each other. The **Auto Re-layout** option in the **cogwheel** menu might be of help.

Click the tool **Column join** under the category **Join, Subtract and Group**. Move the box near to the **Cut** boxes. Click **Add new Additional Input** to create a third input arrow for the new tool.

Connect the output of the three **Cut** boxes to the inputs of the **Column Join** box.

Type **c1** in the first option box (as the **hinge** column). Type **c1,c2** in **incluce these columns.** Select **Yes** for Fill empty columns. Select **Single fill value**, and type **0** (the number) as the **Fill value**.

Save the workflow and exit the editor.

Before running the workflow on the Chow data, you must carry out the **Column join** step on the HCHF data manually. Just be sure to exit the workflow editor, click **LifePortal** on top of the screen to see the tool meny. Then, select the tool, and choose the correct input as described above, and click **execute**.

Rename the resulting history item by clicking the **pencil** icon, type "HCHF_count_proper" in the **Name** box, and click **Save.**

Click the **eye** icon on the history item to look at the results.

Now, run the complete workflow using the 3 Chow datasets. Make sure that the correct adapter sequence is selected for all three files. Change the name of the last result to "Chow_count_proper"

Note that many of the jobs now run in parallel, meaning that the complete workflow should now take about the same time as when only one dataset was used.

8. Run a differential expression analysis with EdgeR

We now have count files for both the HCHF samples and the Chow samples. The differential expression analysis will be done with the help of an r-script.

Click on the **R programming language for statistical computing** under **R**.

Input files should be **One R script and data files**.

The **R-script** is called **r_script_Lifeportal.txt**.

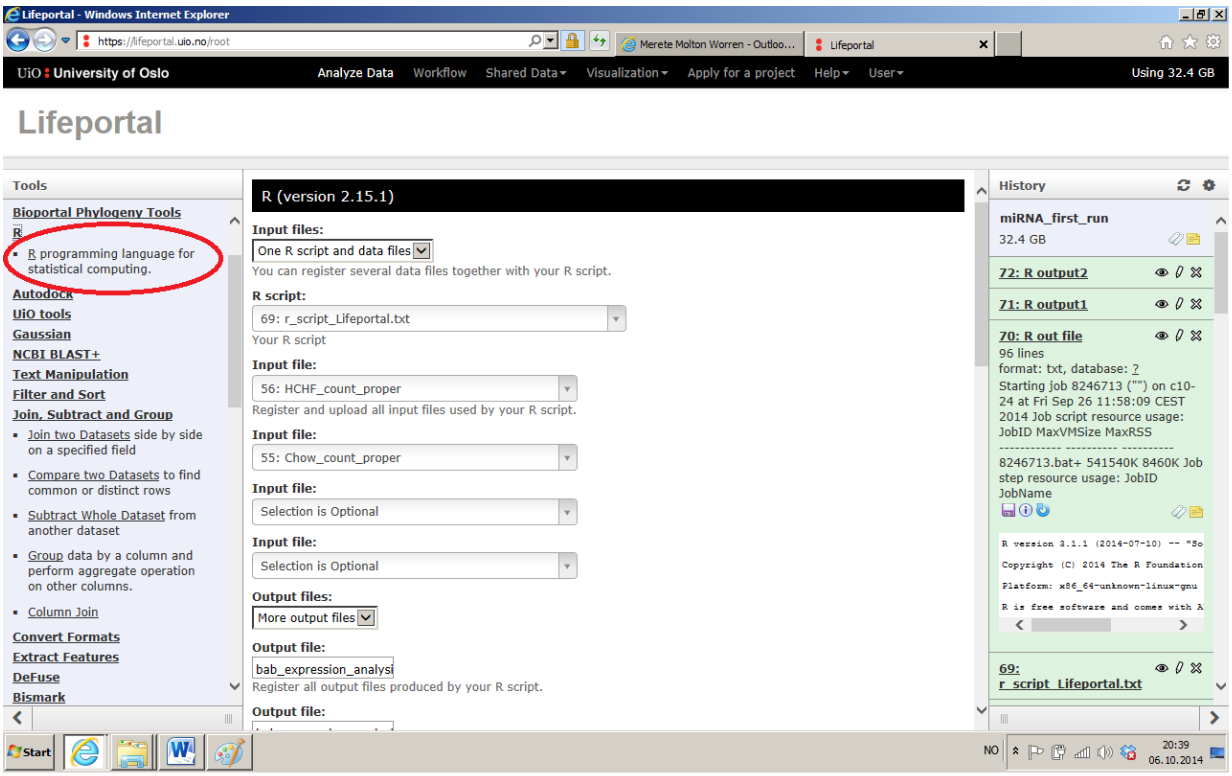
The first input file is the HCHF count file, the next is the Chow count file, the others are left blank.

Output files should be set to **More output files**.

The script produces two output files, and you need to write in their names.

The first is called: bab_expression_analysis_top_results.txt

The second is called: bab_expression_analysis_top_counts.txt



9. Make a Galaxy page

A Galaxy page is a public page explaining the reasoning and details behind you analysis, together with a presentation of you main results. Most importantly for *reproducibility* is the inclusion of previously run histories, with all related *metadata*, in addition to workflows that other researcher can use to run the analysis on their own datasets.

To create a page select **Saved Pages** from the **User** meny at the top of the LifePortal screen. Then click **Add new page**.

Give the page a meaningful **title** and description (**page annotation**), and click **submit**.

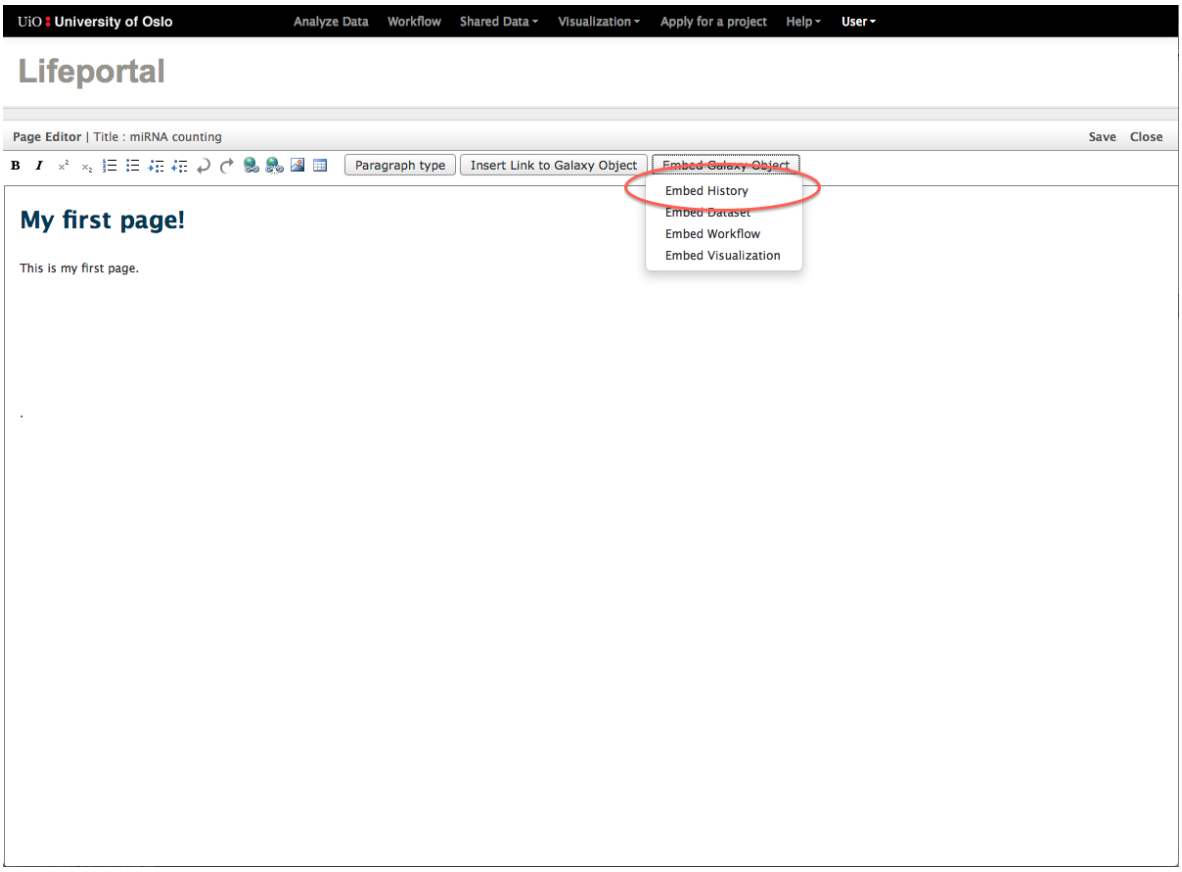
In the list of pages, click the name of the new page and select **Edit contents**.

Create a nice looking page with information about the analysis you have just carried out. Use the **Paragraph type** menu to create headings as needed.

Embed your history and the last workflow using the choices under **Embed Galaxy object*.**

*Note: The implementation of Galaxy Pages contains a bug complicating the creation of pages (resulting in content not being displayed in the output). In our experience, some simple measures avoids the issue: First, when beginning a page, add several blank lines and end the document with a period: "." After having

embedding histories or other elements, make sure to manually jump to the next pre-entered line (by clicking) instead of creating a new line with the newline (return) key on the keyboard.



When you are finished, click **Save** and **Close**.

Click the page title and select **Edit** to view the finished results.

Lastly, share your page by clicking the page title selecting **Share or Publish**.

Click **Make Page Accessible via Link** to get a **URL** (internet address) you can share with people on email or put in your article.

Click **Share with a user** to share the page to the teachers of this course (meretewo@uio.no and sveinugu@uio.no).