Running RNA-Seq Pipeline on WEXAC

Refael Kohen

I will show you an interface of pipelines for analyzing RNA-seq and Mars-seq data. You have learned how to analyze RNA-seq and Mars-seq data step by step. But in standard cases, you can run a pipeline that it will perform for you all steps of the analysis with a click of the button.

This is the address of the website: ngsbio.wexac.weizmann.ac.il. We need to login to the system. You can use your regular username and password of Weizmann. For the course, we are login with class username. For example class1.

Now you can see the file system of your lab on wexac server.

But, you need to be sure that your lab, has a collaboration folder with the biological services unit. This a folder the IT of WEXAC can make with special permissions of writing for our bioinformatics unit and your group on “wexac” server, the web interface can recognize only the folders and files under this Collaboration folder.

For example, for the “testing“ lab, to which “class1” user belongs, there is this folder:

\home\labs\testing\Collaboration

Now, you need to navigate in the file system and select the input folder. The input folder contains the fastq files. Valid input folder is a folder that is built in a specific format. Beside valid input folder, there is “select” icon, click on the icon for selecting the input folder.

Let's look at an example, what a valid input folder looks like:

Under the main folder, there is a subfolder for each sample. The names of the subfolders are the sample names. Within each subfolder, there is one fastq file in case of a single-read sequencing or two fastq files in case of a paired–end sequencing. The filename needs to start with sample name, continue with an underscore, R1 or R2, and .fastq.

For example:

In folder Sample_name1 we have a file Sample_name1_R1.fastq

The web interface can recognize also compressed files in gz format, so the file name can be also with a suffix of .gz:

Sample_name1_R1.fastq.gz

If you do the sequencing in sandbox unit of the biological services, you will get the fastq files in this format.

Now, we need to write the path of the output folder. The default path of the output is in the same level of the input folder in the file system. You can change it, just verify that the output folder is

under the Collaboration folder. If you write path outside of the collaboration folder, you will get an error message.

Click next and select the pipeline, MARs-seq or RNA-seq. Our example is data of RNA-seq, so we are selecting RNA-seq. Now we need to select the type of the RNA-seq protocol: a stranded or not stranded. In a stranded, the reads come only from one strand of the genome. If you select the "find automatically" option, the pipeline will check to which strand most of the reads were mapped.

Now we can select adapters. The default adapters are from the "true-seq" kit of Illumina.

Now we have the option to run the "Deseq" software for finding differential expression. We can select "no deseq" and continue to the next steps. If we select "run deseq" we will need to specify the levels (or categories) of the samples, for example, knockout vs control. In our example, we have 3 levels: "mir23a", "mir24a" and "control".

Now we need to drag and drop each sample to the level to which it is belongs.

We select the reference genome and the gtf file. For now there are only Human and Mouse genome. If you want run on other genomes, send me the genomes and I will upload them to the system.

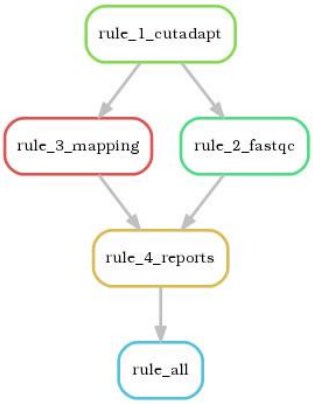
Now we can select the queue of the computer cluster. The jobs of the pipeline are short so you can use the "new-short" queue unless you know that "new short" queue is overloaded. In our course, you can use the "bio-guest" queue that run the pipeline on stronger machines.

Now you can click on "start processing". In this screen, you can follow the progress of the run. In the end of the run, you will get an email with a link to the results.

The results of the pipeline:

Example of report:
==========
/home/labs/testing/Collaboration/course_2017/GUI_Pipeline_output/RNA-seq-output-example/report.html
http://ngsbio.wexac.weizmann.ac.il/runs/OnIYGUHxrMgvwmChaOawlKJB/report.html
---

Following are the steps of RNA-seq pipeline:

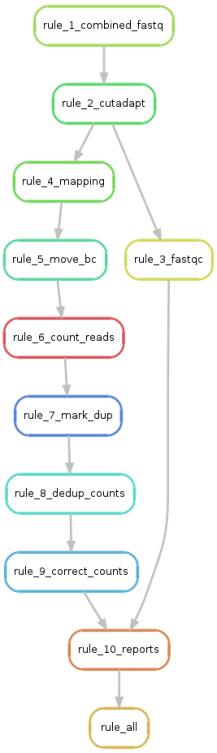


Cutadapt step: remove the adapters from the fastq files.

Fastqc step: creates QC statistics on fastq files.

Mapping step: maps the reads to the reference genome with STAR software, and count the reads that mapped on genes.

In Mars-seq pipeline there are another steps as you have learned.



In the output folder:

/home/labs/testing/Collaboration/course_2017/GUI_Pipeline_output/RNA-seq-output-example

For each step, there is folder with its output files.

The pipeline is built with Snakemake package of python. In each step, the pipeline receives the input files from the output of the previous step and create its output files, and so on.

In the last step 4_reports you can find the final report file: report.html

You get also link to this file in the email.

In the report has three sections: sequencing and mapping QC, Sample QC and Differential expression (only if you run the "DESeq" software).

sequencing and mapping QC section:

First graph: shows the number of reads for each sample in raw data.

Second graph: shows the **percent** of short reads that are discarded after trimming the adapters.

Third graph: shows the number of reads for each sample in each step of the pipeline.

Fourth graph: shows the average quality of the bases on the reads. Quality of 30 and up is good.

Fifth graph: a. **percent** of reads that mapped uniquely and not uniquely per sample.

b. percent of the uniquely mapped reads that mapped to genes with count of above 5.

Biological QC section:

First heatmap: show the top highly-expressed genes. For example the expression of gene RN45S in sample SRR3112243 is 15% out of all genes.

Second heatmap: show the correlation between samples according to the gene expression.

Clustering: show the clustering of the samples according to the gene expression.

PCA of the samples. And how much each component explains the variability between the samples.

Differential expression section:

In this section we show the top 100 differential expressed genes in each condition and a table with the number of genes differentially expressed in each category (up/down).