WORKBOOK FOR NUCLEOTIDE SEQUENCE ANALYSIS COURSE

-2024/2025-

At the end of this course you will know how to:

- download the sequencing data from SRA database
- check the quality of the sequencing data
- perform filtering based on quality parameters
- align reads to a reference genome
- obtain a count matrix, i.e. number of aligned reads per gene which could be used for gene expression analysis
- *de novo* genome assembly of bacteria

Instructions for filling in the worksheets:

- Worksheets include questions (some are for refreshing the knowledge, whereas some will probably require a little bit of literature review) and practical work, where you carefully insert all codes and as much as possible comments.
- Each step of the bioinformatics analysis should be well documented and explained why it was performed and why each parameter/option of the command was used.
- Bioinformatics tools which you will use have several options. I encourage you to explore why they are used for. Make sure that you add this to the report as well.
- Comments regarding the worksheets, how can be improved, what should be added, etc. are welcome.

During this course you will work on the RNA-Seq dataset, linked to the following article:

"Wolbachia pipientis modulates germline stem cells and gene expression associated with ubiquitination and histone lysine trimethylation to rescue fertility defects in Drosophila"

Wolbachia pipientis are maternally transmitted endosymbiotic bacteria commonly found in arthropods and nematodes. These bacteria manipulate reproduction of the host to increase their transmission using mechanisms, such as cytoplasmic incompatibility, that favor infected female offspring. The underlying mechanisms of reproductive manipulation by W. pipientis remain unresolved. Interestingly, W. pipientis infection partially rescues female fertility in flies containing hypomorphic mutations of bag of marbles (bam) in Drosophila melanogaster, which plays a key role in germline stem cell daughter differentiation. Using RNA-seq, we find that W. pipientis infection in bam hypomorphic females results in differential expression of many of bam's genetic and physical interactors and enrichment of ubiquitination and histone lysine methylation genes. We find that W. pipientis also rescues the fertility and germline stem cell functions of a subset of these genes when knocked down with RNAi in a wild-type bam genotype. Our results show that W. pipientis interacts with ubiquitination and histone lysine methylation genes which could be integral to the mechanism by which W. pipientis modulates germline stem cell gene function.

SOFTWARE INSTALLATION

Command line cheat sheet

https://github.com/RehanSaeed/Bash-Cheat-Sheet

With Conda package manager installation of bioinformatics software is easier. Bioconda is a channel (package repository) for the conda package manager specializing in bioinformatics software.

Installation instructions for **Miniconda (or mamba - GitHub - conda-forge/miniforge, Troubleshooting — documentation, strict to flexible,...)**
Linux installer: Installing Miniconda - Anaconda

Instructions: Installing on Linux — conda 25.3.1 documentation

Bioconda repository:
Installation instructions can be found here: Biococnda
Set channels: https://bioconda.github.io/user/install.html

Conda cheat sheet: Cheatsheet — conda 25.3.1 documentation

For the beginning, create a new environment (it is easy to create an environment with a particular python version, if necessary).

You need to install miniconda or anaconda or micromamba.

As always use the cheatsheet, if you don't remember commands: CONDA CHEAT SHEET

Other tools for setting up bioinformatics tools that it is recommended to check are Docker and Singularity (it can import Docker images too).

INFO ABOUT RESEARCH AND SEQUENCING DATASET

If you don't have already esearch , create a conda environment for that. Run this command to get some data about the sequencing datasets:

```
$ conda create -n esearch
$ conda activate esearch
$ conda install bioconda::entrez-direct
$ esearch -db sra -query PRJNA1166928 | efetch -format runinfo > runinfo.csv
```

Questions:
What is the aim of the study, described in the scientific article?
Provide some info about the dataset you will work with and which sequencing technology was used.
Which kit was used for sequencing library preparation? Does this kit preserve strand information (stranded library) or not?
What is the advantage of stranded mRNA library preparation compared to non-stranded library?

A Guide to RNA-Seq: Stranded vs Non-Stranded RNA-Seq by GENEWIZ from Azenta Life Sciences

DOWNLOADING THE DATA

Instructions how data can be downloaded from SRA database are available here:
https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/

Note that there is an option to download public data (SRA format) and original submitted files.

Practical work:
Sequencing data can be obtained from the SRA database with the SRA Toolkit. Provide SRAtoolkit code for downloading SRR dataset. Be careful if single or paired end sequencing was performed.

https://github.com/ncbi/sra-tools/wiki/HowTo:-fasterq-dump (2.10)
conda install -c bioconda sra-tools

example of a command for downloading the sequences linked to SRR5831049
```
fastq-dump --origfmt --skip-technical --read-filter pass --clip
--split-3 -X 500000 SRR5831049
```

fasterq-dump!

Or using the following link: https://sra-explorer.info/
You can select the needed SRAs, put them in the chart and copy a bash script for downloading data.

Create a new environment where you should install curl tool
`conda install -c conda-forge curl`

Go on the terminal, open the text editor **nano name_of_the_file.sh** , and paste what you have copied before.
SAVE IT with CTRL+ o

EXIT with CTRL + x

Launch the command:

bash **name_of_the_file.sh**

If after the download, the file ends with *gz, this means it's compressed. So, you have to decompress it using **gunzip.**

CHECKING QUALITY THE SEQUENCES

Check FASTQ format example:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Quality score

https://help.basespace.illumina.com/files-used-by-basespace/quality-scores

Check the quality of the downloaded dataset with FastQC.
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Practical work:

a) Run FastQC over your dataset. Explain the results.
b) Exchange FastQC results with your colleagues and run MultiQC to get a joined report for all datasets.

conda install -c bioconda fastqc
conda install -c bioconda multiqc

- Installing multiqc on Conda produces "UnsatisfiableError:" - Stack Overflow

run FastQC

```
fastqc SRR5831049_pass_1.fastq
# or fastqc *.fastq to run fastQC on all files
```

run MultiQC in a folder with FastQC results for the forward and reverse reads and examine the multiqc_report.html

```
multiqc .
```

Questions:

a) How is a fastq file composed?
b) How can I count the number of reads in a fastq file? Describe different ways to perform that.
c) What about the quality of your reads?
d) Describe your fastqc and/or multiqc and interpret the results.