article
general-instructionsGeneral instructions
Instructions for filling in the worksheets:
Worksheets include questions (some are for refreshing the knowledge, whereas some will probably require a little bit of l
Each step of the bioinformatics analysis should be well documented and explained why it was performed and why each
Bioinformatics tools which you will use have several options. I encourage you to explore why they are used for. Make su
Comments regarding the worksheets, how can be improved, what should be added, etc. are welcome.

setting-up-your-environmentSetting up your environment
Virtual environments are great, because they let you have separate environments for separate projects. This is a
condaConda
An advantage that **Conda** provides is not only for managing Python libraries, but also command line tools. This c
I recommend installing Conda with these instructions: docs.conda.io/projects/conda/en/latest/user-guide/install/i
biocondaBioconda
To use certain bioinformatics tools, we need to use the **Bioconda** channel. No installation is needed, only this 3 co
[] conda config --add channels bioconda conda config --add channels conda-forge conda config --set channel_priority
virtual-environmentVirtual environment
Finally, we can create an environment for this project using:
[] conda create --name ans
Answer the prompt with yes to create an environment and then **activate** the environment with `conda activate a`
finding-our-dataFinding our data
During this course we were tasked to work on the RNA-seq dataset linked to the article titled *"Wolbachia pipientis*
bioproject-accession-numberBioProject accession number
First thing to do is to find this article on **NCBI** or **PubMed**. This can be done with a quick Google search: https
In the article, find the section *Data availability.* There will be an accession number for a **BioProject** (a BioProjec
sra-accession-numberSRA accession number
We can search for sequencing data related to this BioProject through the command line using **NCBI Entrez Dire**
`esearch` queries the database and returns all accession numbers that match that query
`efetch` fetches the data linked to those accession numbers
[] esearch -db sra -query PRJNA1166928 — efetch -format runinfo > runinfo.csv
`-db` specifies an NCBI database to search
`-query` specifies the search query
| pipes the output of the previous command as input for this command
`-format` specifies the output format
If we take a quick look at this file, it has many different columns, while we are only interested in the SRA (sequenc
[] tail -n +2 runinfo.csv — cut -d ',' -f1 > SraAccList.txt
`-n +2` tells `tail` to start displaying from the second line
The `-d` flag is used to specify a delimiter and the `-f1` tells `cut` to extract the first field.
Each of us students had to choose one accession number. I choose **SRR30833097**. Save this into a file with `echo`
questions-regarding-research-and-sequencing-datasetQuestions regarding research and sequencing dataset
a-what-is-the-aim-of-the-study-described-in-the-scientific-articlea) What is the aim of the study, described in the sci
To study how *Wolbachia pipientis* influences **gene expression** (particulary the gene expression of genes associated
b-provide-some-info-about-the-dataset-you-will-work-with-and-which-sequencing-technology-was-used.b) Provide som
[[Pasted image 20250421105830.png]]

Name: - This dataset contains reads from ovarie tissue of infected and uninfected *Drosophiila menogaster* organism
Sequencing technology: - ILLUMINA (NextSeq 500).
Run statistics: - The data comes from a single sequencing run with the mentioned technology. - 4.5 million reads
Library: - Source material are transcripts. - Selection (filtering of RNA) by polyA tail was used. - SINGLE layout
c-which-kit-was-used-for-sequencing-library-preparation-does-this-kit-preserve-strand-information-stranded-library-o
The kit is not listed, so I cannot answer this question.
d-what-is-the-advantage-of-stranded-mrna-library-preparation-compared-to-non-stranded-libraryd) What is the adva
Stranded libraries are prepared in a way to contain information about the strand of cDNA from which the transc
This kind of library: - allows you to **distinguish between the sense** and **antisense** strands of cDNA - is useful f
source: https://youtu.be/yp9A5E-Y49Y?si=qzVTqlXUrEowuIKG, https://lcsciences.com/why-is-strand-specific-lib
downloading-dataDownloading data
prefetchPrefetch
Sequencing data is obtained from the SRA database with the SRA Toolkit. It can be installed with `conda install`
[] prefetch --option-file OurAcc.txt
Prefetch is used to obtain *Runs* (sequence files in compressed SRA format). The `--output-file` flag is used to us
The prefetched runs can be converted into FastQ format using `fasterq-dump`, that takes the created directory as in
[] fasterq-dump --skip-technical SRR30833097/
`--skip-technical` returns only biological reads.
Since we have only single-end sequences, it should output a single `.fastq` file in the current directory. You can chec
[] wc -l SRR30833097.fastq
generating-a-quality-reportGenerating a quality report
a-run-fastqc-over-your-dataset.-explain-the-resultsa) Run FastQC over your dataset. Explain the results
command-line-fastqcCommand line FastQC
To get a **full report** on all sequences in our dataset, we can use **FastQC tool**. Install it with `conda install -c`
[] fastqc SRR30833097.fastq
The quality report is the generated `.html` file. To view the rendered file, you can open it with your browser (e.g. `op`
gui-fastqcGUI FastQC
Alternativelly, you can open a **graphical user interface** of FastQC tool by executing only `fastqc`. This will open
[[Pasted image 20250418203153.png]]
Click on `File > Open` and select your file. When it's finished, you should see the report. Click on `File > Save r`
explaining-the-resultsExplaining the results
FastQC shows a summary of modules that were run on our data. On the left there is a **green tick** if the module s