01-Basic_bioinformatic_finding_virus_in_plant_with_conda

September 3, 2020

1 Preparing your data

If you are planning an analysis, the following consideration will facilitate it (for fasta input):

- Save you data in a specific folder
- Verify that the newiline character is consistent and preferably the UNIX type `\n`
- Verify that sequences are consistent: check there are no weird characters
- Avoid long FASTA headers. Include only the sequence name, *without spaces*

2 Preparing your analysis

Before starting the analysis, you need to check your working environment: - Where are you on your computer *pwd* - check what tool are already present in your environment *env* - if needed, go to the wanted working directory

```
cd /mnt/c/Users/johan/OneDrive/Bureau/bioinfo/Training/Conda
```

Conda is an environment manager allowing the user to install tools in specific environment. It is very usefull in bioinformatic because it allow to start an analysis with tools that required specific requirement without any conflict (python2/python3 for example). There is two main toolkit that use the conda management system.Anaconda, the data science toolkit that contain all the basic tool ready to run. Miniconda that have only the minimal amount of tool installed, which mean that miniconda is much lighter than anaconda but required additional installation.

If needed, download the latest version of miniconda:

```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
-O Miniconda3-latest-Linux-x86_64.sh
```

Then install it

```
bash Miniconda3-latest-Linux-x86_64.sh
```

The installation will require several validation from you side.

Once the installation is done, you can start to play with the different conda option and environment !

[7]:
```
conda env --help
```

```
usage: conda-env [-h] {create,export,list,remove,update,config} ...
```

```
positional arguments:
  {create,export,list,remove,update,config}
    create              Create an environment based on an environment file
    export              Export a given environment
    list                List the Conda environments
    remove              Remove an environment
    update              Update the current environment based on environment
                        file
    config              Configure a conda environment

optional arguments:
  -h, --help            Show this help message and exit.

conda commands available from other packages:
  env

Note: you may need to restart the kernel to use updated packages.
```

3 Create your environment

you can use the create option to build an environment from scratch, or use a configuration file. Here is an example of a configuration file:

```
name: Basic_bioinformatic
channels:
  - conda-forge
dependencies:
  - python=3.6
  # scientific python
  - numpy
  - scipy
  - matplotlib
  ## Code edition
  - notebook
  - jupyter_contrib_nbextensions
```

Those instructions are in the *REQUIREMENTS_conda.yml* file. You can create an environment named Basic_bioinformatic that use python3.6 with several python libraries and a jupyter notebook by runing this command:

```
conda env create -f REQUIREMENTS_conda.yml
```

4 Adding tools to your environment

First, you need to activate your environment

```
conda activate Basic_bioinformatic
```

When your environment is active, everything you are doing is inside it, if you install a tool, it will be limited to that specific environment.

We will try to analyze some of the data presented in this article following the same step (kind of). So the analyse that we are going to do is : - Fastqc (read checking) - Trimmomatic (read trimming) - (Meta)spades (assembly) - Blast (identification)

Let's install the tools:

```
conda install -c bioconda fastqc

conda install -c bioconda trimmomatic

conda install -c bioconda spades

conda install -c bioconda blast
```

[14]:
```
%%bash

    which fastqc


    #fastqc --help
```

```
/home/jrollin/miniconda3/envs/Basic_bioinformatic/bin/fastqc
```

5 Make the analysis

First, you need to activate your environment

```
fastqc &


trimmomatic PE -threads 1 -trimlog testlog -summary sumlog
data/SRR10715671-1.fastq data/SRR10715671-2.fastq trimming_result_1.fastq
trimming_result_unpaired1.fastq trimming_result_2.fastq
trimming_result_unpaired2.fastq SLIDINGWINDOW:4:15 MINLEN:36

metaspades.py --only-assembler -o assembly_result/ -1 data/SRR10715671-1.fastq
-2 data/SRR10715671-2.fastq
```

The blast will require too much computing power and takes too much time to run on every computer. We can do it on Blast web interface using only the first contig and limited to Moroccan watermelon mosaic virus (taxid:167129) to save time.