Snaq Documentation

Contents

1 Background

This software is a great snakemake tool to run the analysis easily and productively.

2 Installation

2.1 Snaq installation

2.1.1 Clone github repository

```
$ git clone https://attayeb/snaq.git
```

or download the zipped release file, [this part will be explained when the repository is made public]

2.1.2 Download the latest release

the latest release of this snakemake pipeline is in github repository: `https://github.com/attayeb/snaq/releases`

```
$ wget https://github.com/attayeb/snaq/releases/tag/untagged-77b9a20481d70db1aff2
```

2.2 Snakemake installation

2.2.1 Linux, Mac and windows using Windows Linux Subsystem

The easiest way to install snakemake is by using Conda, however, other ways are mentioned in details in snakemake website `https://snakemake.readthedocs.io/en/stable/getting_started/installation.html`

Mamba is a lighter conda interface that is recommended by Snakemake developers to be installed.

```
# install mamba using conda-forge
$ conda install -n base -c conda-forge mamba
$ conda activate base
$ mamba create -c conda-forge -c bioconda -n snakemake snakemake
$ conda activate snakemake
```

2.2.2 Using docker

Docker can be used in Windws, linux and MAC, You need to have docker client installed in your system, and you need to pull snakemake image from dockerhub, snakemake image is maintained by the same guys who created snakemake, for more details you can check this link `https://hub.docker.com/r/snakemake/snakemake`

```
$ docker run -it -v C:\snaq:/work -w /work \
     snakemake/snakemake {command}
```

`docker run`: docker main command

`-it`: This means interactive and allow sending command at the end.

`-v`: will map [Host directory]:[inside container directory]

`-w`: sets the working directory inside the container, and in that directory snakemake commands will be executed.

`snakemake/snakemake`: The docker image that is going to run.

`{command}`: Whatever command you write here will be executed inside the container after it is built.

Basically docker will run an ubuntu system inside a container, and it will map the user folder in which he has snaq installed to the work folder inside the container system, will run the analysis using the container system and save the output, including the installation of qiime, other tools and databases in the host directory, then the user can use the results in the host without any issue.

3 Preparing data for analysis

3.1 Input data structure

Snaq can analyze paired enda fastq files from illumina sequncers. They should be saved in new folder inside `results/` folder in the main snaq folder

3.2 Data set concept

Because this pipeline allows automatic installation of applications and downloading of datyabases, the size of the whole pipeline is rather large, therefore, we decided that it is better to give the pipeline the ability to anayze multiple data sets, which should be stored in separate folders in `/results` folder.

4 Snakemake commandline

All commands are executed by snakemake Snakemake requires two important parameters `--cores {Number of cores}` and `--use-conda` then the target file name to be produced. For more details of snakemake command line tools please check: `https://snakemake.readthedocs.io/en/stable/`

If the user has already all the requirements for the pipelines installed, he can ommit `--use-conda` and run the pipeline in the environment that he prepared. however, number of cores

```
$ snakemake --use-conda --cores 10 results/AB/AB.qza
```

`--use-conda`: To tell snakemake to use the rule specific conda environment.]

`--cores`: number of allowed parallel jobs by the user.]

in case of docker

```
$ docker run -it -v C:\snaq:/work -w /work snakemake/snakemake \
      snakemake --use-conda --cores 10  results/AB/AB.qza
```

Finally the submitted command should be like this:

5 Stages

5.1 Import data

Importing data is taking two steps, first step is creating a manifest file to read all the file names in the specified folder and list them in a manifest file following QIIME2 requirement. This manifest file is used in the second step to identify the files that need to be imported to form QIIME zipped artifact (QZA) of the data.

fastq files have to be saved in `data/AB/` folder.

5.2 Trimming

The input of this step is qza of a sequence data, `results/AB/AB.qza`, then we add the steps needed depending on the users' requirement using + sign. for example if the user wants to use fastp to crop 10 nucleotides from R1 and 20 from R2, then the command should be added to the file name +`fp-f10-r20`, the target becomes: `results/AB/AB+fp-f10-r20crop.qza`. The output will be a qza file with sequence data inside. if another step need to be done then it should follow the same way of giving the command. if bbduk is need to be used for quality trimming, then + and the second command need to be add in form of `bb-t{threshold}`. If the threshold is 18 then the command should be `bb18t`, which makes the target: `results/AB/AB+fp-f10-r20crop+bb18t.qza`.

These two steps produce QZA files, fastq-sequences.

5.3 DADA2

Next step applies DADA2 algorithm using QIIME2 DADA2-plugin, to run this, we need to add a command `dd` after + sign. DADA2 produce 3 output files, `dd_seq.qza`, `dd_table.qza`, `dd_stats.qza`. if this is the last step of the analysis then the target can be any one of them. Example:

`results/AB/AB+fp-f10-r20+bb-t18+dd_stats.qza`.

5.4 Taxonomy assignment

There are 3 classifiers already attached to this pipeline, and will be downloaded automatically when they called for the first time. Taxonomy assignment happens after DADA2 step. To do it `cls-gg` for greengenes, `cls-silva` or `cls-silvaV34`.

5.5 Diversity measurements

This stage computed alpha and beta diversity, for alpha diversity, "simpson", "chao1", "shannon_entropy", "observed_features", for Beta diversity: "unifrac", "braycurtis" and "jaccard" distances are measured. No need to run taxonomy assignment before calculating diversity, however, rarefaction step is required.

```
$ snakemake --use-conda --cores 10 \
↪  results/AB/AB+bb-t16+fp-f17-r21+dd+rrf-d10000+alphadiversity.tsv
```

5.6 Summary

Summary step will produce all the required output

6 Tutorial example

let's assume that we have a data set of Fastq files, and we want to analyze it using Snaq. Snakemake, Snaq and docker (in case of windows).

6.1 Step 1

Create a new folder in `data/` with the name AB

6.2 Step 2

6.2.1 Windows

```
docker run -it -v C:\snaq:/work -w /work snakemake/snakemake \
      snakemake --use-conda --cores 10  \
       ↪  results/AB/AB+bb-t18+fp-f17-r21+dd+cls-gg+rrf-d10000.zip
```