Installing the Workflow; DNA2A-seq-analyzer

This document was created by Dustin F.

Contents

1 Install Miniconda and Mamba

Workflow was tested on a Linux [Ubuntu] environment!

First you have to install Miniconda3!
Tested with: Conda 23.11.0

```
> wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
> chmod +x Miniconda3-latest-Linux-x86_64.sh
> ./Miniconda3-latest-Linux-x86_64.sh
```

If you do not have Mamba installed in your conda environment!
Tested with: Mamba 1.5.6

```
> conda install -c conda-forge mamba
```

2 Create environment and install workflow

Create a folder for your workflow.

```
> mkdir -p path/to/project-workdir
> cd path/to/project-workdir
```

Download the Snakemake workflow from GitHub.
If you don't have the workflow yet, you can download it from GitHub. Otherwise, if you already have the data, you can skip the following step.

Download and unzip the workflow from GitHub. You can also download the workflow as a ZIP file and then unpack it in your project folder.

Clone repo from GitHub

```
> git clone https://github.com/dusti1n/DNA2A-seq-analyzer.git
```

To install the environment with all packages, navigate to the project folder. It is important that you are in the parent folder. This folder contains, for example, the folder: config and workflow. The file (environment.yaml) is also located in the folder.

Create an env and install all required packages.

```
> conda env create -n snakenv -f environment.yaml
> conda activate snakenv
```

If you have Mamba you can also use the following command.

```
> mamba env create -n snakenv -f environment.yaml
> mamba activate snakenv
```

3 Load samples and start the workflow

Open and set configfile (config.yaml) parameters.

```
config.yaml

# Set the path to the folder where your sample folders are located!
# The sample folders must have the following structure!

# Example:
; smpl_01/ont/ontfile.fastq.gz; smpl_01/illumina/illuminafile_1.fastq
; smpl_01/illumina/illuminafile_2.fastq

# The exact folder structure is shown in the image (folder_structure.jpg).
# It is important that the path to the folder (example_data) is set.

# All sample folders are then located in this folder
CONFIG; set sample_path: /path/to/example_data/

# Set the project name for your samples; Example: drosophila_samples;
# This will create a subfolder in 'results' with the project name you entered.
CONFIG; set project_name: saccharomycetes_smpls

# Set save_dict: true to save a JSON file (Dictonary) with your samples!
CONFIG; set save_dict: true

# Canu; Set genomeSize; genomeSize={params.canu_genome_size}
# Determine the size of your genome!
# Example "12m" for Saccharomyces cerevisiae
CONFIG; set canu_genome_size: "12m"

# BUSCO; lineage; extra=config["busco_lineage"];
# For a specific eukaryotic organism:
# https://busco.ezlab.org/list_of_lineages.html
CONFIG; set busco_lineage: "saccharomycetes_odb10"

# FASTQC; Set memory; 8GB should be used as a minimum!
# Example(use of 8GB): fastqc_memory: "8192";
CONFIG; set fastqc_memory: "8192"


# THIS PART IS IMPORTANT IF YOU ALSO WANT TO USE ILLUMINA DATA!

# Set illumina_data: true; If you want to use Illumina data!
# Set illumina_data: false; If you don't want to use Illumina data!
CONFIG; set illumina_data: true

# Pilon; Set Memory for JAVA heap Space; 32G = 32 Gigabyte;
# 8GB should be used as a minimum! Example(use of 8GB): pilon_memory: "8G";
CONFIG; set pilon_memory: "32G"
```

First load all your samples with the Python script, then you can start the workflow!
The Python script creates an automatic database for all your samples.

```
> python workflow/scripts/import_samples.py
```

Start the workflow with all available cores and install all packages.

```
> snakemake --dag | dot -Tpng > dag.png # Optional; Create a flowchart
> snakemake --cores all --use-conda
# Workflow creates a results folder with all analyzed samples.
```