LCS

Lineage deComposition for Sars-cov-2 pooled samples.

Supporting material for the paper "A mixture model for determining SARS-Cov-2 variant composition in pooled samples".

Running the pipeline

The pipeline was written with snakemake: https://snakemake.readthedocs.io/en/stable/.

To get started, clone this repository and use it as a template:

```
git clone https://github.com/rvalieris/LCS.git
cd LCS
```

1. Create the conda env

All the software used on the pipeline can be installed with conda.

If you don't already have conda installed in your machine, you can follow this guide for instalation according to your operational system.

On **Linux**, you can execute the command below to create an environment, install all dependencies to run LCS and activate the new environment:

```
conda env create -n lcs -f conda.env.yaml
conda activate lcs
```

On **MacOS**, you can use another environment file to install all required dependencies:

```
conda env create -n lcs -f conda.env.macosx.yaml
conda activate lcs
```

> We have successifuly tested LCS on a MacOS version 11.5.2 with python 3.8 and ray 1.9.0.

2. Markers source choice

The markers table contains the list of all mutation markers found in each of the variant-groups defined in `data/variant-groups.tsv`.

You can either generate a new table or use a pre-generated one.

> **You will need to generate a new table if you want to change the variant-groups definition.**

Choose one of these 3 options:

1. **Use a pre-generated table**:

   Pre-generated tables are provided to shorten the time required to run the pipeline, simply choose which table you want to use and copy it to the appropriate place:

   1. pango-designation:

      ```
      mkdir -p outputs/variants_table &&
      cp data/pre-generated-marker-tables/pango-designation-markers-
      v1.2.60.tsv outputs/variants_table/pango-markers-table.tsv
      ```

   2. ucsc:

      ```
      mkdir -p outputs/variants_table &&
      cp data/pre-generated-marker-tables/ucsc-markers-2021-08-19.tsv
      outputs/variants_table/ucsc-markers-table.tsv
      ```

2. **Generate a new table using pango-designation as a source**:

   To do this you need to have a fasta file in `data/gisaid.fa.gz` containing all GISAID genomes listed in the `lineages.csv` file from pango-designation repository.

   You must register on the GISAID website to gain access to these sequences.

   The variable `PANGO_DESIGNATIONS_VERSION` on `rules/config.py` controls which version of pango-designation to use.

   You can run `snakemake --config markers=pango dataset=x -j1 repo` to download the appropriate pango-designation repository to `data/pango-designation`.

3. **Generate a new table using sequences tree generated by UCSC as a source**:

   This data, gathered by the UShER team, includes only public sequences, as such they are downloaded by the pipeline automatically.

   The variable `PB_VERSION` on `rules/config.py` controls which version of UCSC data to use.

3. Prepare your pooled sample dataset

Place your raw-fastq files pooled samples in `data/fastq`, and create a tags file listing your samples name. It should look like this:

```
$ ls data/fastq/
sample1.fastq.gz
```

```
sample2.fastq.gz
sample3.fastq.gz

$ cat data/tags_pool_mypool
sample1
sample2
sample3
```

4. Run the pipeline

To execute the pipeline run the command:

```
snakemake --config markers=pango dataset=mypool --cores <C> --resources
mem_gb=<M>
```

The `markers` config indicates which markers table you are using (*pango* or *ucsc*) ahd the `dataset` config should match your tags file `data/tags_pool_mypool` describing your samples.

You also need to indicate how many cores and memory you have available to run the analysis, snakemake will parallelize the pipeline accordingly.

5. View the results

After the pipeline completes, the results should be in `outputs/decompose`.

Generate plots and tables

Plots can be generated by running the notebook:

- results.ipynb

Citing

If you use this software please consider citing:

```
@misc{valieris2021mixture,
      title={A mixture model for determining SARS-Cov-2 variant composition
in pooled samples},
      author={Renan Valieris and Rodrigo Drummond and Alexandre Defelicibus
and Emannuel Dias-Neto and Rafael A. Rosales and Israel Tojal da Silva},
      year={2021},
      eprint={2110.01117},
      archivePrefix={arXiv},
      primaryClass={q-bio.GN}
}
```