SUBTYPING PIPELINE

Contents

- Quick Guide
- Supplementary
- Conda
- GitHub Repo

Quick Guide

Pipeline Directory

Locate to `Pipeline` directory:

```
$ cd ~/rki_subtyping/Pipeline
```

Open IDE

```
$ code ../
```

Input Folders

Make sure you have 5 directories:

```
$ tree  -d
```

```
├── AllSeqsCO20
├── InputFasta
├── ManualRega
├── References
└── Scripts
```

AllSeqCO20 Folder

Provide this folder with .xlsx files as listed (from NGS pipeline):

```
AllSeqsCO20/
├── MS95_Seqs_ENV_CO20_V5.xlsx
├── MS95_Seqs_INT_CO20_V5.xlsx
└── MS95_Seqs_PRRT_CO20_V5.xlsx
```

InputFasta Folder

Provide this folder with files as listed (from NGS pipeline):

```
InputFasta/
├── MS95_ENV_20.fasta
├── MS95_INT_20.fasta
└── MS95_PRRT_20.fasta
```

Conda Environment

Activate `subtyping_pipeline` environment.

```
$ conda activate subtyping_pipeline
```

Be sure you have change in prompt:

```
(subtyping_pipeline) beast2@Beast2:~/rki_sybtyping/Pipeline$
```

Pipeline with --outdir parameter

Parameter `--outdir` determines a name of an ouput folder. The command will generate four enumerated output folders within `Results` folder. Without specifying an output folder you can get a warning message.

```
$ nextflow Scripts/subtyping_pipeline.nf --outdir Results
```

ManualRega Folder (1)

Provide the folder with `.csv` files generated by Rega using marked `.fasta` files from the folder procuded by the pipeline:

`~/rki_subtyping/Pipeline/Results/1_marked_fasta`

These `.fasta` files have M at the end of the file name:

```
1_marked_fasta/
├── MS95_ENV_20M.fasta
├── MS95_INT_20M.fasta
└── MS95_PRRT_20M.fasta
```

ManualRega Folder (2)

Name Rega-produced files, using the pattern as in the example below:

```
ManualRega/
├── manual_rega_MS95_ENV_20M.csv
├── manual_rega_MS95_INT_20M.csv
└── manual_rega_MS95_PRRT_20M.csv
```

Pipeline with --fullpipeline parameter

Repeat the previous command with `--fullpipeline` parameter and `-resume` flag. The latter allows for generating an output up to `12_mafft` folder. The complete processes are cached.

```
$ nextflow Scripts/subtyping_pipeline.nf --outdir Results
--fullpipeline -resume
```

Pipeline with --iqtree parameter

Parameter `--iqtree` allows for running the iqtree process that produces `13_iqtree` folder within `Results`. The folder contains `.iqtree`, `.treefile`, and `.log` files. The parameter can be added at this point, as the last command with report and plot outputs being produced or not added at all (no `13_iqtree` folder then).

```
$ nextflow Scripts/subtyping_pipeline.nf --outdir Results
--fullpipeline --iqtree -resume
```

Decision

Manually modify files (see below) which contain `Manual` tag in PRRT_Subpype, INT_Subtype, and ENV_Subtype columns. Save changes and close `.xlsx` files.

```
9_joint_with_tags/
├── full_MS95_ENV_20M.xlsx
├── full_MS95_INT_20M.xlsx
└── full_MS95_PRRT_20M.xlsx
```

Report and Plot

```
$ nextflow Scripts/subtyping_pipeline.nf --outdir Results
--fullpipeline --iqtree -resume
```

Repeating the command above generates `14_report` folder with `MS95_subtype_uploads.xlsx` report file.

Repating it again generates a `MS95_subtype_counts.png` plot and adds it to the `14_report` folder.

Processes Overview

```
[1b/f2f10a] process > mark_fasta (2)        [100%] 3 of 3, cached: 3 ✓
[73/a28f41] process > stanford (3)          [100%] 3 of 3, cached: 3 ✓
[e6/e4af1d] process > json_to_csv (3)       [100%] 3 of 3, cached: 3 ✓
[65/e0eb90] process > clean_rega (3)        [100%] 3 of 3, cached: 3 ✓
[97/70bbdd] process > comet (3)             [100%] 3 of 3, cached: 3 ✓
[62/e59285] process > join_prrt (1)         [100%] 1 of 1, cached: 1 ✓
[87/4d2fcf] process > join_env (1)          [100%] 1 of 1, cached: 1 ✓
[34/36991e] process > join_int (1)          [100%] 1 of 1, cached: 1 ✓
[a9/dd644a] process > get_tags (3)          [100%] 3 of 3, cached: 3 ✓
[8d/dad394] process > make_decision (1)     [100%] 1 of 1, cached: 1 ✓
[d8/983216] process > join_with_tags        [100%] 1 of 1, cached: 1 ✓
[e6/ceaa42] process > fasta_for_mafft (2)   [100%] 3 of 3, cached: 3 ✓
[f7/9e1ccf] process > prrt_concat_panel (1) [100%] 1 of 1, cached: 1 ✓
[a4/b7aaee] process > int_concat_panel (1)  [100%] 1 of 1, cached: 1 ✓
[54/89322b] process > env_concat_panel (1)  [100%] 1 of 1  cached: 1 ✓
[c0/786bcd] process > mafft (3)             [100%] 3 of 3, cached: 2 ✓
[68/72f0eb] process > iqtree (3)            [100%] 3 of 3, cached: 3 ✓
[3c/0fb71f] process > report                [100%] 1 of 1, cached: 1 ✓
[c5/462a18] process > countplot (1)         [100%] 1 of 1 ✓
```

Supplementary

Example of .fasta within InputFasta

```
>20-02955_ENV_20
GGAATTAGGCCAGTGGTGTCAACCCAACTATTGTTAAATGGCAGCCTAGCAGAAGAAGAT
GTGGTCATTAGATCTGAAAATTTCACAAACAATGCTAAAACCATAATAGTACAGCTTAAT
GAAACAGTAGTGATTAATTGTACAAGACCCGGCAACAATACAAGAAAAAGTATACATATA
GGACCAGGAAAAGCATGGTATGCAACAGGAGAGATAATAGGAGATATAAGACAAGCACAT
TGTAAACTTAATAAAACACAATGGGAAAAAACTTTAAAAAGGGTAGCTAGTAAATTAAGG
AAACAATCCAACCTTACAACAGTAATCTTTAAGAACTCCTCAGGGGGGGACCCAGAAATT
GTAATGCACAGTTTTAACTGTGGAGGGGAATTTTTCTATTGTAACACAACACAGTTGTTC
AATAGTATTTGGAATGACACTACTAATAGTACTGACACAAATGAAACTATCACACTCCCA
TGCAGAATAAAACAAATTATAAATAGATGGCAGGAAGCAGGAAGGG
```

Example of .xlsx within AllSeqsCO20

```
Scount          Fragment    Cutoff  Header              Lauf   NGS-ID   Index GenBank-ID Sequenz
20-02944        PRRT        20      20-02944_PRRT_20    95              1                CCCCT...
20-02945        PRRT        20      20-02945_PRRT_20    95              2                CCCCT...
20-02947        PRRT        20      20-02947_PRRT_20    95              3                CCCCT...
20-02949        PRRT        20      20-02949_PRRT_20    95              4                CCCCT...
20-02950        PRRT        20      20-02950_PRRT_20    95              5                CCCCT...
```

References Folder

This folder contains reference panels and does not need any change unless reference panels should be replaced.

```
References/
├── Reference_ENV_Panel_Stanford.fas
├── Reference_INT_Panel_Stanford.fas
└── Reference_PRRT_Panel_Stanford.fas
```

Scripts Folder

This folder contains the scripts and does not need any change.

```
Scripts/
├── comet_rest.py
├── decision.py
├── fasta_for_mafft.py
├── full_join.py
├── json_parser.py
├── nexflow.config
├── plot.py
├── rega_cleanup.py
├── repeat_marking.py
├── report.py
├── subtyping_pipeline.nf
└── tag_parser.py
```

Conda

Conda Info

List available conda environments.

```
$ conda info --envs
# conda environments:
#
base                  *  /home/beast2/anaconda3
subtyping_pipeline       /home/beast2/anaconda3/envs/subtyping_pipeline
```

Conda Version

Pipeline's version of conda `4.14.0`

```
$ conda --version
```

Deactivation of Environment

This command is used to deactivate the current invironment.

```
$ conda deactivate
```

Be sure you have change in prompt:

```
(base) beast2@Beast2:~/rki_sybtyping/Pipeline$
```

GitHub Repo

Repo Link

The project is hosted here. Use this link to clone the repo in case of data loss.

How to Clone

Locate to home directory

```
$ cd
```

Clone the repo

```
$ git clone https://github.com/vera-rykalina/rki_subtyping
```

Modify path of `ProjectDir` within `subtyping_pipeline.nf`

```
projectDir = "/home/beast3/rki_subtyping/Pipeline"
```