SUBTYPING PIPELINE

Contents

Quick Guide

Important!

> Before you start the pipeline, make absolutely sure that all file names, sequence names, structures of input tables are correct and **follow exactly the patterns** as described in examples.

Pipeline Directory

Locate to `Pipeline` directory:

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

Open IDE (Visual Code Studio)

```
$ code ../
```

Input Folders

Make sure you have 5 directories:

```
$ tree  -d
```

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

AllSeqsCO20 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
```

Refer to Supplementary part of the guide, if you need more information (Page 21).

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
```

Refer to Supplementary part of the guide, if you need more information (Page 22).

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 (separator: comma) 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)

Rega online tool always generates files with the same name **results**, e.g. `results.csv`.
Rename these files accordingly, 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
```

Refer to Supplementary part of the guide, if you need more information (Page 23).

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
```

In the absence of the ENV-related files, also use the parameter **--noenv**

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

Check the output of `12_mafft` folder before you run iqtree analysis (msa files - multiple sequence alignments)!

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 output being produced or not added at all (no `13_iqtree` folder then).

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

You can monitor the `.log` file while running iqtree within `work` folder using respective process ID (68), e.g [68/72f0eb].

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.

Clean Up

Once the pipeline has generated `Results` folder with all desired output files (**save** all needed outputs first), the input files can be removed from the input folders.

```
$ rm -rf InputFasta/* AllSeqsCO20/* ManualRega/* Results/
```

The same is true for dot and nextflow temp files/folders:

```
$ rm -rf .nextflow work .nextflow.log .nextflow.log.*
```

Processes Overview

```
[1b/f2f10a] process > mark_fasta (2)          [100%] 3 of 3, cached: 3 ✓
[a9/dd644a] process > get_tags (3)            [100%] 3 of 3, cached: 3 ✓
[97/70bbdd] process > comet (3)               [100%] 3 of 3, cached: 3 ✓
[73/a28f41] process > stanford (3)            [100%] 2 of 2, cached: 2 ✓
[e6/e4af1d] process > json_to_csv (3)         [100%] 2 of 2, cached: 2 ✓
[65/e0eb90] process > clean_rega (3)          [100%] 3 of 3, cached: 3 ✓
[87/4d2fcf] process > join_env (1)            [100%] 1 of 1, cached: 1 ✓
[34/36991e] process > join_int (1)            [100%] 1 of 1, cached: 1 ✓
[62/e59285] process > join_prrt (1)           [100%] 1 of 1, cached: 1 ✓
[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 ✓
[54/89322b] process > env_concat_panel (1)    [100%] 1 of 1  cached: 1 ✓
[a4/b7aaee] process > int_concat_panel (1)    [100%] 1 of 1, cached: 1 ✓
[f7/9e1ccf] process > prrt_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 .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...

Example of .fasta within InputFasta

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

Example of .csv within ManualRega

An example of a `.csv` file produced by Rega online tool (names of columns and only one sample for demonstration)

```
"name","length","assignment","rule","support","begin","end","type","pure",
"pure_support","pure_inner","pure_outer","scan_best_support","scan_assigned_support",
"scan_assigned_nosupport","scan_best_profile","scan_assigned_profile","crf",
"crf_support","crf_inner","crf_outer","crfscan_best_support",
"crfscan_assigned_support","crfscan_assigned_nosupport","crfscan_best_profile",
"crfscan_assigned_profile","major_id","minor_id"
"20-02944_PRRT_20","1026.0","HIV-1 CRF 06_CPX","4","98.0","1823.0","2848.0","Human
immunodeficiency virus 1","HIV-1 Subtype G","93.0","0.0","93.0","0.5","0.357","0.643",
"G K A1 A1 A1 A1 G A1 A1 G G G G G","G - - - - - - - - G G G G -","HIV-1 CRF 06_CPX",
"98.0","0.0","98.0","1.0","1.0","0.0","06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX
06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX","06_CPX 06_CPX 06_CPX 06_CPX
06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX 06_CPX","",""
```

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.git
```

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

```
projectDir = "/home/beast2/rki_subtyping/Pipeline"
```

Notes

Keep in Mind (1)

- The pipeline does not take into account subsubtypes. If there are subsubtypes they are converted to subtypes. For instance, A1 is converted to A, F2 is converted to F etc.
- The pipeline does not perform a full quality check of `.fasta` sequences. Illegal characters should be excluded (the pipeline takes care only of underscores so far). Sequences with illegal characters are not accepted by Rega online tool.

Keep in Mind (2)

- Make sure that sample names do not exceed 30 characters in length. Long sample names get shorten by Rega online tool that can cause issues. E.g, this sequence name **PK105_F482_23_MiS84_S86_20consensus_PRRT_20** is too long and gets shortern by Rega to **PK105_F482_23_MiS84_S86_20cons**. In such cases a manual change is necessary.

Keep in Mind (3)

- Be sure you are connected to the Internet
- You can always delete the whole `Results` folder or individual subfolder/subfolders within `Results` and repeat the commnand with `-resume`.

Pipeline Updates

- Created: September 2022
- Updated: January 2023
  A parameter **--noenv** has been added; should be used for the cases when ENV fragment is not sequenced. This info is reflected in the final report table.
- Updated: February 2023
  Slurm congif file added. Sierrapy client -> v.0.4.1
- Updated: March 2023
  Sierrapy client -> v.0.4.2 (--no-sharding)