Snakemake pipeline PanTools v3

This document guides you through the required steps to reproduce the five pangenomic analyses and two scalabilty use cases of the publication *"PanTools v3: functional annotation, classification, and phylogenomics"*.

Step 1. Download the input data

Download and uncompress the input data from the public repository at https://doi.org/10.4121/19874485. All required files are compressed into one zip file of ~4.5Gb, containing the Snakemake pipeline plus genome and annotation data.

```
unzip pantools_v3_usecase_data.zip
cd pantools_v3_usecase_data
unzip a_thaliana/\*.zip -d a_thaliana/
unzip drosophila/\*.zip -d drosophila/
unzip s_cerevisiae/\*.zip -d s_cerevisiae/
unzip pectobacterium/\*.zip -d pectobacterium/
unzip sars_cov2/\*.zip -d sars_cov2/
```

The datasets required for the **scalability** use cases must still be downloaded due to the large genome sizes.

```
cd tomato
sh download_tomato_genomes.sh
cd ../..
```

```
cd human
sh download_human_genomes.sh
cd ../..
```

Step 2. Install Conda

If you don't have Anaconda or Miniconda, follow the download and install the instructions on: https://docs.conda.io/projects/conda/en/latest/user-guide/install/download.html.

After the installation, also install Mamba into the Conda base environment to enable much faster dependency solving

```
$ conda install mamba -n base -c conda-forge
```

Step 3. Create a Snakemake Conda environment

Create a Conda environment for Snakemake

```
$ mamba create -c conda-forge -c bioconda -n snakemake snakemake
```

Activate the environment

```
$ conda activate snakemake
$ conda config --set channel_priority strict
```

Step 4. Download PanTools

Clone the most recent version of PanTools from Gitlab and checkout to the v3.4 release branch. Make sure to clone the repository inside the `pantools_v3_usecase_data` directory as the Snakemake pipeline assumes PanTools is located here.

```
$ cd pantools_v3_usecase_data
$ git clone https://git.wur.nl/bioinformatics/pantools.git
$ cd pantools
$ git checkout pantools_v3.4
```

Optional: Download the InterPro database in the `pantools/addons/` directory

```
$ cd addons/
$ wget https://ftp.ebi.ac.uk/pub/databases/interpro/interpro.xml.gz
$ gzip -d interpro.xml.gz
```

Step 5. Run the pangenome analysis

Go back to the `pantools_v3_usecase_data` directory, open 'Snakefile' and uncomment the config file (by removing '#') for the desired analysis.

Run the snakemake pipeline using the command below. As the pangenome construction and analysis can take a considerable amount of time, please run the pipeline within a screen. The first time the pipeline is executed, all dependencies are installed automatically on the location specified by the `--conda-prefix` argument. Use the same path when running the pipeline another time to avoid creating a new conda environment.

Certain analyses can be disabled by changing the 'True' statement into 'False' in the config files.

```
$ cd pantools_v3_usecase_data
$ snakemake -rp --cores 27 --use-conda --conda-prefix PT_dependencies/
```

The pipeline executes the pangenome analysis in a directory called `output`. A log file of every step in the pipeline is written to the `output/logs` subdirectory. To run the pipeline with another dataset, first move or remove the previous `output` directory.

Snakemake does not report intermediate progress of PanTools's functionalities but only the finished steps. As the construction runtimes for the human and tomato scalability use cases are considerably longer, open the log file to check the current status.

Step 6. Visualize phylogeny in iTOL

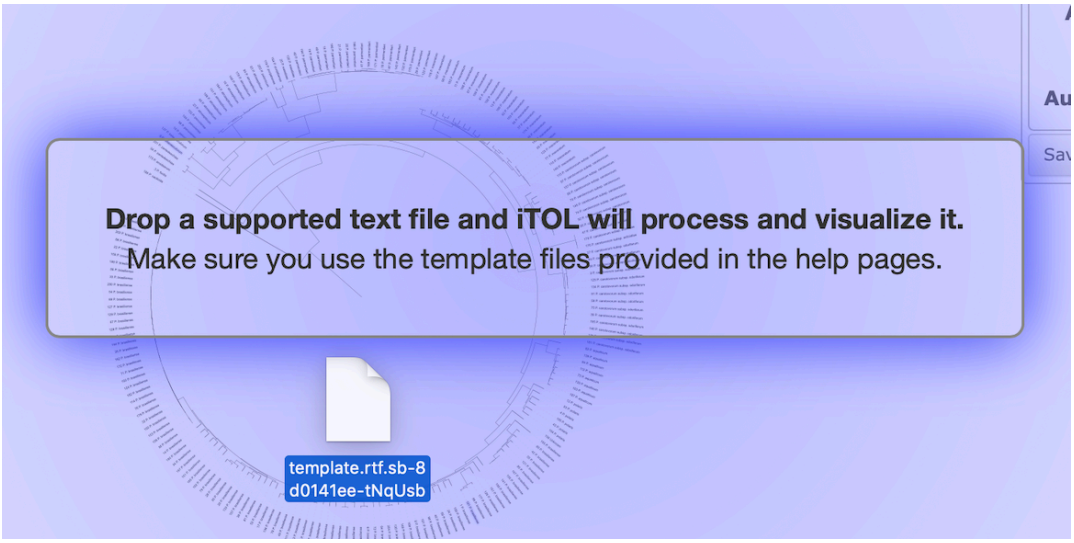
In case you want to visualize the phylogenies similar to our paper; Upload a phylogenetic tree to iTOL (https://itol.embl.de) and include a template file by draggging it into the tree visualization webpage. All phylogenies have an additional renamed version with strain or accession information incorporated. These updated trees are recognized by 'RENAMED' in the filename.

Locations of phylogenetic trees:

- Core SNP tree - core_snp_tree/informative_RENAMED.treefile
- K-mer distance tree - kmer_classification/genome_kmer_distance_tree_RENAMED.tree
- Gene distance tree - gene_classification/gene_distance_tree_RENAMED.tree
- MLSA - mlsa/output/mlsa.fasta_RENAMED.treefile
- ANI distance tree - ANI/MASH/ani_RENAMED.newick

Color the tree using a ***label*** or ***ring*** template file located in `output/pangenome_DB/tree_templates/`.

- *A. thaliana* - feature2.txt
- *S. cervisieae* - feature2.txt
- *Pectobacterium* - species.txt
- SARS-CoV-2 - feature1.txt



Whenever you are done, close the Conda environment.

```
$ conda deactivate
```