Pipeline Documentation

Dansk version

Hej!

Her finder du dokumentation for mit arbejde ved Retsgenetisk Afdeling, Københavns Universitet.

Alle relevante filer er placeret i min brugermappe under følgende sti:
`/mnt/ngs/projects/wgs_pipeline_testing_6180/users/wmj412/20250523_summary_script_pipeline`

Der er to hovedmapper:

1) summary_python_script

Indeholder et Python-script til opsummering af resultater fra WGS-pipeline-output.

Sådan bruges scriptet:

```
python extract_statistics.py /path/to/all/result/directories/from/analysis/
    sampleIDs_list.txt > summary_of_experiment.tsv
```

Forklaring:

- `extract_statistics.py` er scriptet, der ekstraherer statistikker.
- `/path/to/all/result/directories/from/analysis/` er stien til mappen med output fra alle samples (mapper skal være navngivet efter sampleID).
- `sampleIDs_list.txt` er en tekstfil med en liste over de sampleID'er, der skal behandles.

Der findes eksempler i de to medfølgende filer i samme mappe som scriptet.

2) snakemake_wo_WGS_pipeline & snakemake_w_WGS_pipeline

Indeholder Snakemake-scripts til at køre valgfri sekundæranalyser – enten uden eller med WGS-basis-pipelinen.

Der findes to versioner:

- Én til en komplet pipeline fra rå sekventeringsdata.
- Én til kun sekundæranalyser, hvis den primære pipeline allerede er kørt.

(`extract_statistics.py` er ikke inkluderet her, da det bør integreres direkte i pipelinen.)

Indhold:

- `Snakefile`: hovedpipeline med alle tilknyttede sekundæranalyser.
- `config.yaml`: konfigurationsfil til både den primære pipeline og aktivering/deaktivering af sekundæranalyser.

Sådan køres Snakemake-pipelinen:

1. Åbn en terminal.
2. Navigér til den ønskede arbejdsmappe.
3. Kopiér `Snakefile` og `config.yaml` (fra den ønskede version) til denne mappe.
4. Redigér `config.yaml`: opdater filnavne, stier og angiv hvilke steps der skal aktiveres.
5. Kør:

```
snakemake --cores 4
```

Angiv antal cores efter behov.

Både `Snakefile` og `config.yaml` indeholder kommentarer, som kan være nyttige ved videreudvikling.

Ekstra note: Det kan være relevant at se nærmere på *Haplocheck* og *Haplocart* til estimering af mtDNA-haplogrupper.

English version

Hi there!

This document contains an overview of my work at the Department of Forensic Medicine, University of Copenhagen.

All relevant files can be found in my user directory at:
`/mnt/ngs/projects/wgs_pipeline_testing_6180/users/wmj412/20250523_summary_script_pipeline`

There are two main folders:

1) summary_python_script

Contains a Python script that summarizes results from WGS pipeline output.

Usage:

```
python extract_statistics.py /path/to/all/result/directories/from/analysis/
    sampleIDs_list.txt > summary_of_experiment.tsv
```

Explanation:

- `extract_statistics.py` is the script that extracts statistics.
- `/path/to/all/result/directories/from/analysis/` is the directory containing output folders for each sample (named by sampleID).
- `sampleIDs_list.txt` is a text file with a list of the sample IDs to be processed.

Example input files are provided in the same directory as the script.

2) snakemake_wo_WGS_pipeline & snakemake_w_WGS_pipeline

These folders contain Snakemake scripts for running optional secondary analyses, either without or with the full WGS base pipeline.

There are two versions:

- One for a full pipeline from raw sequencing data.
- One for secondary analyses only, if the primary pipeline has already been executed.

(`extract_statistics.py` is not included here, as it should be integrated into the pipeline itself.)

Contents:

- `Snakefile`: the main pipeline including configuration of all connected secondary analysis tools.
- `config.yaml`: configuration file used to define input paths and select which steps to run.

To run the Snakemake pipeline:

1. Open a terminal.
2. Navigate to the working directory of your choice.
3. Copy the appropriate `Snakefile` and `config.yaml` into this folder.
4. Edit `config.yaml`: set file names, paths, and enable/disable specific steps.
5. Run:

```
snakemake --cores 4
```

You may adjust the number of cores as needed.

Both `Snakefile` and `config.yaml` include comments that can be useful for future development.

Additional note: Consider exploring *Haplocheck* and *Haplocart* for mtDNA haplogroup estimation.