ribosomal_snakemake

Release 01/04/2021

LCrossman

Aug 13, 2021

CONTENTS:

CHAPTER

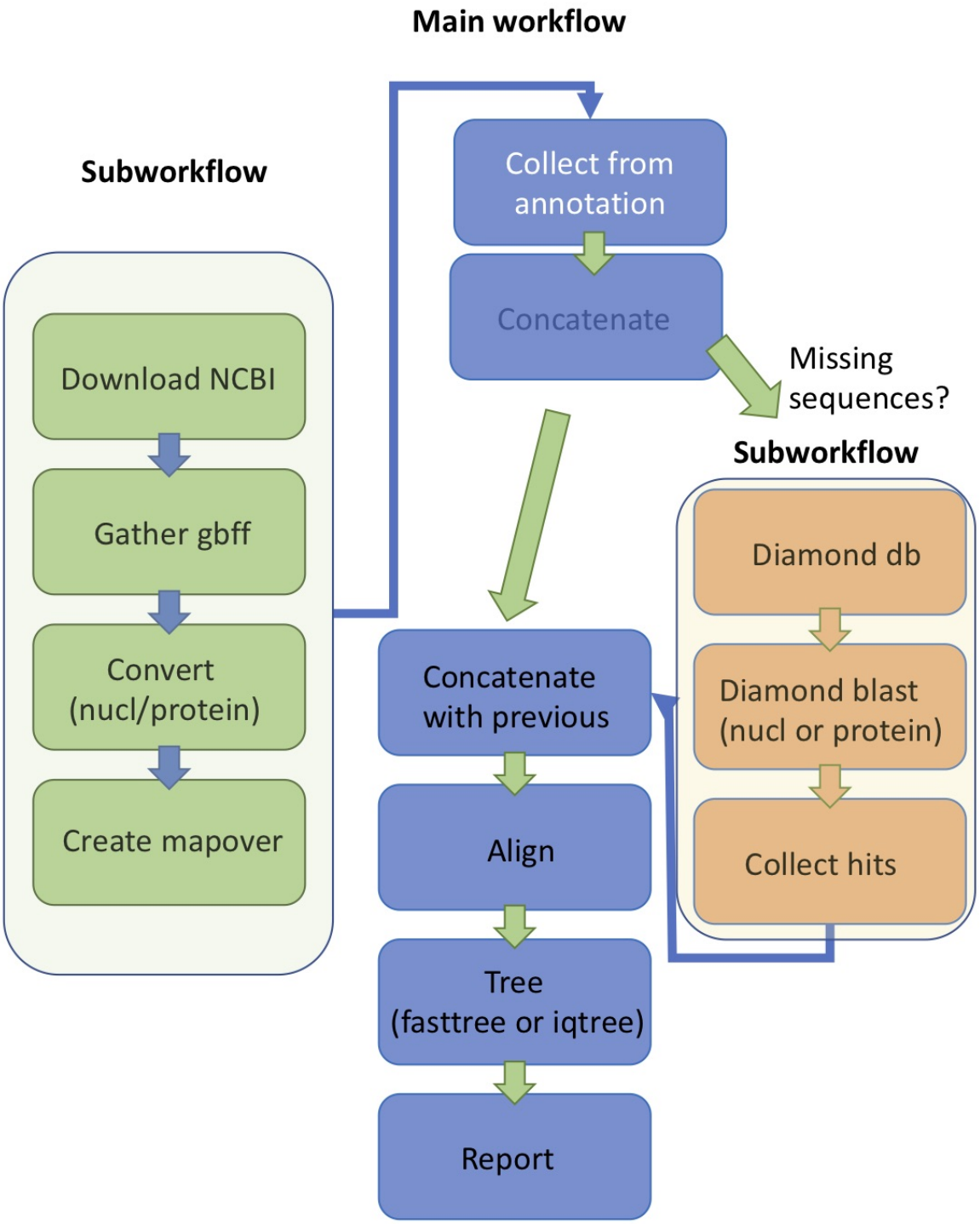
ONE

RIBOSOMAL_SNAKEMAKE

1.1 Ribosomal tree generation for DNA or protein sequences

A workflow to generate ribosomal protein phylogenetic trees, using 15 ribosomal protein sequences, after the tree of Hug *et al*. (2016), "A new view of the tree of life". Nat. Microbiol. 1, 16048. Either protein or DNA sequences can be used in this workflow to build the tree. It may be of benefit to use DNA sequences for more closely related genomes, and protein sequences for those that are more divergent.

The idea behind creating trees from 15 ribosomal sequences with potential in typing, is to guide a high level taxonomic clustering which is available for all life (these sequences are found within eukaryotes, archaea and prokaryotes. Clustering can be carried out on the trees and alignment files using hierBAPS (here adapted for protein sequences: self github) and here for DNA sequences: https://github.com/gtonkinhill/rhierbaps or fastBAPS (https://github.com/gtonkinhill/fastbaps) available for DNA sequences.


Main workflow
Subworkflow
Collect from annotation
Concatenate
Missing sequences?
Download NCBI
Gather gbff
Convert (nucl/protein)
Create mapover
Subworkflow
Diamond db
Diamond blast (nucl or protein)
Collect hits
Concatenate with previous
Align
Tree (fasttree or iqtree)
Report

CHAPTER

TWO

GETTING STARTED

2.1 Requirements

snakemake>=3.5
python>=3.3
biopython>=1.77
diamond BLAST version>=0.9.32
mafft>=7.429
fasttree>=2.1.10 and/or:
iqtree>=1.6.11
ncbi-genome-download>=0.3.0

Python scripts from this github folder

2.2 Installation

Use conda to install snakemake to a Linux or Mac OSX environment:
To install conda:

Install conda here https://conda.io/en/latest/miniconda.html
Install dependencies *via* conda

Install snakemake:

conda install -c conda-forge mamba
mamba create -c conda-forge -c bioconda -n snakemake snakemake

conda activate snakemake
snakemake –help

Install the ribosomal protein tree workflow from github:
git clone XXX

cd XXX

CHAPTER

THREE

QUICKSTART

Execute workflow: Test your configuration by performing a dry-run *via* snakemake -n

Execute the workflow locally *via*

- snakemake –cores num_cores

where num_cores are the number of available cores on your machine *e.g.* 4

or run it in a cluster environment such as snakemake –use-conda –cluster qsub –jobs 100

Further information on running snakemake in different cluster environments can be found on the snakemake website https://snakemake.readthedocs.io/en/stable/

CHAPTER

FOUR

USAGE

4.1 Configure workflow

1. Configure the workflow according to your needs by editing the file config.yaml. For use without any sequence database downloads, you need to provide the files and pathnames in cleannames.txt and atccs.txt.
2. Gather your ribosomal protein sequences as protein sequence fasta files in 15 separate files and place in folder named "seqs". These will be used for either protein or DNA sequence trees as per your choice laid out in the config.yaml. They should be named within the file as L14_rplN, L16_rplP, L18_rplR, L2_rplB, L22_rplV, L24_rplX, L3_rplC, L4_rplD, L5_rplE, L6_rplF, S10_rpsJ, S17_rpsQ, S19_rpsS, S3_rpsC and S8_rpsH, respectively. *Staphylococcus* sequences are included in the github folder and for examples. Other sets of sequences are coming soon!
3. Gather all of your genome sequences as annotated genbank files in the same directory as you are working in.
4. Create a file, "genus.txt" containing the name of the genus and species you want to build a tree for, one per line. Note that the genus or species name **must** be enclosed with quotes.
5. If you do not want to download any genomes from the sequence databases, create a file, "cleannames.txt" containing the name of each genome file you want to include in your tree, one per line, and place "no" in the config.yaml under download_genbank options. To use a mixture of download and provided genomes place "yes" in the config.yaml and do **not** provide a cleannames.txt file.
6. To update any previous trees, collect any files generated as concatenated deduplicated fasta files that you want to update, and edit the config.yaml file with the filenames as appropriate. NOTE: If you are not updating any previous trees, you will need to include an empty file named "previous.fasta" or filename as specified in the config.yaml
7. Add Genbank files, ribosomal protein sequence files and a list of the files: If you do not want to download any genomes, add the names of all provided genbank format files to cleannames.txt on a one-line per file basis. The suffix of the genbank files should end in .gbff (following the genbank download nomenclature), .gbk, .gb or .genbank. Add the names of your ribosomal sequence files on a one-line per file basis to atccs.txt. An easy way to generate these text files in unix is with: ls *gbff > cleannames.txt

- Example files are contained in example_files folder with files for a testrun in the testfiles folder

CHAPTER

FIVE

TUTORIAL

1. Edit config.yaml to include the following:

threads: 4

download_genbank: options: “no”

tree_type: options: “fasttree”

protein_dna: options: “protein”

report: “report.html”

previous_files: “previous.fasta”

2. Create genus.txt to include only the following (do not omit the quotes):

“Staphylococcus argenteus”

3. Assure all the files are available in the correct folder and include an empty file named “previous.fasta”
4. run:

   snakemake –cores 4 (if 4 cores, 8 threads, are available)
5. Have a hot drink and wait for the results!

CHAPTER

SIX

INDICES AND TABLES

- genindex
- modindex
- search