Test pipeline. 18S sequences

Pipeline contains **11** steps, each of which corresponds to a code in the files. The code from the first **2** steps must be copied into the terminal, the steps **3-11** are started with the ‘python test_pipeline_step**№**.py’ command.

The working directory should contain a file with reference ‘18S_reference.fasta’, a folder with the original transcriptome assemblies ‘./fastas’, files with 11 steps to play them. In the folder ‘example’ you can see the files that appeared as a result of performing all steps. The final file is called ‘final_file.fasta’.

The 11 sequential steps are listed below:

1. Making BLAST databases (“test_pipeline_step1.txt”)

2. Reference alignment (“test_pipeline_step2.txt”)

3. Taking 5 best hits by e-value with Python code (“test_pipeline_step3.py”)

4. Create fasta files with 5 best sequences (for further search in case the 1st sequence does not fit) (“test_pipeline_step4.py”)

5. Taking the 1st best sequence from a fasta with the top 5 (“test_pipeline_step5.py”)

6. Variant with mafft+trimal + Blast check identity + PairwiseAligner alignment (from Bio.Align) for a final identity check of at least 90%. (“test_pipeline_step6.py”)

7. Local BLAST identity check (“test_pipeline_step7.py”)

8. Make union file with seqs that passed checks (“test_pipeline_step8.py”)

9. Final identity check with global PairwiseAligner (“test_pipeline_step9.py”)

10. Make union file with seqs that passed **all** checks ("test_pipeline_step10.py")

11. Totals. UGENE variant. Identity check with reference (with global PairwiseAligner) ("test_pipeline_step11.py")

A total of 25 sequences for rnaspades and 18 trinity were checked, after removing sequences with missing coordinates and selecting the best sequence for those that passed in both assemblies, the final file contained 26 sequences. The headers and identifiers in the final fasta files were moved to match the headers of the geographic coordinates table ("./geographic_coordinates")

It was found that many sequences are similar to the reference and when blasted at NCBI many of them gave a better hit with the reference sequence.

It was therefore decided to perform a multiple contig alignment with the best hits for these 26 sequences and the reference, with trimming at the boundaries of the reference using UGENE.

Comparison table of identity to the reference (result files "union_nogaps.fasta" and "18S_ugene.fasta" are in "./final_seqs" directory):

Sample	Identity_ugene	Identity_nogaps
SRR3467039_Oxyacanthus_sowinskii_18S	99.5	99.5
SRR3467090_Pentagonurus_dawydowi_18S	99.5	99.5
SRR3467037_Oxyacanthus_curtus_18S	99.5	99.5
SRR3467095_Pallaseopsis_kessleri_18S	99.4	99.4
SRR3467063_Eulimnogammarus_cruentus_18S	99.4	99.5
SRR3467052_Sluginella_kietlinskii_18S	99.0	99.0
SRR3467065_Eulimnogammarus_messerschmidtii_18S	98.9	98.9
SRR3467066_Eulimnogammarus_maackii_18S	99.3	99.3
SRR3467055_Eulimnogammarus_violaceus_18S	99.3	99.4

SRR3467067_Eulimnogammarus_sp._18S	99.3	99.3
SRR3467057_Eulimnogammarus_cyaneus_18S	99.3	99.3
SRR3467086_Ommatogammarus_flavus_18S	99.2	99.2
SRR3467048_Carinurus_bicarinatus_18S	99.5	99.5
SRR3467081_Micruropus_parvulus_18S	97.8	97.8
SRR3467077_Macrohectopus_branickii_18S	97.8	97.8
SRR3467083_Micruropus_wahlii_18S	97.5	97.5
SRR3467071_Gmelinoides_fasciatus_18S	97.3	97.4
SRR3467068_Eulimnogammarus_verrucosus_18S	99.0	99.1
SRR3467075_Hyalellopsis_setosa_18S	91.7	91.7
SRR3467073_Hyalellopsis_costata_18S	91.1	91.6
SRR3467089_Homalogammarus_brandtii_18S	99.1	99.2
SRR3467043_Boeckaxelia_carpenterii_18S	99.3	99.3
SRR3467072_Hyalellopsis_carinata_18S	99.1	99.1
SRX1736878_Gammarus_lacustris_18S	97.7	97.8
SRR3467059_Eulimnogammarus_testaceus_18S	94.8	97.0
SRR3467047_Asprogammarus_rhodophthalmus_18S	79.6	94.3

% of identity to the reference was calculated using the code (“test_pipeline_step11.py”)