Setting up a portable metagenome assembly pipeline for CBIO - Bug #46

Read binning with CONCOCT

04/30/2018 11:42 AM - Katie Lennard

Status:	New	**Start date:**	12/04/2017
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour

Description

CONCOCT “bins” metagenomic contigs. Metagenomic binning is the process of clustering sequences into clusters corresponding to operational taxonomic units of some level.
CONCOCT is a whole pipeline in itself but Ulas uses only the binning part of CONCOCT.

- Before using concoct, the input files (QCed read files from trimmomatic) and contigs file (named final.contigs.fa from megahit) need to be prepared for concoct as follows:
- The contigs file needs to be Indexed, using the 'bowtie2-build' command (produces a number of files with same name as input file but ending in extension .1 .2 .3 etc)
- Individual trimmed read files need to be aligned to the indexed contigs file, using the 'bowtie2' command (script 'prepare_for_concoct.single.sh' and prepare_for_concoct.batch.sh) - can be run with batch script on hex
- Index the original contigs file with samtools faidx command (this is needed to get file in the rigth format for the next step)
- Then we can use the 'samtools view -bt' command to convert the output from 2. from sam to bam format
- Sort bam file from 4. with 'samtools sort' so that reads occur in genome order
- Index output from 5.
- Locate, tag and removes duplicate reads from 6. (need MarkDuplicates.jar for this:download from here https://repo.jbei.org/users/mwornow/repos/seqvalidation/browse/tools/Picard-NERSC_version/MarkDuplicates.jar?at=master)
- After removing duplicates sort output from 7. (again using 'samtools sort')
- Index output from 8. ('samtools index')
- Compute coverage profile of 9. using 'genomeCoverageBed' from bedtools2]
- The next step is to take the coverage profiles (.smds.coverage files) from 10. and create a coverage table for input with concoct (python /opt/exp_soft/CONCOCT-0.4.0/scripts/gen_input_table.py)
- Now we can run concoct (with coverage table and contigs file as input) > module load python/anaconda-python-2.7 > source activate concoct_env > /opt/exp_soft/CONCOCT-0.4.0/bin/concoct > For concoct you need to specify a) the max number of clusters (default=400) b) number of cores to use (according to this https://bitbucket.org/berkeleylab/metabat/wiki/ concoct uses 10 threads regardless of the number specified..so set to 10 currently: ppn=10) > The following warning/errors were produced but didn't seem to affect output: /opt/exp_soft/anaconda/python2.7/envs/concoct_env/lib/python2.7/site-packages/Bio/Seq.py:341: BiopythonDeprecationWarning: This method is obsolete; please use str(my_seq) instead of my_seq.tostring(). BiopythonDeprecationWarning) python: symbol lookup error: /opt/exp_soft/anaconda/python2.7/lib/python2.7/site-packages/numexpr/../../../libmkl_vml_def.so: undefined symbol: mkl_serv_getenv python: symbol lookup error: /opt/exp_soft/anaconda/python2.7/lib/python2.7/site-packages/numexpr/../../../libmkl_vml_def.so: undefined symbol: mkl_serv_getenv python: symbol lookup error: /opt/exp_soft/anaconda/python2.7/lib/python2.7/site-packages/numexpr/../../../libmkl_vml_def.so: undefined symbol: mkl_serv_getenv python: symbol lookup error: /opt/exp_soft/anaconda/python2.7/lib/python2.7/site-packages/numexpr/../../../libmkl_vml_def.so: undefined symbol: mkl_serv_getenv >The final step in the binning process is to visually evaluate the output using the R script ClusterPlot_KL.R which produces a sort of color coded PCA of the clusters NB: concoct documentation recommends splitting larger contigs before running concoct so as to give more weight to larger contigs (I have not tested this yet)

History

#1 - 04/30/2018 04:00 PM - Katie Lennard

- Description updated

- Start date changed from 04/30/2018 to 12/04/2017

- % Done changed from 0 to 100