**ycConsensus.sh**

**Author(s)**

Angela Crabtree

**Dependencies**

Java, BWA, Samtools, Python3, BCFtools, Muscle

**Description**

ycConsensus.sh allows the user to align raw reads to a reference sequence and returns a lot of great alignment information including a consensus sequence and a graph of read depth across the reference sequence.

**Usage (MacOS)**

sh ycConsensus.sh -s sample\_name -1 for\_read.fastq.gz -2 rev\_read.fastq.gz -r reference\_genome.fasta -m

**Options**

|  |  |  |
| --- | --- | --- |
| -s | [ARG] | (required) sample name |
| -1 | [ARG] | (required) forward read file |
| -2 | [ARG] | (required) reverse read file |
| -r | [ARG] | (required) reference genome file |
| -o | [ARG] | (required) destination of output folder |
| -m |  | operating system is MacOS |
| -w |  | operating system is Windows, using Ubuntu |
| -x |  | operating on an external server, such as the IBEST core cluster |
| -t |  | tests program with test files (ensures required CL apps are working) |
| -h |  | help (print options) |

**Input**

Sample name (-s) – The strain name of your sample, preferably without spaces in it

Forward read file (-1) – File contains raw NGS reads in forward orientation; “R1” should be in the filename somewhere (ex: “NCYC190\_S3\_L001\_R1\_001.fastq.gz”); can be .fastq or .fastq.gz

Reverse read file (-2) – File contains raw NGS reads in reverse orientation; “R2” should be in the filename somewhere (ex: “NCYC190\_S3\_L001\_R2\_001.fastq.gz”); can be .fastq or .fastq.gz

Reference genome file (-r) – File contains the nucleotide sequence of your reference; must be in fasta format; you don’t have to worry about creating index files since the script automatically creates them

Destination folder (-o) – The script creates a folder called ‘output” within this folder where your results will be stored

**Output**

|  |  |
| --- | --- |
| SAMPLE\_REF\_aligned.fasta | A fasta-formatted pairwise alignment using MUSCLE of the sample and the reference. |
| SAMPLE\_REF\_consensus\_trunc.fasta | A consensus sequence that has been truncated on the 5' and 3' ends where there is 0 coverage. CAUTION: my python script does not cut out areas of zero coverage in the middle of the sequence. |
| SAMPLE\_REF\_muscle\_input.fasta | This is a fasta-formatted file containing the truncated sample consensus sequence and the reference sequence (before alignment) |
| SAMPLE\_REF\_read\_depth.jpeg | A graph of sample read depth across the reference genome |
| SAMPLE\_REF\_read\_depth.txt | A tab-delimited list containing the read depth at each nucleotide position (if >0 read depth) |
| SAMPLE\_REF\_sorted.bam | A file that contains information on where each read is aligned along the reference genome, sorted from start of reference genome to end of reference genome, in a binary format (not human readable) |
| SAMPLE\_REF\_sorted.bam.bai | An index file for the corresponding bam file; this makes it easier for some programs to utilize the information in the bam file |
| SAMPLE\_REF.bam | A file that contains information on where each read is aligned along the reference genome, in a binary format (not human readable) |
| SAMPLE\_REF.sam | A file that contains information on where each read is aligned along the reference genome, in a human-readable format |
| troubleshooting | A folder for storing troubleshooting information. This will have several text files containing output from different command line apps. If there are errors or the output is messed up, try browsing these files. |

**Examples**

*# view all the options (“-h”)*

sh ycConsensus.sh -h

*# makes sure program and dependencies are working right (“-t”)*

sh ycConsensus.sh -t -m

*# minimum input (make sure you use full pathnames when you do this)*

sh ycConsensus.sh -s NCYC190 -1 NCYC190\_S3\_L001\_R1\_001.fastq.gz -2 NCYC190\_S3\_L001\_R2\_001.fastq.gz -r M1.fasta -m

*# use ‘-w’ option if you’re using windows*

sh ycConsensus.sh -s NCYC190 -1 NCYC190\_S3\_L001\_R1\_001.fastq.gz -2 NCYC190\_S3\_L001\_R2\_001.fastq.gz -r M1.fasta -w