**ycDeNovo.sh**

**Author(s)**

Angela Crabtree

**Dependencies**

Java, BBMap, SPAdes, R, fastp, Python

**Description**

ycDeNovo.sh allows the user to clean and filter raw Illumina reads and assemble de novo contigs. The script returns a graph of contig length and contig “coverage” as calculated by SPAdes assembler as well as filtered read files. It also prints number of raw reads, number of final cleaned & filtered reads, % of reads mapping to the human genome, % of reads mapping to S. cerevisiae, and % of reads mapping to S. paradoxus. Note that this script assumes you are working on an external server so it will need to be adapted for use locally. The paths to BBSplit and fastp must be either edited or moved so that they are “~/bin/bbmap/bbsplit.sh” and “~/bin/fastp/fastp”, respectively.

**Usage (MacOS)**

sh ycDeNovo.sh -s sample\_name -1 for\_read.fastq.gz -2 rev\_read.fastq.gz -o output\_folder

**Options**

|  |  |  |
| --- | --- | --- |
| -s | [ARG] | (required) sample name |
| -1 | [ARG] | (required) forward read file |
| -2 | [ARG] | (required) reverse read file |
| -o | [ARG] | (required) destination of output folder |
| -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

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

**Output**

|  |  |
| --- | --- |
| SAMPLE\_R1\_filt.fastq.gz | A fastq file with reads in forward orientation “R1” that remained after filtering with fastp |
| SAMPLE\_R1\_filt.fastq.gz | A fastq file with reads in reverse orientation “R2” that remained after filtering with fastp |
| SAMPLE\_contigs.fasta | This is a fasta-formatted file containing the de novo assembled contigs generated by SPAdes assembler. |
| SAMPLE\_contigs.jpeg | A graph of contig length and contig “coverage” as calculated by SPAdes assembler |
| clean1.fastq.gz | A fastq file with reads in forward orientation “R1” that remained after filtering with fastp, and filtering out all designated contaminants (in this case, human, S. cerevisiae, and S. paradoxus). |
| clean2.fastq.gz | A fastq file with reads in reverse orientation “R2” that remained after filtering with fastp, and filtering out all designated contaminants (in this case, human, S. cerevisiae, and S. paradoxus). |
| dirty\_reads\_Hsap.fastq.gz | A fastq file containing all reads (forward and reverse) mapping to the human genome. |
| dirty\_reads\_Scer.fastq.gz | A fastq file containing all reads (forward and reverse) mapping to the S. cerevisiae genome. |
| dirty\_reads\_Spara.fastq.gz | A fastq file containing all reads (forward and reverse) mapping to the S. paradoxus genome. |
| SAMPLE\_spades\_output | A folder containing all the SPAdes output. The contigs file is copied into the main “denovo” folder as “SAMPLE\_contigs.fasta” so you don’t really need to open this spades folder unless you’re doing some weird behind-the-scenes stuff. |
| stats | A folder containing SAMPLE\_fastp.html, a fastp report that is opened with a web browser and contains a lot of nice information about the read quality before and after filtering with fastp. |

**Examples**

*# view all the options (“-h”)*

sh ycDeNovo.sh -h

*# 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 -o ~/Downloads