

QCB Workshop W5a Quiz

Arjun Bhattacharya

11/18/2021

This quiz is due February 28, 2021.

Feel free to use any materials you like to help you answer these questions. Email your responses to me at abtbhatt@g.ucla.edu. There are 20 total points. There are bonus points along the way too, but you are capped at a total of 20 points.

1. (1 point) There are two files at this Google Drive link: <https://drive.google.com/drive/u/1/folders/1HZUWZBzr9pCUNXFzJcNMCjmvOkIZ2VHw>. They are named `KD2.fastq.gz` and `SC2.fastq.gz`. Download them to your local computer and upload them to `hoffman2`. What software did you use to upload them to `hoffman2`? For 1 bonus point, what kind of client did you use (four words)?
2. (3 points) Using UNIX commands, how many lines are in `KD2.fastq.gz`? How many reads does this correspond to? What commands did you use? For this last question, there is no one right answer.
3. (3 points) Run quality evaluation on both `fastq.gz` files using `FastQC`. What kind of files does `FastQC` output? What commands did you run? What is the sequence length for `SC2.fastq.gz`, according to the `FastQC` report?
4. (4 points) Using `trimmomatic`, trim both `FastQC` files with the following parameters: minimum leading and trailing base qualities of 15 and a minimum final read length of 60 basepairs. How many reads were trimmed from both `fastq` files? For 1 bonus point, what command(s) did you use?
5. (2 points) You already have a reference genome for Chromosome 18 built. What are the two types of files that are required to build a reference genome sequence database using `STAR`?
6. (3 points) Align the cleaned and trimmed reads for both samples to this reference genome. What commands did you use? What proportion of reads mapped unambiguously to the genome database?
7. (3 points) Count the number of reads that map to each gene using `htseq-count`. At the end of the run, there is some information that the program outputs to the terminal. What commands did you use? For each sample, how many alignments were processed? For 1 bonus point, `HTSeq` is a software package written in what language?
8. (1 point) What R package can you use to read in the counts from `htseq-count` output?