***Submit a zip folder having all the files as Solution_LastName_FirstName_HW1.zip to assemblyteam1submission@gmail.com**

****Contact assemblyteam1submission@gmail.com for any issues with the Subject Line: HW1_GA_Query**

Klebsiella spp (bonus: 10 points).

https://emedicine.medscape.com/article/219907-overview#a4

1. Genus of *Klebsiella*
2. Gram (+/-)
3. Explain the different mechanisms of developing antibiotic resistance

Additional Reading:
Comparative Genomics of *Klebsiella*: http://aac.asm.org/content/55/9/4267.full

T1G1_Genome Assembly HomeWork (70 Points)

Step by step Sample questions

(1) Retrieve raw data (10 Points)

- a. Download SRR3982229 with SRA Toolkit (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/)
- b. Convert the file into FASTQ file with *fastq-dump*.
- c. Split the data.

(2) Check quality of raw data (10 Points)
Install FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc).
Perform quality examination on the FASTQ file you gained from problem

- a. What is %GC?
- b. Briefly interpret graphs gained from FASTQC. Save short answer as ***rawqc.txt***
- c. Attach FastQC report for Forward and Reverse Reads as **2forward.pdf** and **2reverse.pdf.**

(3) Trimming data (10 Points)
Download Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic). Carefully read through manual.

- a. Trim adapter sequence and perform quality trimming.
- b. Write parameters used in this step with reasoning behind the parameter selection. Also provide the command that was run. Save the short answer as ***trim.txt***.
- c. Check your trimming performance through FASTQC.

- d. Attach the FastQC file as **3forward.pdf** and **3reverse.pdf.**

(4) De Novo assembly (15 Points)

Download SPAdes (http://bioinf.spbau.ru/spades). Carefully read through the instruction from the website.

- a. What trimmed data should we use for the assembly? Write down your reasoning.
- b. Perform assembly on the sample based on the selection you made.
- c. Write down k-mer size selection.

Save the short answer as ***DeNovo.txt***

(5) Reference based assembly (15 Points)

Download and install BWA (https://sourceforge.net/projects/bio-bwa/files/), SAMtools, BCFtools (http://www.htslib.org/download/) and SeqTK (https://github.com/lh3/seqtk).

- a. Download the reference genome of Klebsiella pneumoniae ASM988v1 from NCBI (https://www.ncbi.nlm.nih.gov/assembly/GCF_000009885.1/) .
- b. Using BWA: Index the file containing the reference genome
- c. Using BWA: map reads against reference genome after indexing is done.
- d. Using SAMtools: sort and index obtained BAM file.
- e. Using "samtools mpileup -v" generate VCF format, pipe it to "bcftools call -c" to run calling and generate VCF format, convert it to FASTQ with vcfutils.pl and finally convert FASTQ to FASTA with seqtk (all this processes can be written as one pipeline).

(6) Final QC (10 Points)

Download QUAST (http://bioinf.spbau.ru/quast). Carefully read through the manual from the website.

- a. Input assembled FASTA files and perform quality assessments.
- b. Concepts explanation: number of contigs, N50 (for De Novo), and NG50 (for reference based assembly).
- c. Is there any unaligned contigs? What are those?
- d. Briefly interpret the results you got from the QUAST (Does the result look good and why is that?)

Save short answer as **finalqc.txt**