***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**

*****Deadline for this homework is March 1st, 1:20pm. Yes, right before homework 2 being announced by the gene prediction group.**

Attention please: We installed all software on the server for you guys to make your life easier.

We also include the command in each question to help check whether the path is updated successfully. (I won't tell you they are also some kind of hints for doing the homework)

Add these lines of code to your .bashrc file in your home directory after you ssh into the server (if you don't know what does ".bashrc" mean and how to add lines to it, don't hesitate to ask your classmates for help):

```
export PATH=$PATH:/projects/data/team1_genomeAssembly/local/bin
export PATH=$PATH:/projects/data/team1_genomeAssembly/bin/SPAdes-3.11.1-Linux/bin
export PATH=$PATH:/projects/data/team1_genomeAssembly/bin
export PATH=$PATH:/projects/data/team1_genomeAssembly/bin/FastQC
export
PATH=$PATH:/projects/data/team1_genomeAssembly/bin/lib/python2.7/site-packages/multiqc-1.5dev-py
2.7.egg/multiqc/modules/quast
export PYTHONPATH=/projects/data/team1_genomeAssembly/bin/lib/python2.7/site-packages
```

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)

- (Use command "which prefetch" and "which fastq-dump" to check whether tools are available)
- a. Download SRR3982229 with SRA Toolkit https://www.ncbi.nlm.nih.gov/books/NBK242621/
- b. Convert the file into FASTQ file with *fastq-dump*.
- c. Split the data.
-

(2) Check quality of raw data (10 Points)

(Use command "which fastqc" to check whether the tool is available)
Practice using FASTQC. Perform quality examination on the FASTQ file you gained from problem (Use "fastqc -help" to see the manual is needed).

- 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)

(use command "java -jar trimmomatic-0.36.jar" to check whether trimmomatic is available. And yes, we install version 0.36 instead of 0.35).
Practice using 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)

(use command "spade.py" to check whehter spade is available).
Practice using 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)

(use "which" to check whether tools are available before proceeding. such as "which bwa")
Practice using: 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).
- attach the resulting vcf file as ***result.vcf***

(6) Final QC **(10 Points)**

(command for checking whether the is available: "quast.py")

Practice using 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**