Extracting information from files

For this first part, we will be using PccAS_v3_genome.fa and PccAS_v3.gff3 located under ~/course_data/rna_seq_pathogen/data

1) What command would I use to count the number of sequences in the fasta file
    a. grep ">" PccAS_v3_genome.fa
    b. grep > PccAS_v3_genome.fa | wc -l
    c. grep ">" PccAS_v3_genome.fa | wc -l
    d. grep ">" -c PccAS_v3_genome.fa

2) What command would I use to extract sequences names in file called PccAS_v3_sequences
    a. grep > PccAS_v3_genome.fa > PccAS_v3_sequences
    b. grep ">" PccAS_v3_genome.fa > PccAS_v3_sequences
    c. grep sequence PccAS_v3_genome.fa > PccAS_v3_sequences |wc -l

3) What would be the output of this command
head -n14 ~/course_data/rna_seq_pathogen/data/PccAS_v3.gff3| tail -n 4
    a. From line 1 to line 5 of PccAS_v3.gff3
    b. From line 10 to line 14 of PccAS_v3.gff3
    c. From line 16 to line 20 of PccAS_v3.gff3

4) What would be the result of this command:
grep mrna ~/course_data/rna_seq_pathogen/data/PccAS_v3.gff3

    a. 0 lines
    b. 5177 lines
    c. 24437 lines

5) What option could you add to the previous command to display all lines describing mRNA features
    a. -c
    b. -v
    c. -i

Sed assessment questions:

For the sed portion of the questions, change into the sed_practical working directory created. Do not copy and paste the commands below due to formatting issues with syntax when using MS Word. Rather type the commands out and make use of other unix commands and pipes to view and interact with the output.

6) Type sed on the command line and look at the options provided. What option is required to obtain the sed version number?

a. sed -v
b. sed --v
c. sed -version
d. sed --version

7) What sed version is present in the VM?
   a. 4.4
   b. 4.6
   c. 4.8
   d. 4.9

8) From the sed_practical working directory, what sed command would I use to to convert all the characters in the sequences.fasta file to lower case?
   a. sed 's/[A-Z]/\l&/g' sequences.fasta
   b. sed 's/[A-Z]/\L&/' sequences.fasta
   c. sed 's/[a-z]/\l&/g' sequences.fasta
   d. sed 's/[A-Z]/\L&/g' sequences.fasta

9) What command would I use to print out only even numbered lines from the formatted_exercises.bed file created in the practical?
   a. sed ‘p:n’
   b. sed ‘n:p’
   c. sed -n ‘p:n’
   d. sed -n ‘n:p’

10) Which sed command would I use to print out lines 8 to 20, 25 to 40, 50 to 60, 89, 93 from the formatted_exercises.bed file created in the practical?
   a. sed '8,20p; 25,40p; 50,60p; 89p; 93p' formatted_exercises.bed
   b. sed -n '8,20p; 25,40p; 50,60p; 89p; 93p' formatted_exercises.bed
   c. sed -n '8,20p; 25,40p; 50; 60p; 89p,93p' formatted_exercises.bed
   d. sed -n '8p;20p;25;40p; 50; 60p; 89p,93p' formatted_exercises.bed

11) How many lines are present in the output from the correct command in question 10? (Hint, pipe the output to a unix command wc -l to count the number of lines).
   a. 38
   b. 42
   c. 47
   d. 52

12) In the formatted_exercises.bed file there is a value of scaffold- that was not changed to scaffold_ in the set of sed pipes completed in the practical. We need to recreate the formatted_exercises.bed from the exercises.bed file by adding another pipe to convert scaffold- to scaffold_ and send the output to a file called formatted_exercises.bed while using the global flag through out each of the sed commands in the pipe. Which command below is the correct one?
   a. sed 's/contig-/contig_/g' exercises.bed | sed 's/scaffold-/scaffold_/g' | sed 's/gene-/gene_/g' | sed 's/repeat/REPEAT/g'

b. sed 's/contig-/contig_/g' | sed 's/scaffold-/scaffold_/g' | sed 's/gene-/gene_/g' | sed 's/repeat/REPEAT/' exercises.bed > formatted_exercises.bed
c. sed 's/contig-/contig_/g' exercises.bed | sed 's/scaffold-/scaffold_/g' | sed 's/gene-/gene_/g' | sed 's/repeat/REPEAT/g' > formatted_exercises.bed
d. sed 's/contig-/contig_/' exercises.bed | sed 's/scaffold-/scaffold_/' | sed 's/gene-/gene_/' | sed 's/repeat/REPEAT/' > formatted_exercises.bed

13) For the next three questions, copy the examples.fasta file from the /data_formats/data/ directory into the sed_practical working directory. Examine the file. How many fasta sequences are present in the file examples.fasta?
a. 8
b. 9
c. 10
d. 11

14) The examples.fasta file has two different types of sequences we need to mask using lower case --- so for use in another program. The first set of sequences to replace are at the start of each sequence comprising of CTT. Which command replaces the CTT as the start of each sequence with --- correctly using the global flag?
a. sed 's/CTT/---/g' example.fasta
b. sed 's/CTT^/---/g' example.fasta
c. sed 's/$CTT/---/g' example.fasta
d. sed 's/^CTT/---/g' example.fasta

15) We would like to replace ACC at the end of each sequence in the examples.fasta file with nnn. Which command below is the correct one?
a. sed 's/$ACC/nnn/g' example.fasta
b. sed 's/^CTT/nnn/g' example.fasta
c. sed 's/ACC$/nnn/g' example.fasta
d. sed 's/ACC/nnn/g' example.fasta

AWK assessment questions:

For the awk portion of the questions, change into the awk_practical working directory created. Do not copy and paste the commands below due to formatting issues with syntax when using MS Word. Rather type the commands out and make use of other unix commands and pipes to view and interact with the output. The aw section will make use of the exercises.bed file

16) What is the correct awk command using the -F and tab delimiter to determine how many columns the exercises.bed file has by piping the awk output to the unix sort -u command?
a. awk '{print NF}' exercises.bed
b. awk '{print NF}' exercises.bed | sort -u
c. awk -F "\t" '{print NF}' exercises.bed
d. awk -F "\t" '{print NF}' exercises.bed | sort -u

17) How many columns do you get from the above correct command in the exercise.bed file?
    a. 3 and 4
    b. 4 and 5
    c. 5 and 6
    d. 6 and 7

18) Which awk command prints out columns 1, 4, 5 and 6 in a tab delimited format:
    a. awk -F "\t" {print $1,$4,$5,$6}' exercises.bed
    b. awk 'BEGIN {OFS="\t"} {print $1,$4,$5,$6}' exercises.bed
    c. awk -F "\t" 'BEGIN {OFS="\t"} {print $1,$4,$5,$6}' exercises.bed
    d. awk -F "\t" 'BEGIN {OFS="\t"} {print $1,$3,$5,$6}' exercises.bed

19) Which awk command extracts all repeats that are part of contig-5 from the exercises.bed file?
    a. awk -F"\t" '$1=="contig-5" && $3=="repeat"' exercises.bed
    b. awk -F"\t" '$1=="contig-5" && $4=="repeat"' exercises.bed
    c. awk -F"\t" '$2=="contig-5" && $4=="repeat"' exercises.bed
    d. awk -F"\t" '$2=="contig-5" && $5=="repeat"' exercises.bed

20) Piping the output of the correct command for question 19 to wc -l provides how many lines?
    a. 6
    b. 5
    c. 4
    d. 3

21) Which awk command correctly filters all genes from contig-4 that have a score greater than 50.0?
    a. awk -F"\t" '$1=="contig-4" && $4=="gene" && $5 > 50.0' exercises.bed
    b. awk -F"\t" '$1=="contig-4" && $4~"gene" && $5 > 50.0' exercises.bed
    c. awk -F"\t" '$1=="contig-4" && $4~"gene" || $5 > 50.0' exercises.bed
    d. awk -F"\t" '$1=="contig-4" && $4=+"gene" && $5 > 50.0' exercises.bed

22) Piping the output of the correct command for question 21 to wc -l provides how many lines?
    a. 34
    b. 35
    c. 36
    d. 37

23) Write an awk command that extracts all the repeats from scaffold-2 and pipe the output to wc -l. How many lines are in the output?
    < DO NOT INCLUDE SOLUTION: awk -F"\t" '$1=="scaffold-2" && $4=="repeat"' exercises.bed | wc -l >
    a. 4
    b. 7

c. 8
d. 11

24) Write an awk command that filters all values greater than 60.0 for contig-3, and pipe the output to wc -l. How many lines are in the output?
< DO NOT INCLUDE SOLUTION: awk -F"\t" '$1=="contig-3" && $5 >60' exercises.bed | wc -l>
a. 11
b. 15
c. 18
d. 19

25) Write an awk command that filters all values greater the 50.0 and less then 90.0 for contig-1 and pipe the output to wc -l. How many lines are in the output?
< DO NOT INCLUDE SOLUTION: awk -F"\t" '$1=="contig-1" && $5 >50 && $5 <90' exercises.bed | wc -l>
a. 24
b. 29
c. 33
d. 35

Bash and command line tools assessment questions:

The working directory for the assessment questions for this section:
/home/manager/course_data/unix/practical/Notebooks/advanced_bash/scripts

26) How do I run the bash script script.sh if I have not added it to my path?
a. /script.sh
b. script.sh
c. ./script.sh
d. bash script.sh

27) What is correct header for a bash script to let the system know it is a bash file?
a. /usr/bin/env bash
b. #!/usr/bin/env
c. #!/usr/bin/env bash
d. #!/bin/env bash

28) What is the correct command to add the scripts directory to your PATH variable?
a. export PATH=$PATH:/home/manager/course_data/unix/practical/Notebooks/advanced_bash/
b. export $PATH=PATH:/home/manager/course_data/unix/practical/Notebooks/advanced_bash/script

c. export PATH=$PATH:home/manager/course_data/unix/practical/Notebooks/advanced_bash/script
d. export PATH=$PATH:/home/manager/course_data/unix/practical/Notebooks/advanced_bash/script

29) What version of the program salmon is installed?

a. v1.3.0
b. v1.4.0
c. v1.5.0
d. v1.6.0

30) What command would I use to get the options for the bedtools intersect command?

a. bedtools intersect
b. bedtools -intersect
c. bedtools --intersect
d. bedtools myfile -intersect