Advanced Bioinformatics (7BBG2016): Practical Bioinformatics Data Skills

Student ID: 24002434

1. Basic Linux and the command Line (20pts – 10% of final mark, each question provides 1 point)

1.1 What does ./../.. stand for ?
A. Current directory
B. Up one directory
C. Up two directories
D. None of Above
C

1.2 What does cd / mean in UNIX? Please explain what the cd command does.
cd/ means we change directory to the root directory.
cd command helps us change directories in UNIX.

1.3 What command would you use to get help about the command cp? (please provide an example command)
man cp

1.4 What does the command pwd do?
The command pwd prints the current working directory.

1.5 How do you display a listing of file details such as date, size, and access permissions in a given directory? (please provide an example command)
ls -l

1.6 How do you print on the terminal the first 15 lines of all files ending by .txt? (please provide an example command)
head -n 15 *.txt

1.7 How do you rename a file from new to old? (please provide an example command)
mv new old

1.8 How do you display the contents of a file myfile.txt? (please provide an example command)
less myfile.txt

1.9 How do you create a new directory called flower? (please provide an example command)
mkdir flower

1.10 How do you change the current directory to /usr/local/bin? (please provide an example command)
cd /usr/local/bin

1.11 How can you display a list of all files in the current directory, including the hidden files? (please provide an example command)
ls -a

1.12 What command do you have to use to go to the parent directory? (please provide an example command)
cd ..

1.13 Which command would you use to create a sub-directory in your home directory? (please provide an example)
mkdir ~/directory

1.14 Which command would you use to list the first lines in a text file? (please provide an example)
head -n 1 textfile

1.15 Which command will display the last lines of the text file file1? (please provide an example)
tail -n 1 file1

1.16 Which command is used to extract a column from a text file? (please provide an example)
cut -f 1 textfile

1.17 How do you copy an entire directory structure? E.g. from Project to Project.backup (please provide an example)
cp -r Project Project.backup

1.18 How would you search for the string Hypertension at the end of the line in a file called diseases.txt? (please provide an example)
grep -w Hypertension diseases.txt | tail -n 1

1.19 How do you see hidden files in your home directory? (please provide an example)
ls -d ~/.*

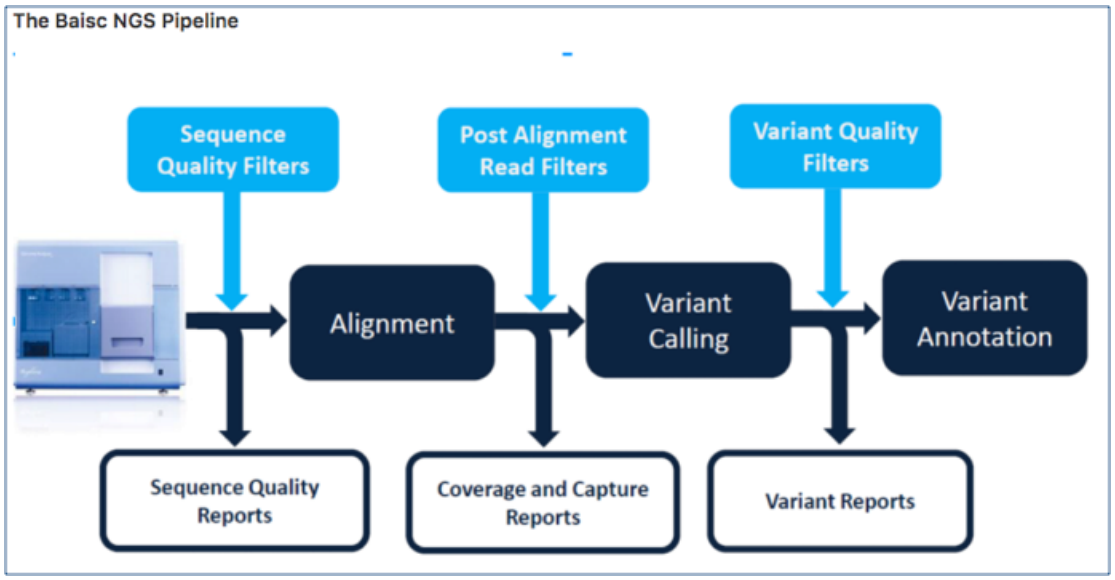
1.20 How do you run a job that will continue running even if you are logged out? (please provide an example)
nohup command &

2. The NGS Pipeline (65pts – 45% of final mark)

2.0 From raw data to alignment and variant calls (20pts)

The assessment is designed to:

- Test your ability to run standard NGS pipeline using the command line on a Linux system.
- Test your ability to create a Bash script that executes your NGS pipeline
- Test your basic knowledge of a standard NGS pipeline.



You have been provided with paired end fastq data and an annotation bed file from an Illumina HiSeq 2500 run. Using the assigned Openstack instance (please contact the module leaders if you have any problems with your Openstack instance), install the necessary tools and execute a standard Bioinformatics NGS pipeline to perform read alignment, variant discovery and annotation as described in the following NGS Pipeline section. **You are required to share a bash script that runs the workflow and takes the provided sequencing data as input (links provided below) with the examiner by uploading it with this report.** If uploading the script via Canvas or KEATS presents technical problems, you can also share the script by uploading it onto your github. **If you do so, please do not forget to provide the link to your github in the assignment and make sure you do not modify the file after the assignment deadline as this will show on github and will make the submission invalid.** Please make sure the bash script lines are adequately commented to provide a clear description of what it is doing. **The script will be evaluated by the examiner and up to 20pts will be given for a fully running and easy to read script.** Based on your pipeline, provide the following information and answer each question.

Fastq Read 1 (~750MB): https://s3-eu-west-1.amazonaws.com/workshopdata2017/NGS0001.R1.fastq.qz

Fastq Read 2 (~750MB): https://s3-eu-west-1.amazonaws.com/workshopdata2017/NGS0001.R2.fastq.qz

Annotation File (10M): https://s3-eu-west-1.amazonaws.com/workshopdata2017/annotation.bed

HINT: Please note that the sequencing data have an "odd" extension. You might consider renaming the files.

In the following questions you will be asked to provide the command lines used to perform the steps of the pipeline and to comment and explain the choice of tools and all options. Please do not forget the latter as copying and pasting the command lines from the bash pipeline will not be sufficient to pass. You will need to demonstrate a clear understanding of your choices. Feel free to provide examples (even graphical/screenshots) if helpful.

2.1 Install the tools and dependencies of your pipeline (using Miniconda when possible) and Download the input files (5 pts)

1. List the command lines to install all dependencies necessary to run the pipeline (3 pts)

2. List all command lines necessary to download the input files (e.g. fastqs, reference genomes, etc) (2 pts)

Implement and run the following NGS Pipeline (please provide the command lines to run the following steps of your pipeline and comment/explain the choice of options):

2.2. Pre-Alignment QC (4 pts)

1. Perform quality assessment and trimming (2pt)
2. Perform basic quality assessment of paired trimmed sequencing data (2pt)

2.3. Alignment (17pts)

- Align the paired trimmed fastq files using bwa mem and reference genome hg19 (edit your bwa mem step to include read group information in your BAM file) (9pts)
- Perform duplicate marking (2pts)
- Quality Filter the duplicate marked BAM file (2pts)
- Generate standard alignment statistics (i.e. flagstats, idxstats, depth of coverage, insert size) (4pts)

2.4. Variant Calling (4pts)

- Call Variants using Freebayes restricting the analysis to the regions in the bed file provided (2pt)
- Quality Filter Variants using your choice of filters (2pt)

2.5. Variant Annotation and Prioritization (10pts)

- Annotate variants using **ANNOVAR** (4pt) and **snpEFF** (4pt)
- Perform basic variant prioritization: filter to exonic variants not seen in dbSNP (2pts)

2.6 Using an alternative tool (5pts)

- Modify the pipeline by replacing either the aligner or the variant caller with an alternative tool, while leaving the rest of the pipeline unchanged. Share a new bash script with the modified pipeline with the examiners by uploading it on Canvas/KEATS with your assignment or via github (3pt)
- Provide below the new commands used to run the alternative tool and comment on your choice of options and how and if using this tool would affect the results (2pt).

```
#Generate the index files
sudo apt install bowtie2
cd ~/7BBG2016/data/reference
bowtie2-build -f hg19.fa hg19_index
#Replace the aligner with Bowtie2
bowtie2 -x hg19_index -1
~/7BBG2016/data/trimmed_fastq/NGS0001_trimmed_R_1P -2
~/7BBG2016/data/trimmed_fastq/NGS0001_trimmed_R_2P -S
~/7BBG2016/data/aligned_data/NGS0001.sam
```

Bowtie2 is more suitable for aligning short sequences, with a relatively high operating speed.
The results of aligning using BWA Mem may be more accurate.

3. R/RStudio assessment (45pts – 45% of final mark)

This R assignment is split into 3 parts. The first part is about the general use of R/Rstudio, the second part about RNAseq and the third about ChIP-Seq. In these parts you will be asked to perform a number of tasks in R/RStudio and report them in your own markdown document.

Initial task: Create a new markdown document in *RStudio*, set the title to "Advanced Bioinformatics 2023 assessment", and insert an "author:" tag below the title, followed by your student id. Share your markdown document and html via your github account.

In the following, for each task, create a new heading called "Task X" for task X, and insert a new R code chunk that holds any code required. Make sure to evaluate the expression before saving to include the output in the html file. If you have multiple lines that produce outputs, you can split them into separate code chunks for increase clarity (but it is not necessary to pass the assessment). Please also explain your steps.

General R/Rstudio assessment (33 pts)

3.1. Using the *sum*() function and : operator, write an expression in the code snippet to evaluate the sum of all integers between 5 and 55. (4pt)

3.2. Write a function called *sumfun* with one input parameter, called *n*, that calculates the sum of all integers between 5 and *n*. Use the function to do the calculation for *n* = 10, *n* = 20, and *n* = 100 and present the results. (4pt)

3.3. The famous Fibonacci series is calculated as the sum of the two preceding members of the sequence, where the first two steps in the sequence are 1, 1. Write an R script using a for loop to calculate and print out the first 12 entries of the Fibonacci series. (4pt)

3.4. With the *mtcars* dataset bundled with R, use *ggplot* to generate a box of miles per gallon (in the variable *mpg*) as a function of the number of gears (in the variable *gear*). Use the fill aesthetic to colour bars by number of gears. (4pt)

3.5. Using the *cars* dataset and the function *lm*, fit a linear relationship between *speed* and breaking distance in the variable *distance*. What are the fitted slope and intercept of the line, and their standard errors? What are the units used for the variables in the dataset? (4pt)

3.6. Use *ggplot* to plot the data points from Task 6 and the linear fit. (4pt)

3.7. Again using the cars dataset, now use linear regression (*lm*) to estimate the average reaction time for the driver to start breaking (in seconds). To simplify matters you may assume that once breaking commences, breaking distance is proportional to the square of the speed. Explain the steps in your analysis. Do you get reasonable results? Finally, use *ggplot* to plot the data points and the fitted relationship. (9pt)

RNA-seq assessment (8 pts)

In this part, we will analyse the RNASeq data used in the RNA-seq tutorial to:

1. create a DESeq2 object,
2. normalize RNA-seq data with DESeq2,
3. perform differential Expression analysis with DESeq2,
4. visualize RNA-seq data using SDM and PCA methods.

You may access to the data that we used during tutorial from here.

3.8. Read in count data and sample description. **(1pts)**

- LMS_RNAseq_short-master-2023-final/course/exercises/data/exercise1_counts.csv
- LMS_RNAseq_short-master-2023-final/course/exercises/data/exercise1_sample_description.info

3.9. Create col_data and check dimensions. **(1 pts)**

3.10 Construct DESeqDataSet object using count data and sample description. **(1 pts)**

3.11. Perform rlog and VST transformation on the data. **(1 pts)**

3.12. Draw a heatmap of count matrix based on the top 40 highly expressed genes using rlog and VST data. **(1 pts)**
3.13. Generate a SDM to see the clustering of count data. **(1 pts)**
3.14. Perform the Principal Component Analysis using rlog method and find out the % significance values of first two principal components. **(1 pts)**
3.15. Repeat the PCA, this time using VST method and compare the plots with the ones obtained using rlog method. **(1 pts)**

ChIP-seq assessment (4 pts)

In this assessment, we will read in two replicate sets of CHIP-seq peaks from the Myc Encode dataset and extract sequences underneath subsets of peaks. We will write these sequences out to a FASTA file and upload the FASTA file to Meme-ChIP to detect motifs underneath of these peaks.

You may access to the data that we used during tutorial from here.

3.16. Read in the two Myc Mel peakset replicates and create the common peakset as we did for our previous exercise. **(1 pts)** The files you need are here:

- LMS_ChIPseq_short-master-2023-final/course/data /MacsPeaks/mycmelrep1_peaks.xls
- LMS_ChIPseq_short-master-2023-final/course/data /MacsPeaks/mycmelrep2_peaks.xls

3.17. Now we can rank them by their fold enrichment, select the top 500 peaks and resize these peaks to 200bp around centre. **(1 pts)**
3.18. Extract the sequences underneath the file and write them to FASTA file in you working directory. Inspect the file in notepad. **(1 pts)**
3.19. Upload the sequences to Meme-ChIP and report the results when complete. **(1 pts)**