TD : Linux and NGS data treatment

In this practical session, you will handle some "classical" files of NGS analysis. In order to succeed, you will need some commands you saw yesterday and you may discover some new ones. The purpose of this session is to be as interactive as possible. This document tells you what to do, but not how to do it, in order to let you find out which command to use. The correct, or at least functionnal, command lines will be added during this session in the *cmd.txt.* You will find this file on the marmadais server, in the directory :
/usr/local/bioinfo/training/rna-seq
This directory also contains all the needed files for this session.

1. Creating your working directory :

Start by creating a working directory in your home directory

Go to this directory

Copy the file *arctikConfigFile.txt* from the directory */usr/local/bioinfo/training/rna-seq*

Create a symbolic link to the file */usr/local/bioinfo/training/rna-seq/reference.fna* that you will name *toto*

Remove this link

Call again the command that generated the link

Rename this link *reference.fasta*

Create a link to the directories *Arctik* and *Arcad* and to the file *adapt_TruSeq.fasta*

Create a directory *Raw_data*

For each files in */usr/local/bioinfo/training/rna-seq/Raw_data/* create a symbolic link in your directory *Raw_data*

Why did we create link and didn't we copy the needed files instead ?

OK, so why did we copy the file *arctikConfigFile.txt* ?

How many reads were there in each fastq file ?

Useful commands for this part:
awk, boucle for, cd, cp, ln, mkdir, mv, unlink, wc, !

2. Reads cleaning and mapping :

Look at the file *arctikConfigFile.txt*

Edit the file *arctikConfigFile.txt* and replace the TO_FILL by the correct value

Create the BWA index for the reference file *reference.fasta*

Create the *.fai* et *.dict* files that we will need later

Launch the perl script *arctikMain.pl* located in *Arctik/pipeline/bin/* . It takes as input an option - -param with as value the configuration file.

Create a *log* directory and move all logs files created by SGE in it

Merge the different mapping files that you will find in the directory *bwaOut_dir*. For this, you will use the program *MergeSamFile.jar* from the picard tools. You can find it in */home/sarah1/src/picard-tool/*
The correct java version to use can be found there : */usr/local/jre/bin/java*

Look at some statistics of your mapping by using the samtools

How many reads were mapped ?

How many reads were mapped on each reference sequence ?

Useful commands for this part :
bwa index, java, more (or less), nano, perl, qsub, picard-tools (MergeSamFile and CreateSequenceDictionary) ,samtools faidx, samtools flagstat, samtools idxstats, samtools view

3. Mapping post-processing and variant calling

Launch the perl script *Arcad/6_bamRealignerRecalibrate.pl*. You can have a look at the expected parameters by launching it without parameters.

Launch the perl script *Arcad/7_genotyper_call.pl.* You can also have a look at the expected parameters by launching it without parameters. Use the filtration options.

Look at your *vcf* file content.

How many variants did you obtain ?

How many variants did pass the quality filters ?

How many variant did pass the quality filter and are located on the first mRNA of the reference file ?

Filter your variant file, such as it will only contains the variants passing the quality filters. (It should keep its header !!)

Useful commands for this part :
grep, more (or less), perl, sed, | operator , > operator