BASh based pipeline for variant calling and annotation

This script is a variant calling pipeline.

I have used for loops in this script to iterate the commands over multiple input files. In these for loops, the filename has been defined as a variable in the for statement,

which will enable you to run the loop on multiple files.

1. I enquired if the necessary packages are installed

2. I moved the dataset and reference genome to the directories they would be called

3. Then I ran fastqc using a for loop

4. I unzipped the zipped dataset

5. I ran fastp using a for loop

6. Finally ran the variant calling by indexing the reference genome after gunzipping then aligned and sorted

The script takes in two arguments

1. The path to the datasets

2. The path to the reference genome

Pls note save your datasets as a "r1.fasta.gz and r2.fasta.gz" file for your forward and reverse sequences respectively.

Also save your reference as a ".fa.gz" file

Datasets

wget https://zenodo.org/record/2582555/files/SLGFSK-N_231335_r1_chr5_12_17.fastq.gz

wget https://zenodo.org/record/2582555/files/SLGFSK-N_231335_r2_chr5_12_17.fastq.gz

wget https://zenodo.org/record/2582555/files/SLGFSK-T_231336_r1_chr5_12_17.fastq.gz

wget https://zenodo.org/record/2582555/files/SLGFSK-T_231336_r2_chr5_12_17.fastq.gz

Reference

Reference: wget https://zenodo.org/record/2582555/files/hg19.chr5_12_17.fa.gz

Software used

FASTP, FASTQC, BWA, SAMTOOLS,