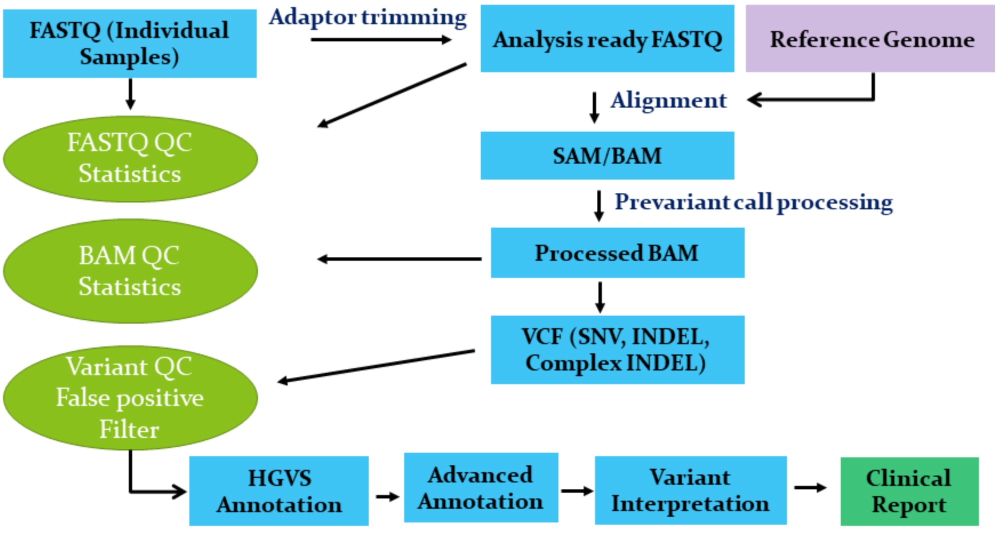
Laboratory Scientist III Coding Assignment – Vanipriyadarsini Ikkurti, PhD

1. a) Based on the information in the preamble, a pipeline is designed implementing gold stand variant calling software(fastp v0.23.2, multiqc v1.12, BWA v0.7.17, deeptools v2.0, GATK v4.2), databases (dbSNP154; snpeff v5.1, Annovar) and human reference genome (GRCh38.p13) to aid clinical diagnosis in a laboratory set up. The pipeline can be accessed at https://github.com/vani-ikk/labscientist-3.

b) The other questions are answered with respective question number.

2. Standard guidelines (ACMG-CAP) for clinical reporting from human NGS data was followed as per the workflow mentioned in figure 1. Accordingly, a skeleton code was designed to perform quality control, alignment, variant calling, annotation and reporting clinically actionable variants (please refer: **question2.sh** in the repository).



Roy et al., 2018 The Journal of Molecular Diagnostics

Figure 1: Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines (WES/CES/WGS).

3. Shell script was used write code for processing BAM file to generate standard VCF file.

a) The pre-processing of BAM file such as de-duplicate, recalibration using dbSNP was performed. The GATK haplotypeCaller was used to generate variants from BAM file. The script for the same is given as question3.sh

b) In question3.sh, script is written to identify clinically actionable variants with comment lines for easy understanding.

4.

a) The data may be transferred through openSSH or FTP protocol periodically.

b) Automated programs may be designed with date and time stamp while data transfer along with md5sumcheck file for each files, such that integrity of the file can be checked while data is received by exported machine.

c) scp –P 22 –r LOCAL_DIR vani@10.0.0.21:/$PATH/300620221202

rsync, ftp etc. may also be used

5. Each machine ID with the date and timestamp can be used to generate a unique ID to recognize machine as as well as data transferred at different time point. Programming languages such as Python or R may be used for the same.

6.

a) Shell script with awk and grep may be used to match variants obtained from NGS data and PCR data.

awk may be used to fetch the desired column and further grep as mentioned below may be used to match the variants in both files. A for loop may be used to read multiple files.

grep –f “variant_list.txt” PCR_variant.txt > matched.txt

In addition, MySQL based databases may be designed for better matching of variants obtained from NGS and PCR experiments.

b) If and else conditions may be used in shell script inside the for loop and echo command may be used in side if and else statement to give feedback on match or mismatch.

7. a) The data should be kept as sql database in hospital information system or laboratory information system with interactive graphical user interface (PHP, MYSQL or PYTHON, MySQL). All processed raw data and intermediate data should be moved to a mounted

storage. The report should be generated in PDF or web accessible through a user account to the patient.

b) Quality control for each variant is assured through three steps. 1. Raw read QC i.e. Average read quality >= 30; 2. Alignment QC i.e. depth and coverage of alignment is more than 10x at each position, 3) variant QC i.e. Depth >=10, Quality >=30, No strand bias, Minor allele frequency should be <1% and the variant should be classified as pathogenic or likely pathogenic as for ACMG guideline. The details of the same is also mentioned in the script **question2.sh**

c) It is preferable to follow automated data transfer with cross checking successful transfer through md5sumcheck files.

8.

a) The sample lab data for same phenotype should have unique laboratory identifier in the database, such that data can be matched irrespective of the time of receiving data.

b) Best approach is to store variant data in a SQL based relational database program after QC pass of variants.

c) If the software is installed in a server, it can be accessed remotely for analysis. Often container programs such as Docker may be used for analysing data. However, the final SQL database which would be accessible to patients would be a web based application so; it is not required to transfer it to a Desktop machine.