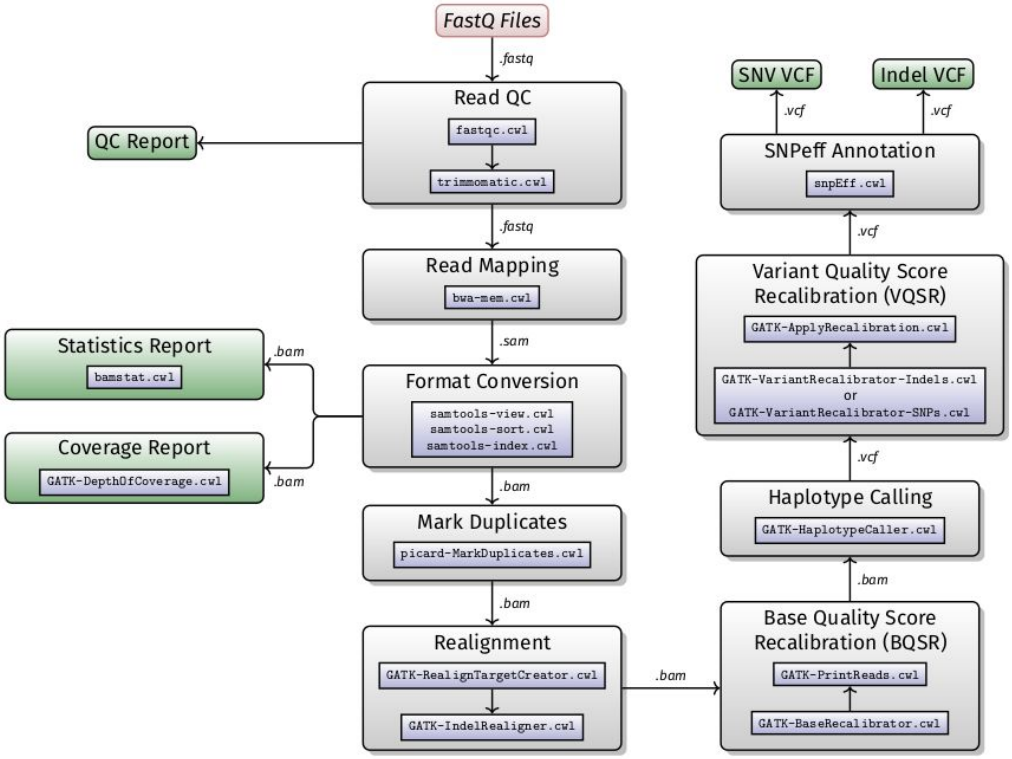
Variant calling work flow using snakemake language

variant calling is the identification of a nucleotide difference in a given reference at a position of an individual genome or transcriptome. Variant calling has become widely accepted in human genetics as a way of identifying variants associated with a specific trait, population or hereditary diseases. It employs next-generation sequencing data to identify two main types of variants, namely single nucleotide variants/polymorphism(SNPs) and INDELs (Small insertion or deletions) within a genome of interest. The H3ABioNet have developed some standard operating procedures (SOPs) for variant calling pipelines for the H3Africa Consortium and interested individuals. Such Large-scale data analyses in bioinformatics involve the chained execution of many command line applications. Workflow helps to automate these pipelines and ensure reproducibility. SnakeMake is a workflow language being developed by Johannes Köster and was first introduced in 2012 SnakeMake takes all the good stuff from make and transports into the 21st century. Since the language is based on the rule and target structure, it's as robust and reproducible as Make. It clicks into the Bash shell but still clearly retains its Python programming background and simplicity. It formulates the jobs your system has to run based on the directed acyclic graph (DAG) it generates from the rules and targets in the Snakemake file. A DAG is a representation of the relationship between the files and tools within your data analysis, it basically shows through what steps your data has to go from start to finish. The DAG can easily be converted to an image to create an overview of what your data analysis is consisting of.



A Workflow generated for analysis of vcf files that determines the SNPs and INDELs

METHODOLOGY

Whole exome sequencing data for human genome chromosome 1 provided for accreditation of the H3bionet in icipe node was used to generate a pipeline. The data was obtained from the H3Africa website for standard operating procedures (SOPs) based on hg19 GATK 2.8 bundle. We generated a pipeline for the variant determination based on the procedures provided. The pipeline was divide in three phases : phase one : preprocessing of the reads (fastqc analysis (fastqc) , adapter removal(trimmomatics), contamination removalor short reads (trimmomatics)) , Phase two:Initial variant discovery(alignment to reference (bwa and samtools) deduplication (sambamba) ,base quality score recalibration (BSQR protocol from GATK), calling the variant and statistical filtering (haplotypecaller from GATK)) and phase three: variant annotation and prioritization (SNP and INDEL variant prioritization(VSQR protocol from GATK)). Once the pipeline was developed from a bash shell script, we were able to generate a work flow from the snakemake language involving (input, output , shell and rules). This methods analysis was based on the workflow described in our github repository.

DISCUSSION

The snakemake pipeline was evaluated with data used for the accreditation. The data was reported to have number of 27921 SNPs and 1589 INDELs indicating the functionality of our workflow we were also able to test the snakemake pipeline in different computers generating the same given data sets. Although the pipeline was capable of generating the different vcf files. we were not able to evaluate the quality of our snakemake with different data sets. We recommended that our workflow be tested by other variant study to demonstrate its ability in different set of conditions .