A snakemake variant calling pipeline for a single sample Illumina short paired-end reads

This is a snakemake pipeline for doing preliminary NGS analysis of converting raw Illumina short paired-end reads to a list of somatic and/or germline variants.

**Snakemake NGS analysis workflow:**

1) Pre-processing using Cutadapt

2) Mapping using BWA

3) Sorting using samtools

4) Removing duplicates using GATK MarkDuplicates

5) Indexing using samtools

6) Base quality score recalibration using GATK BaseRecalibrator and ApplyBQSR

7) Germline variant detection using GATK HaplotypeCaller

8) Somatic variant detection using GATK Mutect2

**Working directory structure for running Snakemake:**

-Paired-end reads (ending in \_R1.fastq and \_R1.fastq, say A\_R1.fastq and A\_R2.fastq) in ‘FastQ’ directory

-CallVars.yml

-Snakefile

-1000G\_phase1.indels.hg19.sites.vcf

- 1000G\_phase1.indels.hg19.sites.vcf.idx

- dbSNP\_20180423.vcf

- dbSNP\_20180423.vcf.idx

- GNOMAD\_hg19.vcf

- GNOMAD\_hg19.vcf.idx

- Mills\_and\_1000G\_gold\_standard.indels.hg19.sites.vcf

- Mills\_and\_1000G\_gold\_standard.indels.hg19.sites.vcf.idx

- Reference Genome Directory ‘UCSCWholeGenomeFasta’ containing below files

* genome.dict, genome.fa, genome.fa.amb, genome.fa.ann, genome.fa.bwt, genome.fa.fai, genome.fa.pac, genome.fa.sa, GenomeSize.xml

**Perform below steps on the Linux command line interface for running the NGS analysis.**

1. **Check working directory and FastQ files:**

Make sure you are in the working directory that contains all the needed files and folders for running snakemake.

Make sure your Fastq files are in ‘FastQ’ directory and they end in ‘\_R1.fastq’ and ‘\_R2.fastq’, say A\_R1.fastq and A\_R2.fastq.

1. **Install miniconda**:

Use below link to install miniconda.

<https://conda.io/projects/conda/en/latest/user-guide/install/linux.html>

1. **Install snakemake:**

conda install -c bioconda -c conda-forge snakemake

1. **Create environment CallVars:**

conda env create –n CallVars –f CallVars.yml

This downloads all the necessary tools (Cutadapt, BWA, GATK and samtools) and dependencies for NGS analysis to run.

1. **Activate CallVars environment:**

conda activate CallVars

1. **Running snakemake:**

Ensure you run the below command in working directory.

snakemake CallVars\_Output/VCF/{your\_sample\_name}\_germline.vcf CallVars\_Output/VCF/{your\_sample\_name}\_somatic.vcf --cores N

Where CallVars\_Output/VCF/A\_germline.vcf is the germline variant file, CallVars\_Output/VCF/A\_somatic.vcf is the somatic variant file and N is number of cores.

For instance, if the names if the FastQ files are A\_R1.fastq and A\_R2.fastq and number of cores available are 6 then run below command.

snakemake CallVars\_Output/VCF/A\_germline.vcf CallVars\_Output/VCF/A\_somatic.vcf --cores 6