# IDENTIFICATION OF SOMATIC AND GERMLINE VARIANTS FROM TUMOR AND NORMAL SAMPLE PAIRS

**Damilola Adegbite**

## ABSTRACT

In sequencing genomic materials from a human tumor, the clinical research question hopes to bring a discovery into the identification of the somatic and germline variants in both tumor and normal pairs that is, what type of mutation distinguishes this tumor from a normal healthy tissue. Questions like: What are the specific challenges in somatic variant calling that set it apart from regular diploid variant calling? How to call variants and classify them according to their presence/absence in/from tumor and normal tissue of the same individual? How to annotate variants and affected genes with prior knowledge from human genetic and cancer-specific databases to generate clinically relevant reports? In view of this, the major objectives are

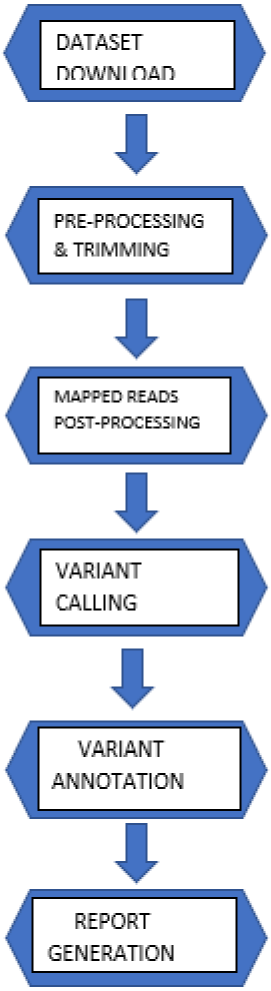
- Call variants and their somatic status from sequencing data
- Annotate variants with human genetic and cancer-specific information extracted from public databases
- Add gene-level annotations and generate reports of annotated somatic and germline variants, loss-of-heterozygosity (LOH) events, and affected genes, ready for interpretation by clinicians
- To perform this task on HackBio Workshop by replicating a pipeline and performing a simple and informative analysis

## INTRODUCTION

Every individual has a unique pattern of somatic and germline variants. This being the beauty of life, however it is not adequate in comparing normal and tumor tissues. Though, the human reference genome is available but it is not the best tool in identifying tumor tissues as there will be a numerous number of variants. Moreover, to be able to determine between these two types of variants, a direct comparison of data from tumor and normal tissue samples is required. In addition to acquiring new variants, tumors can also lose or gain chromosomal copies of variants found heterozygously in an individual's germline. This phenomenon is termed loss of heterozygosity (LOH). We are going to identify somatic and germline variants, as well as variants affected by LOH, from a tumor and a normal sample of the same patient. Our goal is to report the variant sites, and the genes affected by them, annotated with the content of general human genetic and cancer-specific databases.

## METHODOLOGY

This graphical illustration below shows the steps taken in the identification of somatic and germline variants in tumor and normal pairs



### DATASET DOWNLOAD

It is important to first do an analysis of the dataset. The sequencing reads to be analyzed are gotten from real-world data from a cancer patient’s tumor and normal tissue. The original data includes human chromosomes 5, 12 and 17.

```
mkdir -p raw_data
cd raw_data

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
```

The first two files represent the forward and reverse reads sequence data from a patient’s normal tissue, and the last two represent the forward and reverse reads sequence data from a patient’s tumor tissue.

**REFERENCE SEQUENCE**

```
mkdir ref
cd ref
wget https://zenodo.org/record/2582555/files/hg19.chr5_12_17.fa.gz

#gunzip reference
gunzip hg19.chr5_12_17.fa.gz
PRE-PROCESSING AND TRIMMING
The reads quality were examined using fastqc and an aggregate report generated with
multiqc
```

**SOFTWARE PACKAGES**

- FastQC
- MultiQC
- Trimmomatic
- BWA
- Samtools
- VarScan Somatic
- BCFtools
- SnpEff
- Gemini

Note: It is necessary to have all commands installed for this project

## PRE-PROCESSING AND TRIMMING

i) Quality check

The reads quality were examined using fastqc and the aggregate report generated with multiqc.

```
#create directory for the fastqc output
mkdir -p Fastqc_Reports


#create a list.txt file
nano list.txt

#copy the normal and tumor dataset name
SLGFSK-N_231335
SLGFSK-T_231336
#Quality check on reads
for sample in `cat list.txt`
do
        fastqc raw_data/${sample}*_r1_chr5_12_17.fastq.gz -o Fastqc_Reports
done
multiqc Fastqc_Reports -o Fastqc_Reports
```

ii) Removing low quality sequences using Trimmomatic

Trimmomatic is a script that enables quality and adapter trimming. After analyzing data quality, removing sequences that do not meet quality standards is next.

```
nano trimmed.sh

mkdir -p trimmed_reads

for sample in `cat list.txt`
do
       trimmomatic PE -threads 8 ${sample}_r1_chr5_12_17.fastq.gz 
${sample}_r2_chr5_12_17.fastq.gz \
               trimmed_reads/${sample}_r1_paired.fq.gz 
trimmed_reads/${sample}_r1_unpaired.fq.gz \
               trimmed_reads/${sample}_r2_paired.fq.gz 
trimmed_reads/${sample}_r2_unpaired.fq.gz \
               ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8:keepBothReads \
               LEADING:3 TRAILING:10 MINLEN:25

       fastqc  trimmed_reads/${sample}_r1_paired.fq.gz 
trimmed_reads/${sample}_r2_paired.fq.gz \
                 -o trimmed_reads/Fastqc_results
done
bash trimmed.sh
multiqc  trimmed_reads/Fastqc_results  -o trimmed_reads/Fastqc_results
```

## MAPPED READS POSTPROCESSING

Mapping of sample sequences against the reference genome is conducted to determine the most likely source of the observed sequencing reads.

- Read mapping

In order to align the data, we need a reference to align against. First, a directory is created for the reference and then copied. The reference is indexed to be able to align the data.

```
#Index reference file
bwa index hg19.chr5_12_17.fa
```

This produces 5 files in the reference directory that BWA uses during the alignment phase. The 5 files have different extensions named amb,ann,bwt pac and sa.

```
#bwa mem for alignment
nano mapping.sh
mkdir Mapping

#Perform alignment
bwa mem -R '@RG\tID:231335\tSM:Normal' ref/hg19.chr5_12_17.fa trimmed_reads/SLGFSK-
N_231335_r1_paired.fq.gz \
      trimmed_reads/SLGFSK-N_231335_r2_paired.fq.gz > Mapping/SLGFSK-N_231335.sam

bwa mem -R '@RG\tID:231336\tSM:Tumor' ref/hg19.chr5_12_17.fa trimmed_reads/SLGFSK-
T_231336_r1_paired.fq.gz \
       trimmed_reads/SLGFSK-T_231336_r2_paired.fq.gz > Mapping/SLGFSK-T_231336.sam
bash mapping.sh
```

- Conversion of the SAM file to BAM file, sorting and indexing

A Binary Alignment Map (BAM) format is an equivalent to sam but its developed for fast processing and indexing. It stores every read base, base quality and uses a single conventional technique for all types of data.

```
nano indexing.sh

for sample in `cat list.txt`
do
        Convert SAM to BAM and sort it
        samtools view -@ 20 -S -b Mapping/${sample}.sam | samtools sort -n -@ 32 >
Mapping/${sample}.sorted.bam

        Index BAM file
        samtools index Mapping/${sample}.sorted.bam
done
bash indexing.sh
```

- Filtering Mapped reads

```
nano filter.sh
for sample in `cat list.txt`
do
        #Filter BAM files
        samtools view -q 1 -f 0x2 -F 0x8 -b Mapping/${sample}.sorted.bam >
Mapping/${sample}.filtered1.bam
```

```
done
bash filter.sh
```

- Duplicates removal

```
#use the command rmdup
nano rmdup.sh

for sample in `cat list.txt`
do
         samtools collate -o Mapping/${sample}.filtered1.bam
Mapping/${sample}.namecollate.bam
        samtools fixmate -m Mapping/${sample}.namecollate.bam
Mapping/${sample}.fixmate.bam
        samtools sort -@ 32 -o Mapping/${sample}.positionsort.bam
Mapping/${sample}.fixmate.bam
        samtools markdup -@32 -r Mapping/${sample}.positionsort.bam
Mapping/${sample}.clean.bam
done
bash rmdup.sh
```

- Left Align BAM

```
nano leftalign.sh

for sample in `cat list.txt`
do
        cat Mapping/${sample}.clean.bam  | bamleftalign -f hg19.chr5_12_17.fa -m 5 -c
> Mapping/${sample}.leftAlign.bam

done
bash leftalign.sh
```

- Recalibrate read mapping qualities

```
nano recalibrate.sh

for sample in `cat list.txt`
do
        samtools calmd -@ 32 -b Mapping/${sample}.leftAlign.bam
ref/hg19.chr5_12_17.fa > Mapping/${sample}.recalibrate.bam
done
bash recalibrate.sh
```

- Refilter read mapping qualities

```
nano refilter.sh

for sample in `cat list.txt`
do
        bamtools filter -in Mapping/${sample}.recalibrate.bam -mapQuality “<=254” >
Mapping/${sample}.refilter.bam
```

```
done
bash refilter.sh
```

## VARIANT CALLING AND CLASSIFICATION

VarScan somatic was used to identify variants in mapped samples. The command reports germline, somatic, and LOH events at positions where both normal and tumor samples have sufficient coverage

- Convert data to pileup

```
nano variants.sh

mkdir Variants

for sample in `cat list.txt`
do
        samtools mpileup -f ref/hg19.chr5_12_17.fa Mapping/${sample}.refilter.bam --
min-MQ 1 --min-BQ 28 \
                > Variants/${sample}.pileup
done
bash variants.sh
```

- Call variants

```
varscan somatic Variants/SLGFSK-N_231335.pileup \
        Variants/SLGFSK-T_231336.pileup Variants/SLGFSK \
        --normal-purity 1  --tumor-purity 0.5 --output-vcf 1
```

- Merge vcf

VarScan generates 2 outputs (indel.vcf and snp.vcf), merge the two into one vcf file using bcftools.

```
#merge vcf
bgzip Variants/SLGFSK.snp.vcf > Variants/SLGFSK.snp.vcf.gz
bgzip Variants/SLGFSK.indel.vcf > Variants/SLGFSK.indel.vcf.gz
tabix Variants/SLGFSK.snp.vcf.gz
tabix Variants/SLGFSK.indel.vcf.gz
bcftools merge Variants/SLGFSK.snp.vcf.gz Variants/SLGFSK.indel.vcf.gz >
Variants/SLGFSK.vcf
```

## VARIANT ANNOTATION

SnpEff is a variant annotator and functional effect predictor. The output ends with a .ann.

```
#download jar file
```

```
wget https://snpeff.blob.core.windows.net/versions/snpEff_latest_core.zip

#unzip file
unzip snpEff_latest_core.zip

#download snpEff database
 snpEff download hg19

#annotate variants
 snpEff hg19 Variants/SLGFSK.vcf > Variants/SLGFSK.ann.vcf
```

### CLINICAL ANNOTATION

```
#installation
wget https://raw.github.com/arq5x/gemini/master/gemini/scripts/gemini_install.py
python gemini_install.py /usr/local /usr/local/share/gemini

#command to use gemini for clinical annotation
gemini load -v Variants/SLGFSK.ann.vcf -t snpEff Annotation/gemini.db
```

## CONCLUSION

After following the step-by-step process to replicate the recommended tutorial as well as identification of the variants, we can conclude that the interpretation of any list of variants (somatic, germline or LOH) depends crucially on genetic and cancer-specific variant and gene annotations. The insights gotten from the reproduction of this workflow can help to track genetic events driving the growth of tumor in patients which are useful in revealing variants known to affect drug resistance/sensitivity, tumor aggressiveness and useful in diagnosis, prognosis, developing therapeutic tactics.