IDENTIFICATION OF SOMATIC AND GERMLINE VARIANTS FROM TUMOUR AND NORMAL SAMPLE PAIRS

Ismat Ghazal

INTRODUCTION

Genetic variation has proven important for evolution and is now known to be important in the development of many complex diseases, including cancer. Between individuals, genetic makeup often differs due to germline mutations that are passed between generations, Within an individual, genetic alterations may arise spontaneously during mitosis or result from environmental factors. These somatic changes account for majority of cancers and as such, tumor cells often carry somatic mutations that are absent in normal tissue.

Understanding these differences and identifying specific mutations and genomic aberrations that differentiate normal from diseased tissues, has great potential for improving our understanding of the aetioparhogenesis of cancers, developing sensitive and specific diagnostic tools and developing targeted therapies based on these unique genetic features.

The aim of this task was to identify somatic and germline variants from a patient's tumor tissue and normal tissue sample pairs through variant calling and extensive variant annotation.

METHODS

Data

Dataset used for this task were imported from Zenodo data repository. The sample data consists of a patient's normal tissue and tumour tissue sequence reads of chromosomes 5, 12, and 17. The reference genome is a hg19 version of human chromosomes 5, 12, and 17. Variant annotation files and gene-level annotation files were also imported from Zenodo.

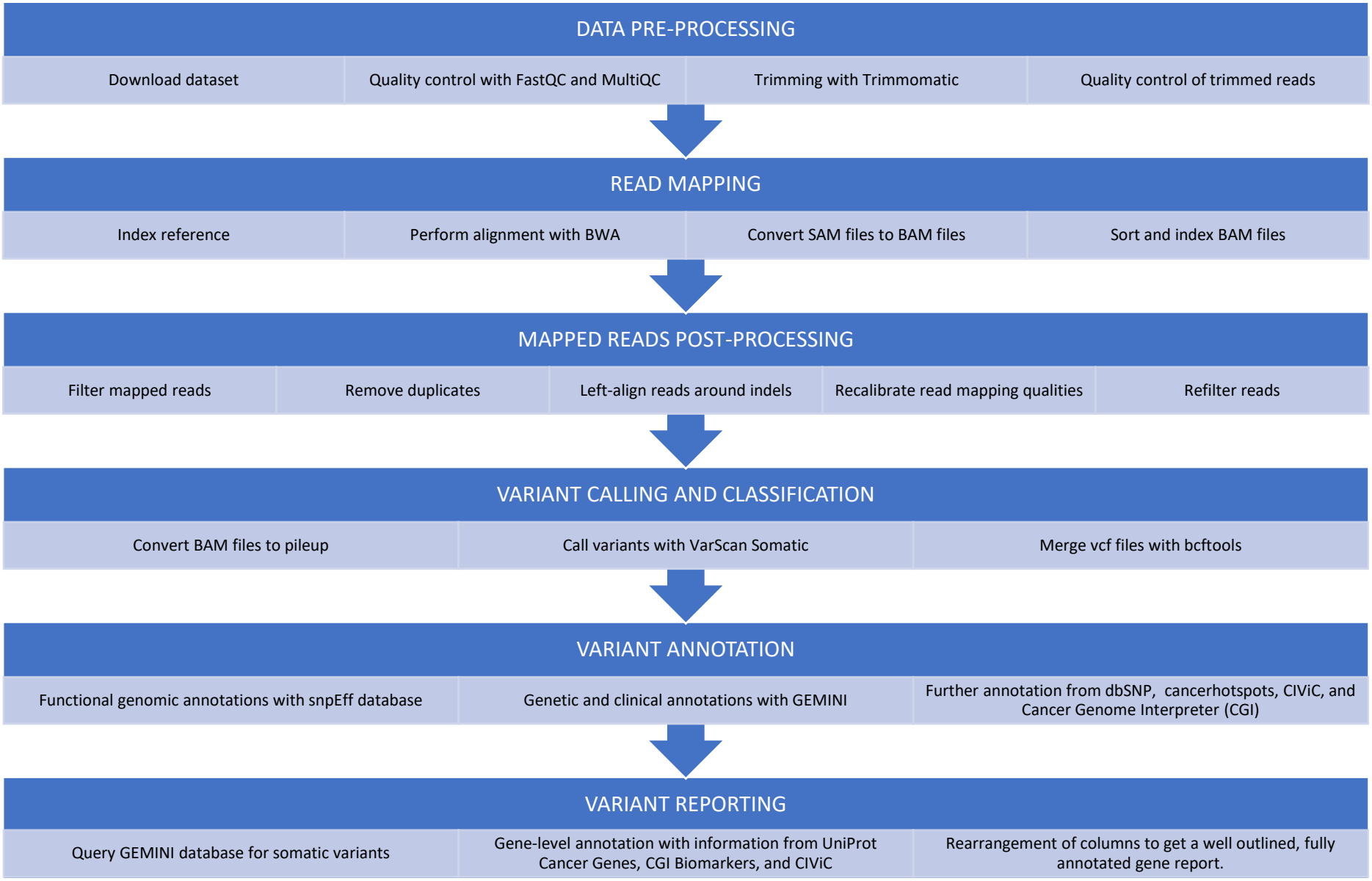
Softwares

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

Procedure

The tutorial was divided into 6 main steps from data pre-processing to variant reporting (see Page 2).

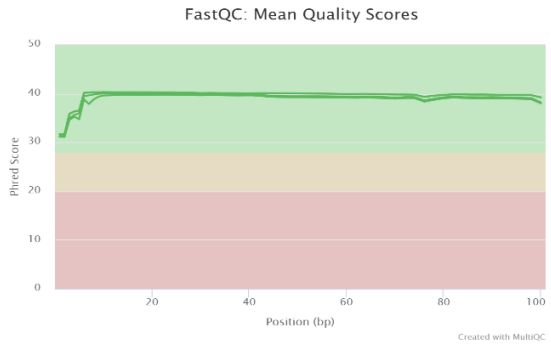
Data pre-processing to variant calling stage was carried out in Linux pipeline on the HackBio server. However due to space constraint, the tutorial was completed on Galaxy Server. The codes for the Linux workflow are published here.

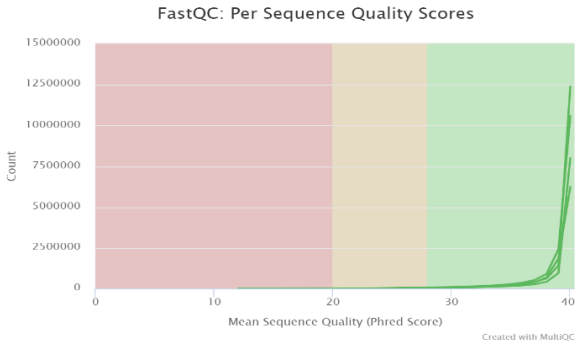

DATA PRE-PROCESSING
Download dataset
Quality control with FastQC and MultiQC
Trimming with Trimmomatic
Quality control of trimmed reads
READ MAPPING
Index reference
Perform alignment with BWA
Convert SAM files to BAM files
Sort and index BAM files
MAPPED READS POST-PROCESSING
Filter mapped reads
Remove duplicates
Left-align reads around indels
Recalibrate read mapping qualities
Refilter reads
VARIANT CALLING AND CLASSIFICATION
Convert BAM files to pileup
Call variants with VarScan Somatic
Merge vcf files with bcftools
VARIANT ANNOTATION
Functional genomic annotations with snpEff database
Genetic and clinical annotations with GEMINI
Further annotation from dbSNP, cancerhotspots, CIViC, and Cancer Genome Interpreter (CGI)
VARIANT REPORTING
Query GEMINI database for somatic variants
Gene-level annotation with information from UniProt Cancer Genes, CGI Biomarkers, and CIViC
Rearrangement of columns to get a well outlined, fully annotated gene report.

RESULTS AND DISCUSSION

Data Pre-processing

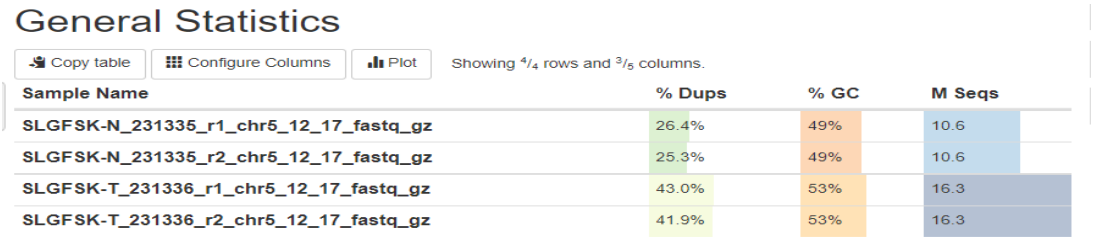
MultiQC analysis of both raw and trimmed reads were of good quality with high mean and per sequence quality scores and zero per base N content.





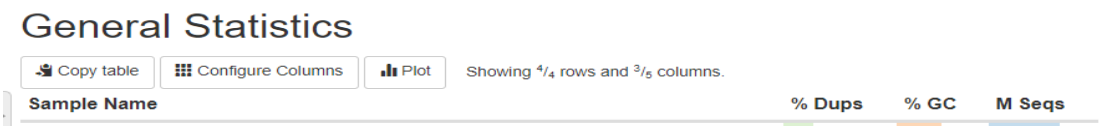
Figures 1 and 2. From MultiQC report on raw data, plots from trimmed data are similar

In both reports tumor cells had a higher proportion of duplicate reads in both forward and reverse strands, as well as a higher percentage GC content.



Sample Name	% Dups	% GC	M Seqs
SLGFSK-N_231335_r1_chr5_12_17_fastq_gz	26.4%	49%	10.6
SLGFSK-N_231335_r2_chr5_12_17_fastq_gz	25.3%	49%	10.6
SLGFSK-T_231336_r1_chr5_12_17_fastq_gz	43.0%	53%	16.3
SLGFSK-T_231336_r2_chr5_12_17_fastq_gz	41.9%	53%	16.3

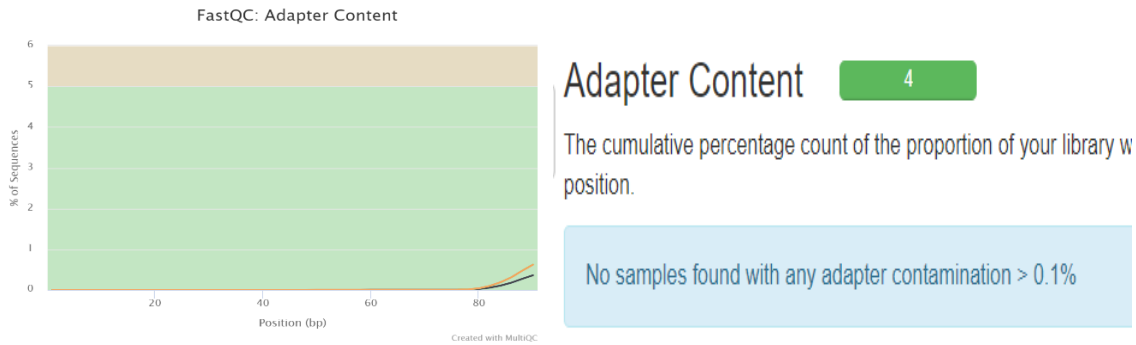
Figure 3. Raw data



Sample Name	% Dups	% GC	M Seqs
Trimmomatic on SLGFSK-N_231335_r1_chr5_12_17_fastq_gz _R1 paired	26.4%	49%	10.6
Trimmomatic on SLGFSK-N_231335_r2_chr5_12_17_fastq_gz _R2 paired	25.3%	49%	10.6
Trimmomatic on SLGFSK-T_231336_r1_chr5_12_17_fastq_gz _R1 paired	42.9%	54%	16.3
Trimmomatic on SLGFSK-T_231336_r2_chr5_12_17_fastq_gz _R2 paired	41.9%	53%	16.3

Figure 4. Trimmed data

The raw dataset had little adapter content while the trimmed data had no adapter contamination.



Figures 5 and 6 . Minimal adapter content in raw data, no adapters in trimmed data.

Download and view full MultiQC reports of raw data and trimmed data, or view plot images.

Read Mapping and Post-Processing

Using BWA, the trimmed reads of both normal and tumor tissue samples were mapped to the Human hg19 genome (reference genome) to determine the specific genomic location of the reads. To improve the quality of the mapped reads, further processing was performed. The dataset were filtered to retain only successfully mapped reads with mapping quality of at least 1 whose mate read was also mapped. Next, duplicate reads were removed with Samtools Markdup (Linux) or rmdup (Galaxy) and only reads with the highest mapping quality scores were preserved. The reads were then left-aligned around indels and recalibrated by capping the mapping quality of reads at 50. Subsequently, the reads were re-filtered to remove reads with poor mapping scores.

Variant Calling and Classification

In this stage, VarScan somatic was used to call and group variants in the high-quality mapped reads of normal and tumor samples. This resulted in identification of variant alleles and classification of these variants as either somatic, germline, or Loss of Heterozygosity (LOH) events. See output file of this stage here.

Variant Annotation and Reporting

Functional genomic annotations were added to the called variants using SnpEff. This added information about predicted effects of the variants (loss of function mutation, nonsense mutation) as well as the genes affected. View SnpEff annotated file.

Next, Gemini Tools suite was used to add genetic and clinical evidence-based annotations from dbSNP, Cancer Hotspot, CIViC, and CGI. These provide additional information such as the frequency of occurrence of the variants in human population and the clinical relevance of the variants. Gemini queries were then used to filter the variant report and most relevant annotations.

The advanced query constructor was used to generate a gene-centred report which contains annotations that are applicable to the whole gene rather than the variant. The advanced query merged information from the gene-detailed table (contains gene information) with the variants table in Gemini database to create the gene-centred report. The report was then further annotated with gene-level information from UniProt Cancer Genes, CGI biomarkers, and CIViC. The columns were subsequently re-arranged to give the final fully annotated gene report.

CONCLUSION

Somatic variant calling is useful for differentiating tumor-specific somatic mutations from germline mutations that are common to both normal and tumor cells, and for identifying LOH events. While variant callers are able to identify these variations, comprehensive variant annotation and gene-level annotation is crucial for understanding any set of variants.

LINKS TO REFERENCE TUTORIALS AND GITHUB REPOSITORY

- Fredrick-Kakembo/Somatic-and-Germline-variant-Identification-from-Tumor-and-normal-Sample-Pairs (github.com)
- Identification of somatic and germline variants from tumor and normal sample pairs (galaxyproject.org)
- LuminIz/HackBio_Genomics_Workshop-Week_3 (github.com)