Identification of somatic and germline variants from tumor and normal sample pairs

About

Stage 2 of the hackbio internship project conducted within team crick, a replication of a Linux/Galaxy pipeline to identify somatic and germline variants genes using tumor and normal samples.

Team members

Halimat- Group leader
@Aanuoluwa, @ Prafful, @Oreoluwa, @Toyosi ,@Seun

1.0 Background

Genomic variation occurs within the same species due to an influx of genetic mutation or recombination. Mutation occurs due to a random change in the order of DNA sequences in form of a deletion, duplication, inversion, or translocation of nucleotide bases. Mutation can be beneficial, neutral, or harmful. When harmful it leads to several diseases such as cancer, blood disorders etc. According to World Health Organisation (WHO), cancer is one of the leading causes of death worldwide, accounting for nearly one in 6 deaths in 2020. Cancer is caused by mutation in crucial genes which proliferate uncontrollable and, in the process, attacks the normal body tissues. These mutated genes can be inherited from the parent cells (germline mutation) or acquired after birth (somatic mutation)**. Cancerous cells often display multiple type of changes, such as Loss of Heterozygosity (LOH) in one or more genes. LOH is the loss of a normal copy of a gene or group of genes inside the affected cells and can be present in both somatic and germline cancer. LOH can occur during replication or cell division when a part of the gene is accidentally deleted. Hence, only one of the original alleles are found in the tumor. In cancerous cell detection, germline mutation is easily detected by comparing with a known sample genome reference. Whereas somatic mutation is not easy to analyse and requires both normal and tumor genes from patient.**

1.1 Aim and Objective

The objective of this task is to replicate a pipeline that reports somatic and germline variants site and genes affecting these sites using both normal and tumor tissues. This task involves annotating the variant sites against a standard human reference genome (HG19) and cancer-specific databases. This is aimed at providing valuable insight into the series of events causing tumor formation and effective therapeutic ways of managing cancers.

2.0 Methodology

2.1 Data Collection

The forward and reverse gene sequences of both the normal and tumor cells (reads from human chromosome 5,12 and 17) in fastq format, were obtained from the zenodo.org and loaded into the Linux terminal using the wget command.

```
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 reference human genome(hg19) used for the comparative mapping was also downloaded from Zenodo. The Zenodo is an open repository maintained by the CERN, which contains datasets, documents, and research materials. Cancer Hotspots & dbSNP (variant annotations), cgi_variants_position & CGI (variant and gene information) files were downloaded from the zenodo open repository for the variant annotation and reporting set

Software Packages & Platform
-Linux
-Galaxy

2.2 Analysis

2.2. 1 Quality Control

Fastqc and Multiqc tools were used to perform the quality check on the DNA sequence to prevent the occurrence of spurious variant calls due to low input quality. Fastqc analysed the tumor and normal dataset and Mutliqc combined the results of the Fastqc report

2.2.2 Trimming of Low-quality sequence

An automated script was created to run Fastp on the forward and reverse sequences of both the normal and tumor cells to remove the sequences that did not meet quality standards. Furthermore, trimming was carried out using trimmonatic. The ILLUMINACLIP function on the trimmomatic was used to remove the adapters sequence with a match accuracy of 10 and set in the PE specific/Palindrome mode. MINLEN was used to remove reads below 25 bases long and leading bases with low quality of 3 were moved.

2.3 Mapping and Postprocessing

The results from trimming were mapped against the reference genome using the BWA-MEM. Firstly, an index was built for the human reference genome (hg19) using the Burrow-Wheeler Aligner tool (BWA). The index allows the reference genome to efficiently search the genome during sequence alignment. After which, the trimmed reads were aligned using the indexed reference genome to generate a bam file (compressed file). The generated bam files were filtered, and PCR duplicates removed using the Rmdup tools. The PCR duplicates may result in a false Single Nucleotide Polymorphism (SNP).

Inconsistencies in the aligned sequences due to the different stages in the mutation process was removed using the BamLeftAlign tool. The reads were recalibrated using the Samtools Calmd tools. The outputs of the recalibrated reads were re-filtered using the bamtools and the map quality set to <=254 to discard the poor reads. During the recalibration, the CalMD may have set some quality scores to 255 which is set for undefined mapping quality allowing for poor reads.

2.4 Variant Calling and Classification

Varscan Somatic tool was used to identify variant alleles in both the tumor and normal mapped reads. The Variants were classified into germline, somatic and LOH events using solid classical statistics. The data obtained from the postprocessing set was converted to pileup before the Varscan somatic tool was applied. The normal purity was set to 1 and tumor purity to 0.5 which generated 2 vcf file (indel.vcf and snp.vcf). These Vcf files were merged into one file using the bcftools command on the linux terminal.

2.5 Variant Annotation

The data obtained from the variant calling step was transferred to GALAXY for further processing. The identified variants were annotated using the variant annotation data obtained in the data collection stages. The first step in the annotation process was functional annotating using the SnpEff eff tool on the output of the VarScan somatic step and the hg19 reference sample. Functional annotation helps us to know which genes are affected by the variant.

The second step involved genetic and clinical-based annotation using the Gemini (annotation and load) tools. The somatic status value was set to identify the germline, somatic and LOH event using 1,2,3 respectively. The output of the **SnpEff eff** was loaded to Gemini to start the annotation of variants and subsequently the output of each annotation step was applied to the variant annotated data.

2.6 Reporting selected subsets of Variants.

The GEMINI query tool was used to filter the somatic variants in the variant report table. The output of the query step was joined using the Join two fills to combine the annotations found

in the uniport cancer genes. Finally, the fully annotated report was rearranged to contain only important columns. The specified columns are indicated in the table below and their meanings

Column names	meaning
Gene	Gene name
Chrom	Chromosome location of the variant gene
Synonyms	Other similar names for the gene name
Hgnc Id	HGNC identifier
Entrez id	Entrez identifier
Rvis_pct	RVIS percentile value
Is_TS	Tumor suppressor gene indicator
Is_OG	Proto-oncogene indicator
In_cgi_biomarkers	Cancer biomarker indicator
Clinivar_gene_phenotype	Type of phenotype associated with gene
Gene_civic_url	URL link to cancer variant database
description	Description of the mutation present

Table 1: Final Annotation Column names. Adapter from The GEMINI database schema

3.0 Result and Discussion

The per base sequence content shows that the forward and reverse strands of the normal and tumor sequence have a length of 101bps and with equal G-C and A-T content at each base length. However, had the end they deviant slightly. Table 2 shows the general statistics, from which we can deduce that the tumor tissues have highly duplicated reads and GC contents. This is also represented in Figure 1 visually.

Sample Name	% Dups	% GC	M Seqs
SLGFSK-N_231335_r1_chr5_12_17	26.4%	49%	10.6
SLGFSK-N_231335_r2_chr5_12_17	25.3%	49%	10.6
SLGFSK-T_231336_r1_chr5_12_17	43.0%	53%	16.3
SLGFSK-T_231336_r2_chr5_12_17	41.9%	53%	16.3

Table 2: Multi-Qc General Statistics on the four reads

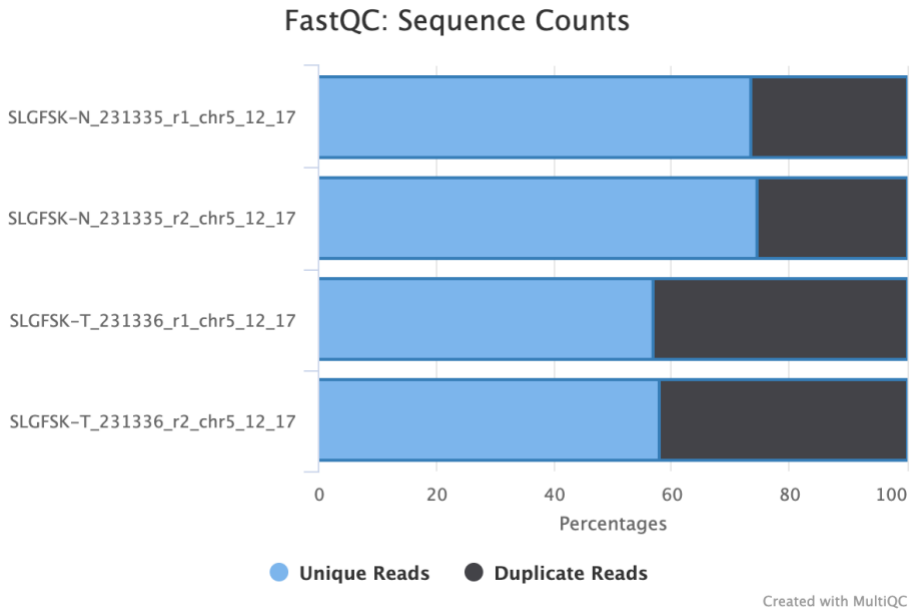


Fig 1: FastQC Sequence Counts

The four reads had quality score of 40 as shown in figure 2 which is greater than 30, hence it is very good quality score which is usually unlikely for most reads.

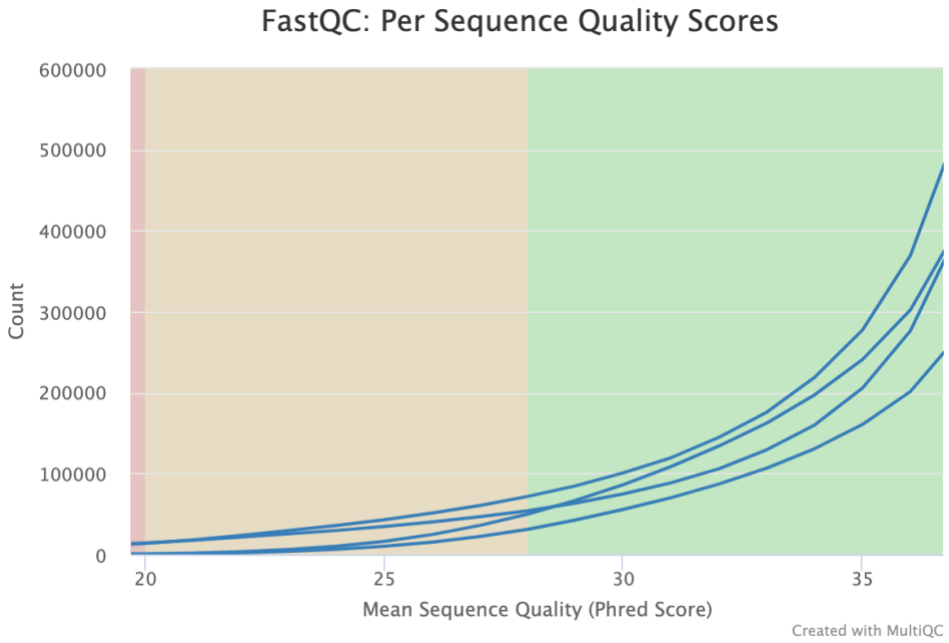


Fig 2: Per Sequence Quality Scores

The GC content plots (fig 3) of the forward and the reverse reads from both samples reveal a bimodal distribution as samples were not random, they were based on an arbitrary selection and had undergone exome enrichment. Before trimming, the normal sequence has an adapter sequence of 0.36% increase at the end of the sequence while the tumor sequence had an adapter sequence of 0.62% (fig4).

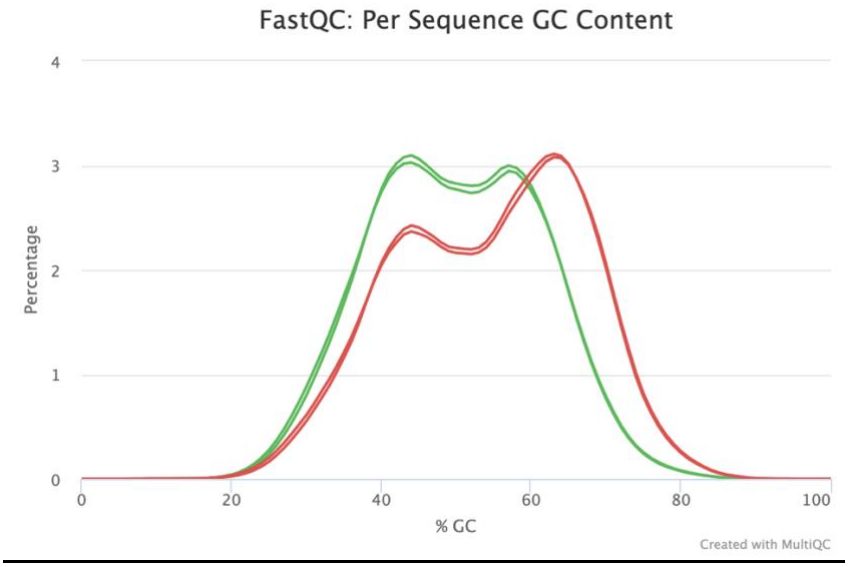


Fig 3: GC content Per Sequence

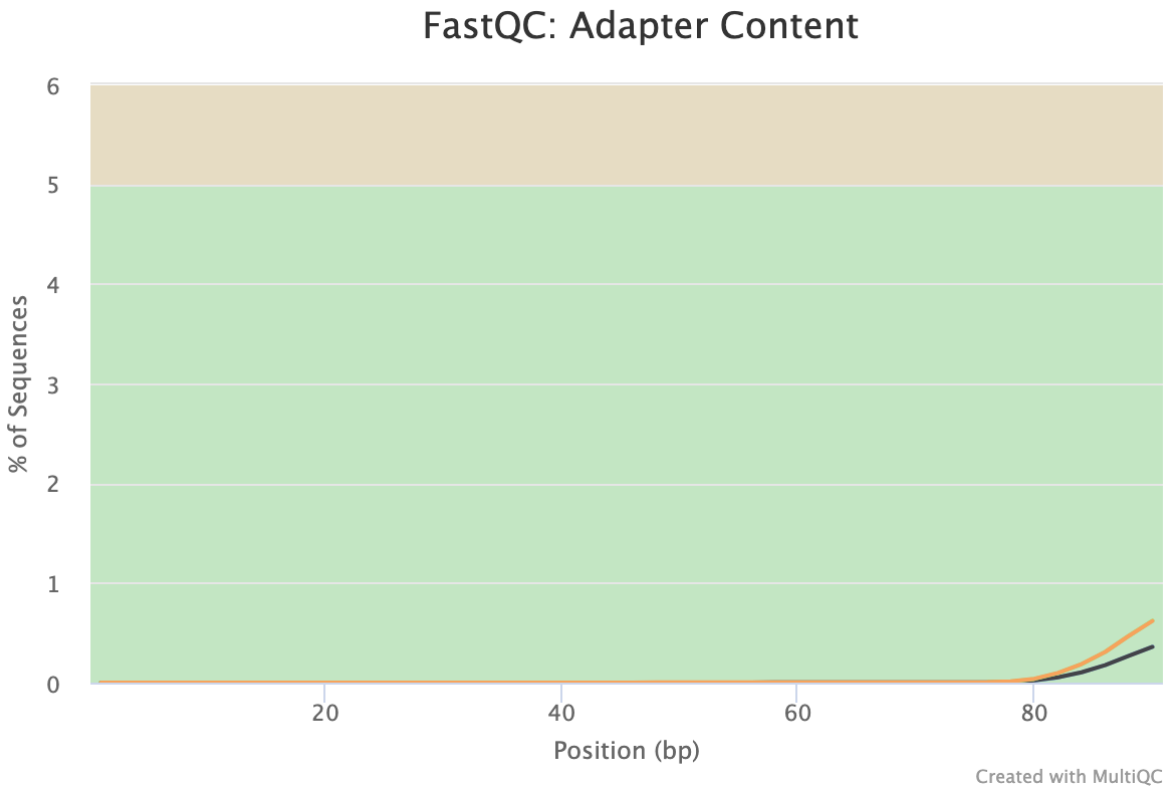


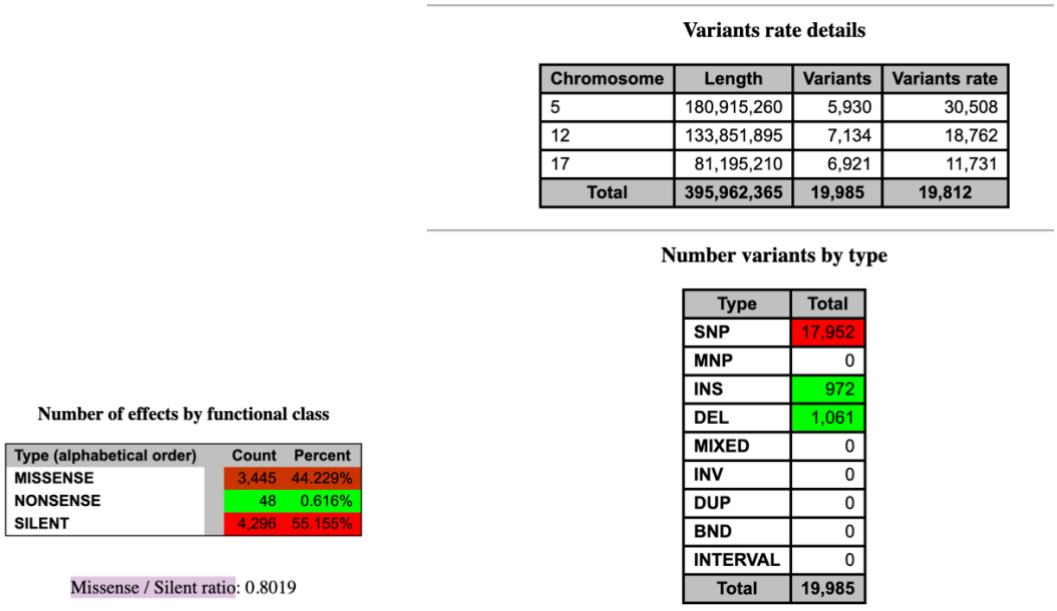
Fig 4: Adapter content, orange line represents tumor sequence and blue line for normal sequence

Trimming did not have significant impact on the average of the high-quality raw reads (Table 3), although there was a slight effect on %GC content and the duplicated reads (Fig5). content. No samples were found with any adapter contamination > 0.10%

Sample Name	% Dups	% GC	Length	M Seqs
Trimmomatic on SLGFSK-N_231335_r1_chr5_12_17_fastq_gz _R1 paired	26.4%	49%	98 bp	10.6
Trimmomatic on SLGFSK-N_231335_r2_chr5_12_17_fastq_gz _R2 paired	25.3%	49%	98 bp	10.6
Trimmomatic on SLGFSK-T_231336_r1_chr5_12_17_fastq_gz _R1 paired	42.9%	54%	98 bp	16.3
Trimmomatic on SLGFSK-T_231336_r2_chr5_12_17_fastq_gz _R2 paired	41.9%	53%	98 bp	16.3

Table 3: Multi-Qc General Statistics after Trimming

For detailed visualization, the quality control analysis report can be found here. The variant analysis result showed that 19,845 variants (Fig 5) were found in the test samples and the variants had a Missense/Silent ratio of 0.8019. With silent mutation (55.155%) accounting for the most type of mutation present in the genome samples. Silent mutations a type does not alter the amino-acid sequence hence it does not have an observable effect on the organism's phenotype. Therefore, we can deduce that majority of the mutations did not have effect on the amino-acid sequence or phenotype. However, Missense mutation (44.229%) appeared as the second most frequent mutation, this type of mutation changes the identity of a codon from one amino acid to another, thereby altering the structure and function of the protein. This shows that while majority of the mutations were silent, there is also a high level of mutation which had effect on the amino-acid and structure. The full snpEff variant analysis report can be found here here.



Variants rate details

Chromosome	Length	Variants	Variants rate
5	180,915,260	5,930	30,508
12	133,851,895	7,134	18,762
17	81,195,210	6,921	11,731
Total	**395,962,365**	**19,985**	**19,812**

Number variants by type

Type	Total
SNP	17,952
MNP	0
INS	972
DEL	1,061
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	19,985

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	3,445	44.229%
NONSENSE	48	0.616%
SILENT	4,296	55.155%

Missense / Silent ratio: 0.8019

Fig 5. Variant Effect, Count and Type

In the final annotation report 42 genes with different phenotypes with 13 clinvar_phenotypes including nonsyndromic_otitis_media, coronary artery spasm, prostate cancer/ oxidative phosphorylation deficiency, acute_myeloid_leukemia, bladder cancer, klippel-feil_syndrome_2. klippel-feil_syndrome_2 etc. These variants are observed across chromosome 5,12 and 17.

Conclusion

This analysis demonstrates a good approach in detecting variants in mutated cells. Thereby necessitating the importance of computational analysis in determine different types of mutations in tumor cells in both somatic mutation and germ-line mutation. This approach helps for easy identification for similarities and differences between the cancer-specific variants in somatic/germline mutation and normal tissue cells. The analysis revealed mutation in the tumor suppressor genes on chromosome 5 (APC genes) and Chromosome 17 (TP53) which is a cancer biomarker. Mutation was also found in ELAC2, RNF213 (chromosome 17) which is a proto-oncogene which causes normal cells to become tumours when mutated. Mutation in RNF213 is associated cerebrovascular diseases and one of causes of stroke in children. Mutations on chromosomes 5 was found on VCAN and TTC37 which causes malignant tumor in the prostrate. TTC37 mutation has been associated with trichohepatoenteric_syndrome which is a condition that affects the kidney, liver, and intestine. While VCAN mutation has been associated with Wagner syndrome.

Links to pipeline reproduced

Galaxy Somatic and Germline Variant Pipeline Somatic-and-Germline-variant-Identification-from-Tumor-and-normal-Sample

Links to Github Repo

https://github.com/Tabetaa/HackBio22-TeamCrick/tree/main/stage_2