Identification of somatic and germline variants from tumor and normal sample pairs - Reproduced by Halimat Chisom (Team Crick)

Introduction

Mutation is a change in the DNA sequence of an organism that occurs in the form of an insertion, deletion, etc as a result of biological and/or chemical activities. Mutations can be beneficial, harmless, or severe in expression. When it is inherited by an offspring from its parent, it is called germline mutation and usually easier to understand by comparing with a reference genome sequence. When it is acquired after birth and only found within certain cells in the body, it is referred to as somatic mutation. Somatic mutation often takes longer to analyse and understand as it requires both normal and tumor genes from the affected individual to be compared with a reference genome. Loss of heterozygosity (LOH) - the loss of one normal copy of a gene or group of genes - is a form of germline or somatic mutation that is common to complex and severe diseases like cancer.

Aim and Relevance of the Task

This task aims to replicate a pipeline or workflow that identifies somatic and germline variants from normal and tumor tissue through data processing and annotation. The results from this analysis will help with early diagnosis of tumors related to this tissue as well as the development of effective therapeutic strategies.

Pipeline of Choice

Galaxy pipeline on `usegalaxy.eu`

Methodology

Data Collection

A new history was created on the Galaxy workspace with the name Bioinformatics analysis. Target genes for this project are located on chromosomes 5, 12, and 17. To start the analysis, 5 datasets were uploaded with respective URLs from the zenodo.org database: the forward and reverse gene sequences of the three chromosomes on the normal tissue, the forward and reverse gene sequences of the three chromosomes on the tumor tissue, and the reference human genome sequence (hg19) for the corresponding chromosomes.

Once the uploading was complete, the normal and tumor datasets were tagged for easy identification.

Data Processing - Quality Control and Mapping

Quality control

FastQC and MultiQC packages/tools on Galaxy were used to check the quality of each dataset. FastQC assessed individual normal and tumor samples and MultiQC aggregated the results generated by FastQC.

Trimming

This was carried out separately on each sequence pair (normal and tumor) by running the Trimmomatic package/tool on Galaxy. The ILLUMINACLIP function was used to remove TruSeq3 adapter sequences with a minimum length of 8, a match accuracy of 10, and a setting to keep both sequence reads.

To remove low quality reads, we also set conditions to remove 3 bases from the start of the read (HEADCROP), maintain the minimum threshold quality at 10 bases (TRAILING), and to drop reads below a minimum of 25 bases (MINLEN). These operations were performed in the specific order for both the normal and tumor sequences.

Quality control checks were performed on the trimmed sequences to confirm that adapter sequences have been completely removed.

Mapping and post-processing

The resulting reads for the normal and tumor data from the trimming stage were aligned separately with the reference human genome using the BWA MEM tool on Galaxy for maximal exact match. This process reads the position of the sample reads on the reference genome.The generated bam files were filtered for mapping quality (mapquality >= 1) and successful paired mapping (isMappped and isMateMapped = Yes). The filtered outputs were subjected to RmDup to remove duplicates that may have occurred during the PCR stage of sequencing the reads.

Left alignment was carried out using BamLeftAlign tool to remove any inconsistencies in the aligned sequences that may arise from insertions and deletions. Recalibration of the deduplicated normal and tumor reads map quality was performed using samtools CalMD which was set to a coefficient of 50 as the quality cap of poorly mapped reads. Finally, the recalibrated outputs of the normal and tumor tissue reads were refiltered with the reading map quality on phred scale set to <=254 to remove poor reads from the data since a score of 255 is reserved for undefined mapping quality.

Variant Calling and Classification

VarScan somatic tool was used for this step to detect variant alleles in the normal and tumor mapped read, classify the mutations into germline, somatic, or LOH, and call sample genotypes at mutation sites. For this process, purity content for the normal read was set to 1 and that of the tumor read was set to 0.5. Minimum base quality was set to 28 because of the relatively high quality of both reads from the quality control checks, and minimum map quality was set to 1. Both normal and tumor samples were run with the reference genome sample at the same time.

Variant annotation and reporting

This step involved adding annotations to the identified variants using variant annotation datasets from cancer hotspots, cancer biomarkers database, cancer variant curation database, NCBI, and UniProt. The variant and gene-level annotation files were imported from zenodo directly into the Galaxy workspace in bed, vcf, and tabular formats.

Functional annotation was added using the SnpEff eff tool which also predicted the effect of each annotated variant. Genetic and clinical evidence-based annotation was added using the GEMINI load and annotate tools. This stage helps to identify the somatic, germline, and LOH variants using customized annotation extraction recipe settings and the imported annotation datasets as reference. The GEMINI Query tool was used to filter out somatic variants based on SQL formatted filters on certain columns in the variant report table. From the results generated by this query, the Join two files command was used to combine the annotations found in the imported annotation data sources.

Finally, the fully annotated gene report was rearranged based on column header names and unwanted columns were dropped.

Results and Discussion

The quality control checks of the raw sequence reads showed that all four reads (normal and tumor) had equal length of 101 bps. The reads from the tumor tissue have high duplicated reads and GC content (Table 1). Figure 1 and 2 show the adapter content level and overall quality status checks of the raw read respectively, while Figure 3 shows the resulting quality (reduced) after trimming adapters. It was also observed that the length of the reads reduced and varied after trimming off the adapters because some reads were lost in the process.

It was also observed that trimming had a slight effect on the %GC content as well as the count of duplicated reads (Table 2). For detailed visualization, the quality control analysis report can be found here.

Table 1: MultiQC summary statistics on 4 sample reads

Sample Name	% Dups (duplicated reads)	% GC (GC content)	M Seqs (Total sequence Millions)

SLGFSK-N_231335_r1_chr5_12_17_fastq_gz (Normal)	26.4%	49%	10.6
SLGFSK-N_231335_r2_chr5_12_17_fastq_gz (Normal)	25.3%	49%	10.6
SLGFSK-T_231336_r1_chr5_12_17_fastq_gz (Tumor)	43.0%	53%	16.3
SLGFSK-T_231336_r2_chr5_12_17_fastq_gz (Tumor)	41.9%	53%	16.3

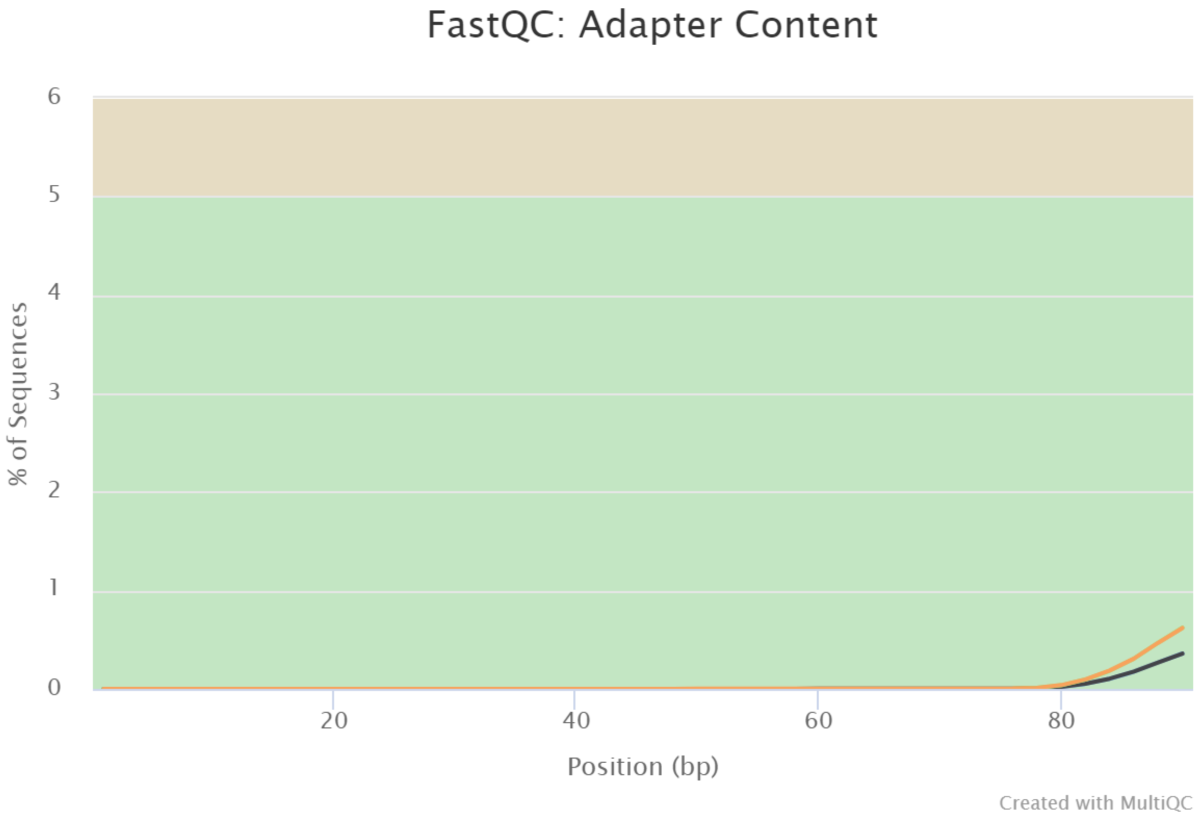

Figure 1: Plot showing adapter content level in each sample. Adapter content for the tumor sample is shown in orange and that of the normal sample is in black.

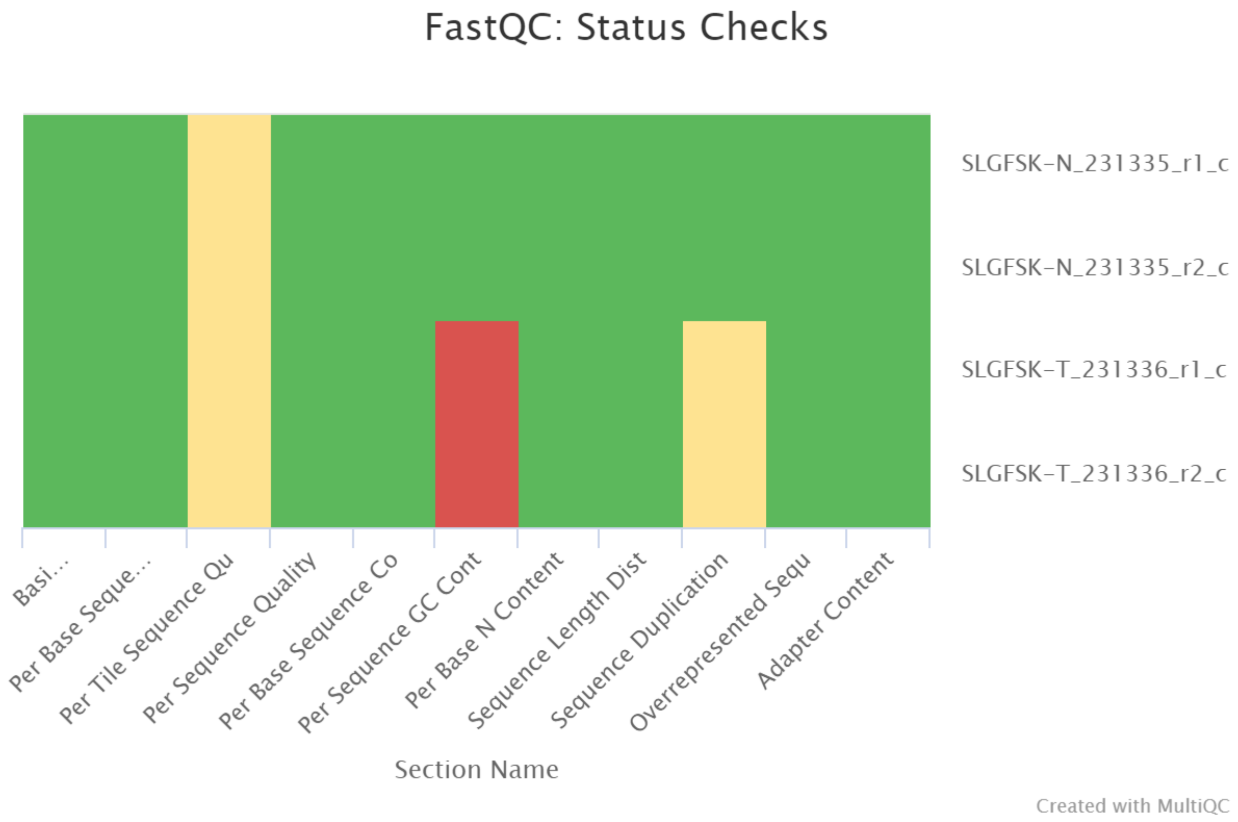

Figure 2: Summarized quality control status checks for untrimmed reads. Green represents entirely normal, yellow for slightly abnormal, and red for very unusual. First two labels represent the forward and reverse reads from the normal tissue and the last two represent the forward and reverse reads from the tumor tissue.

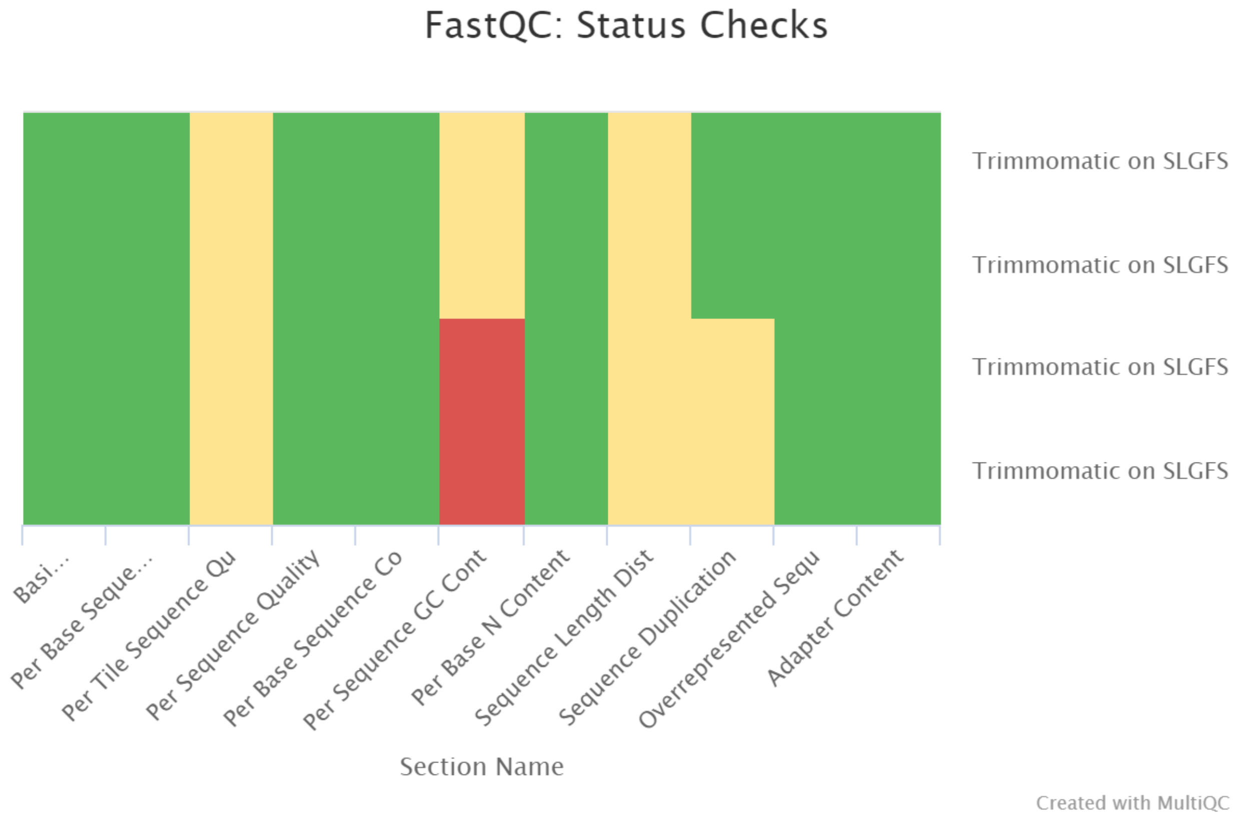

Figure 3: Summarized quality control status checks for trimmed reads. Green represents entirely normal, yellow for slightly abnormal, and red for very unusual. First two labels represent the forward and reverse reads from the normal tissue and the last two represent the forward and reverse reads from the tumor tissue.

Table 2: MultiQC general stats on trimmed sample reads

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

The results generated from the variant calling and annotation process can be found here: final annotation result. Each column has a shortened title for brevity. The specified columns and their interpretations are: gene (gene name), chrom (chromosome location of the gene variant), synonym (other gene names), 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), clinvar_gene_phenotype (phenotype associated with gene), gene_civic_url (url to cancer variant database), and description (description of the mutation).

Source of the abbreviation meanings: GEMINI documentation database

41 genetic variants with different phenotypes, including coronary artery spasm, deficiency in oxidative phosphorylation, leukaemia, and prostate cancer. For the mutations that had a solid description, missense mutation appeared to be the frequently occurring type of mutation. Also, chromosome 5 had the majority of the variant genes (19), followed by chromosome 12 (14 variants) and 17 (8 variants). Two tumor suppressor genes, APC on chromosome 5 and TP53 on chromosome 17 were observed to have undergone mutation, of which TP53 was a cancer biomarker.

Conclusion

This analysis was carried out completely on the Galaxy workspace. It shows the importance of computational tools in biomedical analysis involving understanding the molecular basis of diseases and deriving better therapeutic approaches to such conditions. Some of the challenges encountered during the analysis was navigating the Galaxy work server and understanding the runtime for each step and the process of accessing generated results.

Analysis on the LINUX pipeline is still ongoing and the script will be available on the stage tow folder of this workshop on GitHub.