Submitted by,

Krithika Subramanian

Task 2: NGS Data Analysis

Abstract:

In this task, I identified somatic mutations in a tumor sample by comparing it against a normal tissue sample using the GATK-Mutect2 tool. A total of 57 genetic mutations were detected across 24 targeted regions, including 32 single nucleotide variants (SNVs). The analysis revealed that somatic mutations required the highest confidence thresholds (0.40) and read depths (~398,000 reads per million) for reliable detection. For SNVs, the tumor median background allele frequency (AF) was 0.0259, while the normal median background AF was 0.0104, requiring confidence thresholds of 0.451 and 451,206 reads per million. These results underscore the importance of setting stringent thresholds for accurate somatic mutation detection.

Dataset description:

I obtained the **Pupil Bio NGS Dataset** from a designated online repository. The dataset comprises four fastq.gz files, representing two distinct biological samples: the Normal/Control sample (PA221MH-lib09-P19-Norm) and the Tumor sample (PA220KH-lib09-P19-Tumor). The dataset is in a paired-end sequencing format, as indicated by the file names, specifying **Read 1 (R1)** and **Read 2 (R2)** for a single sequencing lane (**L001**) generated using an Illumina platform. Additionally, the dataset reflects the pooling or sequencing of multiple samples, identified by **"S1"** for the Normal sample and **"S2"** for the Tumor sample.

I used a reference CSV file containing 100 targeted regions from the sequences of 41 genes to create a corresponding reference file in fastq format.

Software:

For the analysis of this dataset, the following tools and software were required:

1. **FASTQC** – For quality assessment of fastq files.
2. **BWA** – To align sequencing reads to the reference genome.
3. **Samtools** – For conversions and sorting sequence alignment/map (SAM) files.
4. **GATK (Genome Analysis Toolkit)** – To perform variant discovery and genotyping.

5. **Bcftools** – For handling variant calling format (VCF) files.

2.1 Quality Control:

Performing a quality check on the fastq files was a critical step in preprocessing the sequencing data. I assessed the quality of each sample's paired-end reads in fastq format using the FASTQC tool, which generated a consolidated summary in an .html file.

This analysis provided me with key insights into the dataset's properties, including read lengths, quality scores (Phred scores indicating the probability of error), GC content, base duplication levels, and adapter content. The results showed 4.77 million reads for the Tumor sample and 5.15 million reads for the Normal sample, both with a consistent read length of 151 bases. Additionally, the GC content percentage was 48% for the Tumor sample and 49% for the Normal sample, as summarized in **Table 1**.

Table 1: Summary of quality check results for Normal and Tumor samples

Sample	**File name**	**Total Sequences**	**Read length (bp)**	**%GC**
Tumor	PA220KH-lib09-P19-Tumor_S2_L001_R1_001.fastq.gz	2,384,174	151	48
	PA220KH-lib09-P19-Tumor_S2_L001_R2_001.fastq.gz	2,384,174	151	48
Normal	PA221MH-lib09-P19-Norm_S1_L001_R1_001.fastq.gz	2,574,922	151	49
	PA221MH-lib09-P19-Norm_S1_L001_R2_001.fastq.gz	2,574,922	151	49

Per base sequence quality check:

I analyzed the per-base sequence quality to evaluate the range of Phred score quality values across all bases at each position. Figure 1A (Tumor sample) and Figure 1B (Normal sample) show the quality graphs for the forward and reverse reads. In these graphs, the x-axis

represents the position of the bases within the reads, while the y-axis displays the Phred scores, ranging from 0 to 36. Higher Phred scores indicate better base calls, with the background colors representing thresholds: Green for very good quality (Phred score above 28), Orange for reasonable quality (Phred score between 20 and 28), and Red for poor quality (Phred score below 20). The blue line indicates the overall median Phred score, which exceeded 30 for all reads in both samples. In the quality graphs, I observed that forward and reverse reads consistently displayed a length of 151 bases with high-quality scores. Although sequencing ends often show lower quality, the median score remained above 30 for reverse reads in both samples. Based on these results, I did not remove any reads and retained all data for downstream analysis.

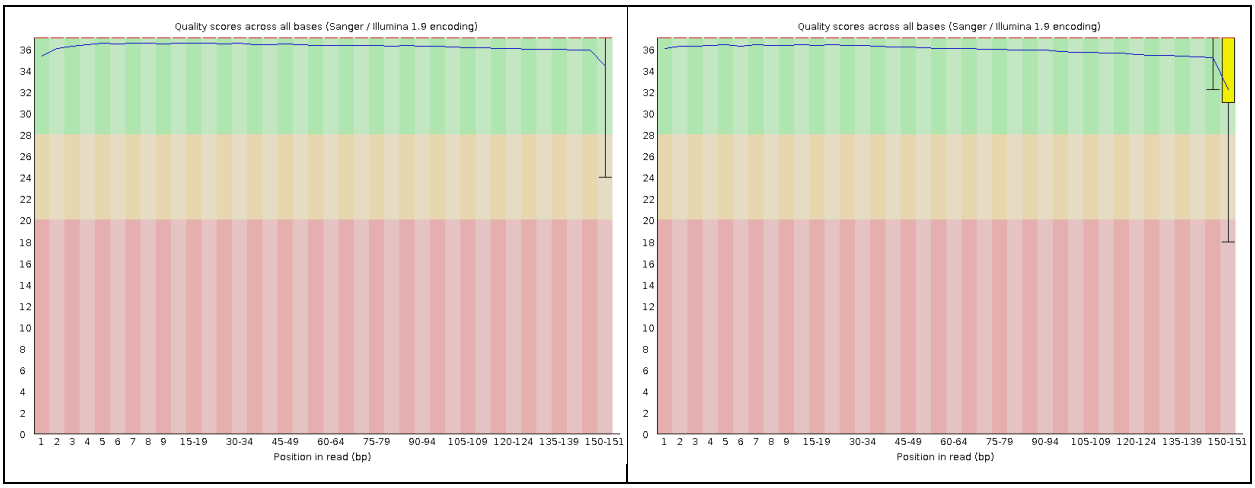

Figure 1A: Per base quality of forward (left) and reverse (right) reads of Tumor samples

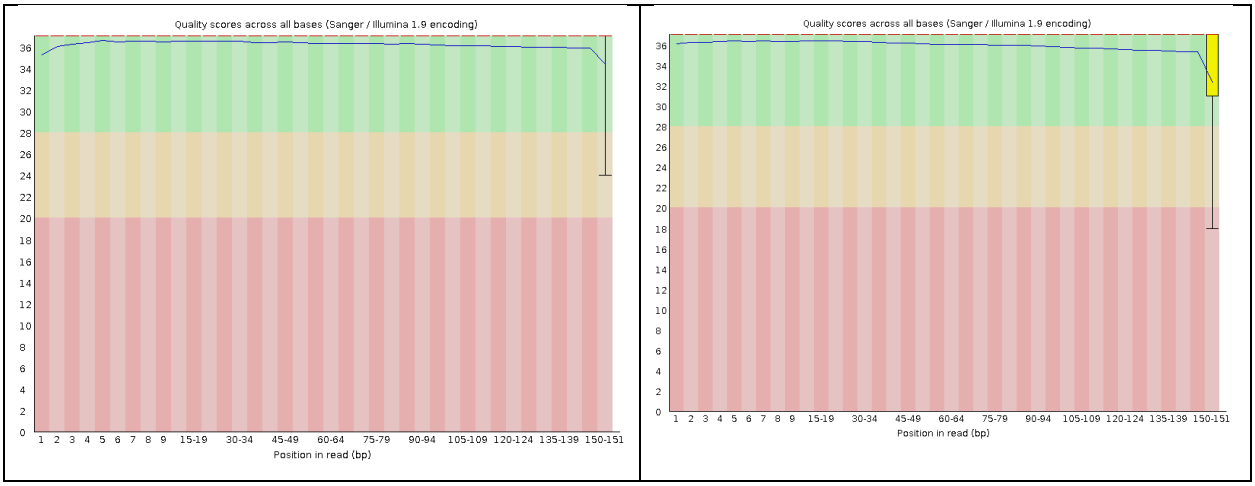

Figure 1B: Per base quality of forward (left) and reverse (right) reads of Normal samples

Per-tile sequence quality:

I analyzed the per-tile sequence quality for both samples (Figure 2), including forward and reverse reads, and found it to be consistent and clear. This analysis allowed me to examine quality variations across different tiles of the sequencing flowcell in detail. The results showed consistent quality across all tiles, with no significant deviations. This uniformity suggests that the sequencing process was stable and free from technical issues affecting specific regions of the flowcell.

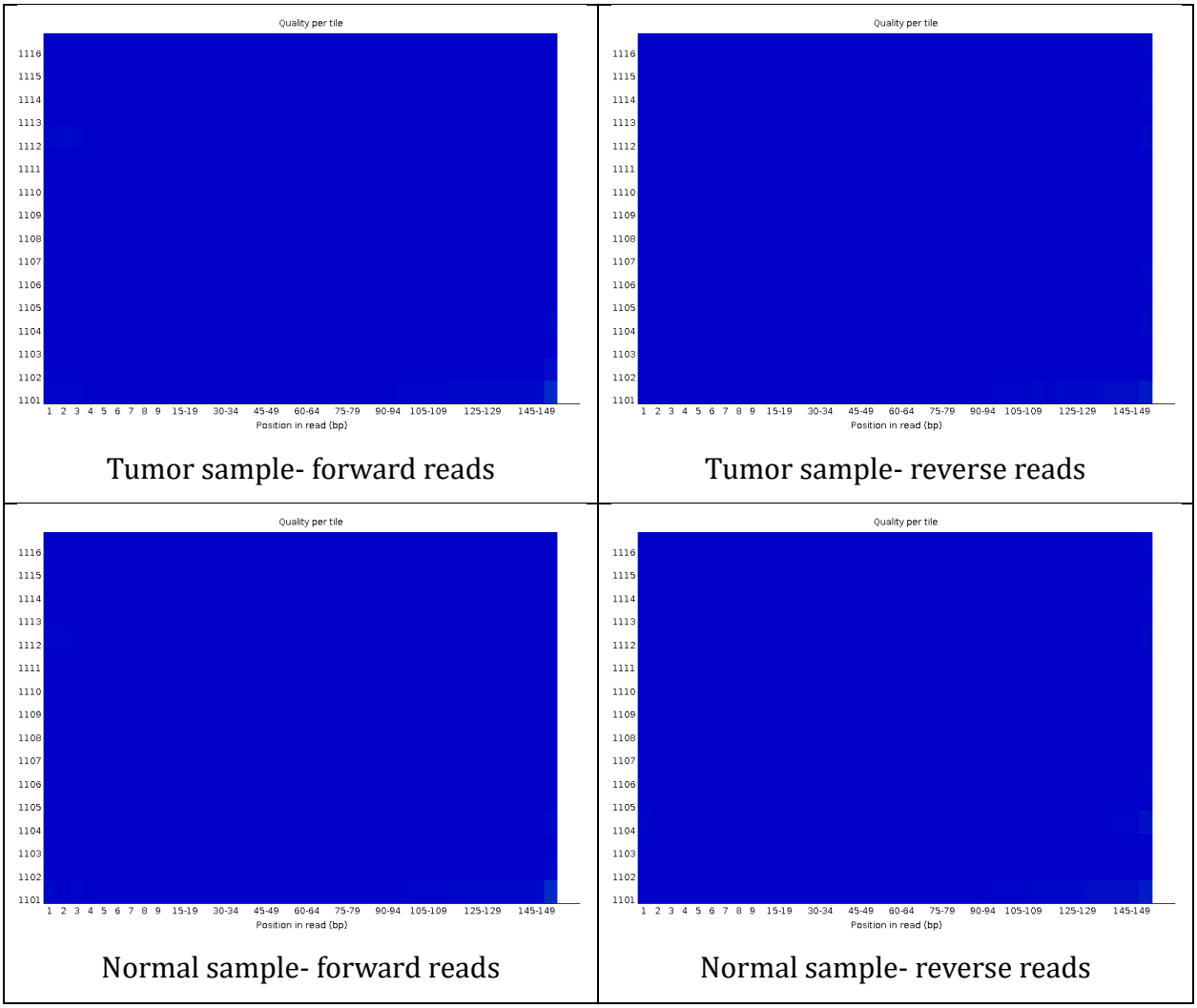

Figure 2: Per tile sequencing quality of tumor and normal samples

The **Sequence duplication level** graph shows that most sequences are unique (duplication level = 1), highlighting high sequencing complexity across both samples. I observed duplication levels of 1.36% and 1.47% for forward reads (R1), while reverse reads (R2)

showed slightly higher levels at 2.57% and 2.67%. These low duplication rates demonstrate minimal redundancy, making the dataset suitable for downstream analysis.

In the analysis of **overrepresented sequences**, I found that less than 5% of the dataset consisted of such sequences. This indicates minimal technical artifacts, such as adapter contamination or library preparation bias. The overall quality checks showed low levels of technical artifacts and sequencing errors. Based on these findings, I used the entire dataset for downstream analysis without exclusions.

2.2 Alignment:

The variant discovery workflow relies on sequence data in the form of reads that are aligned to a reference genome. I converted the CSV file into FASTA format by using the ID and Sequence columns, where the ID served as the header. I then indexed this FASTA file using the "bwa index" software.

Next, I mapped the sequencing reads to the reference genome using the "BWA MEM" software, producing a file in SAM (Sequence Alignment Mapping) format. I converted the SAM file into a BAM (Binary Alignment Mapping) file and sorted it using the SAMTOOLS software.

To evaluate alignment metrics, I used "SAMTOOLS flagstat", which provided detailed insights such as the number of reads that passed quality control (QC), the number and percentage of properly paired reads, and the number of singletons (reads mapped only in forward or reverse orientation). The results, summarized in the table, revealed that more than 97% of reads in both samples were properly paired and mapped to the reference target regions (tumor genes). Additionally, I observed no duplicate reads in the dataset.

Table 2: Alignment metrics for normal and tumor samples

Samples	**Tumor**	**Normal**
QC-passed reads	4810831	5202434
Duplicates reads	0	0
Mapped %	98.50%	97.90%
Properly paired %	97.87%	97.04%
Singletons reads %	0.15%	0.15%

2.3 Mutation calling:

To call somatic mutations, I used the **GATK-Mutect2** software. This process required a reference dictionary file, which I generated using the **gatk CreateSequenceDictionary** tool. The GATK-Mutect2 caller identifies somatic short mutations through local haplotype assembly, including **single nucleotide alterations (SNVs)** and **small insertions and deletions (indels)**.

The Mutect2 caller employs a **Bayesian somatic genotyping model**, which differs from the original MuTect algorithm developed by Cibulskis et al. (2013) and incorporates the assembly-based machinery of **HaplotypeCaller**. Notably, **Mutect2 v4.1.0.0** and later versions enable joint analysis of multiple samples, enhancing the functionality for somatic mutation detection.

Mutation Analysis Results

I identified **57 genetic mutations** in **24 targeted regions**, spanning **20 known tumor genes**. Of these mutations:

- **32** were **single nucleotide variants (SNVs)**,
- **19** were **insertions or deletions (INDELs)**, and
- **6** were **multiallelic variants**.

All 57 mutations were present exclusively in the tumor samples, with alternate alleles occurring either in one copy (**0/1**) or two copies (**1/1**). The average depth for these mutations exceeded **970**, indicating that these variants are true mutations supported by high sequencing depth.

Median Background Mutation Level

To calculate the **median background mutation level**, I assessed the dataset for sequencing errors and biases that could mimic true mutations. I performed separate calculations for the **entire dataset**, **SNVs**, **INDELs**, and **multiallelic variants**, ensuring detailed insights into each mutation type.

Table 3 summarizes the results:

- **SNVs**: Tumor median background AF was **0.0259**, while the normal median background AF was **0.0104**. SNVs required the highest tumor confidence threshold of **0.451**, corresponding to **451,206 reads per million** for reliable detection.

- **INDELs**: Tumor median background AF was **0.0311**, and the normal median background AF was **0.0236**, requiring a tumor confidence threshold of **0.361** and **361,126 reads per million**.
- **Multiallelic variants**: Tumor median background AF was **0.0273**, and the normal median background AF was the highest at **0.0266**, requiring **111,650 reads per million** with a tumor confidence threshold of **0.112**.

Category	Tumor Median Background AF (%)	Normal Median Background AF (%)	Tumor Confidence Threshold (%)	Normal Confidence Threshold (%)	Tumor Reads per Million	Normal Reads per Million
Whole Data	2.61	1.56	39.84	6.40	398396.49	63954.79
SNV	2.59	1.04	45.12	4.77	451206.57	47737.04
Indel	3.11	2.36	36.11	7.48	361126.00	74771.72
Multiallelic	2.73	2.66	11.17	9.11	111650.04	91083.64

Conclusion

The analysis reveals that SNVs require the highest confidence thresholds and reads per million, particularly in tumor samples, reflecting their variability and complexity in detection. INDELs show the highest background allele frequency in both tumor and normal samples, indicating greater sequencing noise and the need for more stringent thresholds. Multiallelic variants exhibit the highest normal background AF, suggesting increased sequencing artifacts or complexity, while their confidence thresholds are relatively lower than SNVs and INDELs. These findings highlight the varying challenges in confidently

detecting different variant types, underscoring the importance of adjusting detection thresholds based on the variant type and background noise