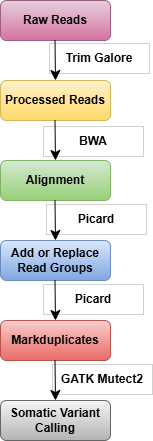
****

**Fig No.1 Somatic Variant calling workflow**

**1. Raw Data Quality Check**

The quality of raw sequencing data was assessed to ensure high-quality reads suitable for downstream analysis. The FastQC tool was used to evaluate metrics such as per-base sequence quality, GC content, sequence duplication levels, and adapter contamination.

**Table No. 1** Raw reads Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample Name** | **Total Sequences** | **Total Bases** | **%GC** | **per-base quality** | **read duplication levels** |
| PA220KH-lib09-P19-Tumor\_S2\_L001\_R1 | 2384174 | 360 Mbp | 48 |  |  |
| PA220KH-lib09-P19-Tumor\_S2\_L001\_R2 | 2384174 | 360 Mbp | 48 |  |  |
| PA221MH-lib09-P19-Norm\_S1\_L001\_R1 | 2574922 | 388.8 Mbp | 49 |  |  |
| PA221MH-lib09-P19-Norm\_S1\_L001\_R2 | 2574922 | 388.8 Mbp | 49 | 36 (PASS) | (>10k) FAIL |

**2. Adapter Trimming**

Adapter sequences and low-quality bases were removed from the raw reads to improve the accuracy of alignment and variant calling. The trimming was done by Trim-galore tool to process ensured that only high-quality reads with a minimum length of 20 base pairs and a quality threshold of 30 were retained for analysis. **Table No. 2** Processed reads Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample Name** | **Total Sequences** | **Total Bases** | **%GC** | **per-base quality** | **read duplication levels** |
| PA220KH-lib09-P19-Tumor\_S2\_L001\_R1 | 2356349 | 328.1 Mbp | 48 |  |  |
| PA220KH-lib09-P19-Tumor\_S2\_L001\_R2 | 2356349 | 326.9 Mbp | 48 |  |  |
| PA221MH-lib09-P19-Norm\_S1\_L001\_R1 | 2543536 | 351.1 Mbp | 48 |  |  |
| PA221MH-lib09-P19-Norm\_S1\_L001\_R2 | 2543536 | 350.1 Mbp | 48 |  |  |

**3. Reference Genome Preparation**

The human reference genome (hg38) was downloaded from the NCBI database. Indexing of the reference genome was performed to facilitate efficient alignment of sequencing reads and reference indexing was done by using the Samtools.

**4. Reference-Based Alignment**

The trimmed reads were aligned to the human reference genome using a BWA alignment tool. This step generated BAM files, which are binary representations of the aligned sequences. The BAM files were sorted to organize the data for subsequent steps.

**5. Mapping Statistics**

Mapping statistics were generated by using samtools flagstat to evaluate the alignment quality. Metrics such as the total number of reads, percentage of mapped reads, properly paired reads, and duplication levels were reviewed. This step confirmed the high mapping efficiency and the quality of the alignment.  
  
**Table No. 3** Mapping Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Name** | **Total Reads** | **Mapped Reads** | **Mapping percentage** | **Reference Used** |
| PA220KH-lib09-P19-Tumor\_S2\_L001 | 4759221 | 4750798 | 99.82 | Homo sapiens (GCF\_000001405.40) |
| PA221MH-lib09-P19-Norm\_S1\_L001 | 5144500 | 5125717 | 99.63 |

**6. Addition of Read Groups**

Read groups were added to the BAM files by using picard tool to assign metadata such as sample identifiers, library information, platform, and sequencing run details. This information is essential for downstream analysis, particularly in distinguishing between tumor and normal samples.

**7. Marking Duplicates**

Duplicate reads, likely arising from PCR amplification during library preparation, were identified and removed. This step done by picard tool to minimized biases in the data, ensuring accurate variant detection.

**8. Dedup BAM File Indexing**

Indexing of dedup BAM files was performed to enable rapid access to specific genomic regions during variant calling. Indexed BAM files allow efficient processing of large datasets.

**9. Creation of Reference Dictionary**

A reference dictionary was generated for the human genome by using gatk CreateSequenceDictionary to standardize its use across the workflow. This dictionary file is required for variant calling tools to interpret the reference genome structure.

**10. Somatic Variant Calling**

Somatic variant calling was performed by using gatk Mutect2 tool designed for identifying mutations in cancer genomes. Tumor and normal samples were analyzed together to detect tumor-specific variants. The output of this step was a Variant Call Format (VCF) file containing somatic variants along with their genomic locations, allele frequencies, and quality metrics.

**Identify somatic mutations present in the cancer sample but absent in the normal tissue** -  
Mutect2 is a powerful somatic variant caller from the Genome Analysis Toolkit (GATK), specifically designed to detect somatic mutations in tumor samples by comparing them to matched normal samples. It identifies single nucleotide variants (SNVs) and small insertions and deletions (indels) in cancer genomes.   
**Interpretation of Mutect2 Output**:  
In the **normal sample**, all genotype calls are 0/0 or 0|0, indicating that the somatic variants are not present. Additionally, the allele frequencies for all variant positions are very low, suggesting noise or sequencing errors.

In the **tumor sample**, all variant positions show genotype calls as 0/1 or 0|1, indicating that the variants are heterozygous. The allele frequencies are higher in the tumor sample compared to the normal sample, confirming the presence of somatic variants in the tumor and their absence in the normal sample.  
  
**Use the normal tissue to calculate the median background mutation level**The background mutation level is an important metric in somatic variant analysis, particularly for distinguishing true somatic mutations from noise or artifacts.Background mutation level calculated by using below formula**:  
background\_mutation\_level = data['Normal\_SM\_AF'].median()**here Normal\_SM\_AF represents Allele frequency for normal variant. In our data background mutation level was 0.0419.   
**Reads per Millon**Calculates **Reads Per Million (RPM)** to normalize the read depth of each genomic position in the normal sample. RPM provides a standardized metric for comparing read coverage between positions or samples, enabling more accurate variant calling.  
Sets thresholds for both allele frequency and sequencing depth to confidently call somatic mutations, ensuring that mutations are both biologically significant (by comparing against background mutation levels) and supported by sufficient data (based on RPM). confidence\_threshold\_AF was 0.0419 and confidence\_threshold\_RPM 1000.  
**Identify confidently called somatic mutations based on both AF and RPM**

data['Confident\_Somatic\_Mutation'] = (

(data['Normal\_SM\_AF'] > confidence\_threshold\_AF) &

(data['Normal\_SM\_RPM'] > confidence\_threshold\_RPM)

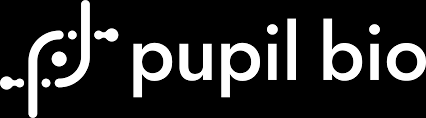
)

**True**: If the mutation passes both the allele frequency and RPM thresholds, indicating a confident somatic mutation.

**False**: If the mutation does not meet either of the thresholds, indicating that it is not confidently called as somatic.

**Report Title:**

**Bioinformatics Coding Challenge Submission**



Submitted To: Pupil Bio

Submitted By: Shikhar Rana

Contact Information: Phone: +91-8929000111  
Email: shikharrana2298@gmail.com  
GitHub: [https://github.com/shikharrana](styles.xml)

Submission Date: JAN.19,2025