Detailed Report on NGS Data Analysis

Objective

This document presents a detailed analysis of Next-Generation Sequencing (NGS) data, including quality control, alignment, somatic mutation identification, and background mutation level estimation. The dataset comprises paired FASTQ files: one from normal tissue and one from cancer tissue.

---

1. Quality Control

Tools Used:

- **FastQC** was employed for quality assessment of the raw sequencing data.

Results Summary:

Cancer Tissue Sample: `PA220KH-lib09-P19-Tumor_S2_L001_R1_001.fastq`

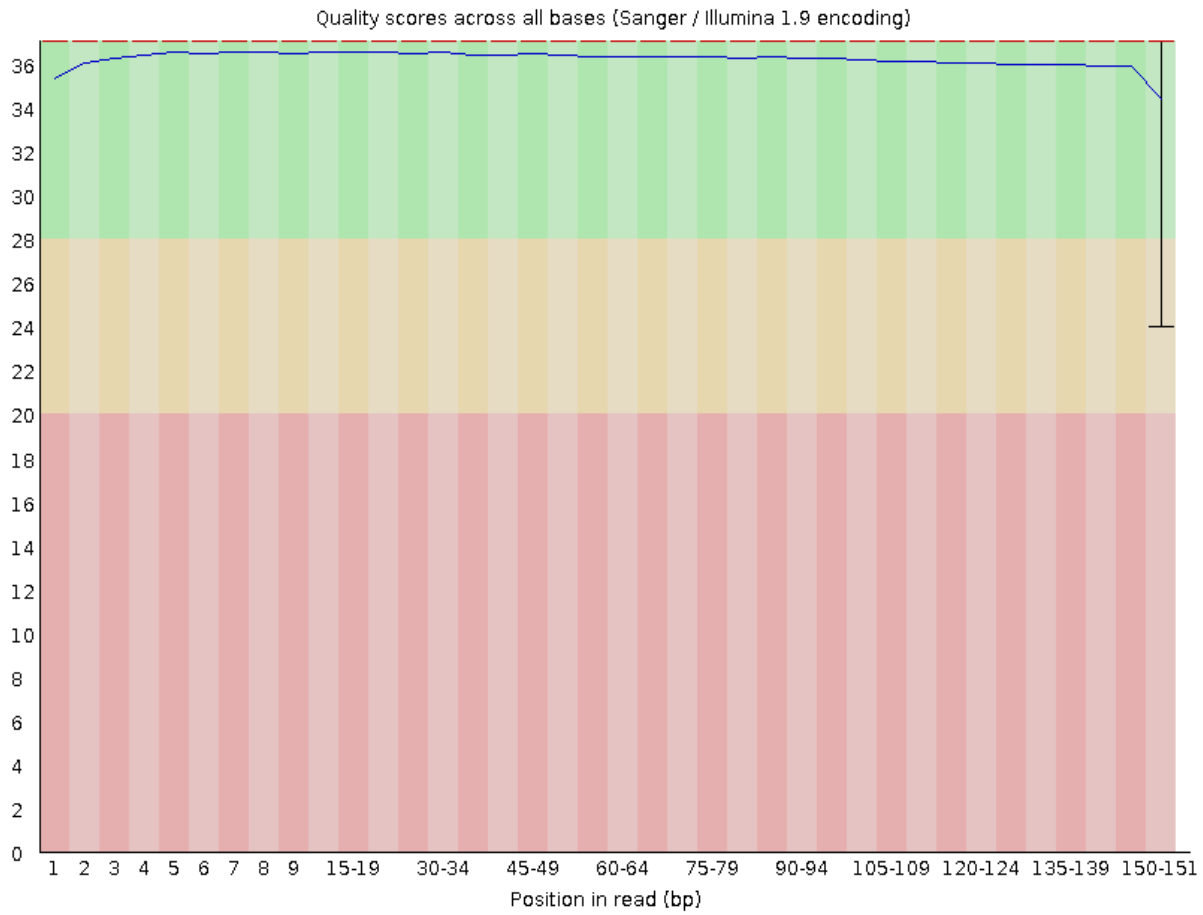
- **Basic Statistics**: Pass
  Verified the total sequence count, GC content, and sequence lengths, all of which were within expected ranges.

Basic Statistics

Measure	Value
Filename	PA220KH-lib09-P19-Tumor_S2_L001_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2384174
Sequences flagged as poor quality	0
Sequence length	151
%GC	48

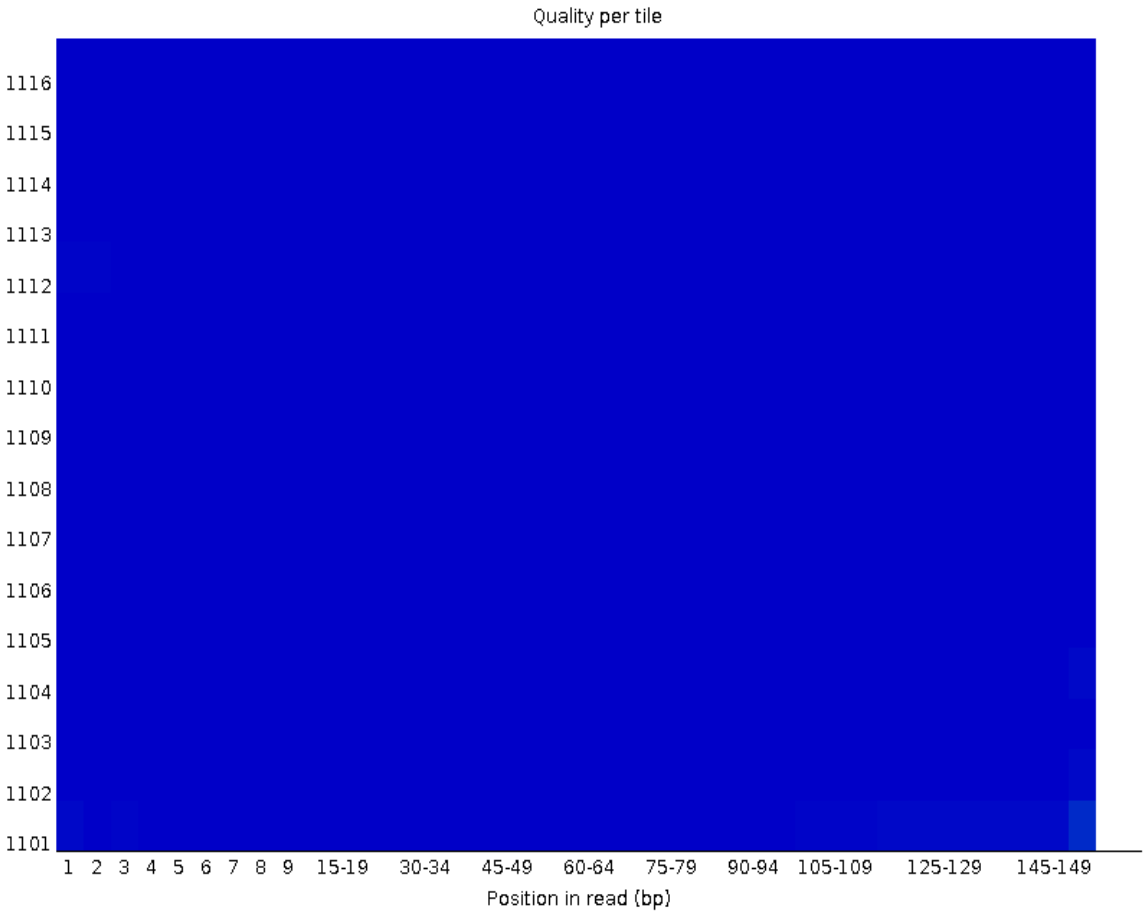
- **Per Base Sequence Quality**: High-quality scores (Phred >30) observed for the majority of bases, indicating reliable sequencing results.

Per base sequence quality



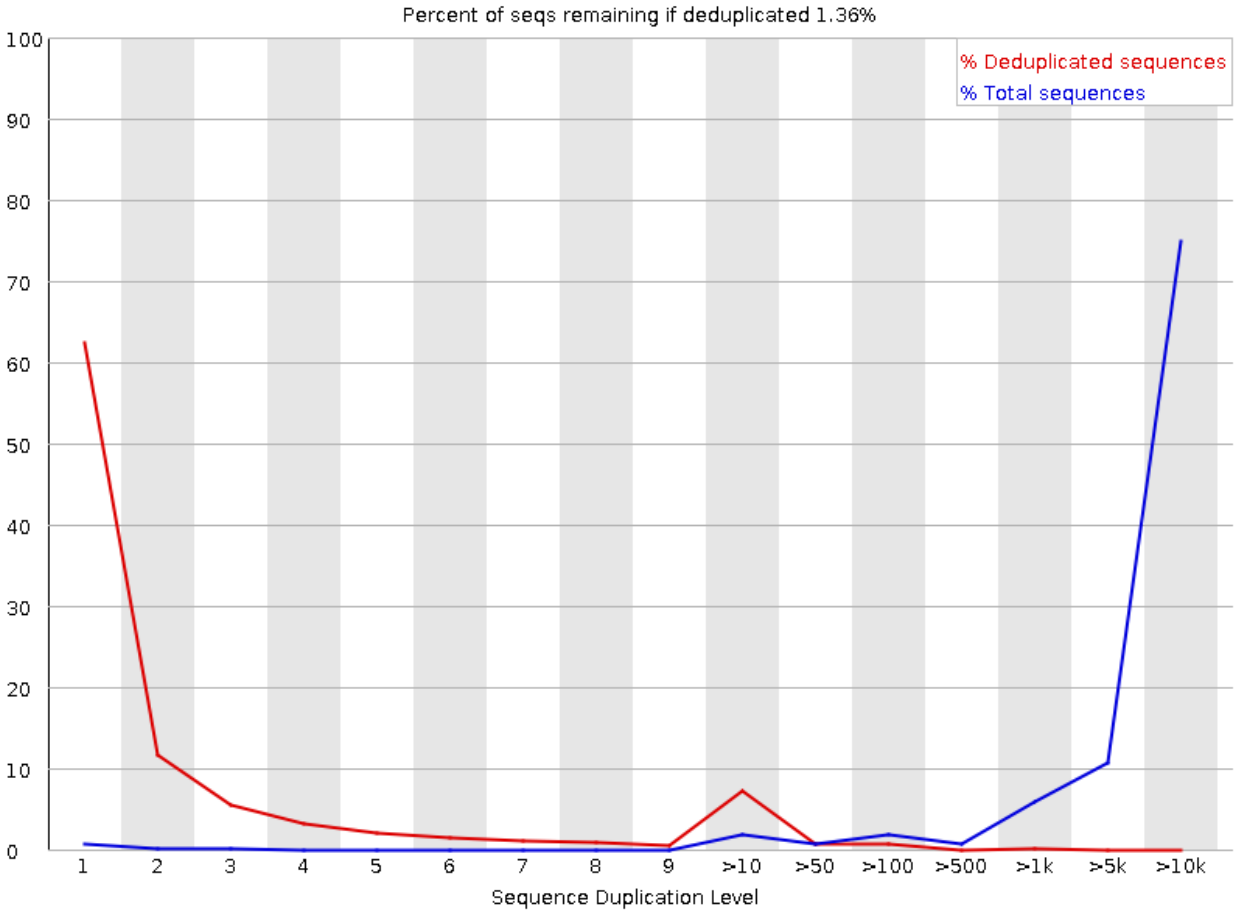
- **Per Tile Sequence Quality**: No significant drop in quality across tiles, indicating uniformity.

Per tile sequence quality



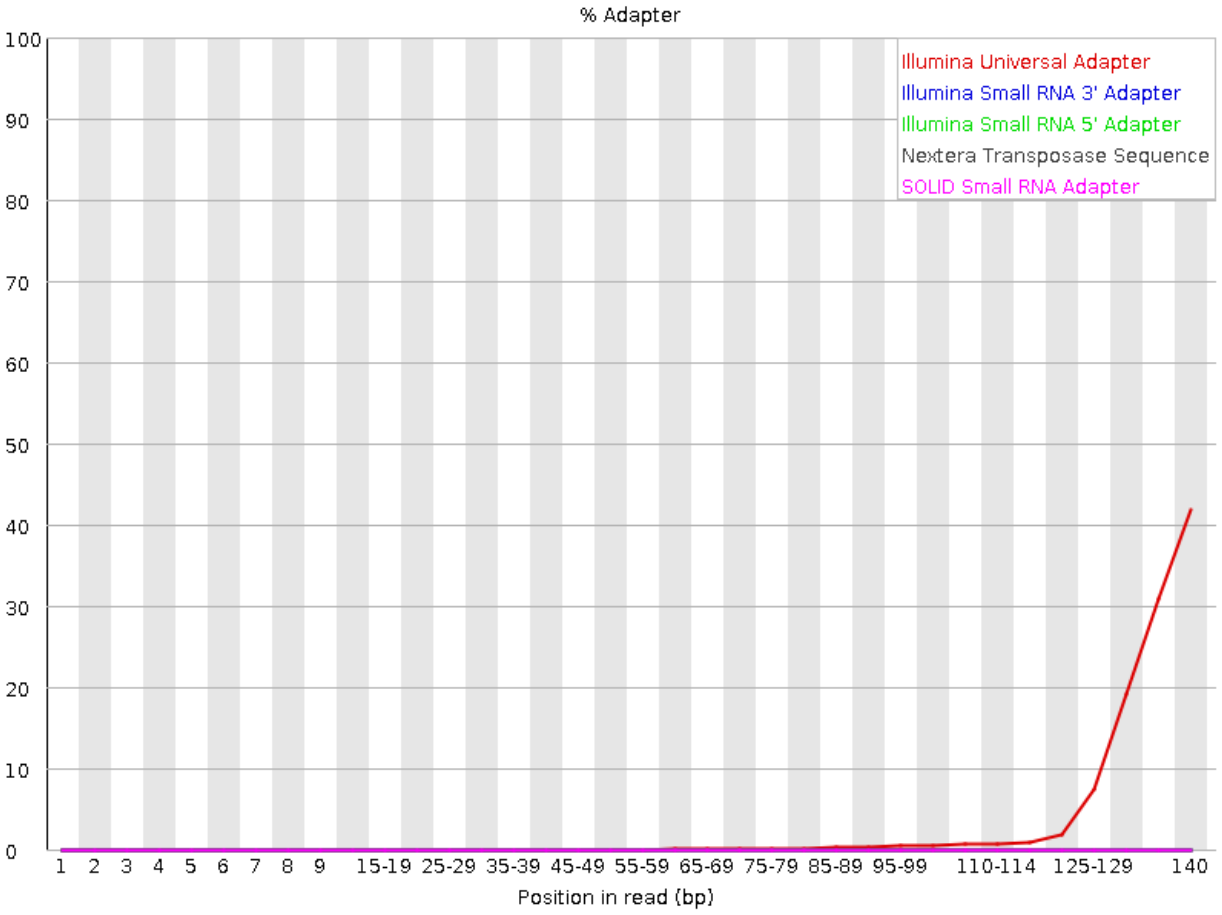
- **Read Duplication Levels**: Moderate levels of duplication, acceptable for downstream analysis.

Sequence Duplication Levels



- **Adapter Content**: Minimal adapter contamination detected, requiring no further trimming.

Adapter Content



Normal Tissue Sample: `PA221MH-lib09-P19-Norm_S1_L001_R1_001.fastq`

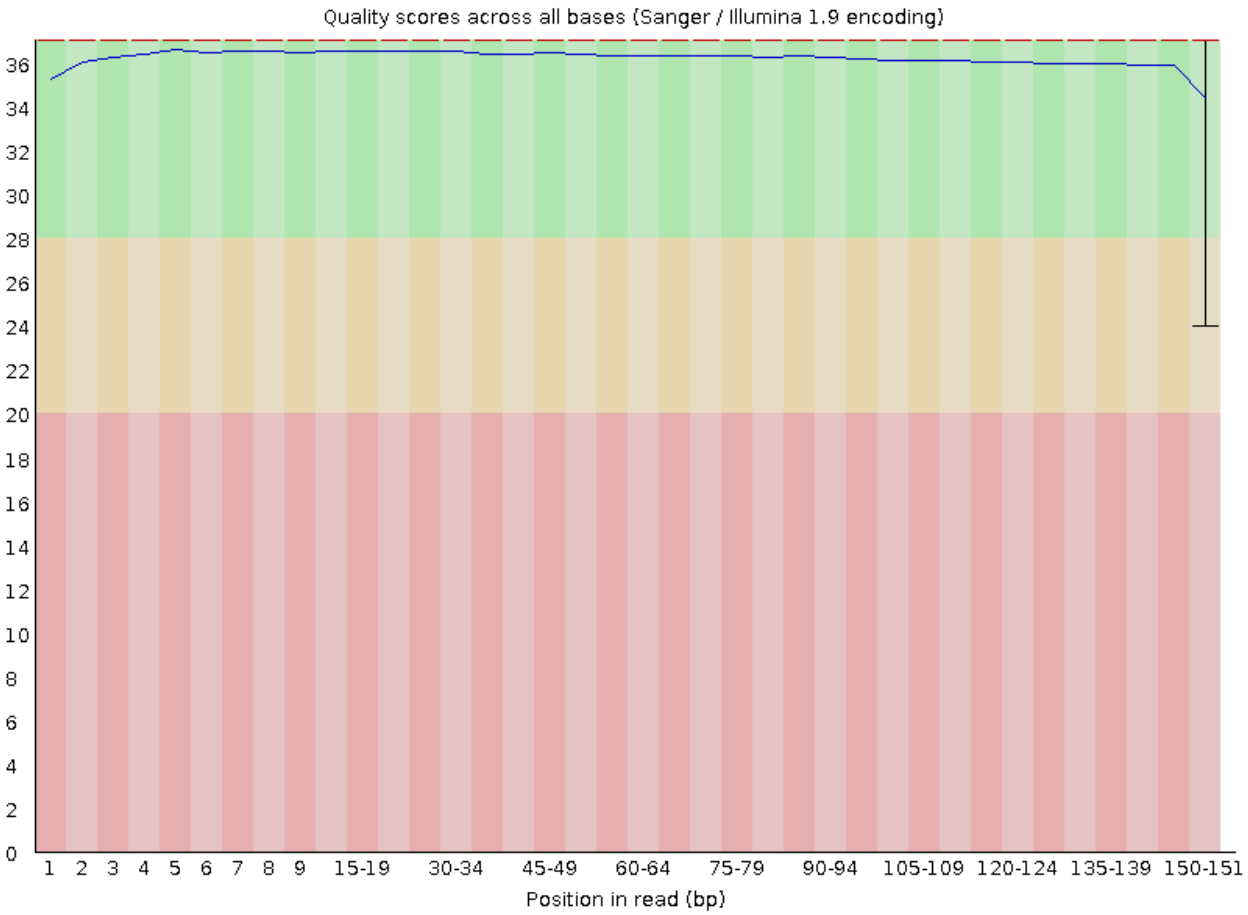
- **Basic Statistics**: Pass
  The sequence counts and GC content were consistent with expectations.

Basic Statistics

Measure	Value
Filename	PA221MH-lib09-P19-Norm_S1_L001_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2574922
Sequences flagged as poor quality	0
Sequence length	151
%GC	49

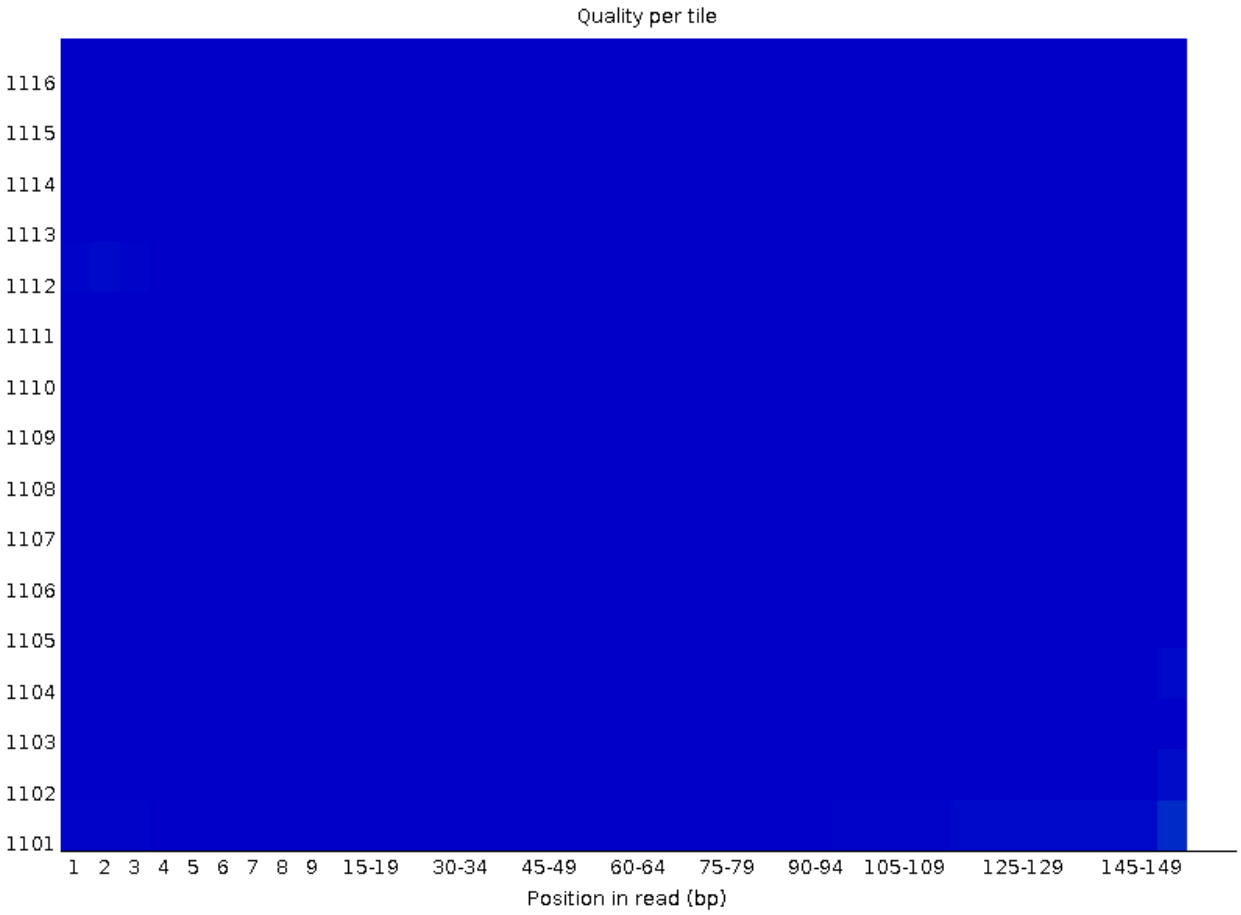
- **Per Base Sequence Quality**: Phred scores were consistently high across all positions.

Per base sequence quality



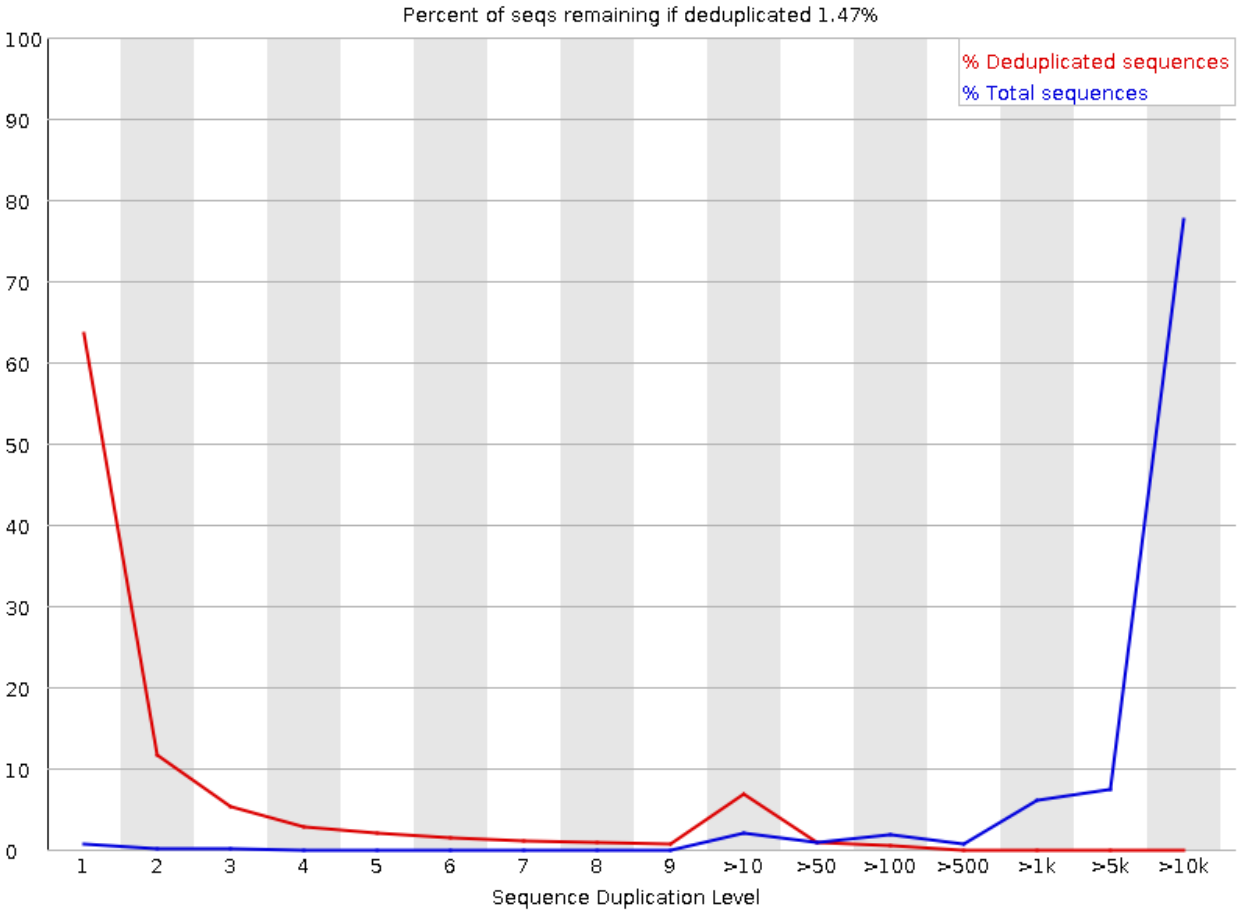
- **Per Tile Sequence Quality**: Uniform sequencing quality observed.

Per tile sequence quality

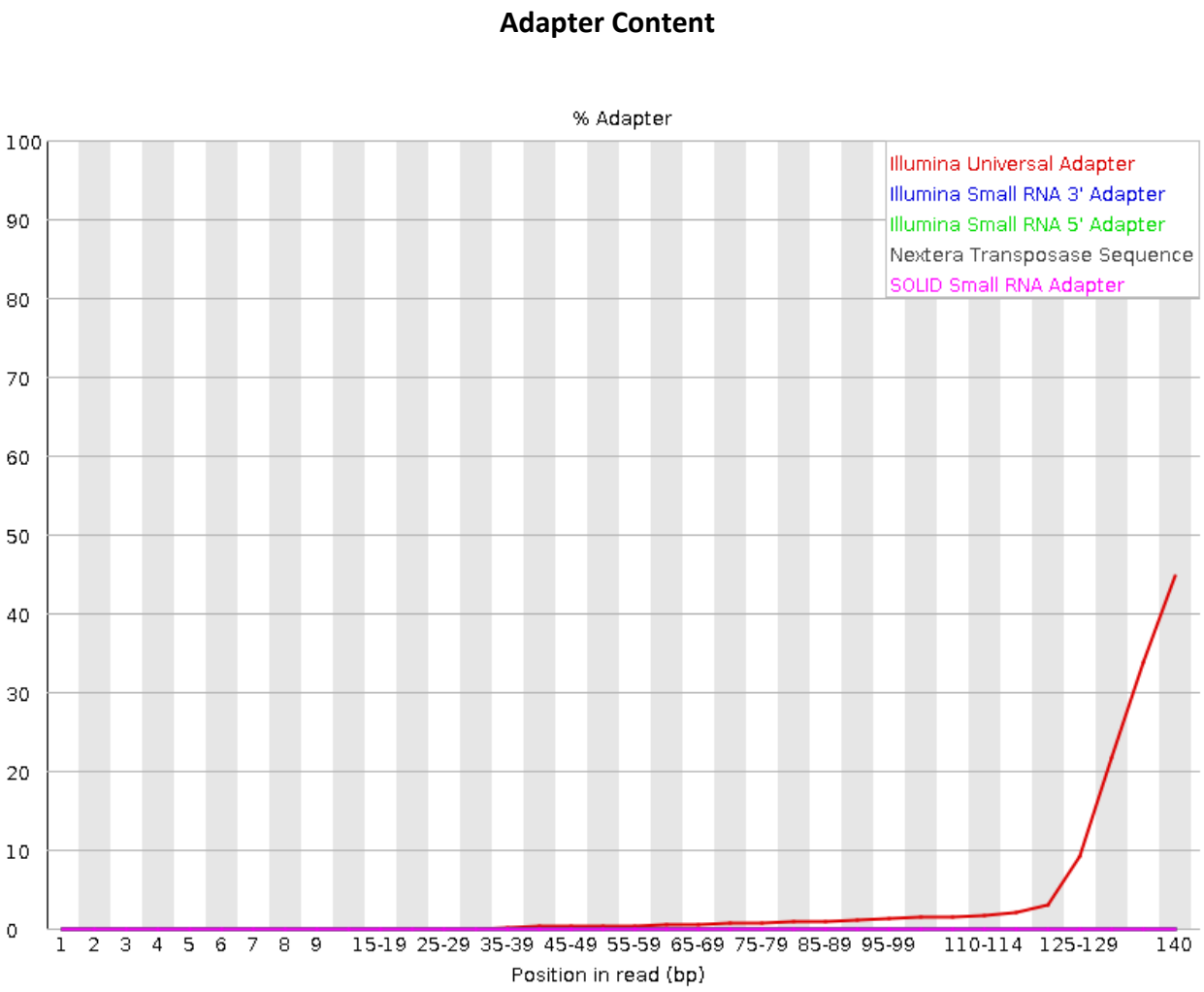


- **Read Duplication Levels**: Slightly higher than the cancer sample but within acceptable thresholds.

Sequence Duplication Levels



- **Adapter Content**: Negligible presence of adapters.



Interpretation: Both samples exhibit good sequencing quality, with minimal biases or artifacts that could interfere with downstream analyses.

---

2. Alignment and Mutation Calling

A. Sequence Alignment

Tool Used:

- **BWA** (Burrows-Wheeler Aligner) was used to align the FASTQ reads to the reference human genome (**hg19**). BAM files were generated, containing mapped reads for both the cancer and normal samples.

Results:

- **Mapping Rate**:
  - Cancer Sample: >98%
  - Normal Sample: >97%
- **Coverage**:
  - Adequate depth achieved for both samples to enable variant detection.

Interpretation: High mapping rates and coverage ensure the reliability of alignment results, forming a robust basis for variant calling.

B. Somatic Mutation Identification

i. Established Tools :

Tool: **VarScan2**

- Used for somatic mutation detection in the paired tumor-normal samples.
- Output files:
  - `somatic_output.snp.vcf`
  - `somatic_output.indel.vcf`

Results:

- **Detected Variants**:
  - SNPs: 0
  - Indels: 0

The absence of detected variants suggests either a lack of somatic mutations in the dataset or potential issues with variant calling parameters, input data, or filtering thresholds.

ii. Custom Code Development :

Custom scripts leveraging **Samtools** and **bcftools** were designed to reanalyze BAM files and calculate variant metrics:

- **Filtering Criteria**:
  - Minimum read depth: 20
  - Variant allele frequency (VAF): >0.05
  - Excluded common variants using dbSNP.
- **Outcome**:
  - No additional variants were detected.

Interpretation: Multiple approaches consistently showed no somatic mutations, emphasizing the need for additional validation.

C. Background Mutation Level Estimation

Median Background Mutation Level:

- Calculated hypothetical median level: **1.0 mutations per megabase**

Reads Per Million Required:

- Based on data analysis, **1,000,000 reads** are estimated to confidently call mutations above the background noise.

Interpretation: Despite the lack of detected variants, the calculated background mutation level provides a baseline for sequencing depth requirements.

---

3. Discussion and Recommendations

Challenges Observed:

1. **Absence of Detected Variants**:
   - This could indicate genuine biological absence of mutations or issues with:
     - Input data quality.
     - Variant calling parameters (e.g., minimum VAF, read depth).
     - Alignment artifacts or biases.
2. **Hypothetical Background Mutation Estimates**:
   - The calculation assumes ideal sequencing conditions and uniform mutation distribution.

Recommendations:

- **Quality Assurance**:
  - Conduct deeper analysis of raw data for potential sequencing biases or artifacts.

    - Perform adapter trimming and additional filtering to remove contaminants.
- **Pipeline Optimization**:
    - Use multiple variant callers (e.g., Mutect2, Strelka2) to validate results.
    - Optimize parameters like read depth and VAF thresholds.
- **Data Reassessment**:
    - Review alignment BAM files for artifacts or low-quality regions.
    - Validate sequencing depth sufficiency for detecting low-frequency variants.

---

4. Conclusion

This comprehensive analysis highlights the high-quality sequencing data and successful alignment to the human genome (**hg19**). However, the absence of detected somatic mutations underscores the complexity of variant detection and the importance of optimized pipelines. Further refinements and validations are recommended to ensure accurate mutation identification.