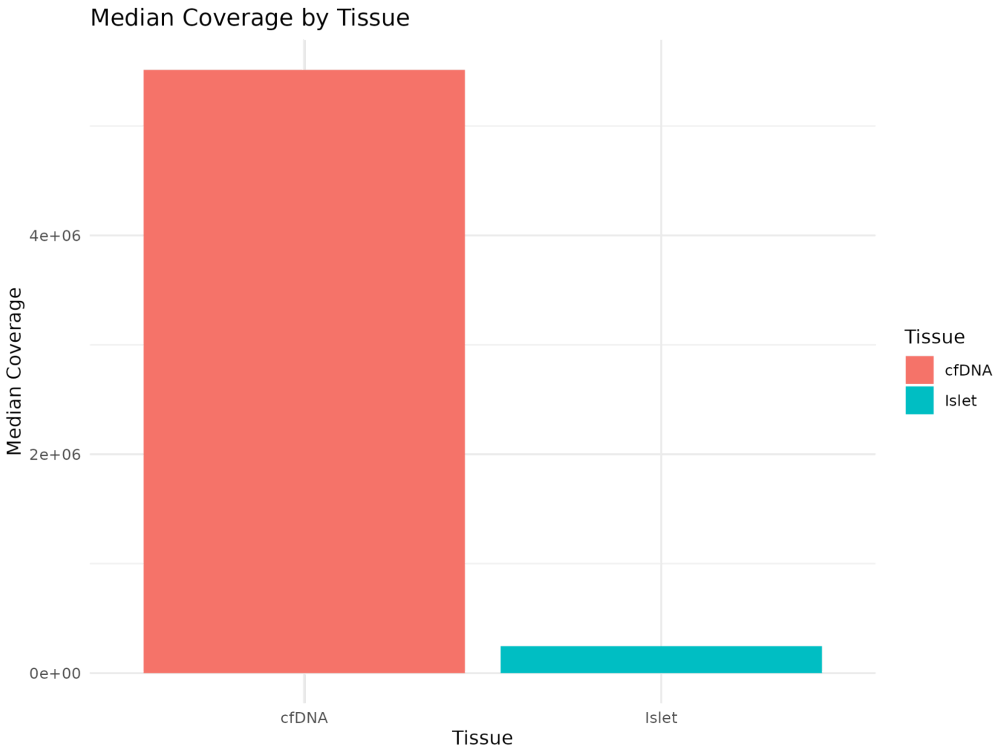
Data Analysis and Statistics

Question 1. Coverage Analysis (10 points):

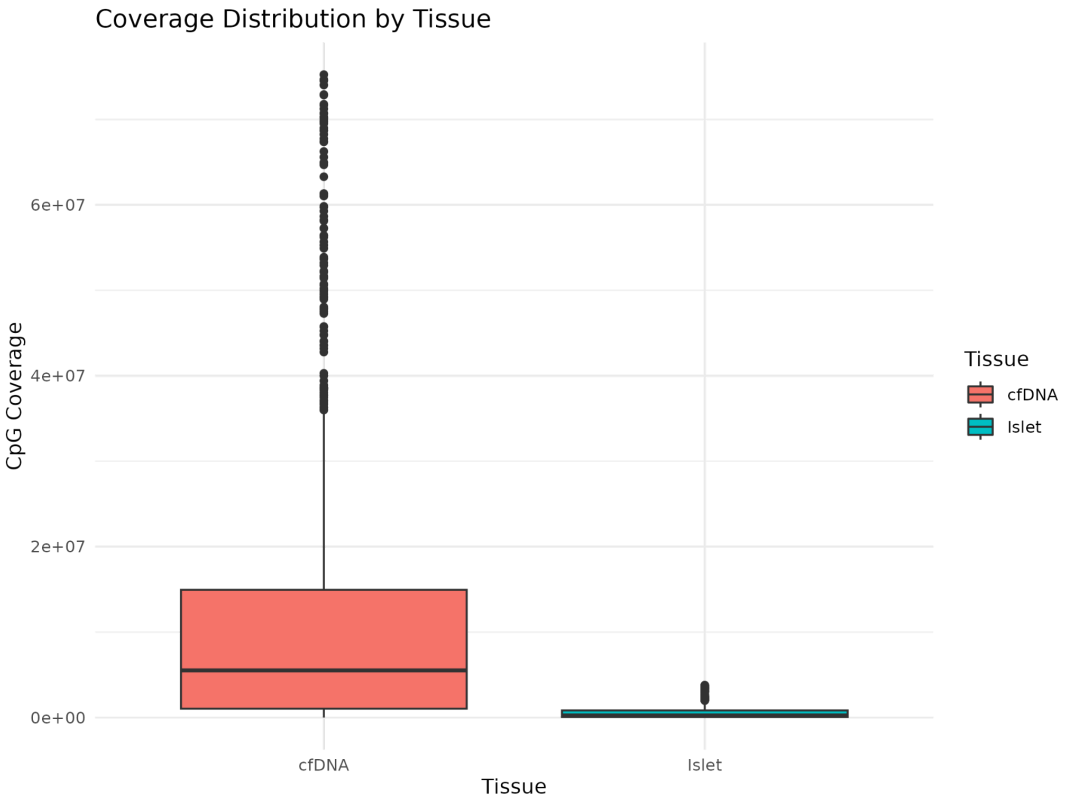
a. Calculate the median and coefficient of variation (CV) for single CpG coverage in each tissue (5 points).

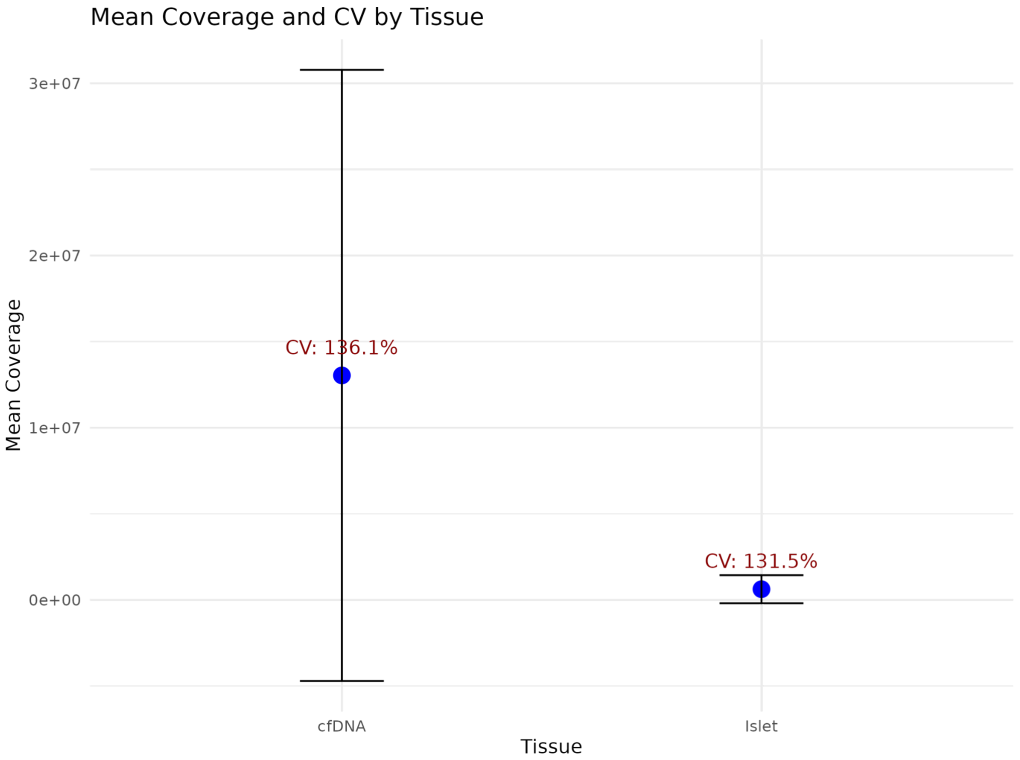
Tissue	Median_Coverage	Mean_Coverage	SD_Coverage	CV_Coverage
Islet	246435.33	619187.59	814204.99	131.49
cfDNA	5512772.16	13036371.07	17747322.92	136.13

b. Generate plots summarizing the coverage statistics (5 points).



Distribution of CpG coverage values for each tissue


Coverage Distribution by Tissue
CpG Coverage
6e+07
4e+07
2e+07
0e+00
cfDNA
Islet
Tissue
Tissue
cfDNA
Islet


Mean Coverage and CV by Tissue
Mean Coverage
3e+07
2e+07
1e+07
0e+00
CV: 136.1%
CV: 131.5%
cfDNA
Islet
Tissue

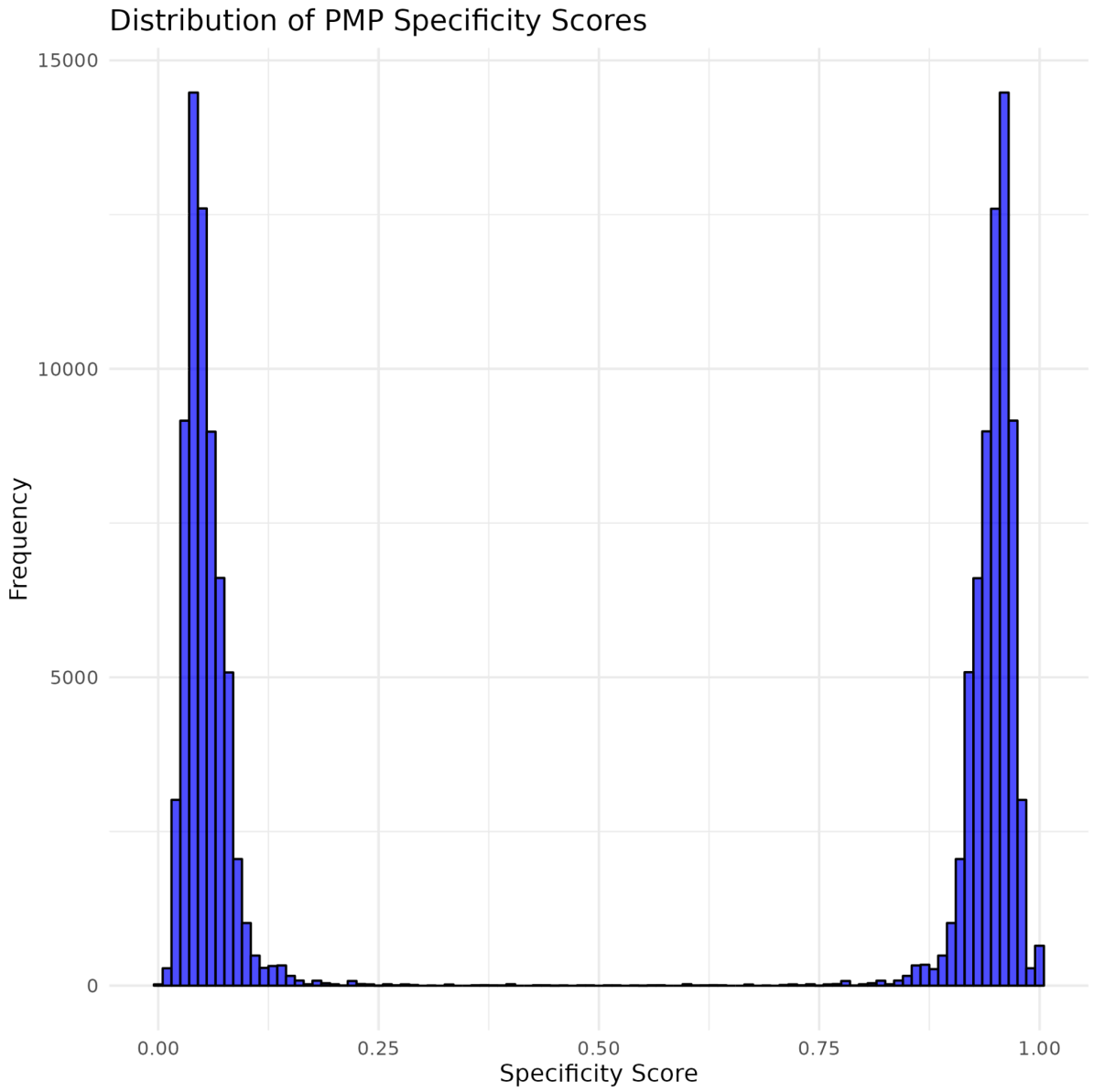
Question 2. Biomarker Identification (20 points):

a. Identify PMPs with high specificity for tissue differentiation, minimizing false positives for Tissue #1 while allowing some false negatives. Use statistical or machine learning approaches to assign confidence (e.g., p-values) to each PMP (15 points).

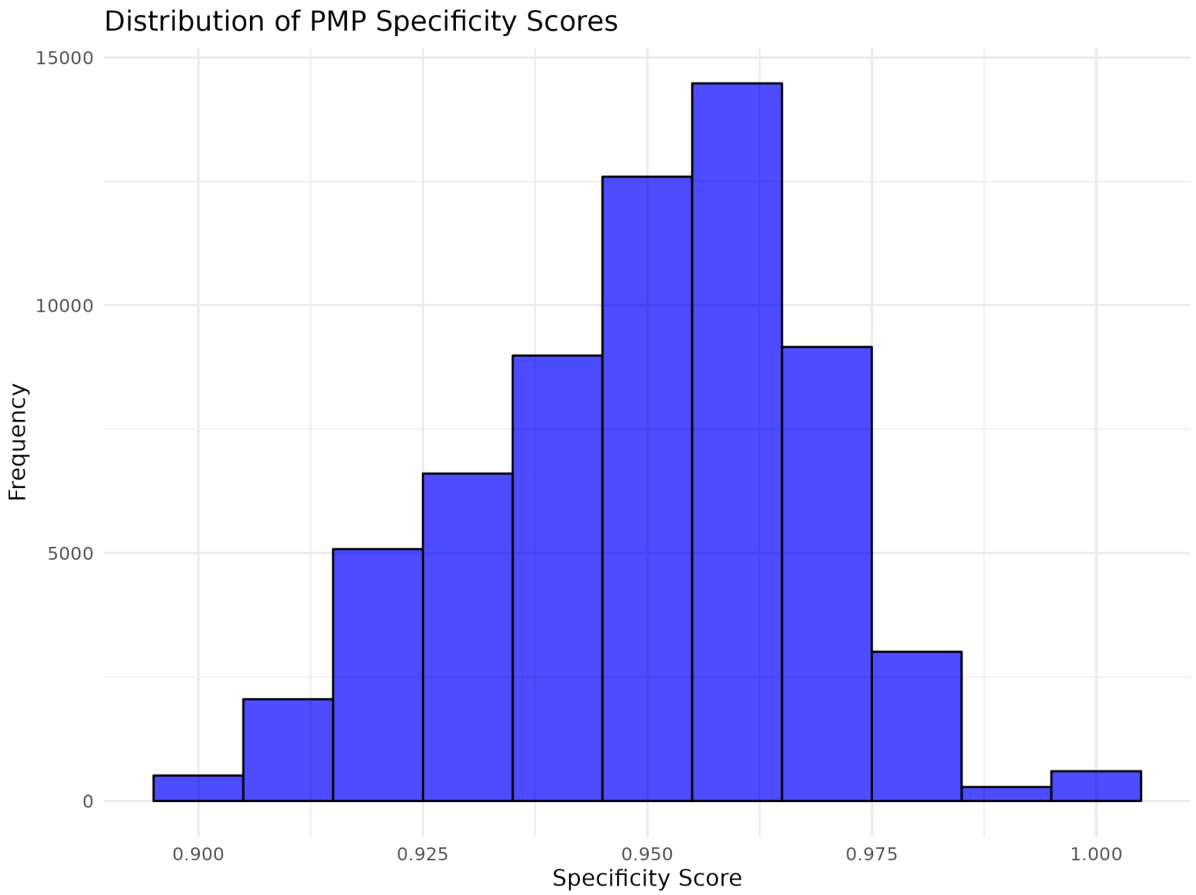
For identification of PMPs specificity, following for

Specificity = True Negatives/True Negatives + False Positives

For this , PMPs with minimal occurrences (low counts) in tissues other than cfDNA were used and then ranked by their occurrence in the cfDNA as compared to Islet.

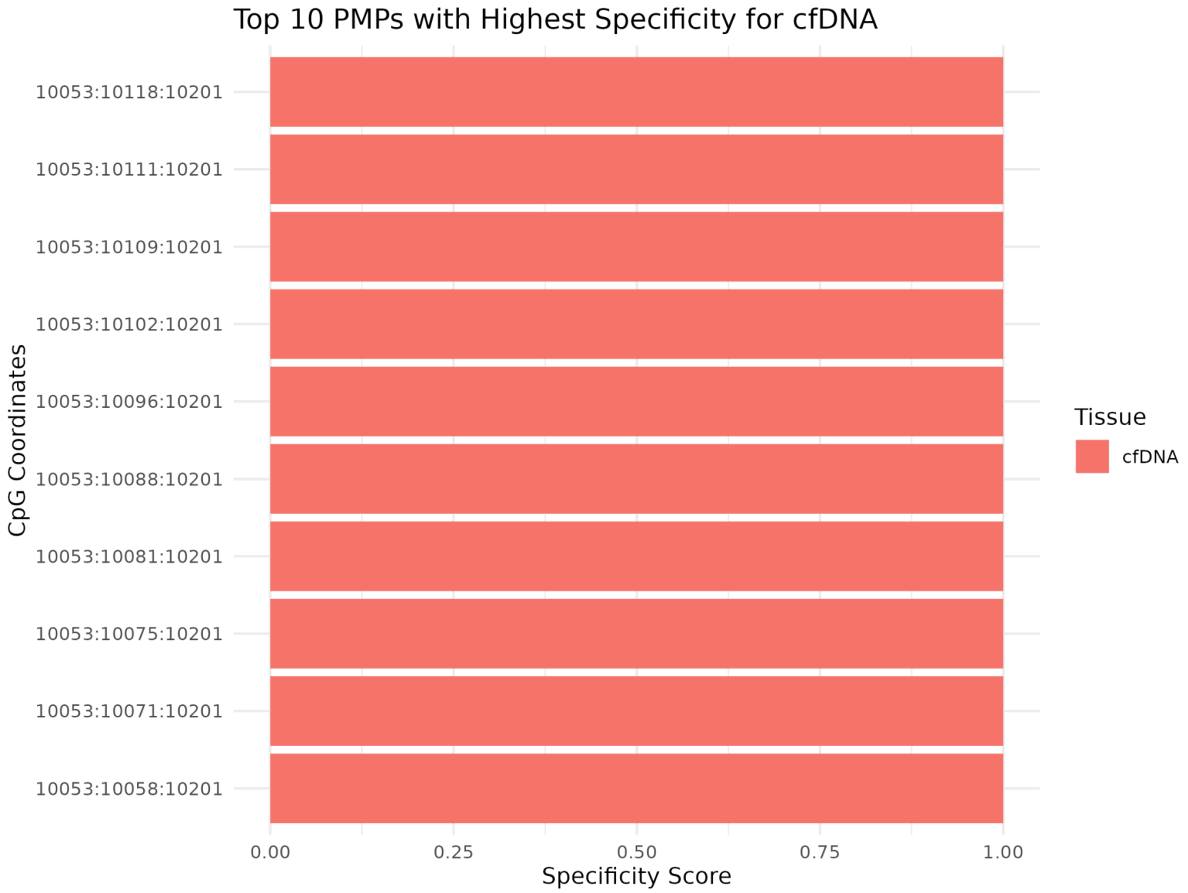


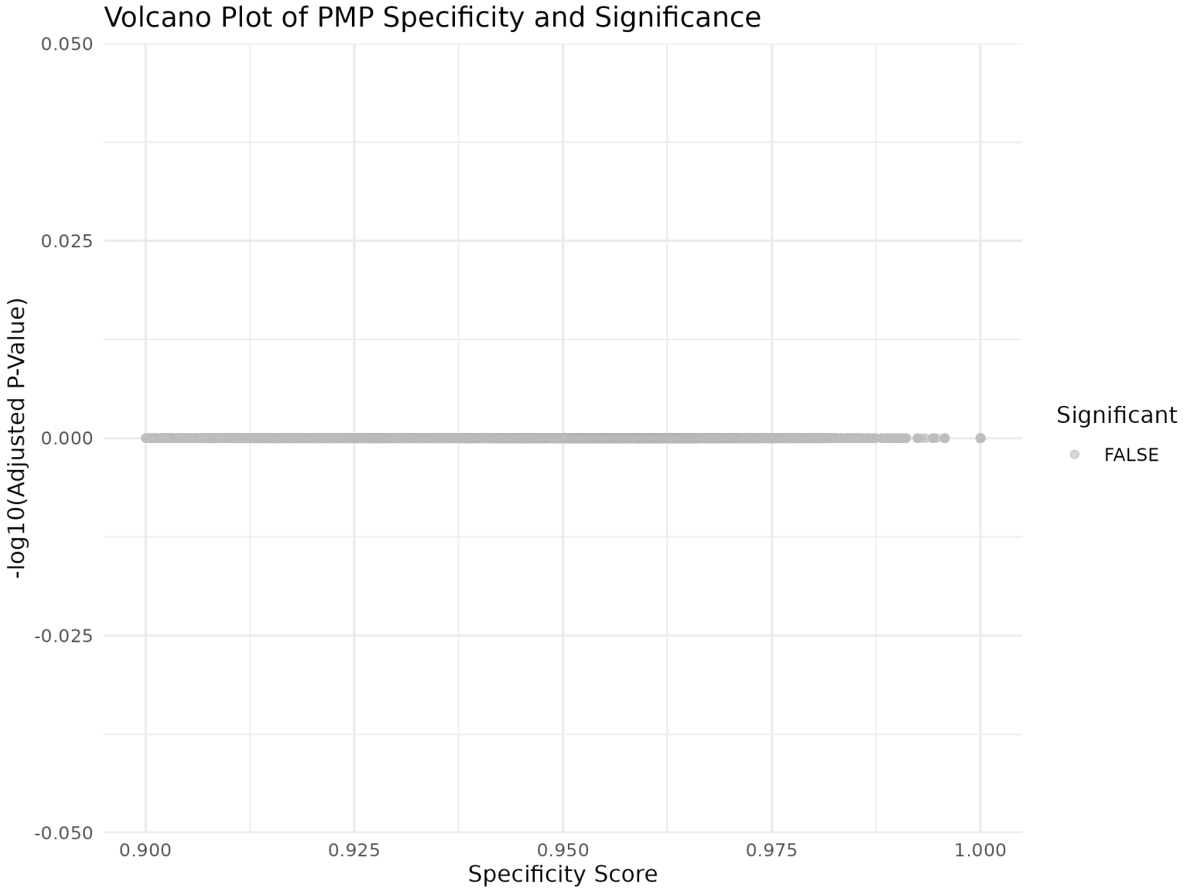
Regions having a specificity value of >0.9 were considered as high specificity PMPs.

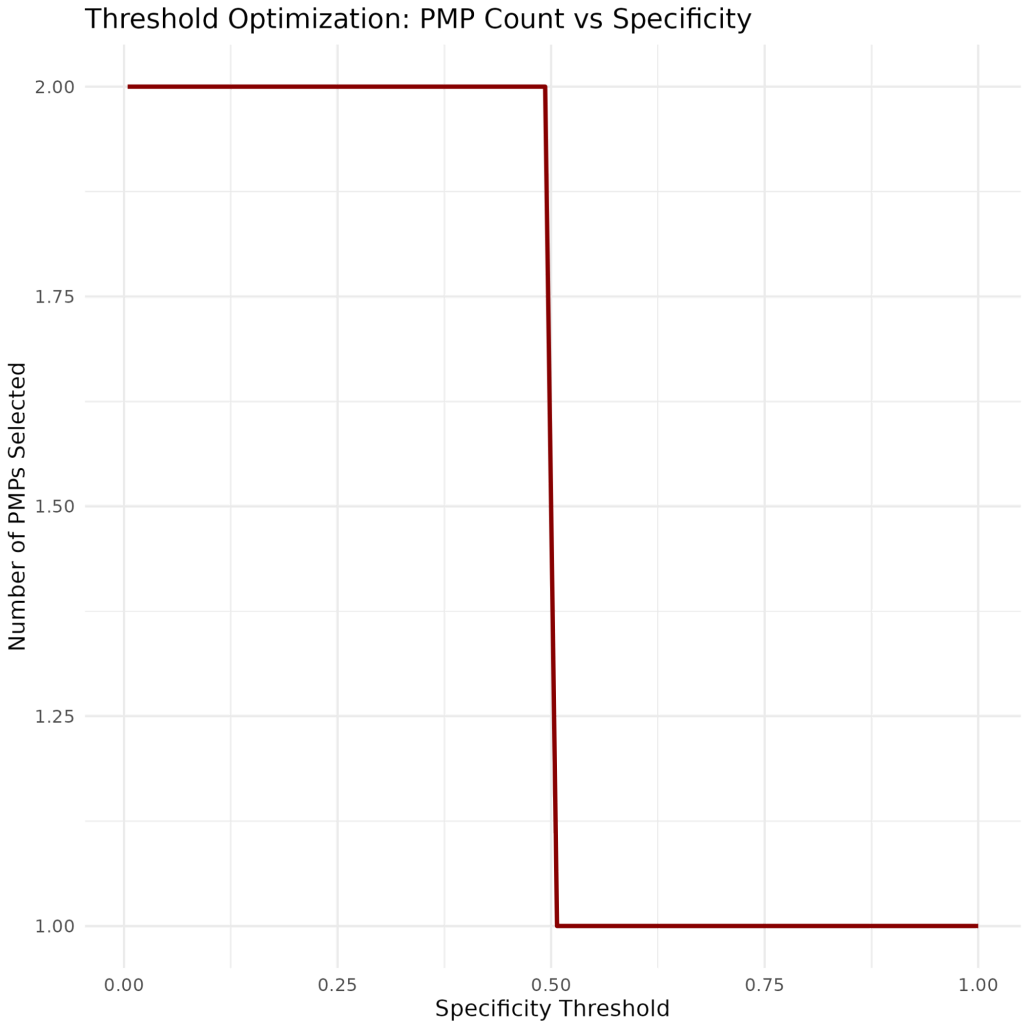


Distribution of High specific PMPs.

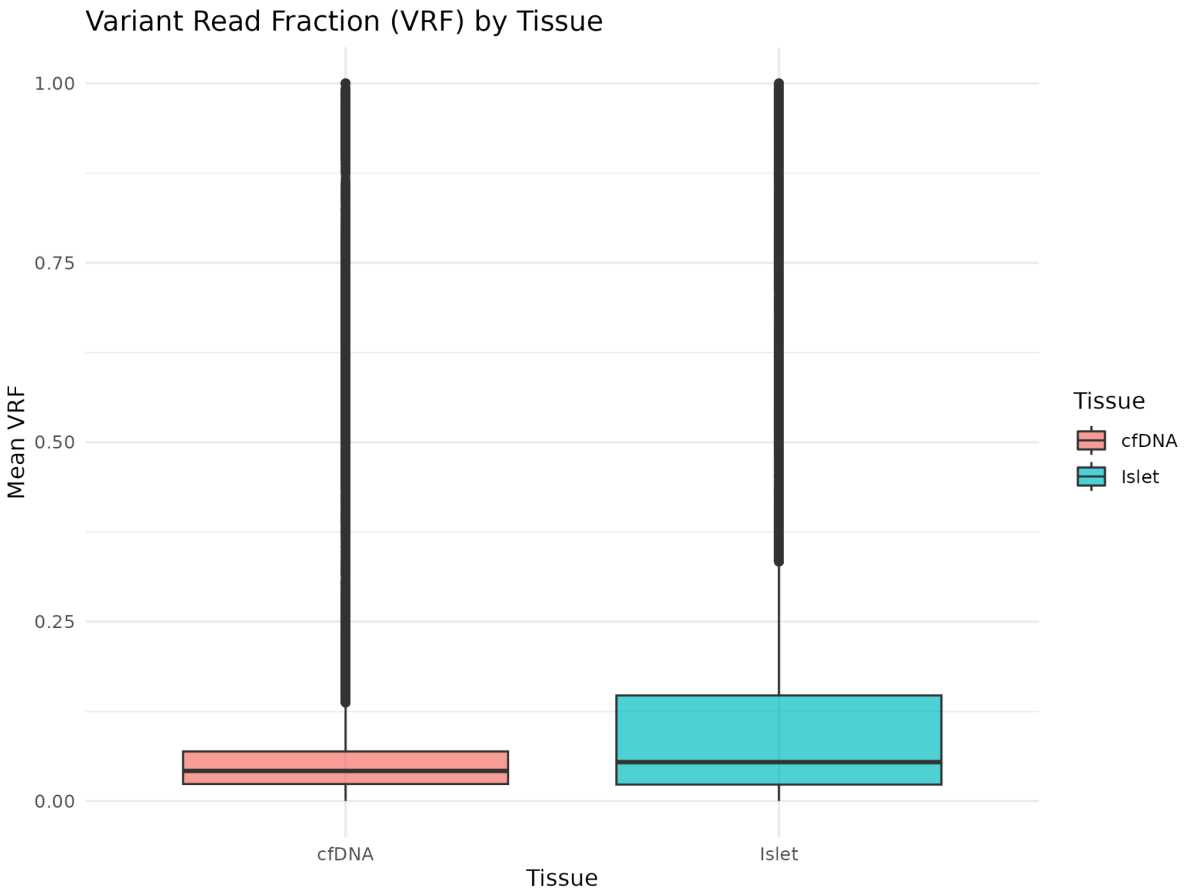
For P-value calculation, both false detection rate and permutation was used, but none of the PMPs were found to be significant at P value 0.05.


Top 10 PMPs with Highest Specificity for cfDNA
10053:10118:10201
10053:10111:10201
10053:10109:10201
10053:10102:10201
10053:10096:10201
10053:10088:10201
10053:10081:10201
10053:10075:10201
10053:10071:10201
10053:10058:10201
CpG Coordinates
0.00
0.25
0.50
0.75
1.00
Specificity Score
Tissue
cfDNA


Volcano Plot of PMP Specificity and Significance
0.050
0.025
0.000
-0.025
-0.050
-log10(Adjusted P-Value)
0.900
0.925
0.950
0.975
1.000
Specificity Score
Significant
FALSE

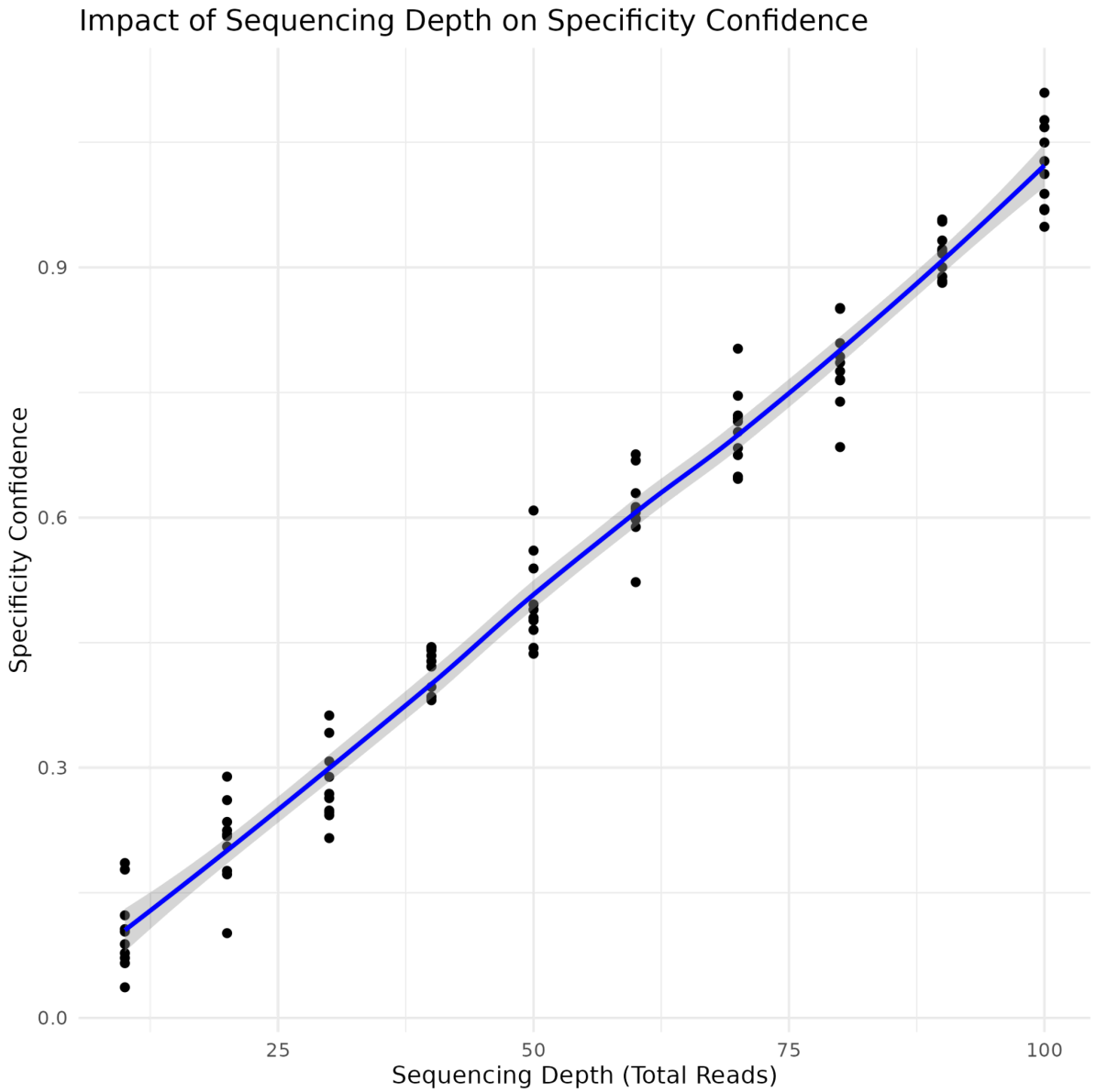


b. Calculate the mean variant read fraction (VRF) for each PMP in both tissues (5 points).

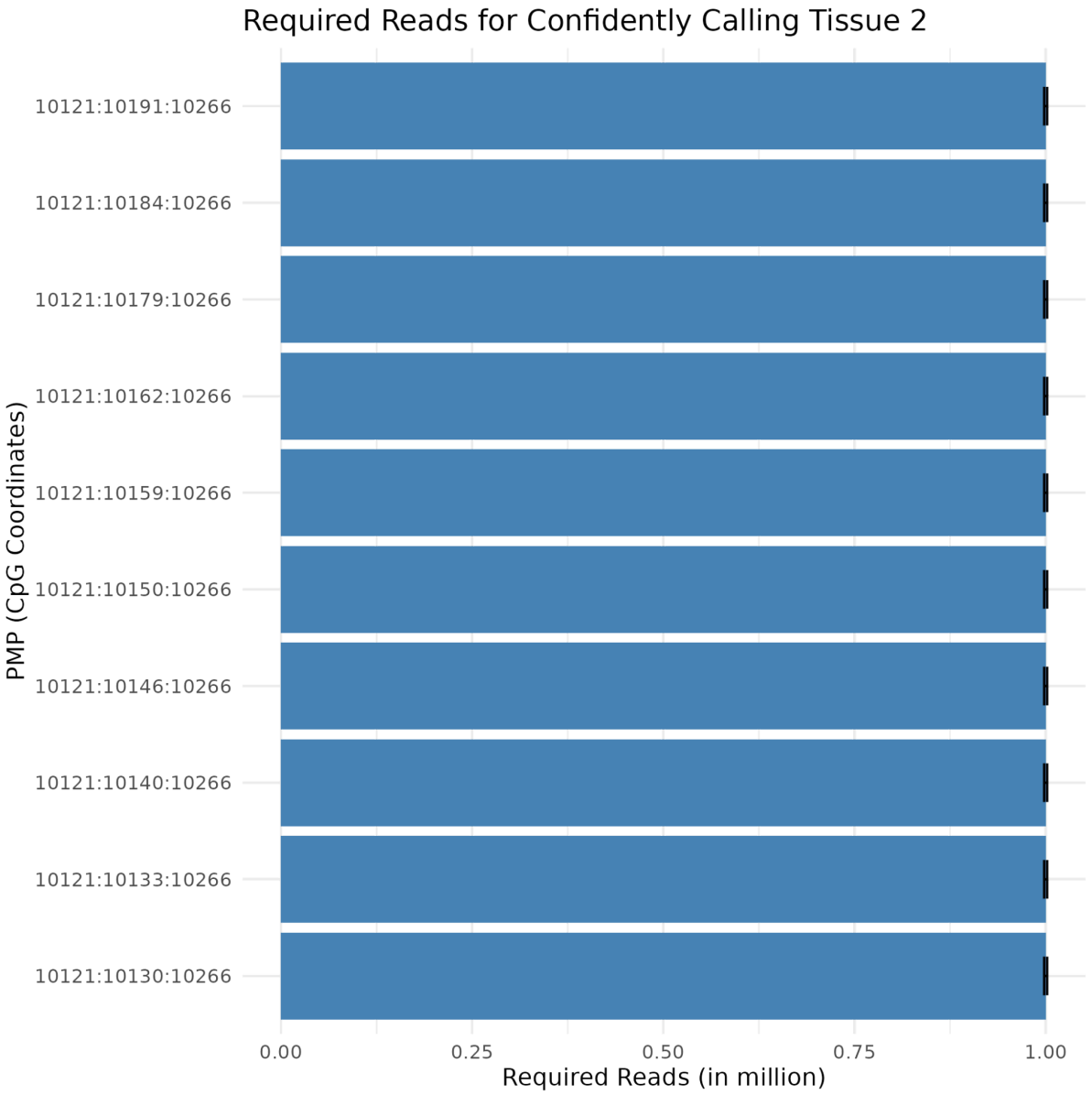


Question 3. Address the following questions (20 points):

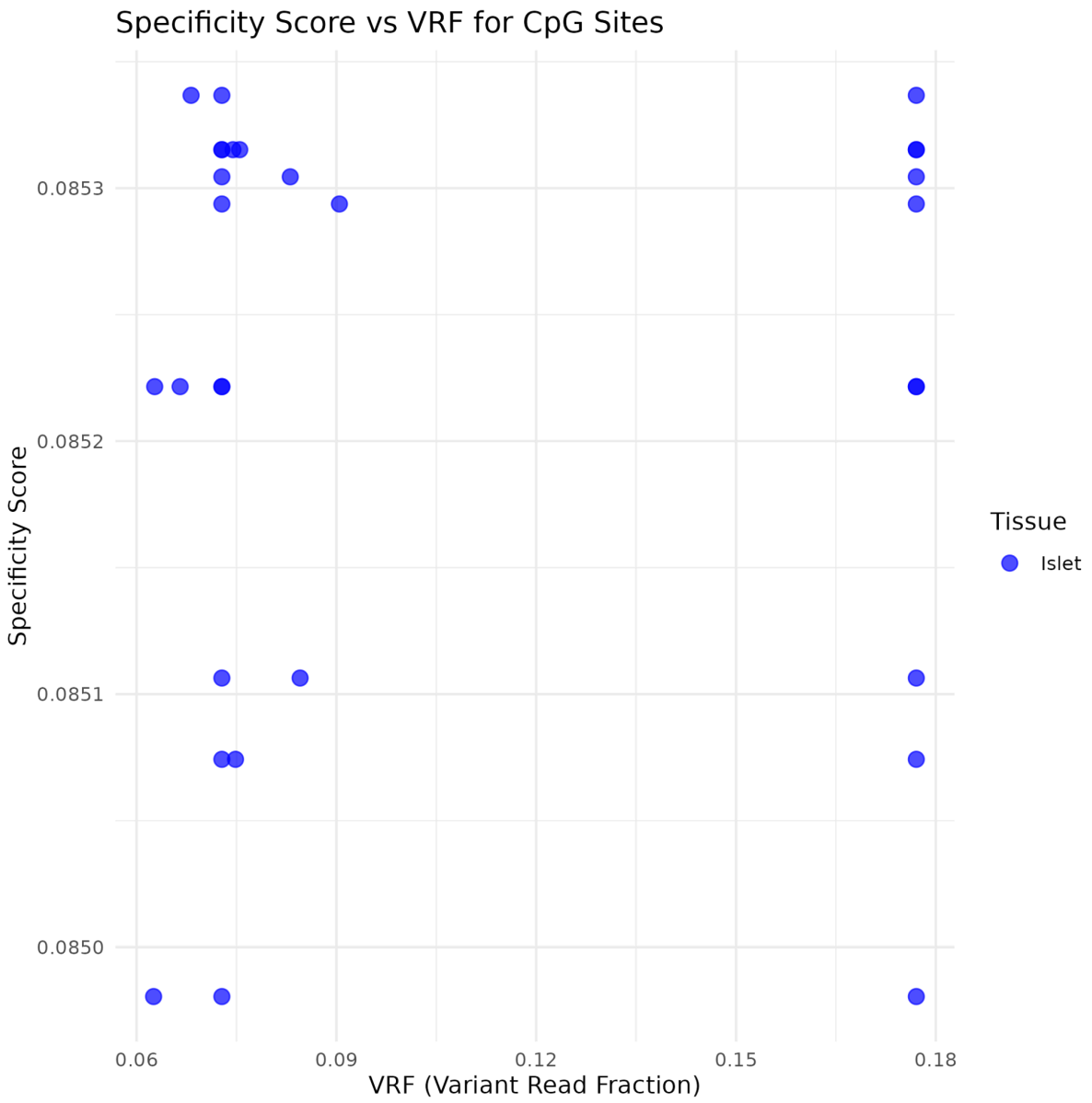
a. How does sequencing depth affect specificity confidence? (5 points).
Sequencing depth is linearly positively correlated with specificity confidence.



b. For the top 10 PMPs, estimate the threshold of reads required to confidently call Tissue #2 at a sequencing depth of 1 million reads. (5 points)



c. Validate the hypothesis by comparing the specificity of the top 10 PMPs against individual CpG sites.(10 points).

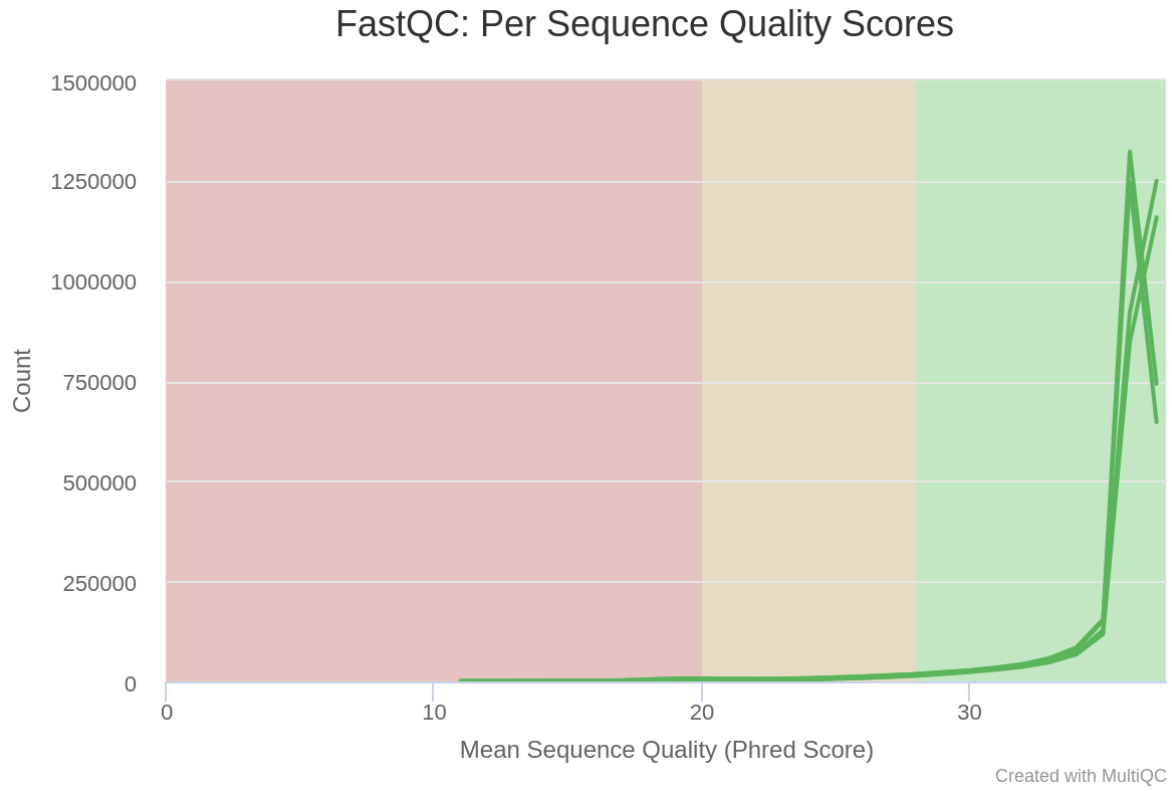

Specificity Score vs VRF for CpG Sites
Specificity Score
0.0853
0.0852
0.0851
0.0850
0.06
0.09
0.12
0.15
0.18
VRF (Variant Read Fraction)
Tissue
Islet

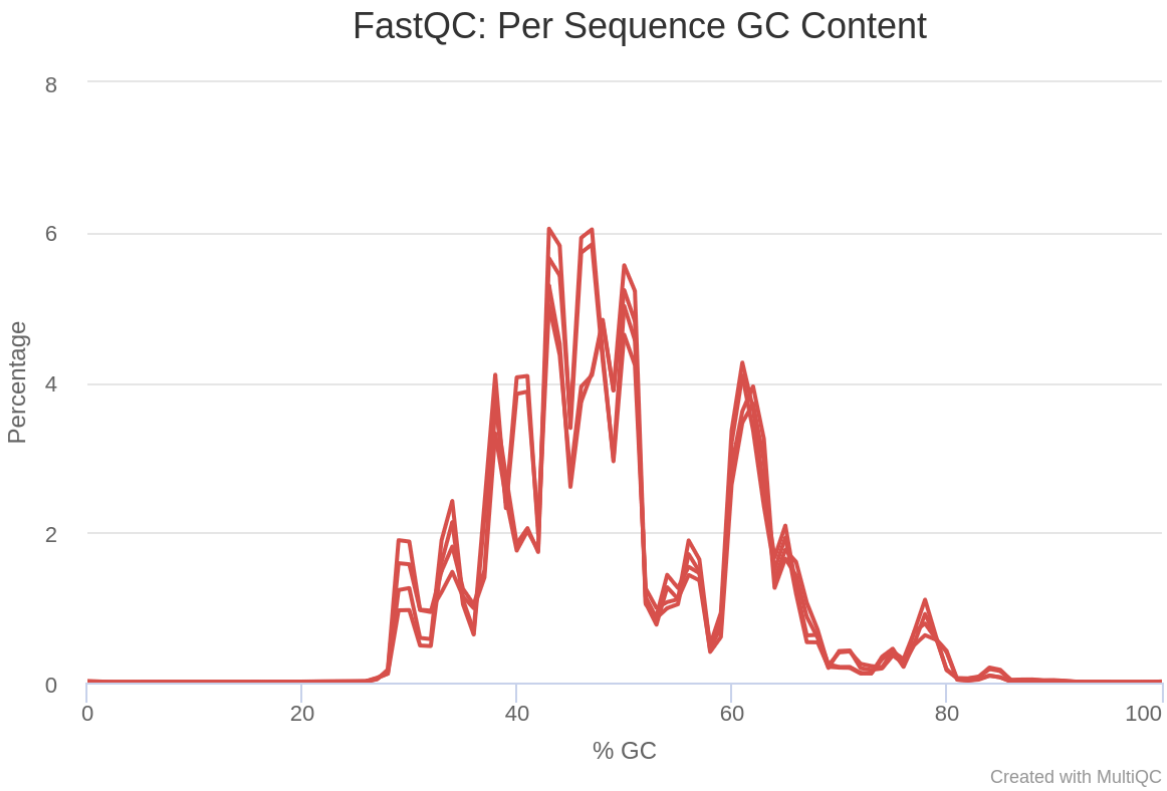
NGS Alignment and Mutation Calling

Question 1. Quality Control (10 points):

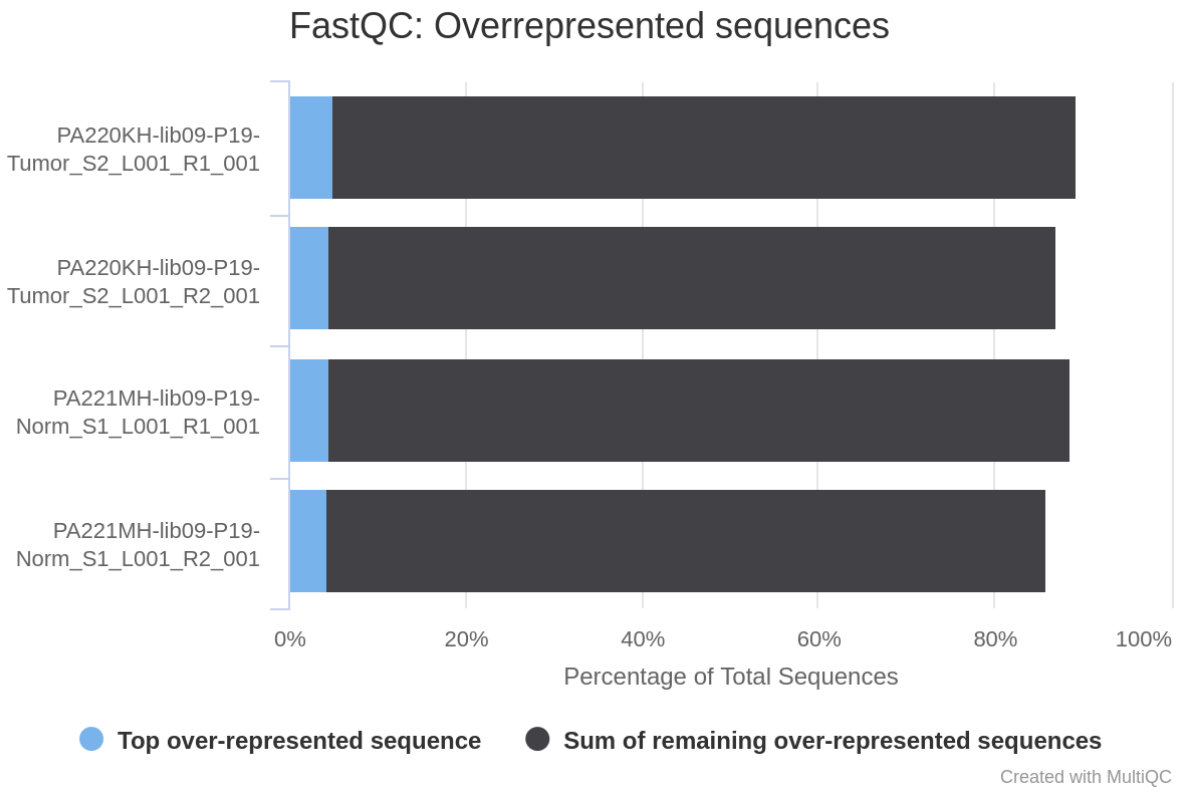
a. Perform quality checks using tools like FastQC and summarize quality metrics (e.g., sequence counts, per-base quality, read duplication levels). (10 points)

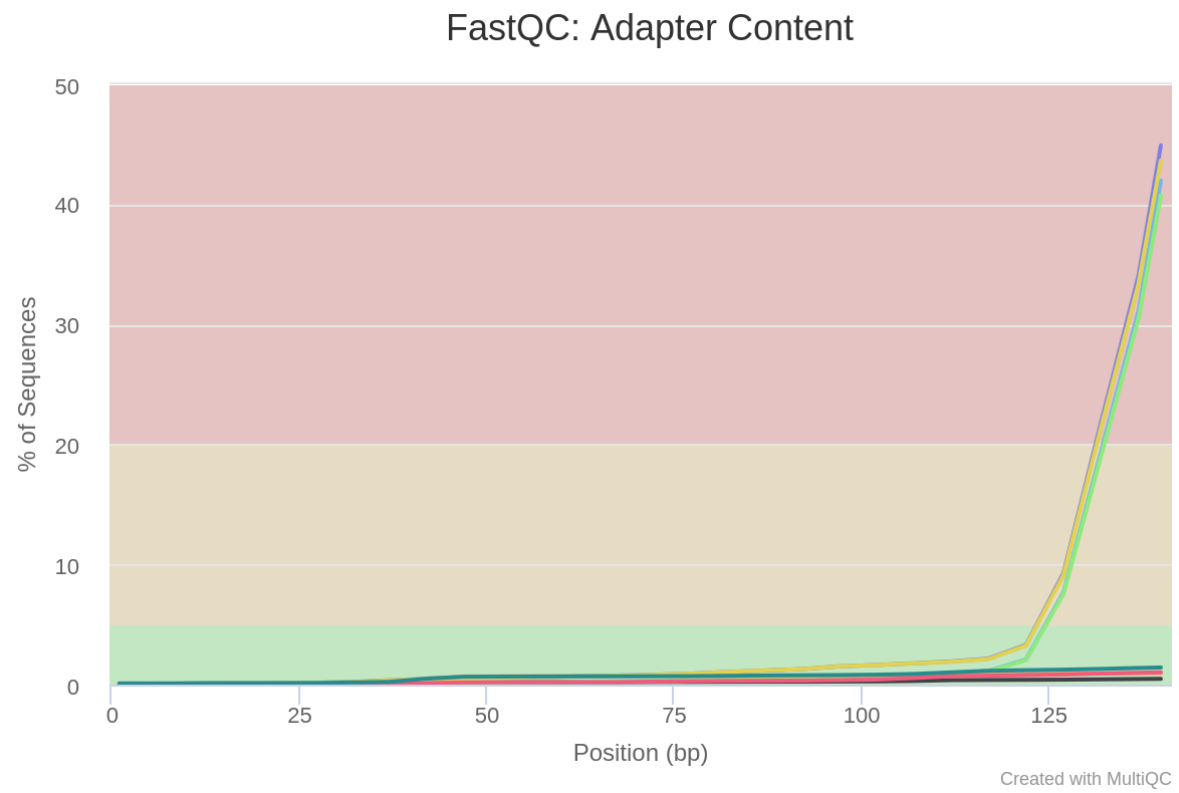
Sample	Total Sequences	Total Bases	Sequences flagged as poor quality	Sequence length	%GC	total_deduplicated_percentage	avg_sequence_length
PA220KH-lib09-P19-Tumor_S2_L001_R1_001	2384174	360 Mbp	0	151	48	1.35	151
PA220KH-lib09-P19-Tumor_S2_L001_R2_001	2384174	360 Mbp	0	151	48	2.56	151
PA221MH-lib09-P19-Norm_S1_L001_R1_001	2574922	388.8 Mbp	0	151	49	1.47	151
PA221MH-lib09-P19-Norm_S1_L001_R2_001	2574922	388.8 Mbp	0	151	49	2.67	151




FastQC: Per Sequence GC Content
Percentage
8
6
4
2
0
0
20
40
60
80
100
% GC
Created with MultiQC


FastQC: Mean Quality Scores
Phred Score
40
30
20
10
0
0
25
50
75
100
125
150
Position (bp)
Created with MultiQC





Question 2.Alignment and Mutation Calling (40 points):

a. Align the samples to the human genome using tools like Bowtie2 or BWA. (10 points).

Mapping on Human genome (GRcH38)

Normal Tissue

Reference size	246,415,288
Number of reads	5,243,876
Mapped reads	566,778 / 10.81%
Supplementary alignments	94,032 / 1.79%
Unmapped reads	4,677,098 / 89.19%
Mapped paired reads	566,778 / 10.81%
Mapped reads, first in pair	282,315 / 5.38%
Mapped reads, second in pair	284,463 / 5.42%
Mapped reads, both in pair	560,933 / 10.7%
Mapped reads, singletons	5,845 / 0.11%
Read min/max/mean length	30 / 151 / 149.03
Duplicated reads (flagged)	461,440 / 8.8%
Clipped reads	429,992 / 8.2%

Coverage

Mean	0.2284
Standard Deviation	139.2513

Mapping Quality

Mean Mapping Quality	14.74

Insert size

Mean	17,180,467.32
Standard Deviation	32,632,289.49
P25/Median/P75	132 / 144 / 2,228,532

Tumor tissue

Globals

Reference size	246,415,288
Number of reads	4,840,672
Mapped reads	470,677 / 9.72%
Supplementary alignments	72,324 / 1.49%
Unmapped reads	4,369,995 / 90.28%
Mapped paired reads	470,677 / 9.72%
Mapped reads, first in pair	234,655 / 4.85%
Mapped reads, second in pair	236,022 / 4.88%
Mapped reads, both in pair	465,514 / 9.62%
Mapped reads, singletons	5,163 / 0.11%
Read min/max/mean length	30 / 151 / 149.36
Duplicated reads (flagged)	389,481 / 8.05%
Clipped reads	352,999 / 7.29%

Coverage

Mean	0.1942
Standard Deviation	118.5248

Mapping Quality

Mean Mapping Quality	10.66

Insert size

Mean	15,784,756.69
Standard Deviation	32,209,876.17
P25/Median/P75	132 / 144 / 150

Mapping was not very good on human reference genome and variant calling alo gave just one variant. Therefore also mapped on the reference gene provided. The summary for this is as follows
Normal Tissue

Reference size	13,772
Number of reads	5,202,434
Mapped reads	5,093,295 / 97.9%
Supplementary alignments	52,590 / 1.01%
Unmapped reads	109,139 / 2.1%
Mapped paired reads	5,093,295 / 97.9%
Mapped reads, first in pair	2,569,916 / 49.4%
Mapped reads, second in pair	2,523,379 / 48.5%
Mapped reads, both in pair	5,085,004 / 97.74%
Mapped reads, singletons	8,291 / 0.16%
Read min/max/mean length	30 / 151 / 150.23
Duplicated reads (flagged)	5,018,811 / 96.47%
Clipped reads	4,432,262 / 85.2%

. Coverage

Mean	50,905.1709
Standard Deviation	74,633.527

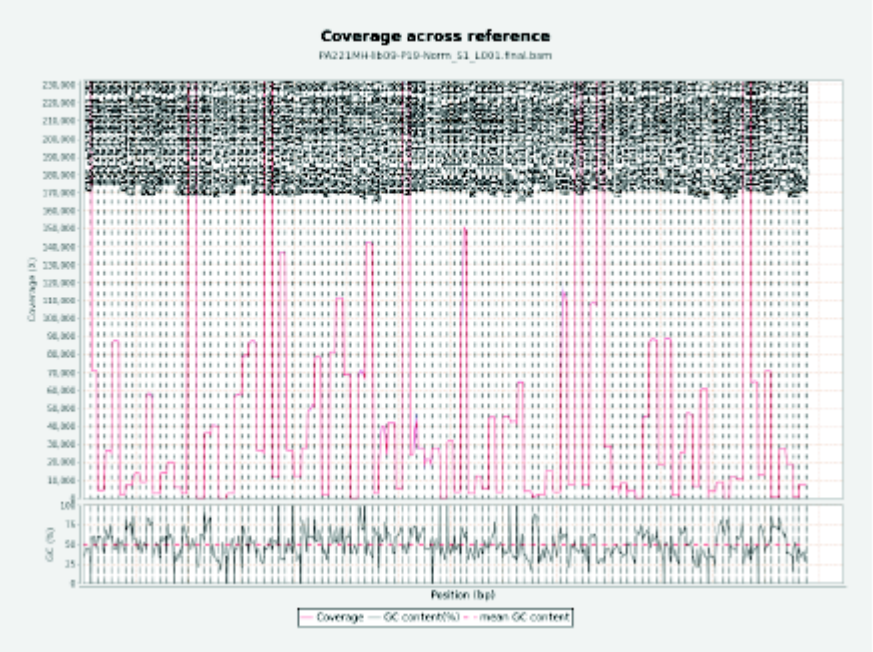
. Mapping Quality

Mean Mapping Quality	59.94

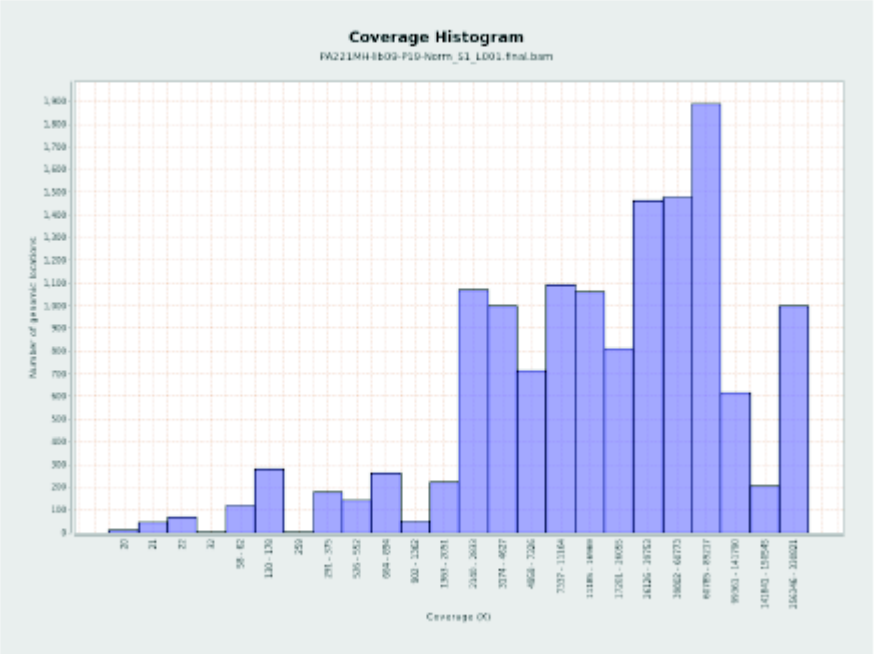
. Insert size

Mean	139.83
Standard Deviation	11.98
P25/Median/P75	131 / 140 / 149

Results : Coverage across reference



Results : Coverage Histogram



Tumor tissue

Reference size	13,772
Number of reads	4,810,831
Mapped reads	4,738,483 / 98.5%
Supplementary alignments	42,483 / 0.88%
Unmapped reads	72,348 / 1.5%
Mapped paired reads	4,738,483 / 98.5%
Mapped reads, first in pair	2,388,430 / 49.65%
Mapped reads, second in pair	2,350,053 / 48.85%
Mapped reads, both in pair	4,731,182 / 98.34%
Mapped reads, singletons	7,301 / 0.15%
Read min/max/mean length	30 / 151 / 150.32
Duplicated reads (flagged)	4,678,340 / 97.25%
Clipped reads	4,065,700 / 84.51%

Coverage

Mean	47,681.2289
Standard Deviation	74,505.4158

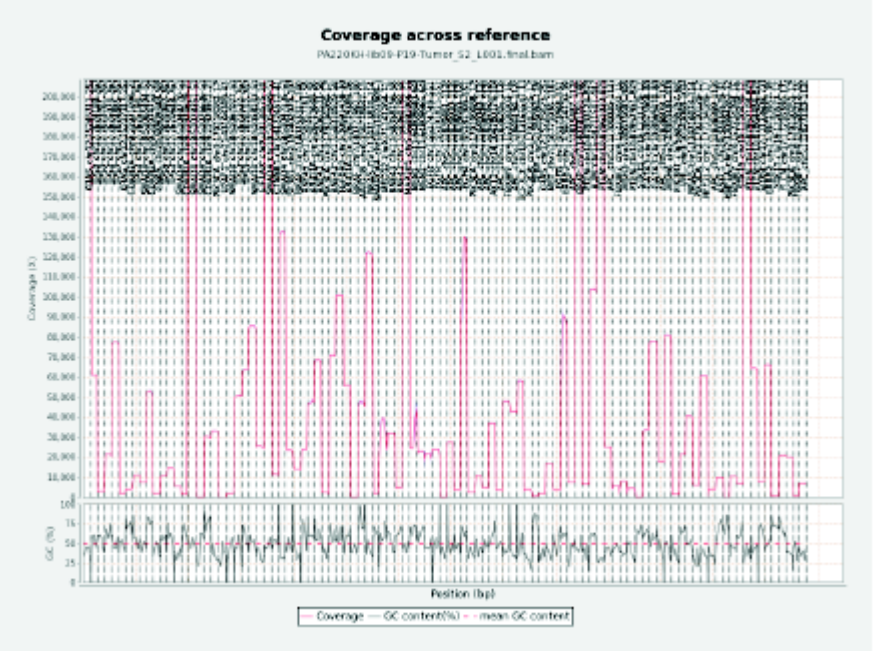
Mapping Quality

Mean Mapping Quality	59.96

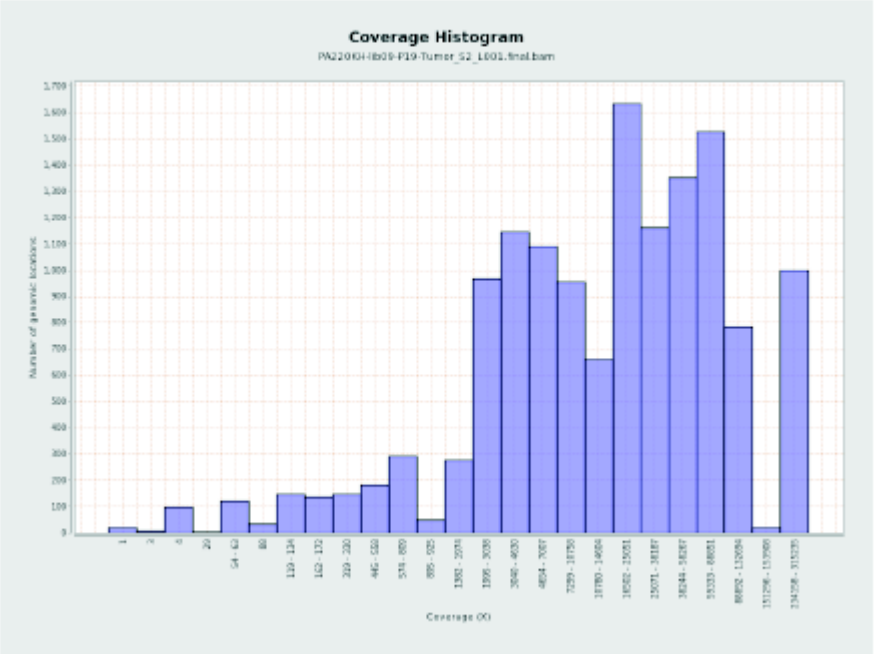
Insert size

Mean	140.8
Standard Deviation	10.96
P25/Median/P75	132 / 141 / 150

Results : Coverage across reference



Results : Coverage Histogram



Further results are from mapping with the gene (provided csv file with fastq)

b. Identify somatic mutations present in the cancer sample but absent in the normal tissue.

i. Benchmark Software: Use established tools such as Mutect2, Strelka2, or VarScan2 for somatic mutation identification and background mutation estimation. (10 points)

Used mutect2 for calling mutations. Script attached.

Summary Numbers

	SNPs			indels		MNPs	others
Callset	n	ts/tv	(1st ALT)	n	frm*		
Pupil	23	1.56	1.56	13	–	1	0

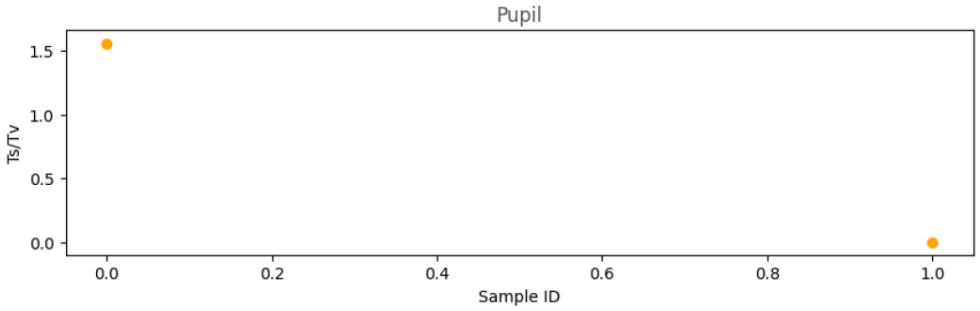
* frameshift ratio: out/(out+in)

	singletons (AC=1)			multiallelic	
Callset	SNPs	ts/tv	indels	sites	SNPs
Pupil	100.0%	1.56	100.0%	0	0

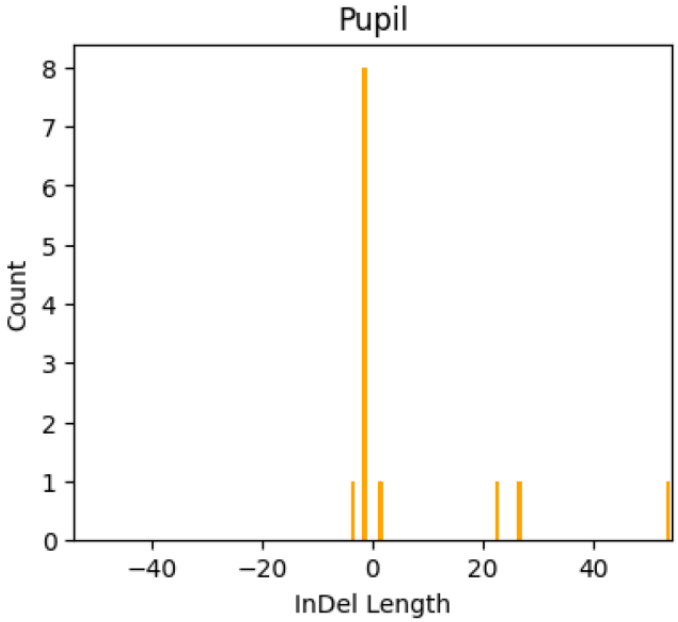
Total Somatic Variants: 37

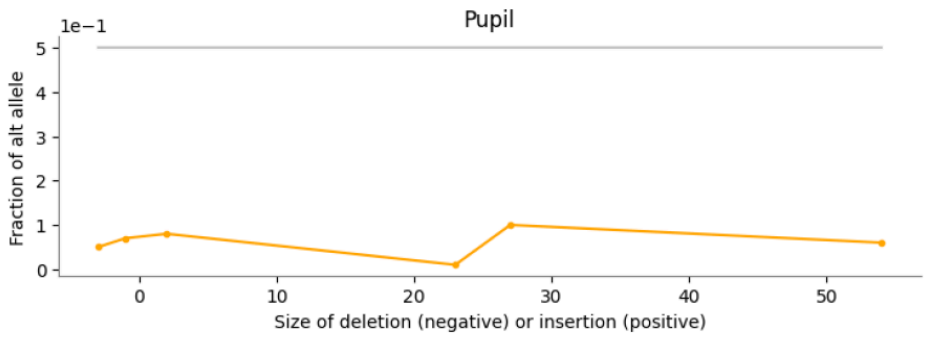
Mutation Rate: 1.23 mutations/Mb (Considereing effective_genome_size_mb = 30 # Example: 30 Mb for exonic regions in WES).

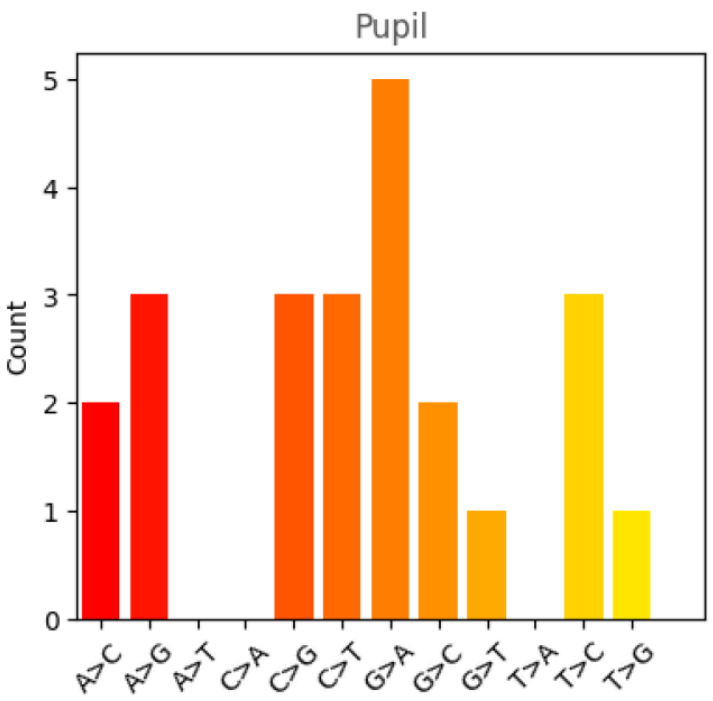
Ts/Tv by sample



Indel distribution







ii. Custom Code Development: Write your own scripts, leveraging tools like Samtools, bcftools, or Python/R libraries, to perform mutation detection and calculate the required metrics. (15 points)

Wrote for bash script for calling variants using samtools and R and bcftools to calculate stats.

NORMAL tissue stats

	SNPs			indels		MNPs	others
Callset	n	ts/tv	(1st ALT)	n	frm*		
norma	25	1.08	1.08	1	–	0	0

* frameshift ratio: out/(out+in)

	singletons (AC=1)			multiallelic	
Callset	SNPs	ts/tv	indels	sites	SNPs
norma	96.0%	1.00	100.0%	0	0

Tumor tissue stats

	SNPs			indels		MNPs	others
Callset	n	ts/tv	(1st ALT)	n	frm*		
tumor	27	0.93	0.93	3	–	0	0

* frameshift ratio: out/(out+in)

	singletons (AC=1)			multiallelic	
Callset	SNPs	ts/tv	indels	sites	SNPs
tumor	92.6%	0.92	66.7%	0	0

Variants private to tumor tissue

	SNPs			indels		MNPs	others
Callset	n	ts/tv	(1st ALT)	n	frm*		
0001.	6	2.00	2.00	2	–	0	0

* frameshift ratio: out/(out+in)

	singletons (AC=1)			multiallelic	
Callset	SNPs	ts/tv	indels	sites	SNPs
0001.	100.0%	2.00	50.0%	0	0

c. Use the normal tissue to calculate the median background mutation level. The background mutation level accounts for sequencing errors or biases that can mimic true mutations. Determine how many reads per million are required to confidently call a given mutation. (5 points)

Background mutation rate "Mutation Rate: 0.13 mutations/Mb"

For Whole-Exome Sequencing (WES):

Mutations Expected=0.13×30=3.9 mutations

Assume 1 read is required per site for every million reads sequenced:

Reads Required Per Million=3.9/30=0.13 reads per mutation detected per million reads.