BIN 508: Next Generation Sequencing & Informatics

Assignment 03

By : Taha Ahmad

Student ID: 2546126

Instructor: Dr Yesim Aydin Son

Galaxy History:

https://usegalaxy.eu/u/taha.ahmad/h/bin508-assignment03-q1

https://usegalaxy.eu/u/taha.ahmad/h/bin508-assignment03-part2

Performed QC first

Reads -> FastQC -> MultiQC -> cutadapt -> Repeated

Multi QC

Before Trimming

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 2/2 rows and 6/6 columns.

Summarize table

Sample Name	Dups	GC	Avg len	Median len	Failed	Seqs
wt_H3K4me3_read1_fastq_gz	0.2 %	57.0 %	51 bp	51 bp	0 %	0.0 M
wt_H3K4me3_read2_fastq_gz	0.9 %	57.0 %	51 bp	51 bp	0 %	0.0 M

After Trimming

General Statistics

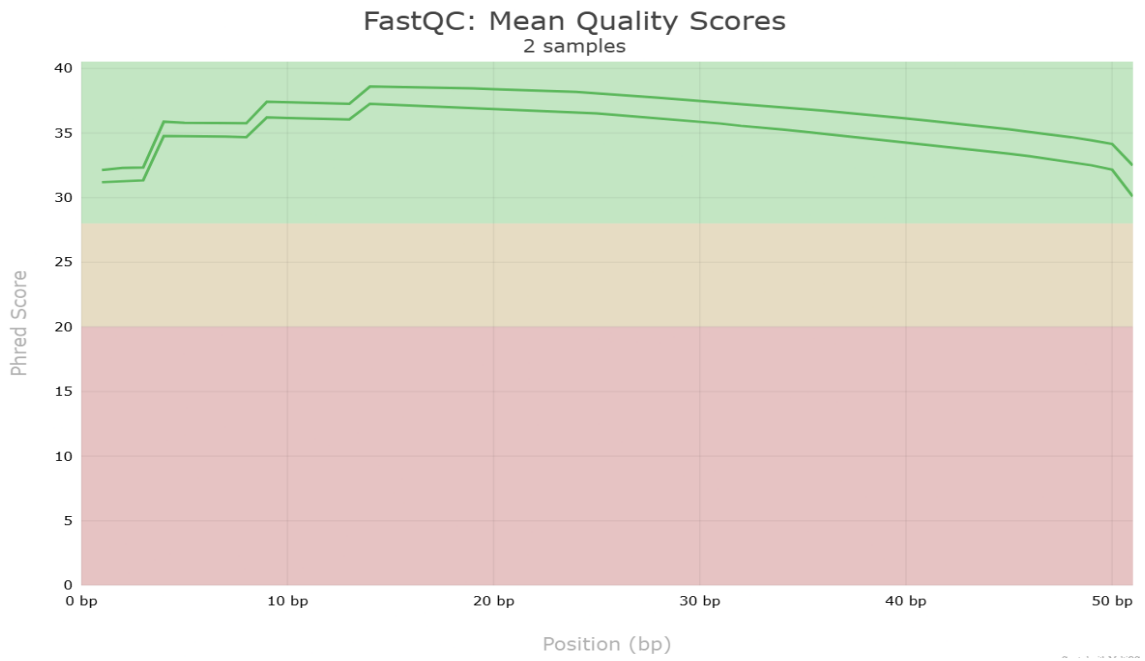
Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 2/2 rows and 6/6 columns.

Summarize table

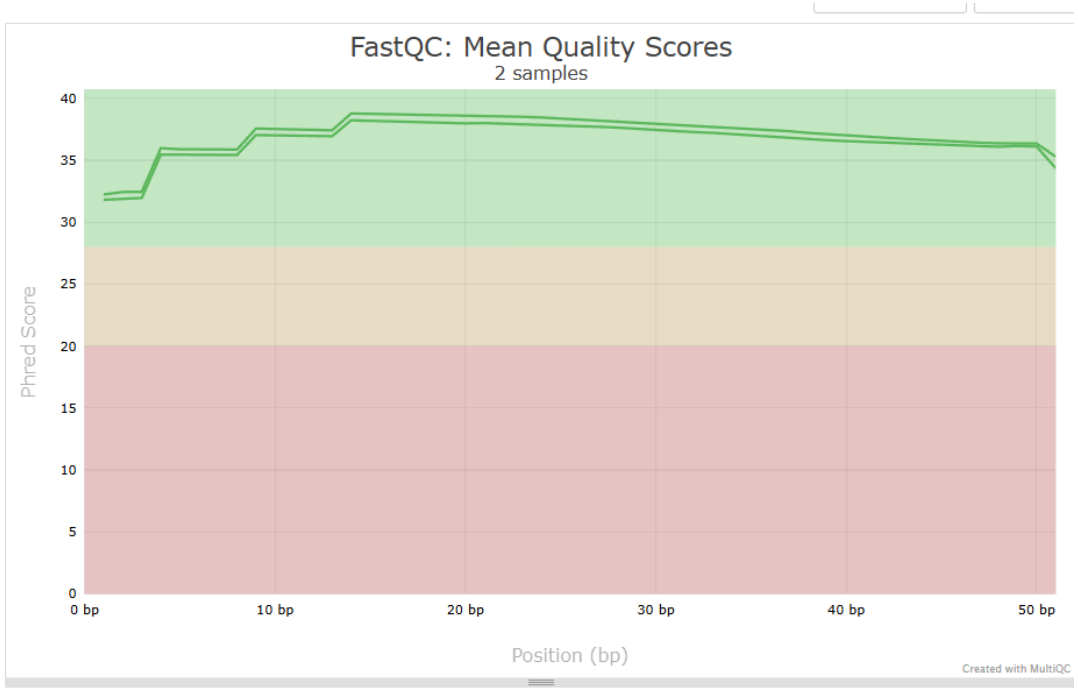
Sample Name	Dups	GC	Avg len	Median len	Failed	Seqs
Cutadapt on data 1 and data 2_ Read 1 Output	0.2 %	57.0 %	50 bp	51 bp	0 %	0.0 M
Cutadapt on data 1 and data 2_ Read 2 Output	0.2 %	57.0 %	50 bp	51 bp	0 %	0.0 M

MultiQC

Before Trimming



After Trimming



Question 1 A)

Samtools Stats

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 12	Column 13	Column 14	Column 15	Column 16	Column 17	Col 18
# This file was produced by samtools stats (1.20+htslib-1.20) and can be plotted using plot-bamstats																	
# This file contains statistics for all reads.																	
# The command line was: stats --ref-seq /data/db/reference_genomes/mm10/seq/mm10.fa -@ 1 infile																	
# CHK, Checksum	[2]Read Names	[3]Sequences	[4]Qualities														
# CHK, CRC32 of reads which passed filtering followed by addition (32bit overflow)																	
CHK	266378d0	c866af15	2df02b7b														
# Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.																	
SN	raw total sequences:	96762	# excluding supplementary and secondary reads														
SN	filtered sequences:	0															
SN	sequences:	96762															
SN	is sorted:	1															
SN	1st fragments:	48381															
SN	last fragments:	48381															
SN	reads mapped:	95449															
SN	reads mapped and paired:	94942	# paired-end technology bit set + both mates														

SN	reads mapped and paired:	94942	# paired-end technology bit set + both mates mapped
SN	reads unmapped:	1313	
SN	reads properly paired:	94030	# proper-pair bit set
SN	reads paired:	96762	# paired-end technology bit set
SN	reads duplicated:	0	# PCR or optical duplicate bit set
SN	reads MQ0:	557	# mapped and MQ=0
SN	reads QC failed:	0	
SN	non-primary alignments:	0	
SN	supplementary alignments:	0	
SN	total length:	4859237	# ignores clipping
SN	total first fragment length:	2418007	# ignores clipping
SN	total last fragment length:	2441230	# ignores clipping
SN	bases mapped:	4793328	# ignores clipping
SN	bases mapped (cigar):	4793328	# more accurate
SN	bases trimmed:	0	
SN	bases duplicated:	0	
SN	mismatches:	22154	# from NM fields
SN	error rate:	4.621841e-03	# mismatches / bases mapped (cigar)
SN	average length:	50	

SN	average length:	50
SN	average first fragment length:	50
SN	average last fragment length:	50
SN	maximum length:	51
SN	maximum first fragment length:	51
SN	maximum last fragment length:	51
SN	average quality:	36.9
SN	insert size average:	201.2
SN	insert size standard deviation:	63.8
SN	inward oriented pairs:	47003
SN	outward oriented pairs:	8
SN	pairs with other orientation:	17
SN	pairs on different chromosomes:	383
SN	percentage of properly paired reads (%):	97.2

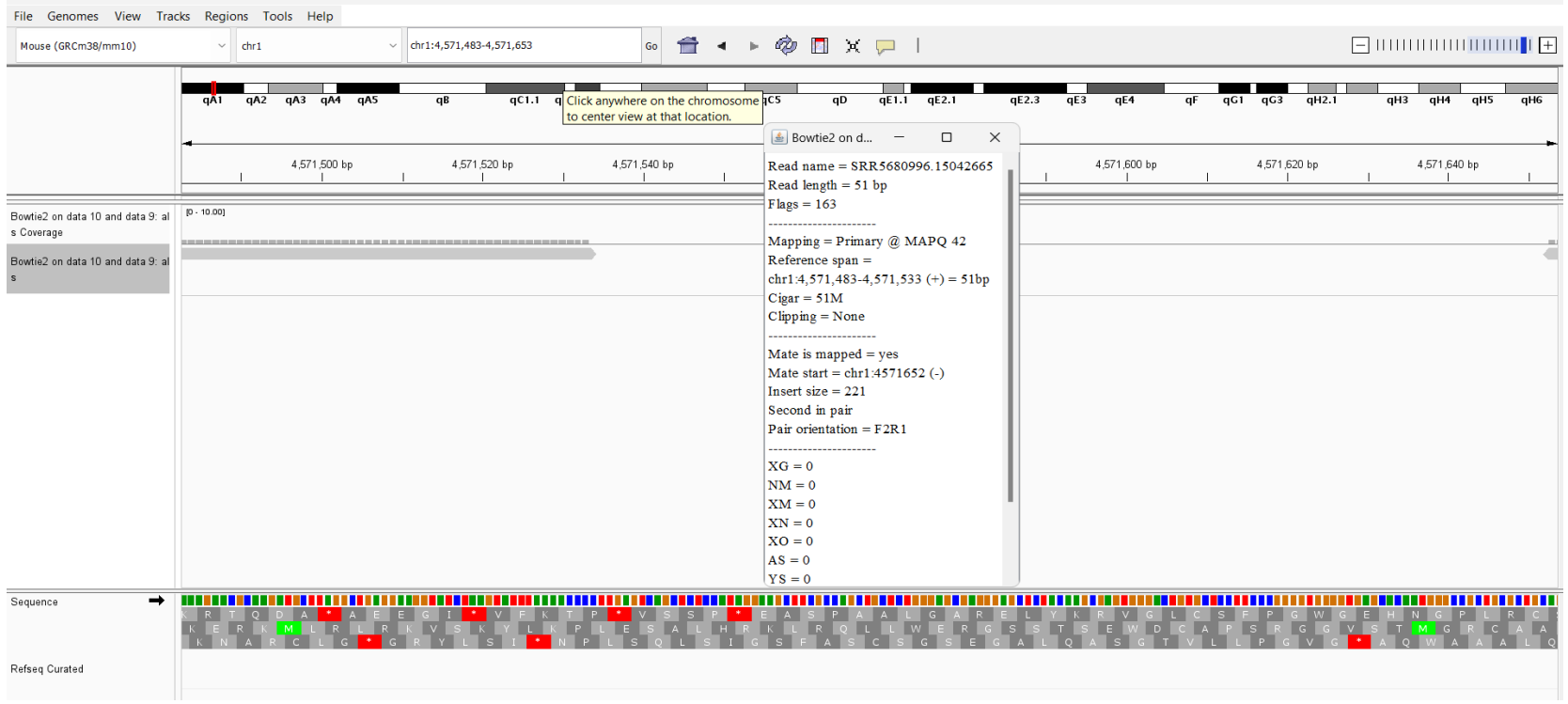
The mapping statistics show strong alignment results, with 96762 reads (98.6%) successfully mapped and only 1316 (1.36%) unmapped which indicates very few failures. A high proportion (97.2%) of reads were properly paired, suggesting good alignment quality. There were 22136 mismatches, resulting in a low error rate of 0.46% per base, which shows reliable sequencing. The average mapping quality score was 36.9 which is a high Phred score (meaning the probability of incorrect mapping is low). Additionally, the average read length was 50 bp, which is typical for short read sequencing data.

Question 1 B)


File Genomes View Tracks Regions Tools Help
Mouse (GRCm38/mm10)
chr2
chr2:98,666,236-98,667,473
Go
qA1 qA2 qA3 qB qC1.1 qC1.2 qC2 qC3 qD qE1 qE2 qE3 qE4 qE5 qF1 qF2 qF3 qG1 qG3 qH1 qH2 qH3 qH4
1,239 bp
98,666,400 bp
98,666,600 bp
98,666,800 bp
98,667,000 bp
98,667,200 bp
98,667,400 bp
Bowtie2 on data 10 and data 9: als Coverage
[0 - 198]
Bowtie2 on data 10 and data 9: als
Sequence
Refseq Curated

Question 1 C)

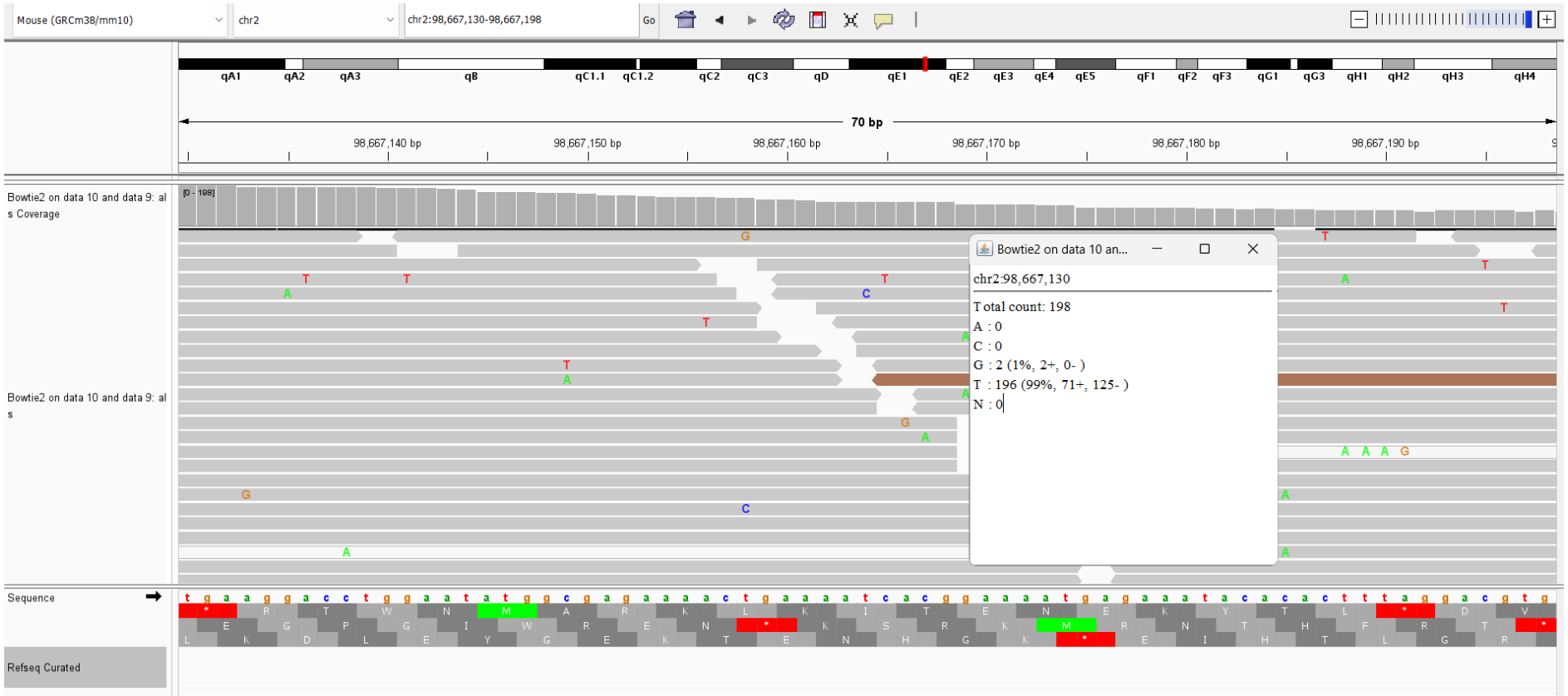
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL
SRR5680996.10098836	163	chr1	3671374	42	51M	=	3671515	191	CTCCTTTATCTGACACCTGTCCCTGCTGTCTCCTCCTCCCGCCACCTCCTC	?8?DDDDFDDDFHHIBGFGHGGG@GG@EBC:
SRR5680996.10098836	83	chr1	3671515	42	50M	=	3671374	-191	TGGGCCCGGCTAGCCGCACGCTCGGGACCCGCAGGAGCGCCGCCTGGCTG	B<85@93C>?B9=BBB@;CCD<6=E=/GEHF>G
SRR5680996.23952614	163	chr1	4484870	1	50M	=	4484971	152	TCAGAAGATGGAAAGATCTCCCATGCTCATGGATTGGCAGGATCAACATT	CCCFFFFDHHFHHIJJGGIIIGGJGFIJJHIGIJ.
S[illegible]5680996.23952614	83	chr1	4484971	1	51M	=	4484870	-152	ATCAAAATTCCAACTCAATTCTTCAACGAATTGGAAGGAGCAATTTGCAAA	GIGIIIIIIGHGC>EIHFEAFHIIIGHGGHFIF<H
SRR5680996.15042665	163	chr1	4571483	42	51M	=	4571653	221	AAAGAACGCAAGATGCTTAGGCTGAGGAAGGTATCTAAGTATTTAAAACCC	CCCFFFFFHHHHHJJJJJJJJJJJJJIJJJDGIJ.



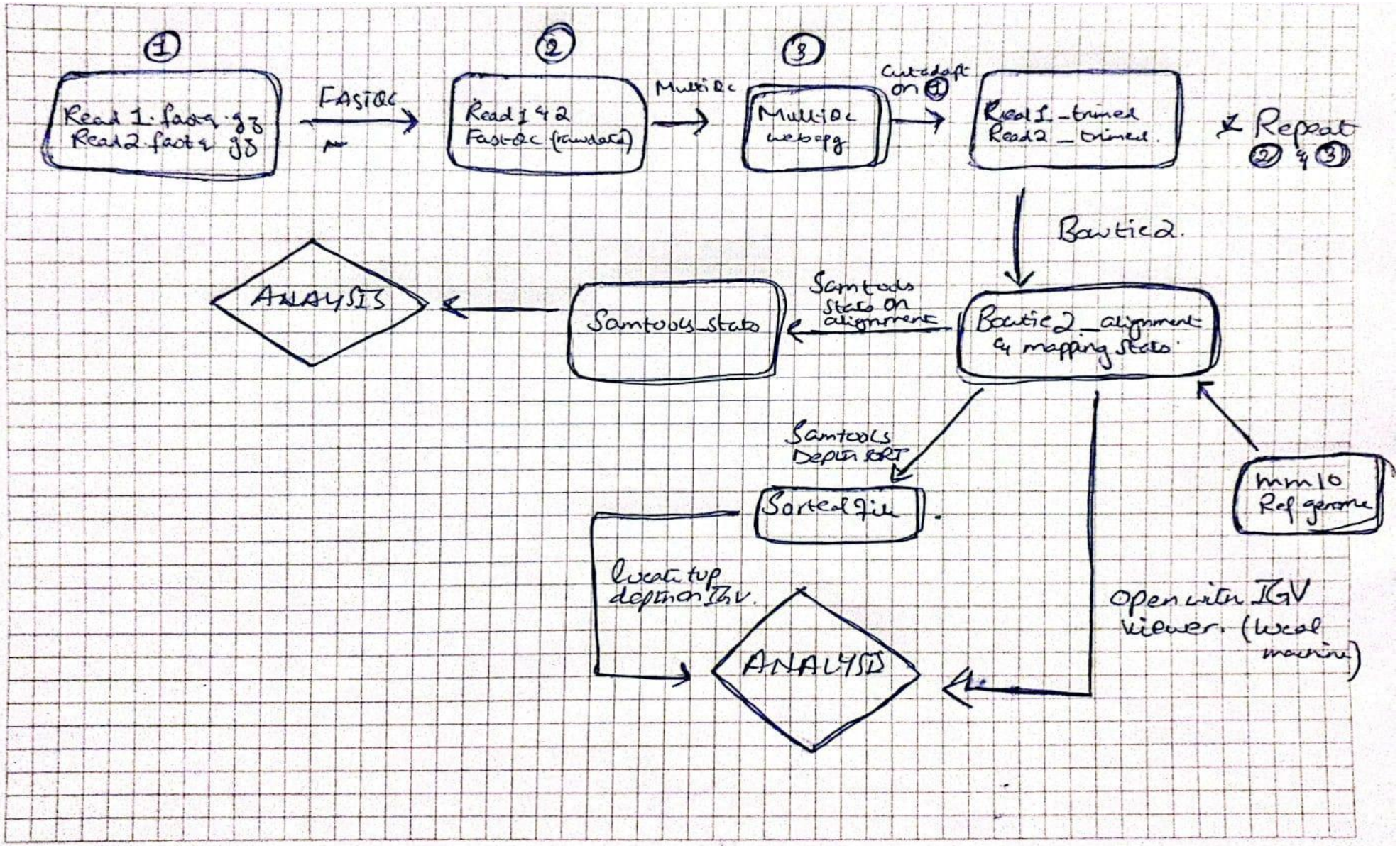
Question 1 D)

Column 1	Column 2	Column 3
chr2	98667130	198
chr2	98667131	198
chr2	98667132	197
chr2	98667128	195
chr2	98667129	194
chr2	98667133	193
chr2	98667135	193
chr2	98667136	193
chr2	98667138	192
chr2	98667134	191
chr2	98667137	191
chr2	98667126	190
chr2	98667127	190
chr2	98667139	190
chr2	98667140	188

The position with the highest read coverage is chr2:98,667,130, with 198 aligned reads. Nearly all reads (196, or 99%) contained the base T, while only 2 reads (1%) had a G, indicating very little variation. The strand distribution was uneven, with 71 reads on the forward strand and 125 on the reverse strand. Since T is overwhelmingly dominant, this suggests strong support for the reference base call at this position, with almost no evidence of a true variant.



WORKFLOW For Question 1



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG002_NA24385_son	HG003_NA24149_father	HG004_NA24143_moth
chr19	617804	.	G	A	0.309944987297	PASS	.	GT	0/1	0/0	0/0
chr19	618159	.	A	G	193.703994751	PASS	.	GT	1/1	1/1	0/1
chr19	619021	.	G	C	352.244995117	PASS	.	GT	1/1	1/1	0/1
chr19	619139	.	G	A	137.481994629	PASS	.	GT	1/1	0/1	0/0
chr19	619408	.	A	G	207.722000122	PASS	.	GT	1/1	0/1	0/1
chr19	619574	.	T	G	594.094970703	PASS	.	GT	1/1	1/1	1/1
chr19	619772	.	G	C	612.351989746	PASS	.	GT	1/1	1/1	0/1
chr19	619913	.	T	C	276.648986816	PASS	.	GT	1/1	0/1	0/1
chr19	620004	.	T	C	0.00524165015668	PASS	.	GT	0/1	0/0	0/1
chr19	620045	.	A	T	15.5733995438	PASS	.	GT	0/1	0/0	0/0
chr19	620381	.	T	G	0.000501392991282	PASS	.	GT	0/1	0/1	0/0
chr19	620728	.	A	G	0.0198510996997	PASS	.	GT	0/1	0/1	0/0
chr19	620807	.	G	A	41.4441986084	PASS	.	GT	0/1	0/1	0/0
chr19	620999	.	CGCCGGGGGAGGGCGCGGGGGT	C	86.6194000244	PASS	.	GT	1|1	0|1	0|0
chr19	621712	.	A	G	424.360992432	PASS	.	GT	1/1	1/1	0/1

Q2 A i) The child has a non reference allele at 15 sites (all)

Tool Parameters

GEMINI database *

7: GEMINI load on data 1 and data 5

accepted formats

Only files with version 0.20.1 are accepted.

Build GEMINI query using

Advanced query constructor

The query to be issued to the database *

SELECT * FROM variants

Formulate your query using SQL syntax. (-q)

Genotype filter expression

1: Genotype filter expression

Restrictions to apply to genotype values *

gt_types.HG002_NA24385_son <> HOM_REF

(--gt-filter)

+ Insert Genotype filter expression

Sample filter expression

+ Insert Sample filter expression

Output format options

Type of report to generate

VCF (simplified)

Add a header of column names to the output

Yes

(--header)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG002_NA24385_son	HG003_NA24149_father	HG004_NA24143_moth
chr19	618159	.	A	G	193.703994751	PASS	.	GT	1/1	1/1	0/1
chr19	619021	.	G	C	352.244995117	PASS	.	GT	1/1	1/1	0/1
chr19	619139	.	G	A	137.481994629	PASS	.	GT	1/1	0/1	0/0
chr19	619408	.	A	G	207.722000122	PASS	.	GT	1/1	0/1	0/1
chr19	619574	.	T	G	594.094970703	PASS	.	GT	1/1	1/1	1/1
chr19	619772	.	G	C	612.351989746	PASS	.	GT	1/1	1/1	0/1
chr19	619913	.	T	C	276.648986816	PASS	.	GT	1/1	0/1	0/1
chr19	620381	.	T	G	0.000501392991282	PASS	.	GT	0/1	0/1	0/0
chr19	620728	.	A	G	0.0198510996997	PASS	.	GT	0/1	0/1	0/0
chr19	620807	.	G	A	41.4441986084	PASS	.	GT	0/1	0/1	0/0
chr19	620999	.	CGCCGGGGGAGGGCGCGGGGGT	C	86.6194000244	PASS	.	GT	1|1	0|1	0|0
chr19	621712	.	A	G	424.360992432	PASS	.	GT	1/1	1/1	0/1

GEMINI query Querying the GEMINI database (Galaxy Version 0.20.1+galaxy2)

Tool Parameters

GEMINI database *

7: GEMINI load on data 1 and data 5

accepted formats ▾

Only files with version 0.20.1 are accepted.

Build GEMINI query using

Advanced query constructor

The query to be issued to the database *

SELECT * FROM variants

Formulate your query using SQL syntax. (-q)

Genotype filter expression

1: Genotype filter expression

Restrictions to apply to genotype values *

(gt_types.HG002_NA24385_son <> HOM_REF AND gt_types.HG003_NA24149_father <> HOM_REF)

(--gt-filter)

+ Insert Genotype filter expression

Sample filter expression

Q2 A ii) There are 12 sites where both father and son have a non reference allele

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG002_NA24385_son	HG003_NA24149_father	HG004_NA24143_mothe
chr19	618159	.	A	G	193.703994751	PASS	.	GT	1/1	1/1	0/1
chr19	619021	.	G	C	352.244995117	PASS	.	GT	1/1	1/1	0/1
chr19	619139	.	G	A	137.481994629	PASS	.	GT	1/1	0/1	0/0
chr19	619408	.	A	G	207.722000122	PASS	.	GT	1/1	0/1	0/1
chr19	619574	.	T	G	594.094970703	PASS	.	GT	1/1	1/1	1/1
chr19	619772	.	G	C	612.351989746	PASS	.	GT	1/1	1/1	0/1
chr19	619913	.	T	C	276.648986816	PASS	.	GT	1/1	0/1	0/1
chr19	620381	.	T	G	0.000501392991282	PASS	.	GT	0/1	0/1	0/0
chr19	620728	.	A	G	0.0198510996997	PASS	.	GT	0/1	0/1	0/0
chr19	620807	.	G	A	41.4441986084	PASS	.	GT	0/1	0/1	0/0
chr19	620999	.	CGCCGGGGGAGGGCGCGGGGGT	C	86.6194000244	PASS	.	GT	1|1	0|1	0|0
chr19	621712	.	A	G	424.360992432	PASS	.	GT	1/1	1/1	0/1

GEMINI query Querying the GEMINI database (Galaxy Version 0.20.1+galaxy2)

Tool Parameters

GEMINI database *

7: GEMINI load on data 1 and data 5

accepted formats

Only files with version 0.20.1 are accepted.

Build GEMINI query using

Advanced query constructor

The query to be issued to the database *

SELECT gts.HG002_NA24385_son, gts.HG003_NA24149_father from variants

Formulate your query using SQL syntax. (-q)

Genotype filter expression

1: Genotype filter expression

Restrictions to apply to genotype values *

(gt_types.HG002_NA24385_son <> HOM_REF AND gt_types.HG003_NA24149_father <> HOM_REF)

(--gt-filter)

+ Insert Genotype filter expression

Sample filter expression

+ Insert Sample filter expression

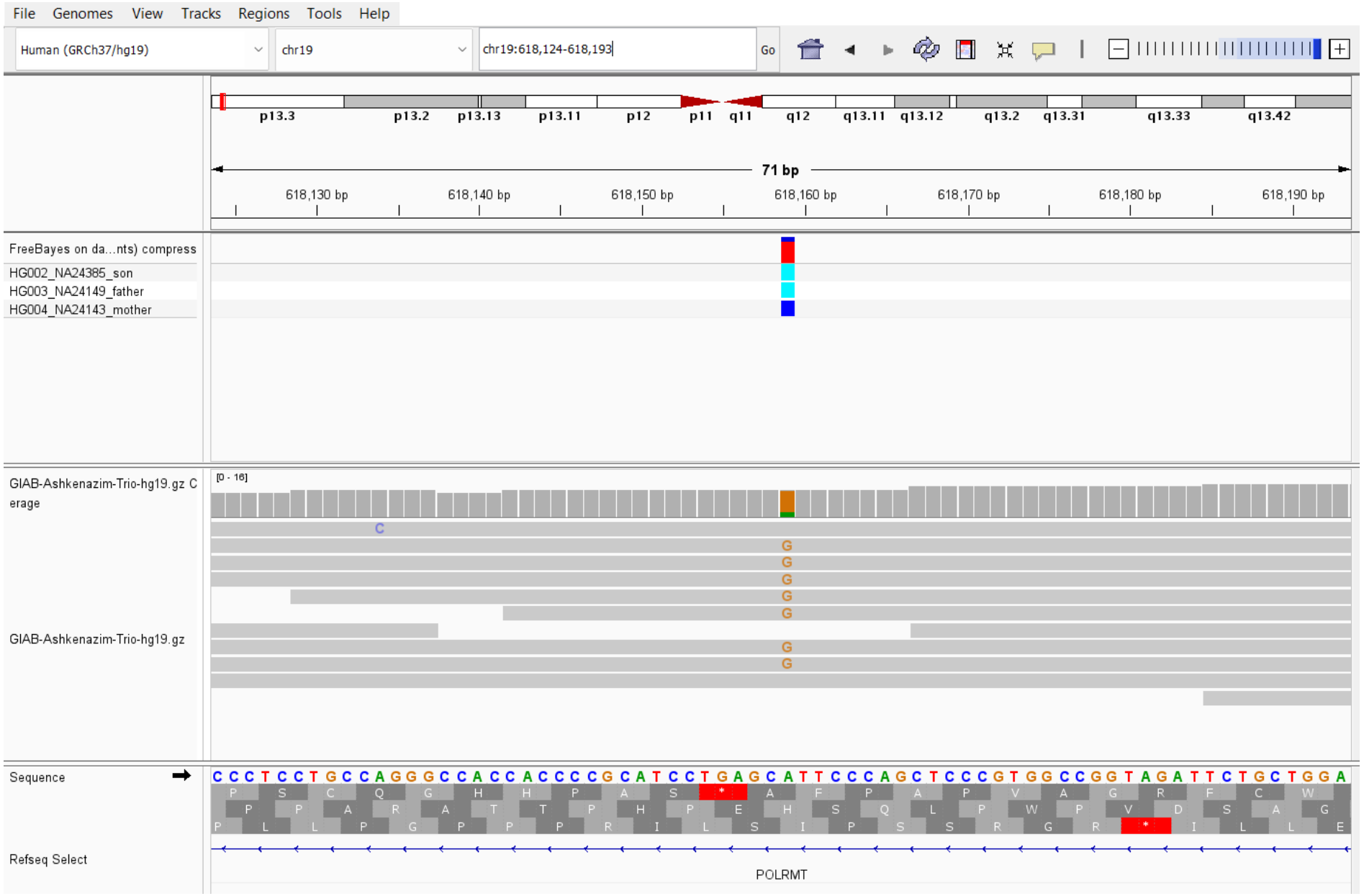
Output format options

Type of report to generate

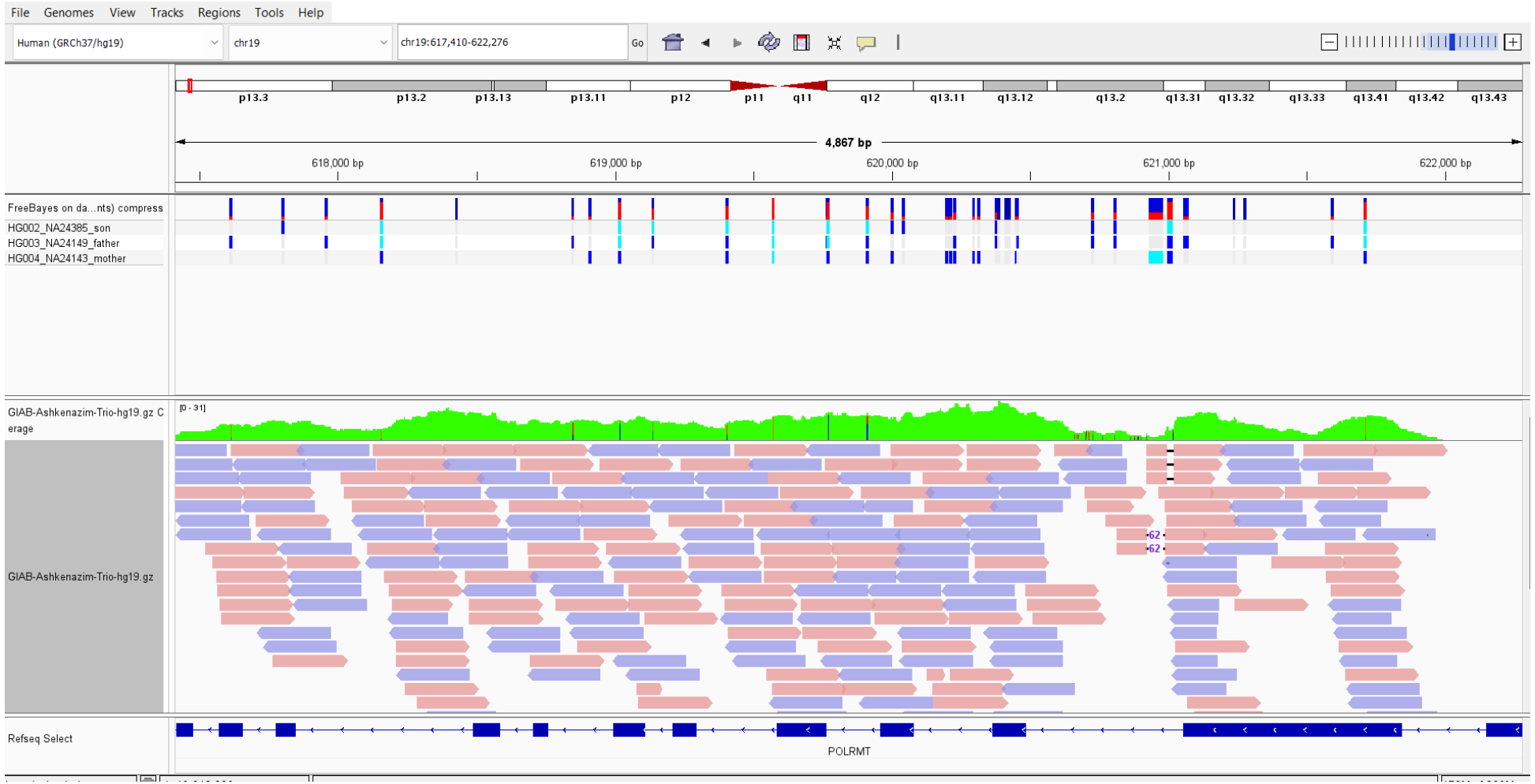
VCF (simplified)

Q2 A iii) A total of 12 genotypes where father and son have non reference alleles

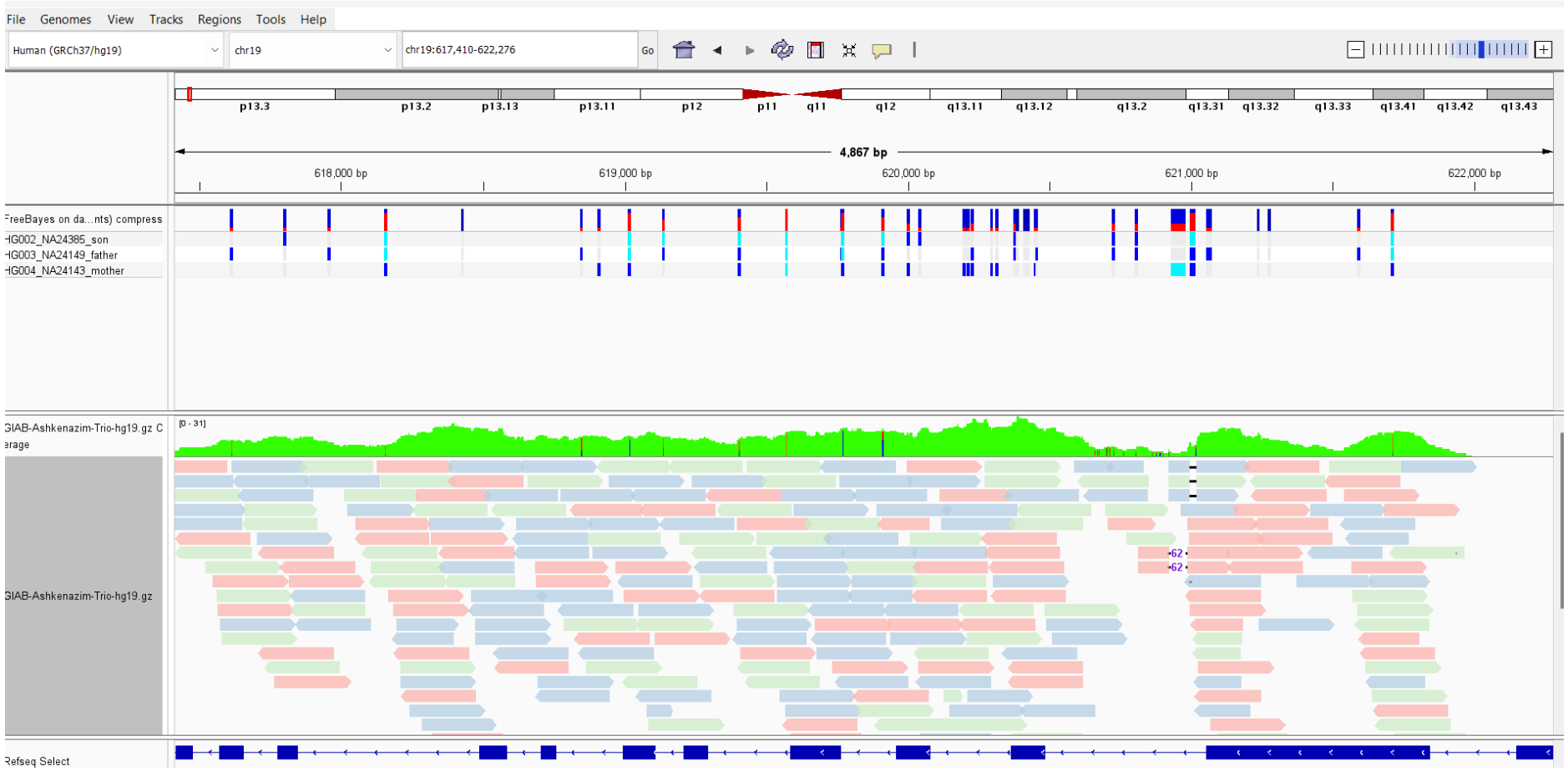
Question 2 B)


File Genomes View Tracks Regions Tools Help
Human (GRCh37/hg19)
chr19
chr19:618,124-618,193
Go
p13.3 p13.2 p13.13 p13.11 p12 p11 q11 q12 q13.11 q13.12 q13.2 q13.31 q13.33 q13.42
71 bp
618,130 bp 618,140 bp 618,150 bp 618,160 bp 618,170 bp 618,180 bp 618,190 bp
FreeBayes on da...nts) compress
HG002_NA24385_son
HG003_NA24149_father
HG004_NA24143_mother
GIAB-Ashkenazim-Trio-hg19.gz Coverage
[0 - 16]
GIAB-Ashkenazim-Trio-hg19.gz
Sequence
Refseq Select
POLRMT

Question 2 B i) Forward: PINK , Reverse: BLUE



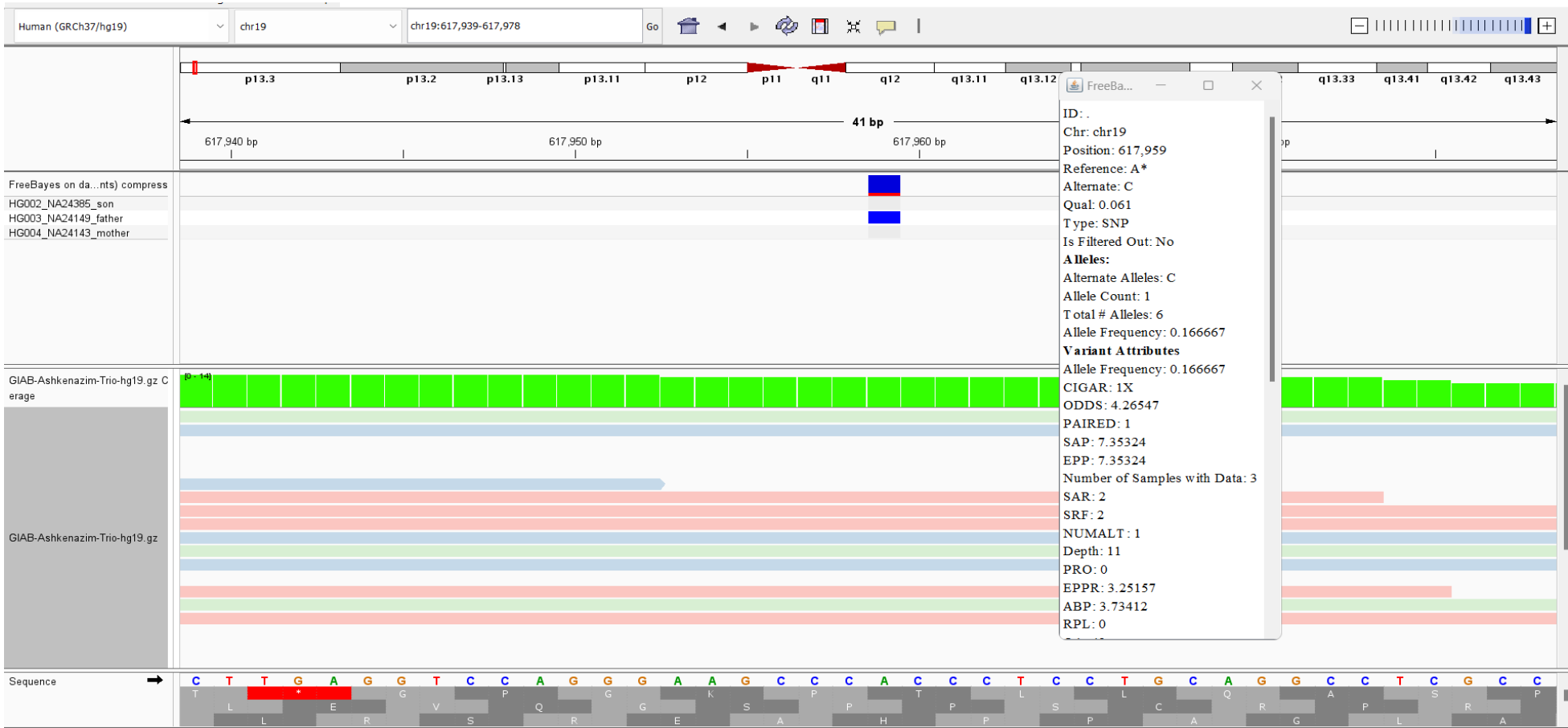
Question 2 ii) Blue: Father , Pink: Son, Green: Mother



Question 2 B iii)

The reference has an A at chr19:617,959. One sample is heterozygous (A/C) (allele frequency ~16.7%), while the other two are likely homozygous (A/A). Depth is 11 reads, with 2 supporting the alternate (C) and 9 the reference (A)

Chrom	Pos	ID	Ref	Alt	Qual	Filter	Info
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr19	617614	.	G	A	39.2648	.	AB=0.6;ABP=3.44459;AC=1;AF=0.166667;AN
chr19	617804	.	G	A	0.309945	.	AB=0.4;ABP=3.44459;AC=1;AF=0.166667;AN
chr19	617959	.	A	C	0.0608339	.	AB=0.666667;ABP=3.73412;AC=1;AF=0.1666
chr19	618159	.	A	G	193.704	.	AB=0.333333;ABP=3.73412;AC=5;AF=0.8333
chr19	618428	.	T	G	1.87498e-05	.	AB=0;ABP=0;AC=0;AF=0;AN=6;AO=2;CIGAR
chr19	618851	.	T	C	74.4613	.	AB=0.555556;ABP=3.25157;AC=1;AF=0.1666 primitives
chr19	618854	.	G	A	74.4613	.	AB=0.555556;ABP=3.25157;AC=1;AF=0.1666 primitives
chr19	618911	.	T	G	0.000288308	.	AB=0.5;ABP=3.0103;AC=1;AF=0.166667;AN=
chr19	619021	.	G	C	352.245	.	AB=0.25;ABP=5.18177;AC=5;AF=0.833333;AI



GIAB-Ashkenazim-Trio-hg19.gz

Read name = D00360:95:H2YWMBCXX:1:1207:10191:101090
Sample = HG004_NA24143_mother
Library = HG004_NA24143_mother
Read group = HG004_NA24143_mother
Read length = 250 bp
Flags = 147

Mapping = Primary @ MAPQ 60
Reference span = chr19:617,726-617,975 (-) = 250bp
Cigar = 250M
Clipping = None

Mate is mapped = yes
Mate start = chr19:617564 (+)
Insert size = -411
Second in pair
Pair orientation = F1R2

NM = 0
AS = 250
XS = 63
Hidden tags: MD, RG

Location = chr19:617,954
Base = A @ QV 40

GIAB-Ashkenazim-Trio-hg19.gz

Read name = D00360:95:H2YWMBCXX:1:2204:3959:9240
Sample = HG002_NA24385_son
Library = HG002_NA24385_son
Read group = HG002_NA24385_son
Read length = 250 bp
Flags = 147

Mapping = Primary @ MAPQ 60
Reference span = chr19:617,837-618,086 (-) = 250bp
Cigar = 250M
Clipping = None

Mate is mapped = yes
Mate start = chr19:617578 (+)
Insert size = -508
Second in pair
Pair orientation = F1R2

NM = 8
AS = 210
XS = 21
Hidden tags: MD, RG

Location = chr19:617,954
Base = A @ QV 14

GIAB-Ashkenazim-Trio-hg19.gz

Read name = D00360:95:H2YWMBCXX:1:2104:7178:14158
Sample = HG003_NA24149_father
Library = HG003_NA24149_father
Read group = HG003_NA24149_father
Read length = 250 bp
Flags = 83

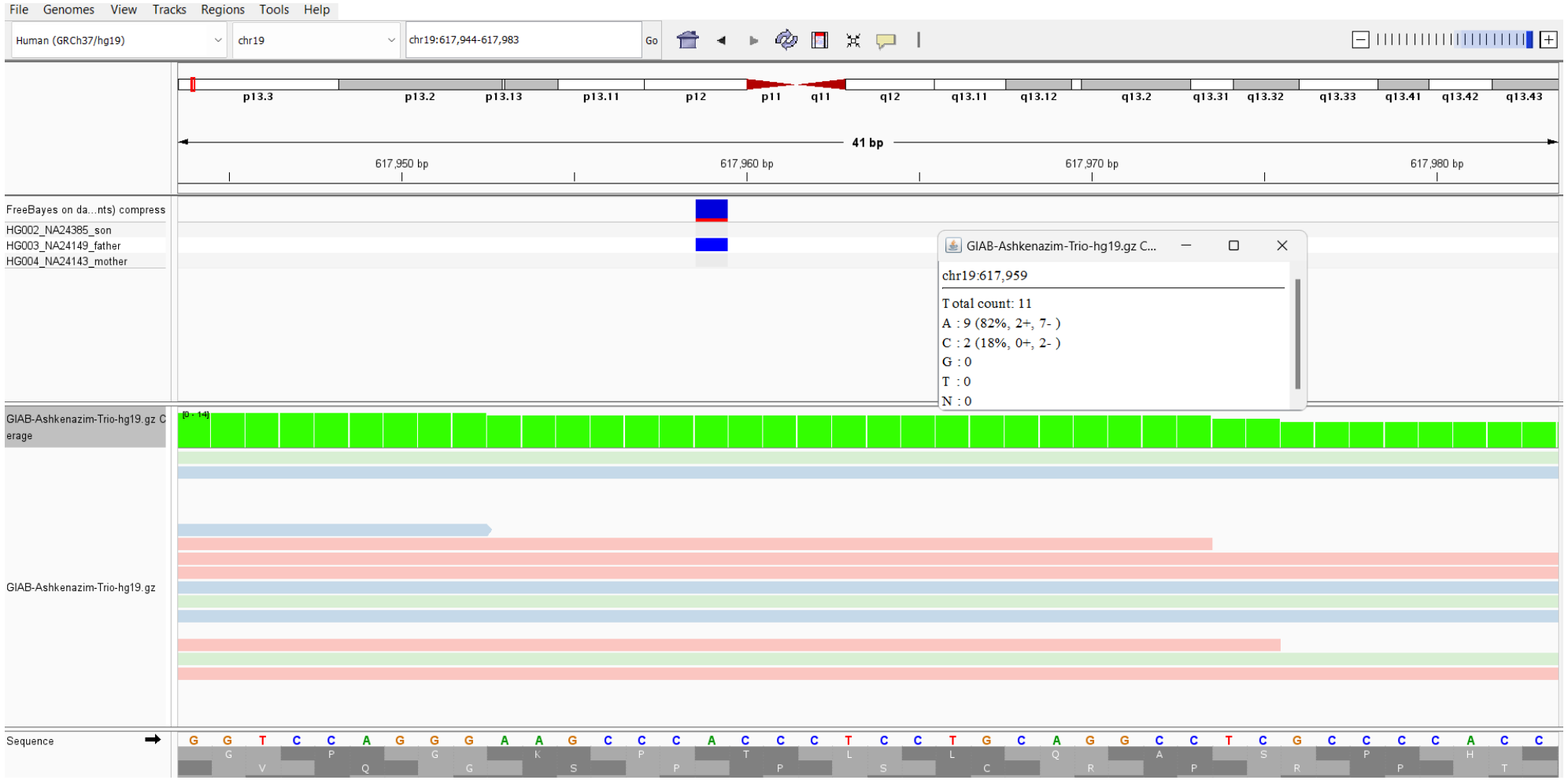
Mapping = Primary @ MAPQ 60
Reference span = chr19:617,856-618,105 (-) = 250bp
Cigar = 250M
Clipping = None

Mate is mapped = yes
Mate start = chr19:617611 (+)
Insert size = -494
First in pair
Pair orientation = F2R1

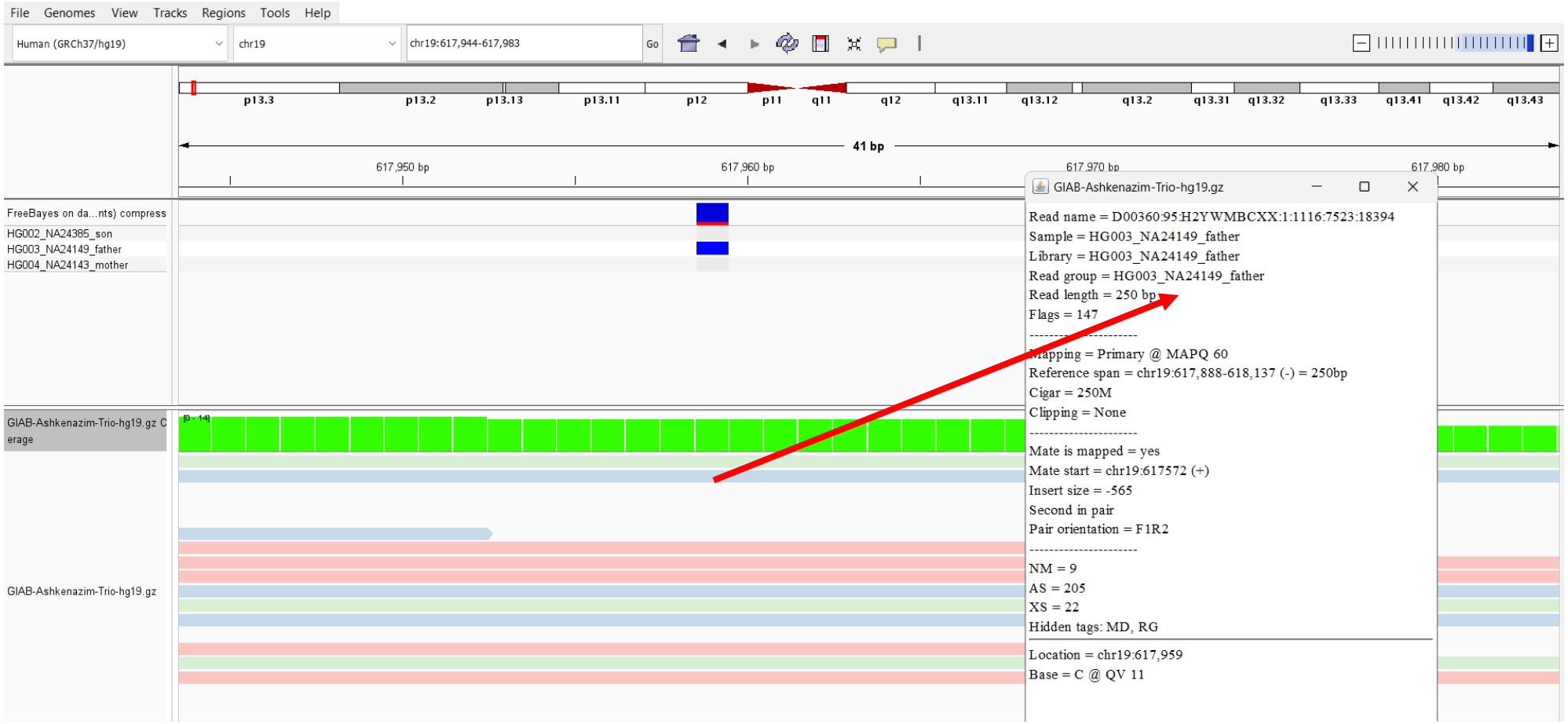
NM = 4
AS = 230
XS = 19
Hidden tags: MD, RG

Location = chr19:617,966
Base = T @ QV 39

Question 2 iv) At position chr19:617,959 the reference nucleotide is A and the alternate allele is C. The son has 18% of his reads supporting the alternate nucleotide C and 82% the reference nucleotide.



Question 2 v) The read belongs to the father with high-confidence alignment (MAPQ=60).
At chr19:617,959, the base is C (quality score 11), differing from the reference (A).
The read has 9 mismatches (NM=9) and an insert size of -565 bp.



Question 2 C) The SnpEff output shows 41 variants in the hg19 genome, averaging 1 variant per ~1.4 million bases. This suggests a relatively low mutation rate.

SnpEff: Variant analysis

Contents

Summary

Genome	hg19
Date	2025-04-20 23:18
SnpEff version	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /data/jwd07/main/082/279/82279080/outputs/dataset_0bc1b1b4-1000-40eb-b956-8ad7f2e9fc73.dat hg19 /data/dnb11/galaxy_db/files/c/a/2/dataset_ca2f0447-ba93-474d-9202-54d6221c0cc2.dat
Warnings	0
Errors	0
Number of lines (input file)	41
Number of variants (before filter)	41
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	41
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of effects	95
Genome total length	3,137,161,265
Genome effective length	59,128,983
Variant rate	1 variant every 1,442,170 bases

All 41 variants are on chromosome 19, mostly SNPs (39/41) with 2 deletions. Functionally, 10% are low impact, 11.6% moderate, and 65.3% modifiers (noncoding but potentially regulatory). Most variants fall in downstream regions (43.1%), while 32% are exonic and 2% near splice sites.

Variants rate details

Chromosome	Length	Variants	Variants rate
19	59,128,983	41	1,442,170
Total	59,128,983	41	1,442,170

Number variants by type

Type	Total
SNP	39
MNP	0
INS	0
DEL	2
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	41

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	12	12.632%
LOW	10	10.526%
MODERATE	11	11.579%
MODIFIER	62	65.263%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	11	61.111%
SILENT	7	38.889%

Missense / Silent ratio: 1.5714

Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
downstream_gene_variant	41	42.268%
intron_variant	23	23.711%
missense_variant	11	11.34%
sequence_feature	1	1.031%
splice_region_variant	2	2.062%
structural_interaction_variant	12	12.371%
synonymous_variant	7	7.216%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	41	43.158%
EXON	30	31.579%
INTRON	21	22.105%
SPLICE_SITE_REGION	2	2.105%
TRANSCRIPT	1	1.053%

The output includes quality statistics, indel counts, and SNP tables detailing base changes

Quality:

Min	0
Max	612
Mean	78.683
Median	0
Standard deviation	156.651
Values	0,6,10,15,39,41,74,86,137,193,207,276,352,424,594,612
Count	24,1,1,1,1,1,2,2,1,1,1,1,1,1,1,1

Insertions and deletions length:

Min	1
Max	1
Mean	1
Median	1
Standard deviation	0
Values	1
Count	2

Base changes (SNPs)

	A	C	G	T
A	0	9	5	1
C	1	0	0	0
G	6	2	0	0
T	0	7	8	0

The following matrix shows changes in codons

Codon changes

How to read this table:
- Rows are reference codons and columns are changed codons. E.g. Row 'AAA' column 'TAA' indicates how many 'AAA' codons have been replaced by 'TAA' codons.
- Red background colors indicate that more changes happened (heat-map).
- Diagonals are indicated using grey background color
- WARNING: This table may include different translation codon tables (e.g. mamalian DNA and mitochondrial DNA).

	AAC	ACC	ATC	ATT	CAA	CAC	CAT	CCC	CCG	CTG	GAA	GAC	GAG	GAT	GCA	GCC	GCT	GGA	GGG	GTG	TAC	TAT	TCC	TCT	TGG	TTG
AAC		1																								
ACC																										
ATC				1																						
ATT																										
CAA						1																				
CAC							1	1																		
CAT																										
CCC									1																	
CCG																										
CTG																										
GAA																		1								
GAC														1												
GAG																			2							
GAT																										
GCA																										
GCC																										
GCT															1	1										
GGA																										
GGG																										
GTG																			2							
TAC																							1			
TAT																								1		
TCC																										
TCT																										
TGG																			1							
TTG										1																

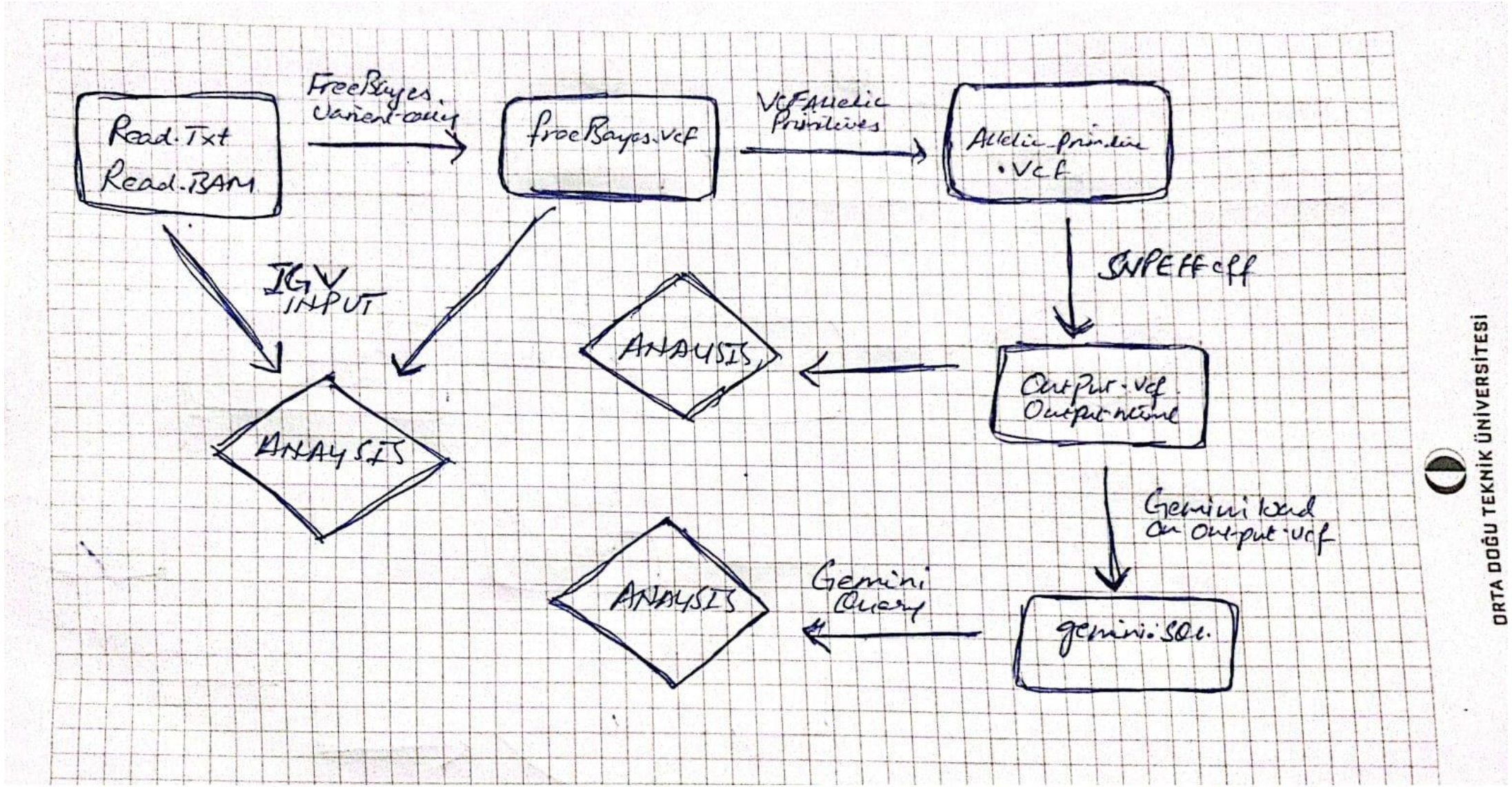
The following matrix shows the changes in amino acids

Amino acid changes

How to read this table:
- Rows are reference amino acids and columns are changed amino acids. E.g. Row 'A' column 'E' indicates how many 'A' amino acids have been replaced by 'E' amino acids.
- Red background colors indicate that more changes happened (heat-map).
- Diagonals are indicated using grey background color
- WARNING: This table may include different translation codon tables (e.g. mamalian DNA and mitochondrial DNA).

	A	D	E	G	H	I	L	N	P	Q	S	T	V	W	Y
A	2														
D		1													
E				3											
G															
H					1				1						
I						1									
L							1								
N												1			
P									1						
Q					1										
S															
T															
V				2											
W				1											
Y											2				

WORKFLOW For Question 2



Question 3:

a) What is the difference between SNV and SNP?

SNV (single nucleotide variant) refers to any single-base change in DNA while SNP (single nucleotide polymorphism) is a common SNV with a population frequency >1%.

b) Please define "haplotype". If there are 2 SNPs close to each other and SNP 1 has two possible alleles, A and T, and SNP 2 has two possible alleles, C and G, what are the possible haplotypes?

A haplotype is a set of alleles inherited together on a chromosome. Possible haplotypes here: A-C, A-G, T-C, T-G.

c) What are heterozygous and homozygous variants?

Heterozygous variants are two different alleles at a locus (e.g., A/T) while Homozygous variants are identical alleles (e.g., A/A)

d) Suppose I have 100 reads aligned to a position on the reference genome. I am investigating whether the individual from whom I obtained the DNA harbours a genetic variation at this specific region, which is known to contain a single nucleotide polymorphism (SNP).

i) 80 of the reads are reverse reads.

ii) 50 of the reads carry the reference nucleotide on the position of the SNP, while the other 50 carry the alternative nucleotide.

iii) All of the alternative nucleotides are observed in the reverse reads.

Can we say that the individual carries a heterozygous variant in this position? Why?

No. The 50/50 allele split could suggest heterozygosity, but all alternative alleles are on reverse reads, indicating potential strand bias (a sequencing artifact). True variants typically appear on both strands.