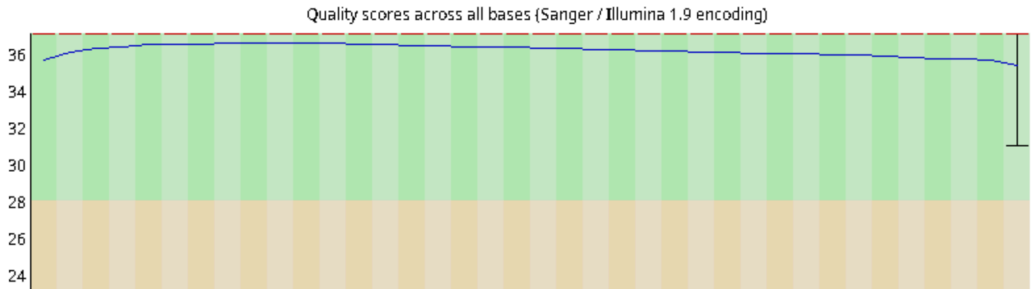
1. Первичный анализ данных (Fastqc до тримминга)

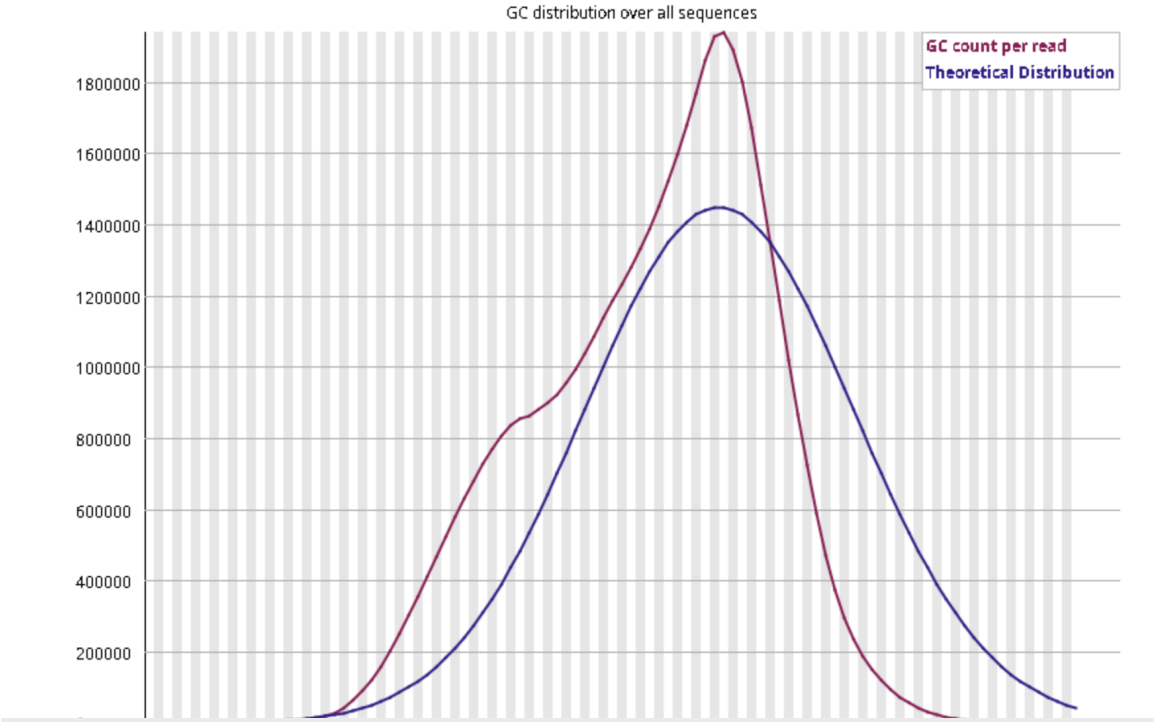
Fastqc прямого прочтения:

Measure	Value
Filename	13_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	52789607
Total Bases	7.9 Gbp
Sequences flagged as poor quality	0
Sequence length	151
%GC	53

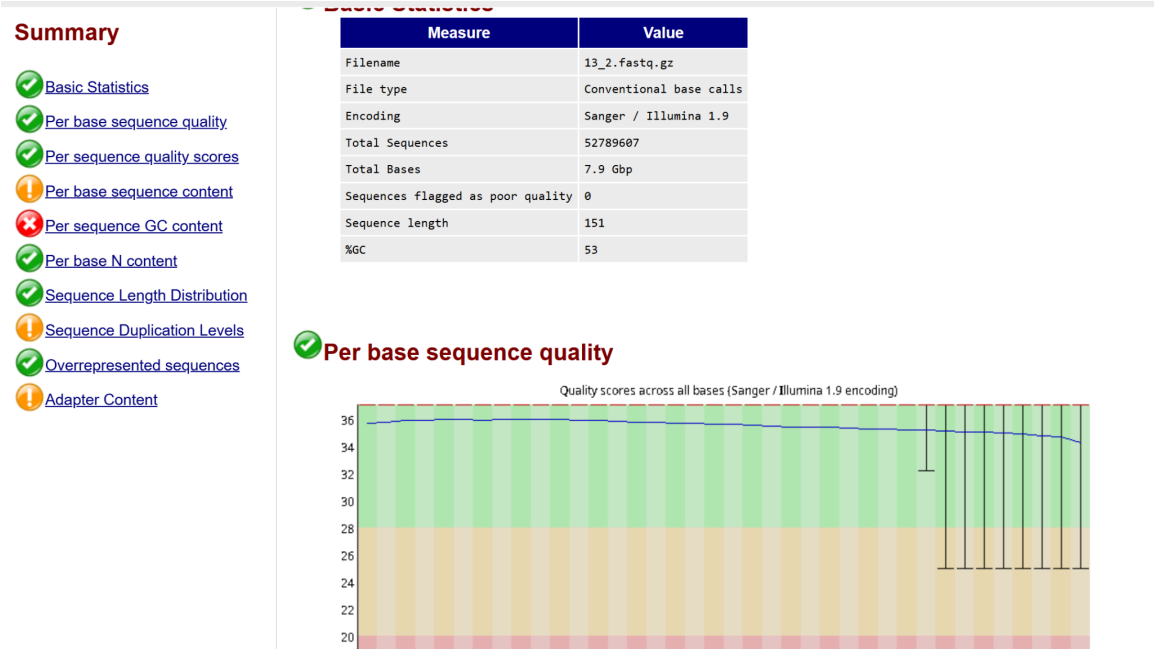
Per base sequence quality



Per sequence GC content



FastQc обратного прочтения:

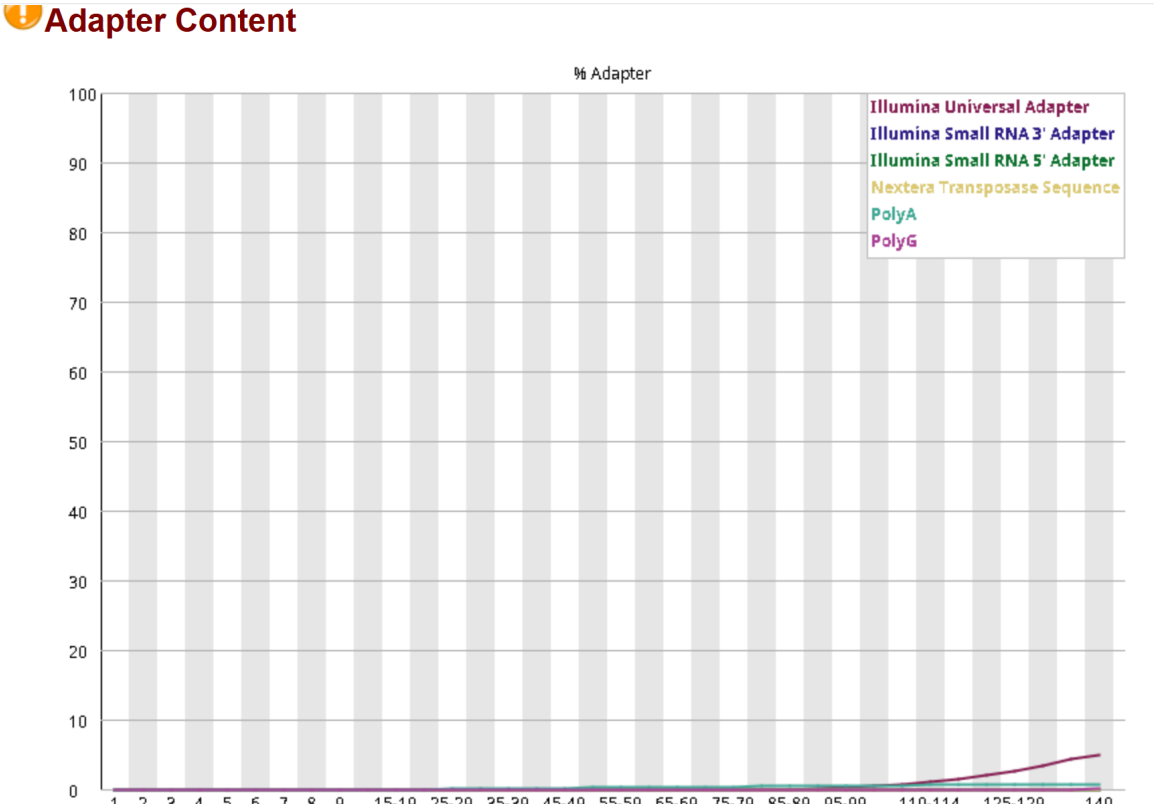


Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Measure	Value
Filename	13_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	52789607
Total Bases	7.9 Gbp
Sequences flagged as poor quality	0
Sequence length	151
%GC	53

Per base sequence quality



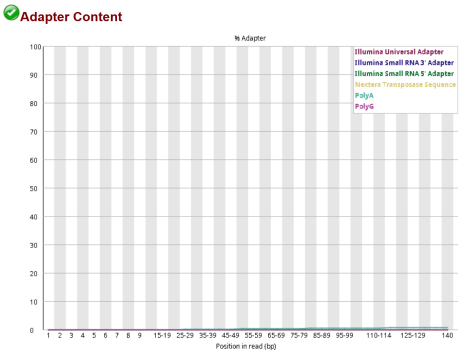
3. Тримминг

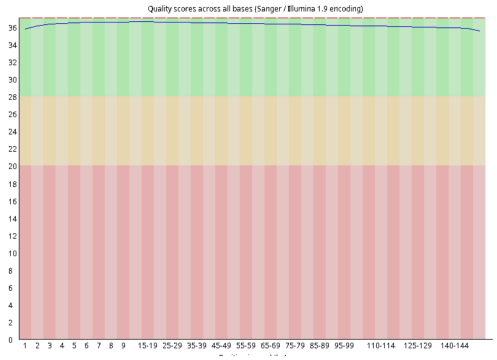
Триммирую с помощью инструмента fastp:

fastp -i 13_1.fastq.gz -I 13_2.fastq.gz -o 13_1_trimmed.fastq.gz -O 13_2_trimmed.fastq.gz

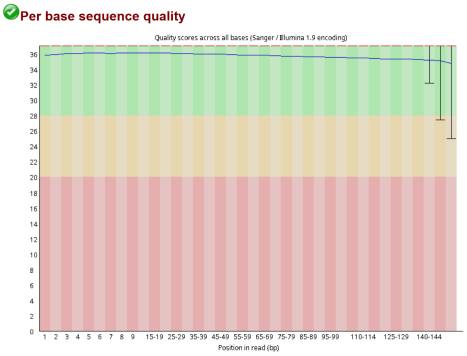
4. Анализ данных после тримминга

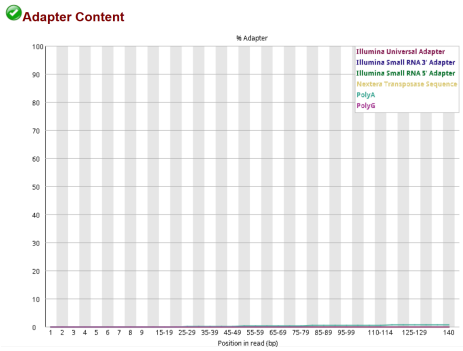
Результаты Fastqc после тримминга для прямого прочтения:





Результаты Fastqc после тримминга для обратного прочтения:





5. Картирование

hisat 3 из таблицы не подходит для картирования прочтений из illumina. Поэтому полный скрипт для картирования я писал для bowtie2.

Скрипт состоит из нескольких этапов:

1. Картирование с помощью bowtie2
2. Сортировка с помощью samtools
3. Построение графиков по bam файлу с помощью plot-bamstats
4. Препроцессинг ридов:
    a. gatk ReadGroups
    b. gatk MarkDuplicates
5. Variant calling с помощью bcftools mpileup | bcftools call
6. Фильтрация vcf файла

7. Аннотация VCF с помощью ANNOVAR по RefGene и ClinVar

Полный скрипт (Blackbox.ai красиво его переписал, чтобы было читаемо):

Step 1: Input Parameters

- *Write the path to bowtie2 index:* `$1`
- `read BOWTIE2_INDEX`
- *Write the path to your fastq file 1:* `read FASTQ_FILE_1`
- *Write the path to your fastq file 2:* `read FASTQ_FILE_2`
- *Write number of threads for bowtie2:* `read THREADS`
- *Write path to output directory:* `read OUTPUT_DIR`
- *Write the name for output SORTED BAM file (without.bam at the end, just name):* `read OUTPUT_FILE_NAME`
- *Write the path to your ref sequence fasta:* `read REFERENCE_FASTA`

Step 2: Create Output Directory

- `mkdir "./${OUTPUT_DIR}/"`

Step 3: Run Bowtie2 and Samtools

- `bowtie2 --threads $THREADS -x $BOWTIE2_INDEX --sensitive -1 $FASTQ_FILE_1 -2 $FASTQ_FILE_2 | samtools view -@ $THREADS -S -b -F 4 | samtools sort -@ $THREADS > "${OUTPUT_DIR}/${OUTPUT_FILE_NAME}_sorted.bam"`
- `samtools index -@ $THREADS "${OUTPUT_DIR}/${OUTPUT_FILE_NAME}_sorted.bam"`

Step 4: Generate BAM File Stats

- `mkdir "${OUTPUT_DIR}/bam_file_stats"`
- `samtools stats -@ $THREADS "${OUTPUT_DIR}/${OUTPUT_FILE_NAME}_sorted.bam" > "${OUTPUT_DIR}/bam_file_stats/${OUTPUT_FILE_NAME}_stats_file.stats"`
- `plot-bamstats -p "${OUTPUT_DIR}/bam_file_stats/${OUTPUT_FILE_NAME}_stats_file_graph" "${OUTPUT_DIR}/bam_file_stats/${OUTPUT_FILE_NAME}_stats_file.stats"`

Step 5: Generate Coverage File

- `samtools coverage "${OUTPUT_DIR}/${OUTPUT_FILE_NAME}_sorted.bam" > "${OUTPUT_DIR}/bam_file_stats/${OUTPUT_FILE_NAME}_coverage_file.txt"`

Step 6: Variant Calling

- `mkdir "${OUTPUT_DIR}/variant_calling"`
- `gatk AddOrReplaceReadGroups -I "${OUTPUT_DIR}/${OUTPUT_FILE_NAME}_sorted.bam" -O "${OUTPUT_DIR}/variant_calling/${OUTPUT_FILE_NAME}_sorted_rg.bam" --RGID 1 --RGLB lib1 --RGPL ILLUMINA --RGPU unit1 --RGSM 20`
- `gatk MarkDuplicates -I "${OUTPUT_DIR}/variant_calling/${OUTPUT_FILE_NAME}_sorted_rg.bam" -O "${OUTPUT_DIR}/variant_calling/${OUTPUT_FILE_NAME}_sorted_markdup.bam" -M "${OUTPUT_DIR}/variant_calling/markdup_info.txt"`
- `bcftools mpileup -f $REFERENCE_FASTA "${OUTPUT_DIR}/variant_calling/${OUTPUT_FILE_NAME}_sorted_markdup.bam" -Ou -@ $THREADS | bcftools call --threads $THREADS --ploidy 2 -mv -Oz -o "${OUTPUT_DIR}/variant_calling/${OUTPUT_FILE_NAME}.vcf.gz"`
- `bcftools filter -sLowQual -g3 -G10 -e'%QUAL<10 || (RPB<0.1 && %QUAL<15) || (AC<2 && %QUAL<15) || %MAX(DV)<=3 || %MAX(DV)/%MAX(DP)<=0.3' "${OUTPUT_DIR}/variant_calling/${OUTPUT_FILE_NAME}.vcf.gz"`

Step 7: Annotate Variants

- `perl convert2annovar.pl "${OUTPUT_DIR}/variant_calling/${OUTPUT_FILE_NAME}.vcf.gz" > output.avinput`
- `table_annovar.pl example/ex1.avinput humandb/ -buildver hg38 -out myanno -remove -protocol refGene, ClinVar -operation gx -nastring. -csvout -polish`

ПОЧЕМУ НЕ ИСПОЛЬЗОВАЛ Octopus?

У меня получилось установить и составить скрипт для запуска Variant Callinga именно в Octopus’e:

С ним бы точно получилось интереснее, но команда для его запуска обязательно требует файл с random_forest. Видимо он необходим для работы нейросети Octopus'а:

```
$ octopus \
     -R data/reference/hs38DH.fa \
     -I data/reads/mapped/CHM1-CHM13.hs38DH.bwa-mem.bam \
     -T chr1 to chrM \
     --sequence-error-model PCRF.X10 \
     --forest resources/forests/germline.v0.8.0.forest \
     -o results/calls/CHM1-CHM13.hs38DH.bwa-mem.octopus.vcf.gz \
     --threads 16
```

Однако в ходе установки самого Octopus'а директория resources/forests у меня вообще была пустая. Сам octopus устанавливается через исполняемый .py файл, который уже сам при запуске подкачивает нужные файлы по ссылкам (запуска установщика прошёл без ошибок). Поэтому я не могу найти forest-файл в исходниках на github, потому что его там просто нет. Отдельно этот файл для запуска загуглить тоже не получилось(

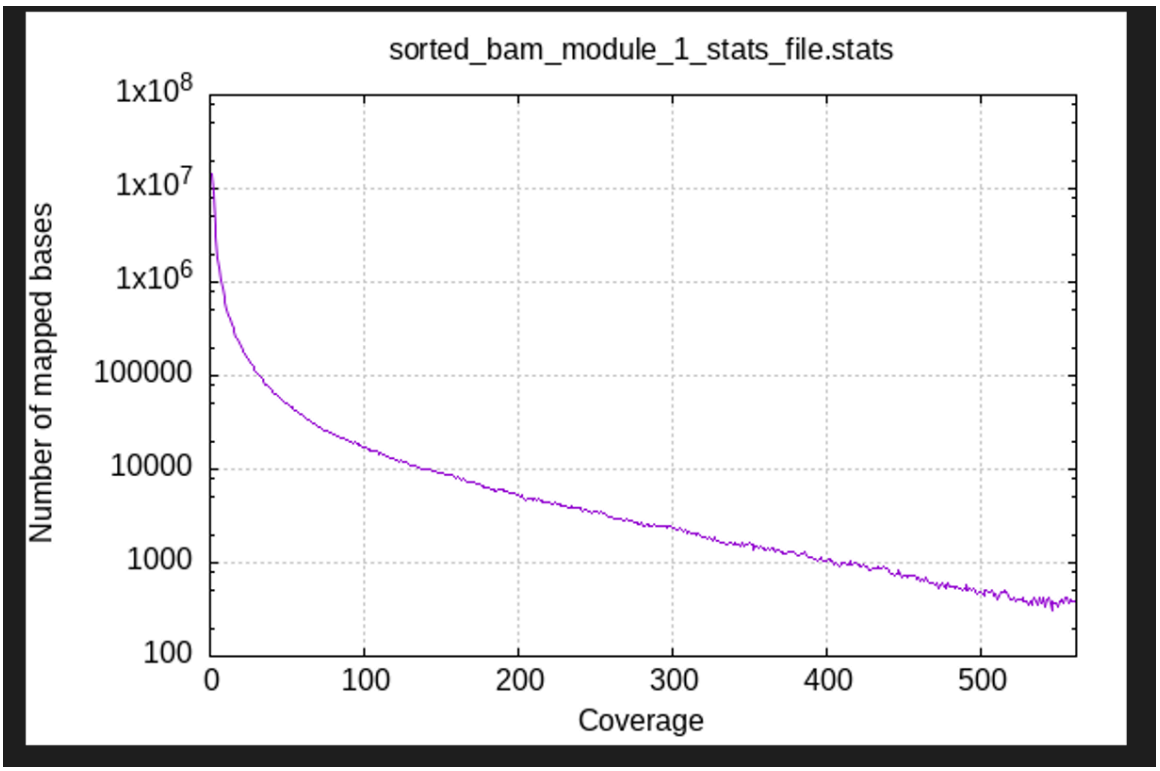
6. Статистика картирования

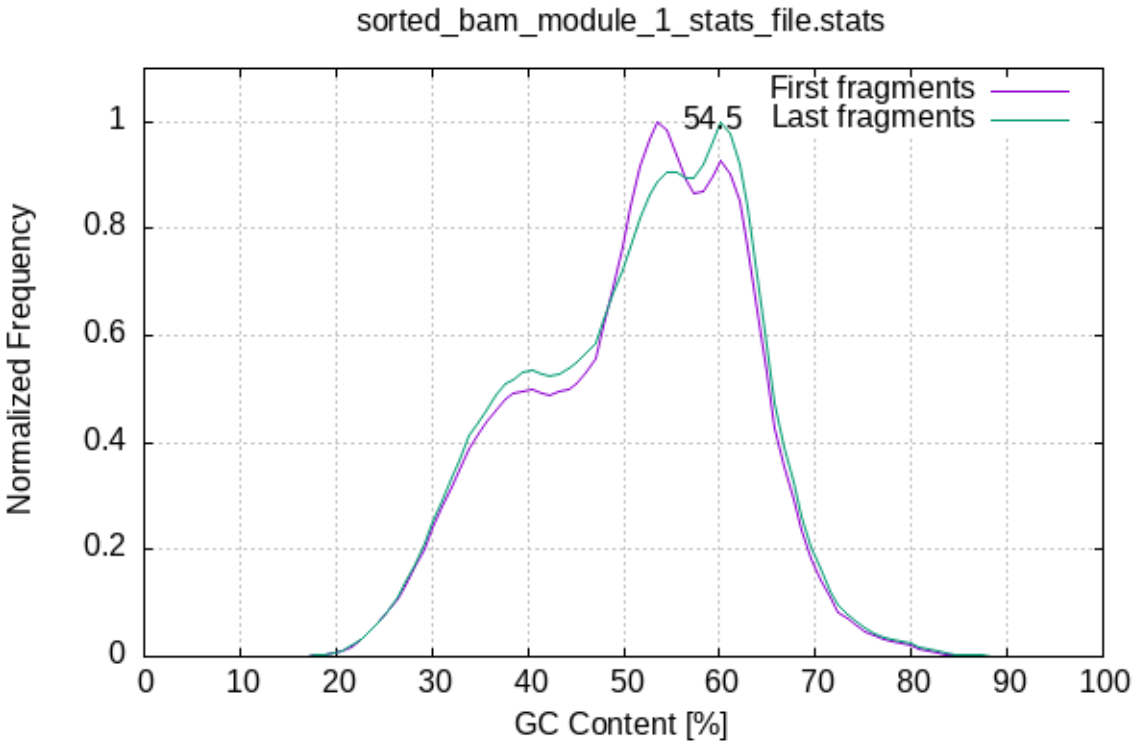
Описание статистики картирования, используемые команды и инструменты.

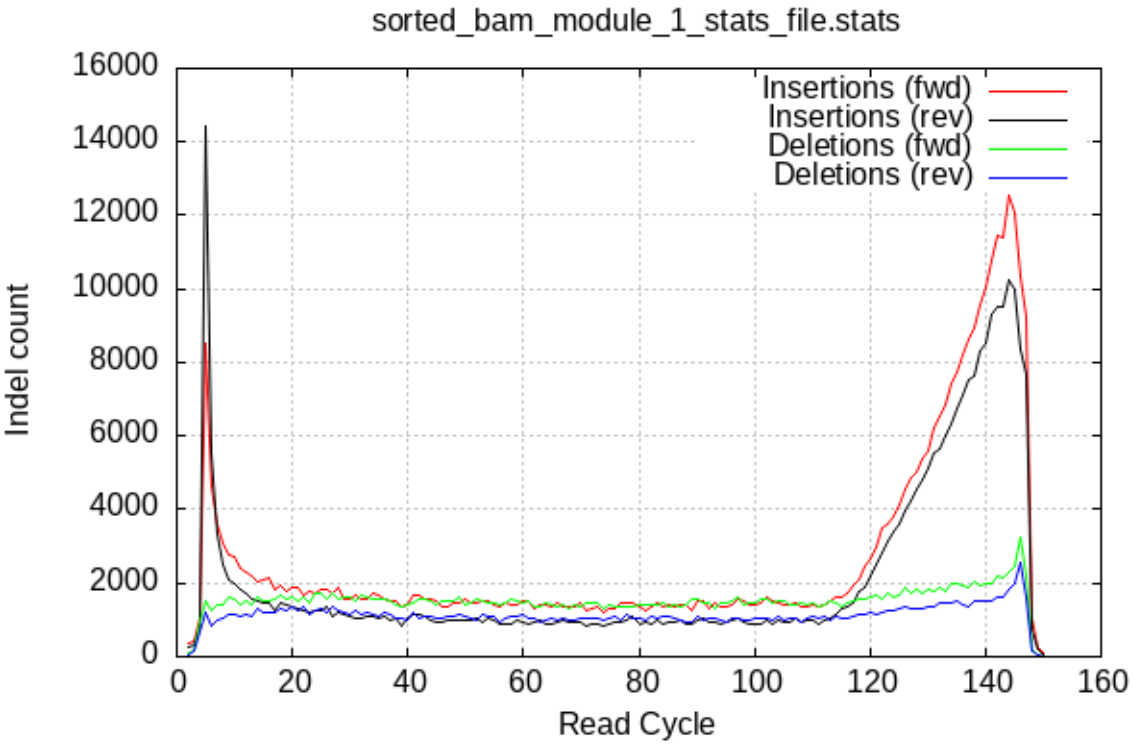
Графики из samtools coverage:

#rname startpos endpos numreads covbases coverage meandepth meanbaseq meanmapq

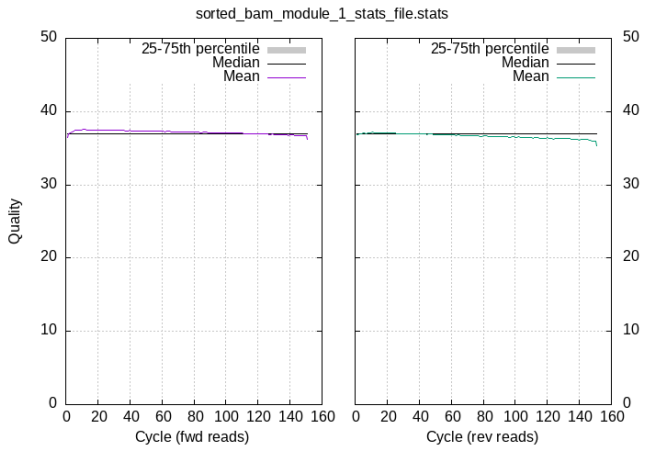
chr6 1 170805979 5406622 46481896 27.2133 4.76502 35.9 30.7


sorted_bam_module_1_stats_file.stats
Number of mapped bases
1x10^8
1x10^7
1x10^6
100000
10000
1000
100
0
100
200
300
400
500
Coverage


sorted_bam_module_1_stats_file.stats
First fragments
Last fragments
54.5
Normalized Frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
GC Content [%]



Видно, что большинство делеций/инсерций находится на концах ридов, что соответсвует необрезанным адаптерам и снижению качетсва прочтения в начале и конце рида



9. Заключение

Annovar выдает следующие строки:

Chr	Start	End	Ref	Alt	Func.refGene
chr6	60113	60113	T	G	intergenic
chr6	60720	60720	C	G	intergenic
chr6	60858	60858	C	T	intergenic
chr6	68927	68927	G	T	intergenic
chr6	70495	70495	A	G	intergenic
chr6	70764	70765	TG	-	intergenic

Gene.refGene	GeneDetail.refGene	ExonicFunc.refGene	AAChange.refGene	CLNALLELEID	CLNDN
NONE;LINC00266-3	dist=NONE;dist=80151	.	.	.	.
NONE;LINC00266-3	dist=NONE;dist=79544	.	.	.	.
NONE;LINC00266-3	dist=NONE;dist=79406	.	.	.	.
NONE;LINC00266-3	dist=NONE;dist=71337	.	.	.	.
NONE;LINC00266-3	dist=NONE;dist=69769	.	.	.	.
NONE;LINC00266-3	dist=NONE;dist=69499	.	.	.	.
NONE;LINC00266-3	dist=NONE;dist=8297	.	.	.	.
NONE;LINC00266-3	dist=NONE;dist=7980	.	.	.	.
NONE;LINC00266-3	dist=NONE;dist=6659	.	.	.	.
NONE;LINC00266-3	dist=NONE;dist=6581	.	.	.	.

CLNDN	CLNDISDB	CLNREVSTAT	CLNSIG
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

По вкладке CLNSIG только 1 мутация оценивается как Pathologic:

11568 chr6 31593133 31593133 C G upstream NCR3 dist=127 . . 15915 Malaria\x2c_mild\x2c_susceptibility_to|Malaria\x2c_severe\x2c_susceptibility_to MONDO:MONDO:0012202\x2cMedGen:C1836721\x2cOMIM:609148|MedGen:C1970029 no_assertion_criteria_provided Pathogenic|risk_factor

Найденная мутация в ClinVar:

[VCV000000876.4 - ClinVar - NCBI (nih.gov)]

NM_001145466.1(NCR3):c.-412G>C

Classification (Last evaluated)	Review status (Assertion criteria)	Condition	Submitter	More information
risk factor (Feb 01, 2007)	☆☆☆☆ Method: literature only	MALARIA, MILD, SUSCEPTIBILITY TO Affected status: not provided Allele origin: germline	OMIM Accession: SCV000021074.1 First in ClinVar: Apr 04, 2013 Last updated: Apr 04, 2013	Publications: PubMed (1)
Pathogenic (Dec 06, 2022)	☆☆☆☆ Method: research	Malaria, severe, susceptibility to Affected status: yes Allele origin: germline	Center for Global Health, University of New Mexico Health Sciences Center, University of New Mexico Accession: SCV002762724.1 First in ClinVar: Dec 17, 2022 Last updated: Dec 17, 2022	

Comment:
CC is wild type in the Luo (Kenya) population, GG is homozygous mutant. Additive model of inheritance shows increased susceptibility to longitudinal (over 36 months) severe malarial anemia (Hb<5.0 g/dL with any density Plasmodium falciparum parasitemia) in children <48 months of age. (less)
Number of individuals with the variant: 753
Age: 1-40 months
Sex: mixed
Ethnicity/Population group: Luo
Geographic origin: Kenya

Скорее всего мутация снижает устойчивость к малярийным горячкам, однако слишком мало исследований по этой мутации. Замечена только ассоциация

Открываю общий файл .vcf в IGV и ищу нужную координату:

ID: .
Chr: chr6
Position: 31 593 133
Reference: C*
Alternate: G
Qual: 222
Type: SNP
Is Filtered Out: No
Alleles:
Alternate Alleles: G
Allele Count: 1
Total # Alleles: 2
Variant Attributes
Allele Count: 1
Mapping Quality: 41
RPB: 0.330443
ICB: 1
DP4: [35, 18, 22, 8]
Depth: 109
Total Alleles: 2
SGB: -0.693097
BQB: 0.893768
VDB: 0.00240868
MQSB: 0.815761
HOB: 0.5
MQ0F: 0
MQB: 0.950834

Видно, что данная позиция имеет хорошее покрытие (109) и качество картирования. За отсутствием других патогенных вариантов в файле, считаю это главным патогенным вариантом