

Анализ экзома.

1. Обработка данных.

Были даны три образца секвенирования белок кодирующих генов человека методом парно-концевых прочтений:

HG00103.final_1.fastq.gz HG03280.final_1.fastq.gz NA18966.final_1.fastq.gz

HG00103.final_2.fastq.gz HG03280.final_2.fastq.gz NA18966.final_2.fastq.gz

Необходимо было:

обработать данные согласно протоколу, аннотировать варианты, описать один потенциально самый опасный вариант для каждого образца, установить Y и MT гаплогруппы.

При первом этапе обработки с помощью gatk FastqToSam данные из формата fastq.gz были преобразованы в формат .bam, который требуется для дальнейших этапов пайплана. Также в данные были добавлены метаданные, необходимые для корректной работы пайплана: название образца, название группы, название платформы и название библиотеки. Полученные три файла:

HG00103.final_1.fastq.bam HG03280.final_1.fastq.bam NA18966.final_1.fastq.bam

На втором этапе проведено выравнивание данных на рефересный геном 38 сборки (GRCh38.p14, GenBank) с помощью gatk BwaSpark. Результат:

HG00103.final_1.fastq.bam HG03280.final_1.fastq.bam
NA18966.final_1.fastq.bam

HG00103.final_1.fastq.bam.sbi HG03280.final_1.fastq.bam.sbi
NA18966.final_1.fastq.bam.sbi

Далее были отмечены дубликаты в данных с помощью gatk MarkDuplicatesSpark.

Получены три файла .txt со статистикой:

HG00103.final_1.fastq.txt

LIBRARY	UNPAIRED_READS_EXAMINED	READ_PAIRS_EXAMINED	SECONDARY_ALIGNMENT_RDS	UNMAPPED_READS	UNPAIR_ED_READ_DUPLICATES	READ_PAIR_DUPLICATES	READ_PAIR_OPTICAL_DUPLICATES	PERCENT_DUPLICATION	ESTIMATED_LIBRARY_SIZE
lib1	17719	3317028	0	17719	2308	336674	43229	0.101575	17153558

HG03280.final_1.fastq.txt

LIBRARY	UNPAIRED_READS_EXAMINED	READ_PAIRS_EXAMINED	SECONDARY_ALIGNMENT_RDS	UNMAPPED_READS	UNPAIR_ED_READ_DUPLICATES	READ_PAIR_DUPLICATES	READ_PAIR_OPTICAL_DUPLICATES	PERCENT_DUPLICATION	ESTIMATED_LIBRARY_SIZE
lib1	11510	2214239	0	11510	1276	146125	19423	0.06611	18271158

NA18966.final_1.fastq.txt

LIBRARY	UNPAIRED_READS_EXAMINED	READ_PAIRS_EXAMINED	SECONDARY_ALIGNMENT_RDS	UNMAPPED_READS	UNPAIR_ED_READ_DUPLICATES	READ_PAIR_DUPLICATES	READ_PAIR_OPTICAL_DUPLICATES	PERCENT_DUPLICATION	ESTIMATED_LIBRARY_SIZE
lib1	20829	3343825	0	20829	2670	362095	47072	0.108349	16133083

А также сами данные:

HG00103.final_1.fastq.bam HG03280.final_1.fastq.bam
NA18966.final_1.fastq.bam

HG00103.final_1.fastq.bam.bai HG03280.final_1.fastq.bam.bai
NA18966.final_1.fastq.bam.bai

HG00103.final_1.fastq.bam.sbi HG03280.final_1.fastq.bam.sbi
NA18966.final_1.fastq.bam.sbi

На следующем этапе проведен подсчет качества с помощью gatk BaseRecalibrator. Получены 3 файла с отчетами и посчитанным качеством:

HG00103.final_1.fastq.table HG03280.final_1.fastq.table NA18966.final_1.fastq.table

Затем с помощью gatk ApplyBQSR проведена рекалибровка файлов .bam и получены финальные файлы .bam с индексами:

HG00103.final_1.fastq.bai HG03280.final_1.fastq.bai NA18966.final_1.fastq.bai

HG00103.final_1.fastq.bam HG03280.final_1.fastq.bam NA18966.final_1.fastq.bam

Они же в IGV:



Далее был проведен коллинг вариантов с помощью gatk HaplotypeCaller, в результате которого были получены следующие файлы:

HG00103.final_1.fastq.gvcf.gz HG03280.final_1.fastq.gvcf.gz
NA18966.final_1.fastq.gvcf.gz

HG00103.final_1.fastq.gvcf.gz.tbi HG03280.final_1.fastq.gvcf.gz.tbi
NA18966.final_1.fastq.gvcf.gz.tbi



Затем три файла были объединены в один с помощью gatk CombineGVCFs:

`chr_all_combin.vcf.gz`

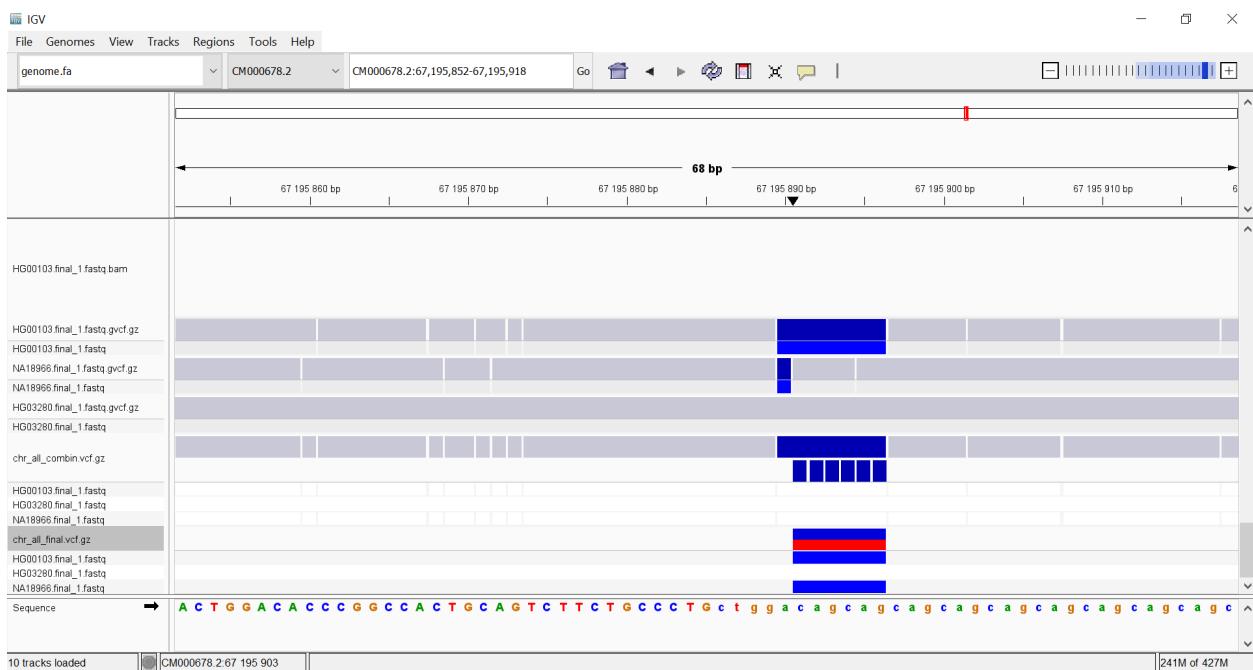
`chr_all_combin.vcf.gz.tbi`



Далее этот файл был преобразован в стандартный .vcf с помощью gatk GenotypeGVCFs, из которого удалены аллели, являющиеся референсными во всех трех образцах:

`chr_all_final.vcf.gz`

`chr_all_final.vcf.gz.tbi`



Для корректной аннотации хромосомы были переименованы с помощью

`bcftools annotate:`

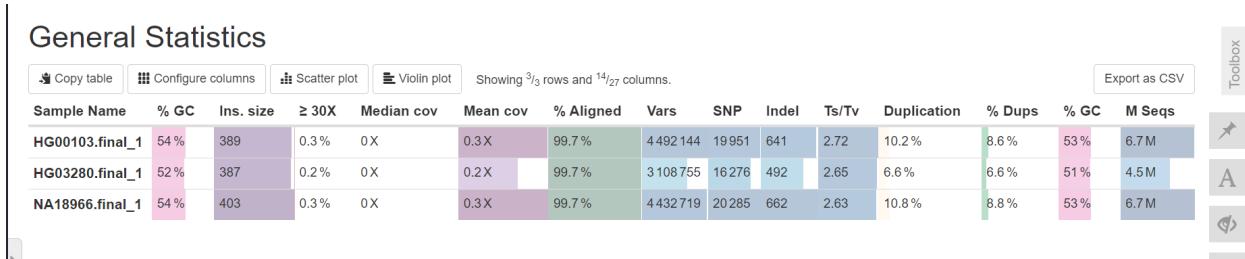
`chr_all_final_renamed.vcf.gz`



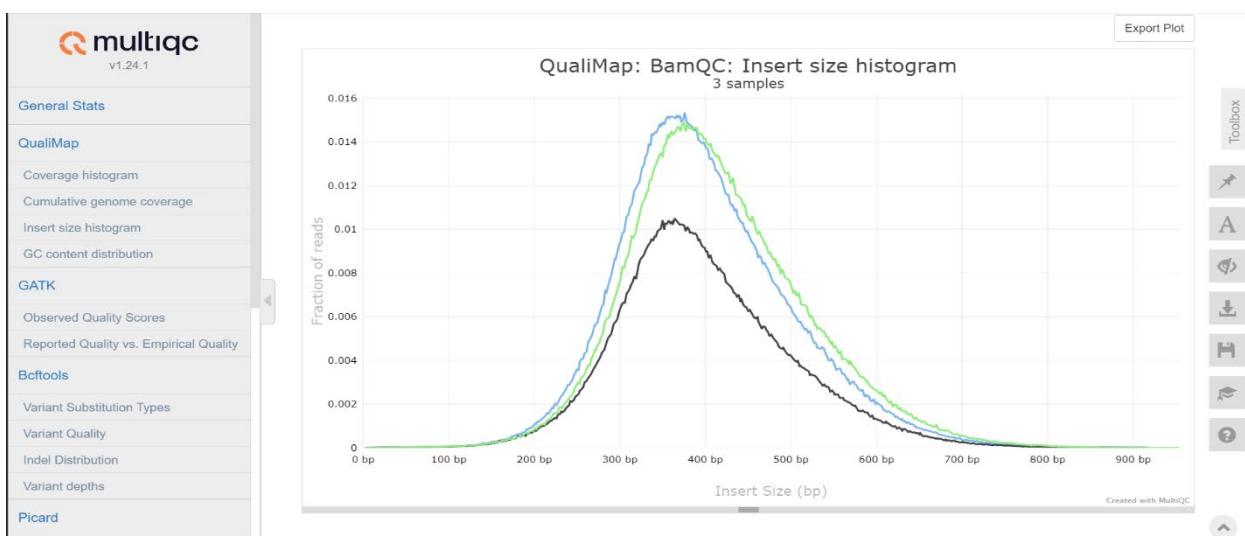
После всей обработки был проведен контроль качества в несколько этапов:

- 1) Получены отчеты о качестве финальных .bam файлов с помощью fastqc:
`HG00103.final_1.fastq_fastqc.html` `HG00103.final_1.fastq_fastqc.zip`
`HG03280.final_1.fastq_fastqc.html` `HG03280.final_1.fastq_fastqc.zip`
`NA18966.final_1.fastq_fastqc.html` `NA18966.final_1.fastq_fastqc.zip`
- 2) Получены отчеты о качестве финальных .bam файлов с помощью qualimap bamqc:
`HG00103.final_1.fastq` `HG03280.final_1.fastq` `NA18966.final_1.fastq`

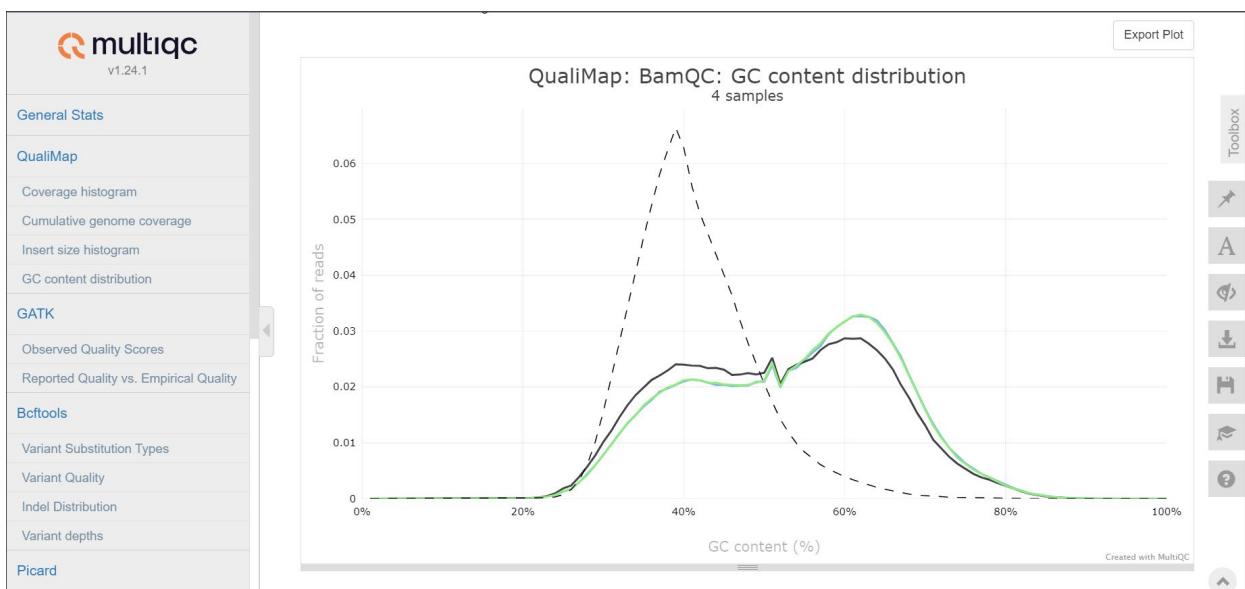
- 3) Получены отчеты о качестве .gvcf файлов с помощью bcftools stats:
- HG00103.final_1.fastq.txt
 - HG03280.final_1.fastq.txt
 - NA18966.final_1.fastq.txt
- 4) Все отчеты объединены в один с помощью multiqc:
- multiqc_report.html



Т.к. не получилось считать файл с референсными регионами, с помощью qualimap статистика расчитана на весь геном, и данные о покрытии неинформативны.



Размер инсерций у всех образцов примерно одинаковый.



GC состав выглядит так, т.к. у нас экзомное секвенирование.

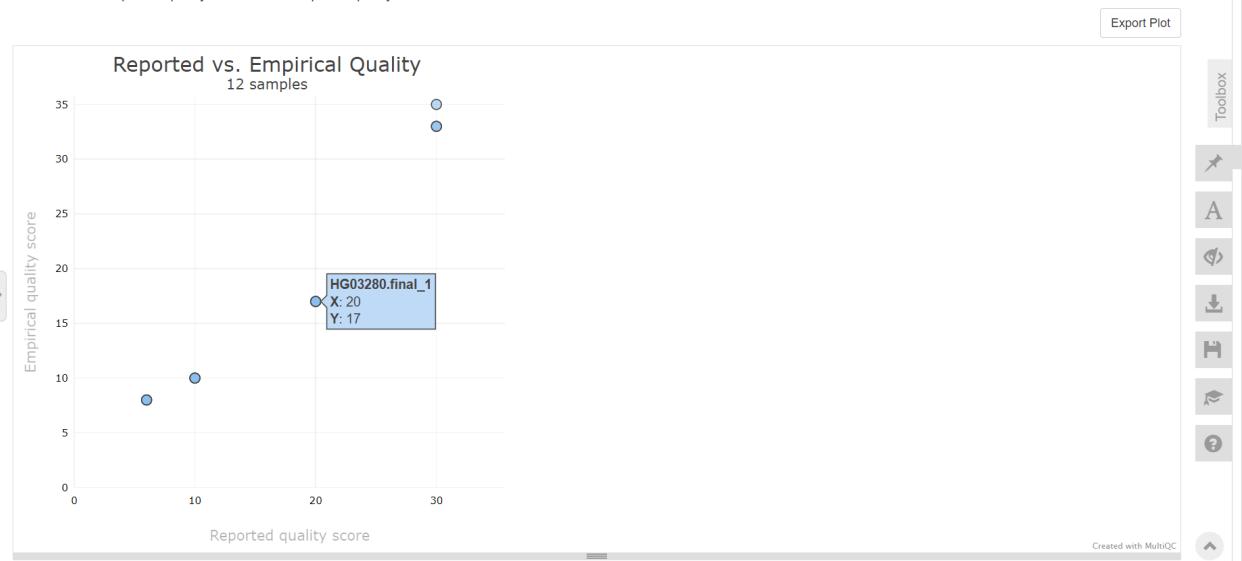
Plot shows the reported quality score vs the empirical quality score.



Plot shows the reported quality score vs the empirical quality score.

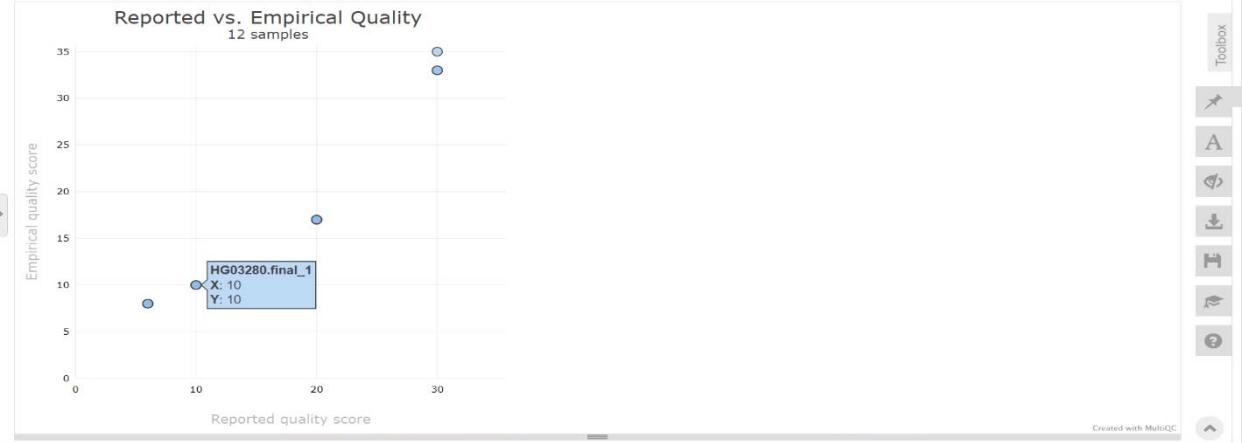


Plot shows the reported quality score vs the empirical quality score.



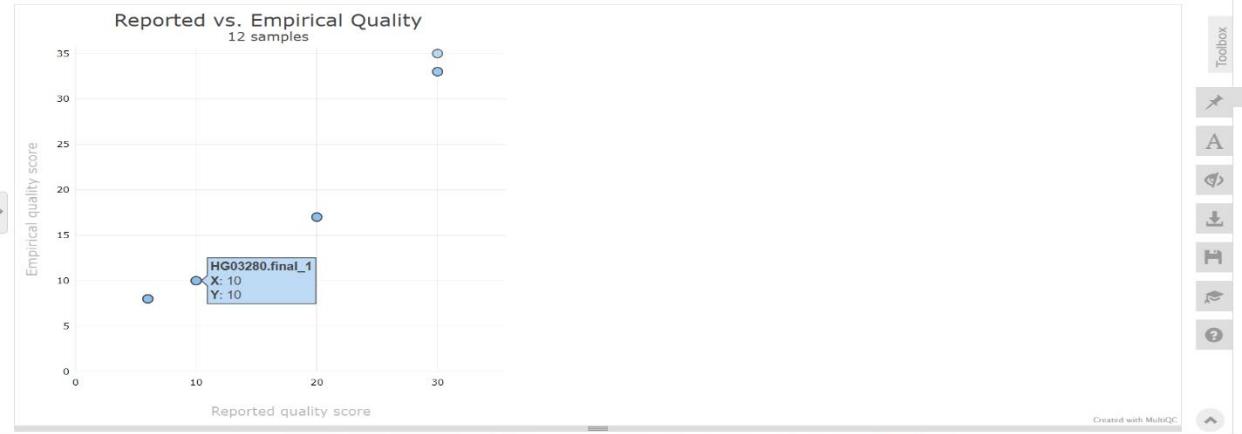
Plot shows the reported quality score vs the empirical quality score.

Export Plot



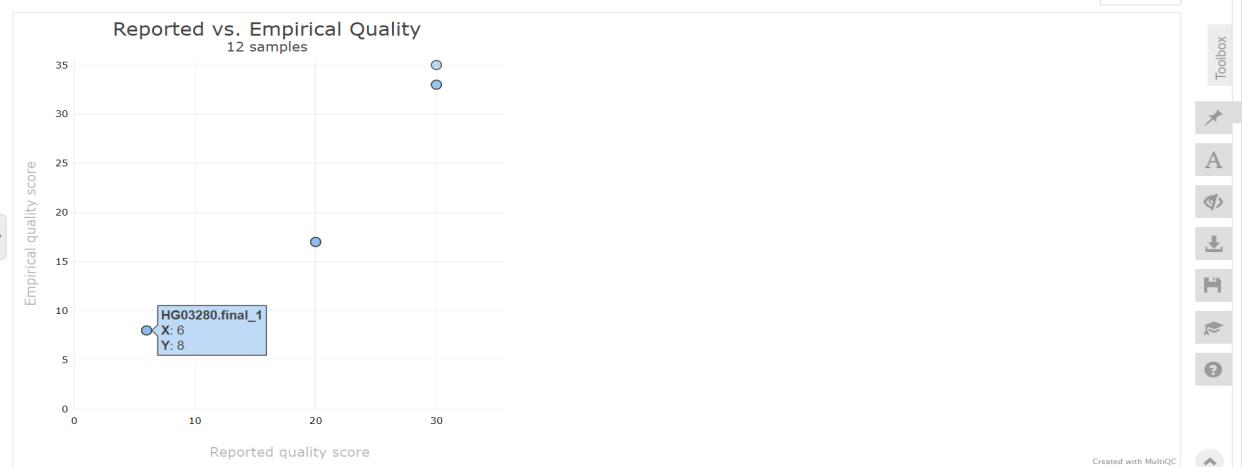
Plot shows the reported quality score vs the empirical quality score.

Export Plot



Plot shows the reported quality score vs the empirical quality score.

Export Plot

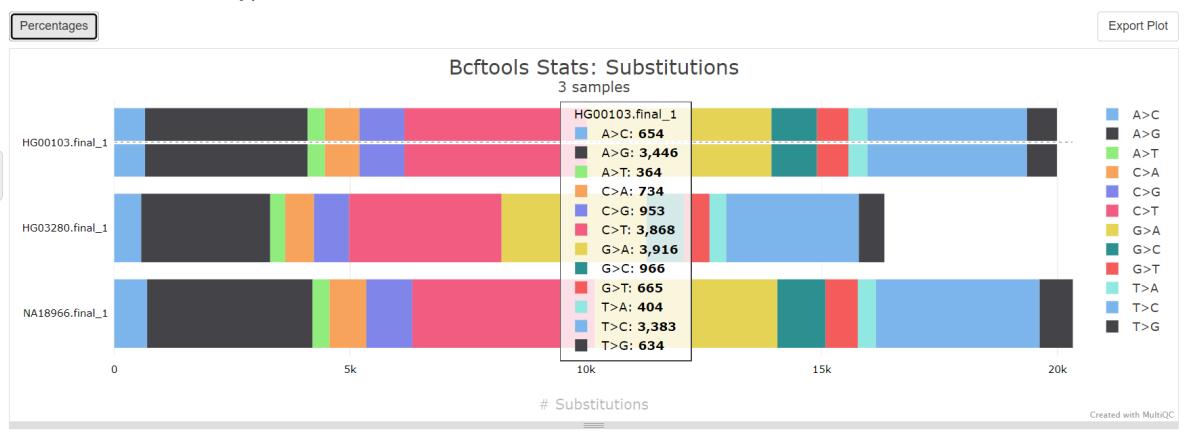


Качество после рекалибровки возросло в тех, случаях, когда прибор показывал качество 6 и 30, осталось тем же при качестве 10 и снизилось при исходном качестве 20.

Bcftools Version: 1.20

Utilities for variant calling and manipulating VCFs and BCFs. URL: <https://samtools.github.io/bcftools> DOI: 10.1093/gigascience/gia008

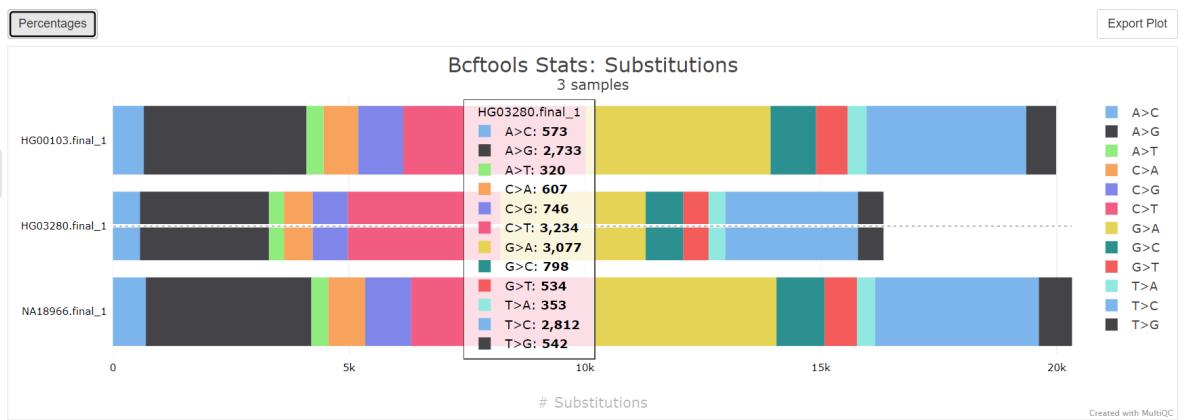
Variant Substitution Types



Bcftools Version: 1.20

Utilities for variant calling and manipulating VCFs and BCFs. URL: <https://samtools.github.io/bcftools> DOI: 10.1093/gigascience/gia008

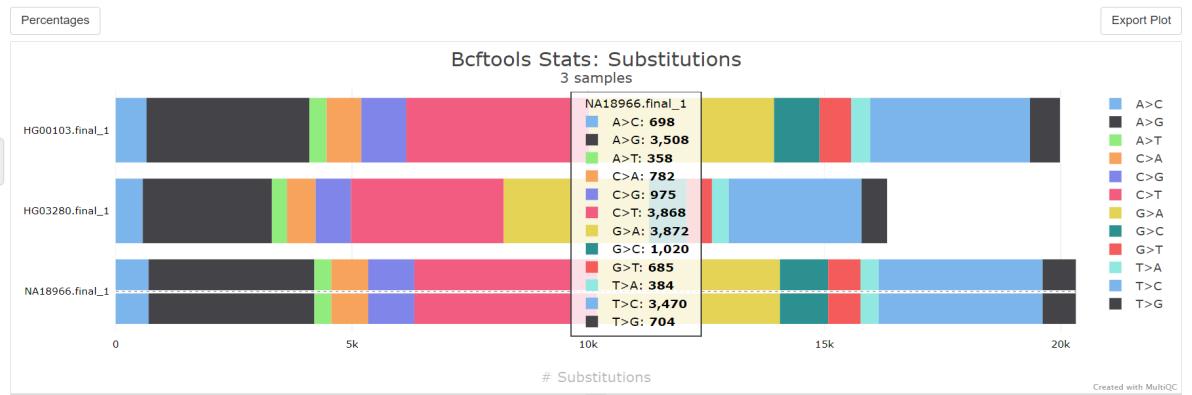
Variant Substitution Types



Bcftools Version: 1.20

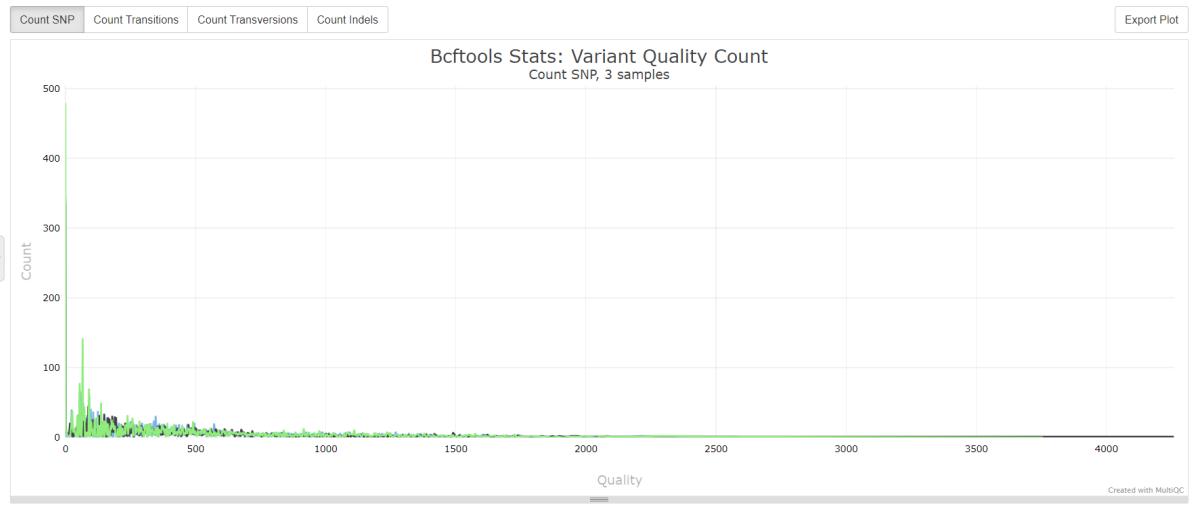
Utilities for variant calling and manipulating VCFs and BCFs. URL: <https://samtools.github.io/bcftools> DOI: 10.1093/gigascience/gia008

Variant Substitution Types

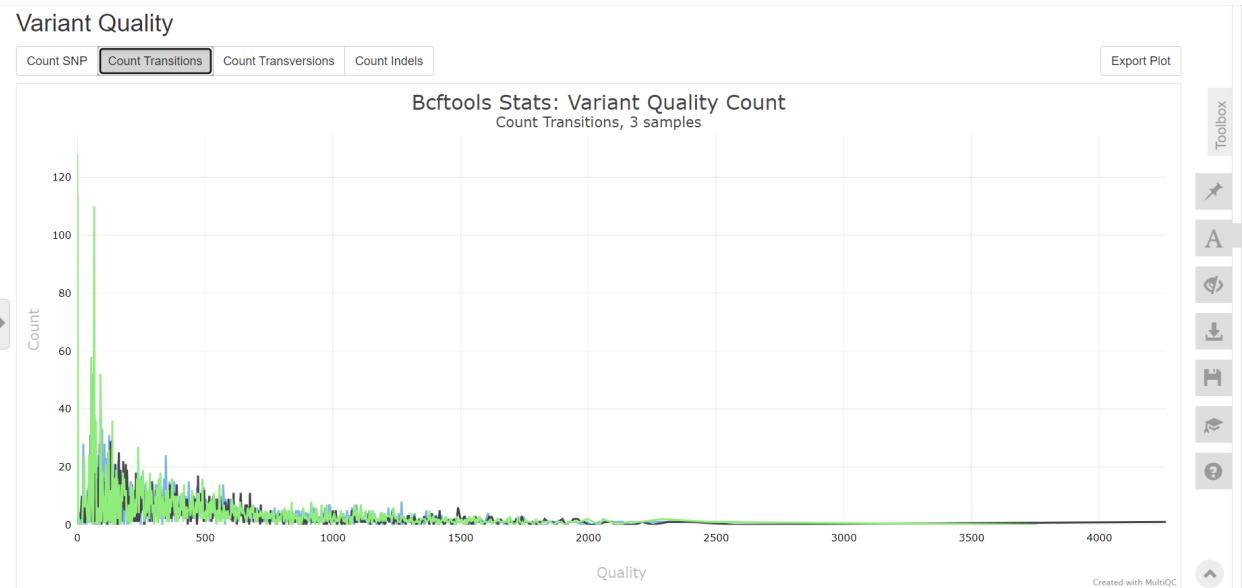


Типы альтернативных вариантов.

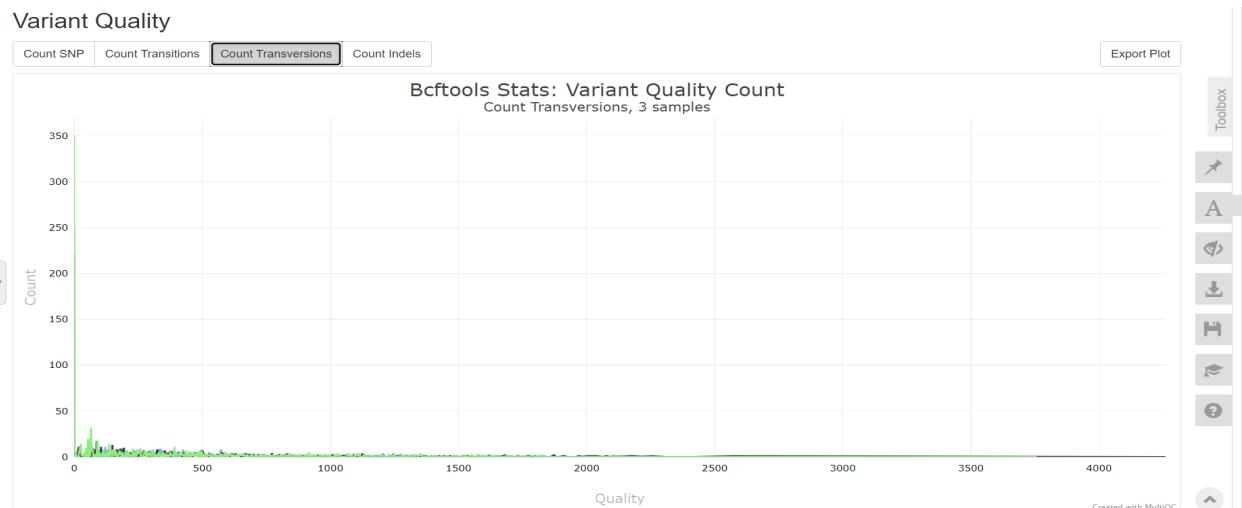
Variant Quality



Количество SNP.

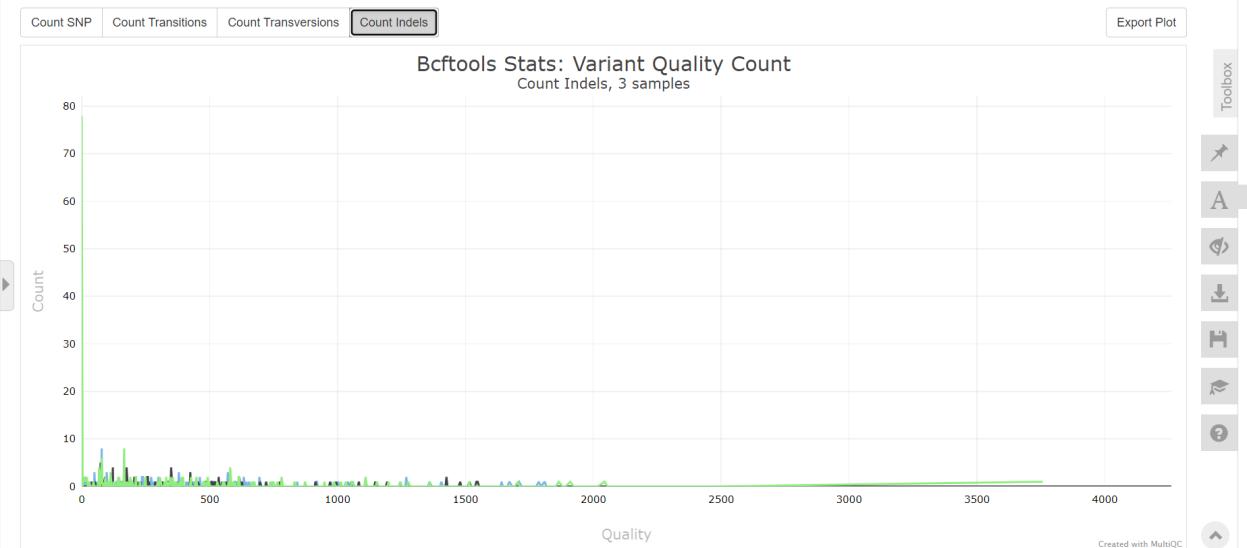


Количество транзиций.



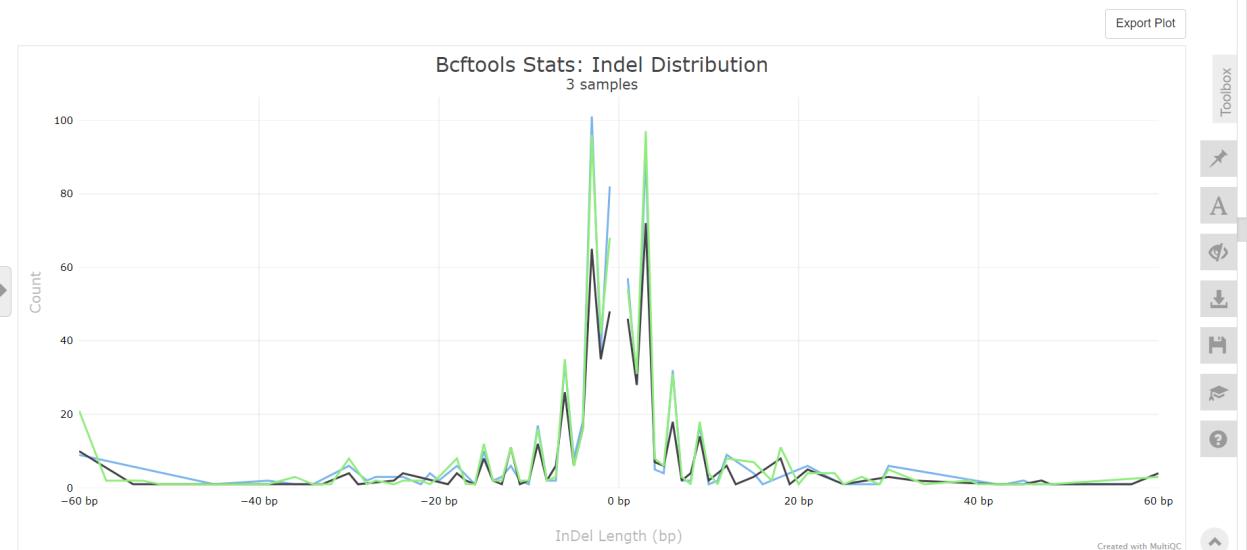
Количество трансверсий.

Variant Quality



Количество инделов.

Indel Distribution



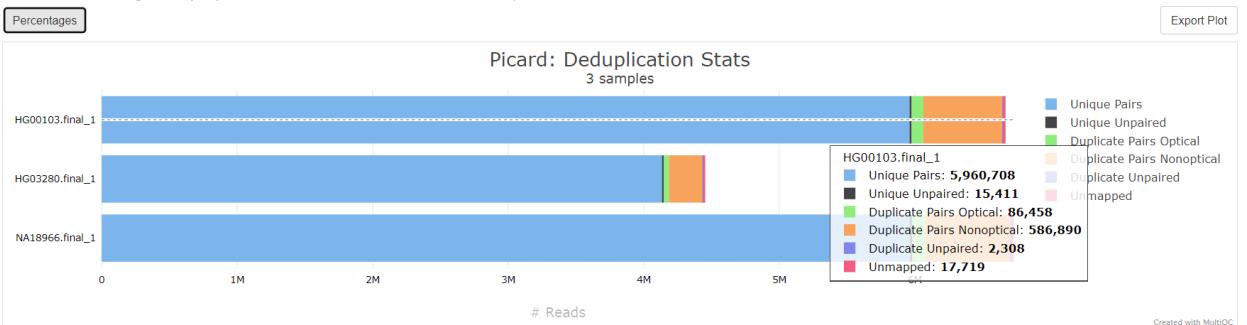
Распределение инделов. Пики характерны для человеческого генома.

Picard

Tools for manipulating high-throughput sequencing data. URL: <http://broadinstitute.github.io/picard>

Mark Duplicates

Number of reads, categorised by duplication state. **Pair counts are doubled** - see help text for details.

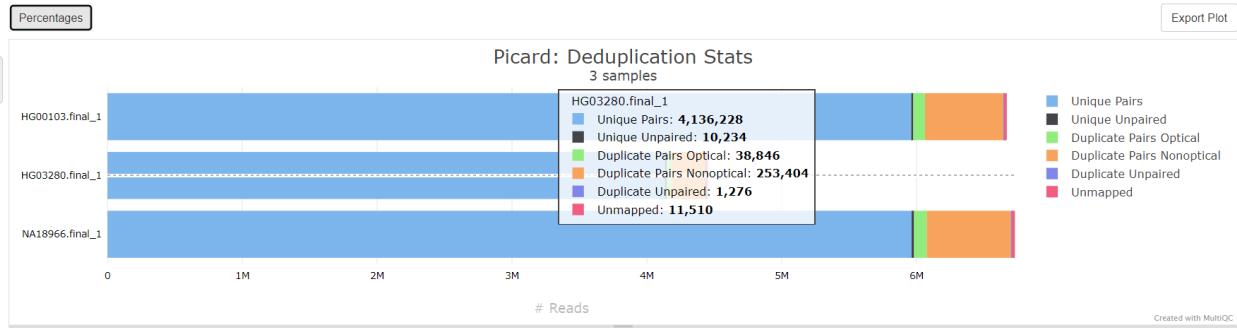


Picard

Tools for manipulating high-throughput sequencing data. URL: <http://broadinstitute.github.io/picard>

Mark Duplicates

Number of reads, categorised by duplication state. **Pair counts are doubled** - see help text for details.

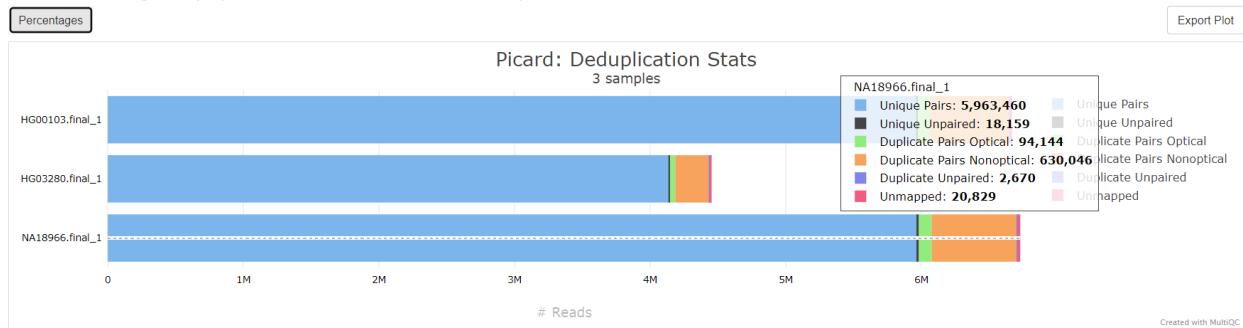


Picard

Tools for manipulating high-throughput sequencing data. URL: <http://broadinstitute.github.io/picard>

Mark Duplicates

Number of reads, categorised by duplication state. **Pair counts are doubled** - see help text for details.



Дубликаты, найденные с помощью Mark Duplicates.

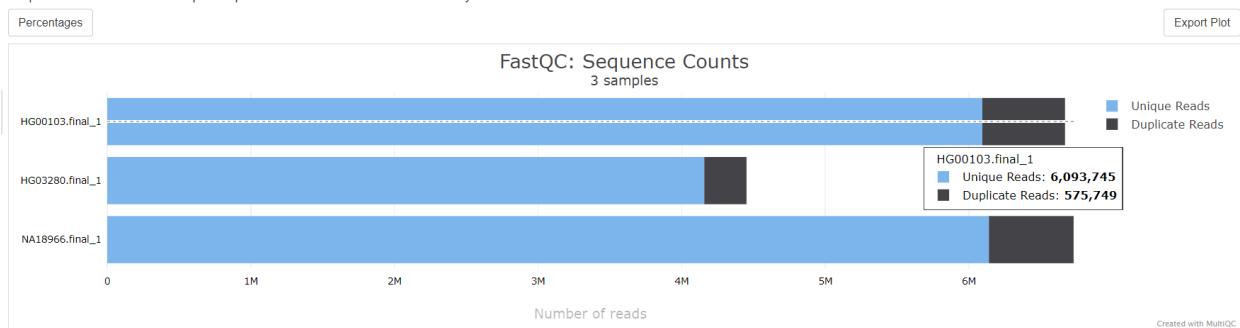
FastQC

Version: 0.12.1

Quality control tool for high throughput sequencing data. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

Sequence Counts

Sequence counts for each sample. Duplicate read counts are an estimate only.

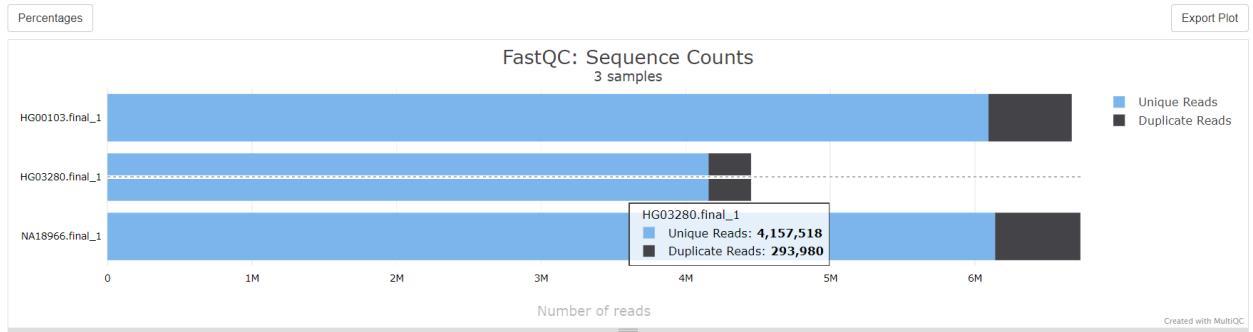


FastQC Version: 0.12.1

Quality control tool for high throughput sequencing data. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

Sequence Counts

Sequence counts for each sample. Duplicate read counts are an estimate only.

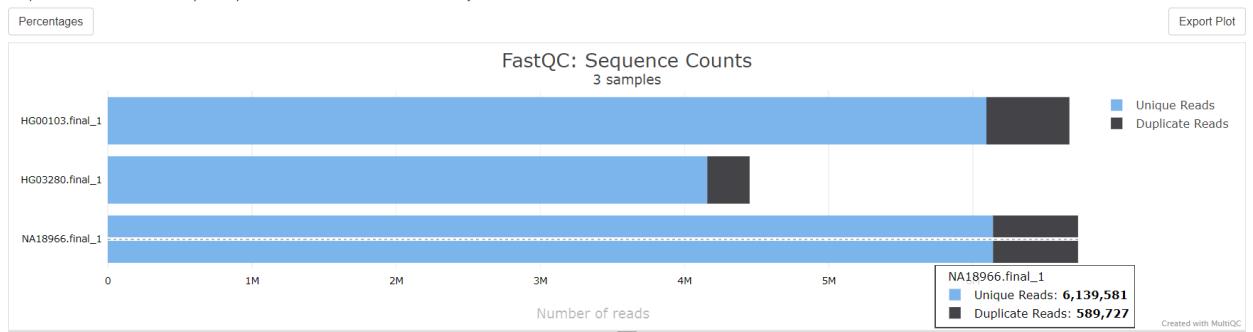


FastQC Version: 0.12.1

Quality control tool for high throughput sequencing data. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

Sequence Counts

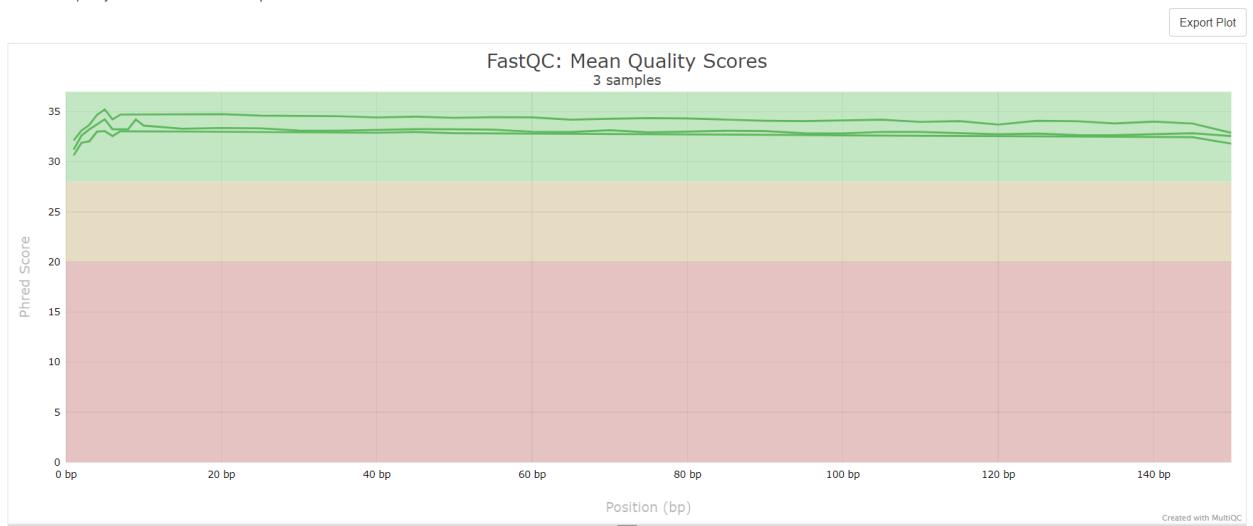
Sequence counts for each sample. Duplicate read counts are an estimate only.



Дубликаты, найденные с помощью FastQC.

Sequence Quality Histograms 3

The mean quality value across each base position in the read.



Среднее качество ридов – очень хорошее.

Per Sequence Quality Scores

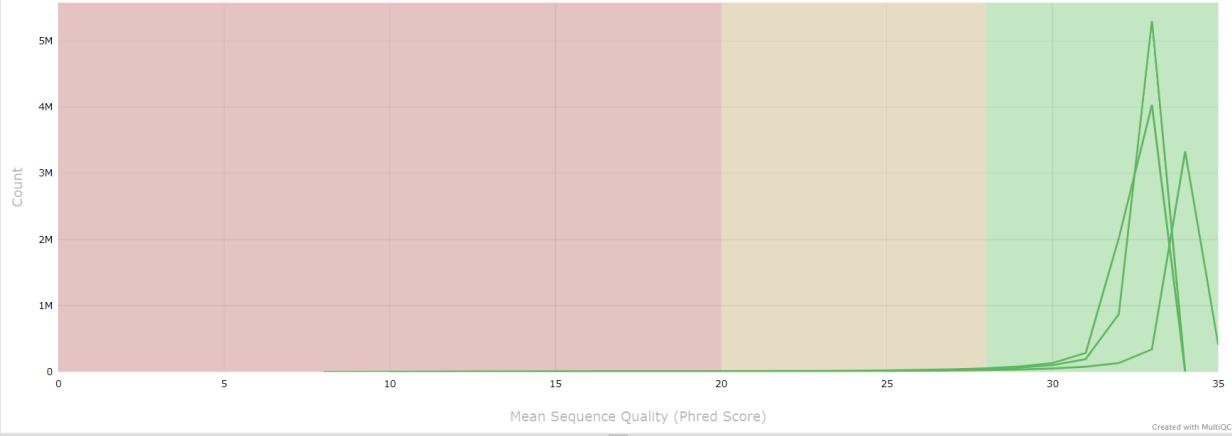
3

Help

The number of reads with average quality scores. Shows if a subset of reads has poor quality.

Export Plot

FastQC: Per Sequence Quality Scores
3 samples



Created with MultiQC

Число ридов со средним качеством.

Per Base Sequence Content

3

Help

The proportion of each base position for which each of the four normal DNA bases has been called.

Click a sample row to see a line plot for that dataset.

HG03280.final_1 Pass

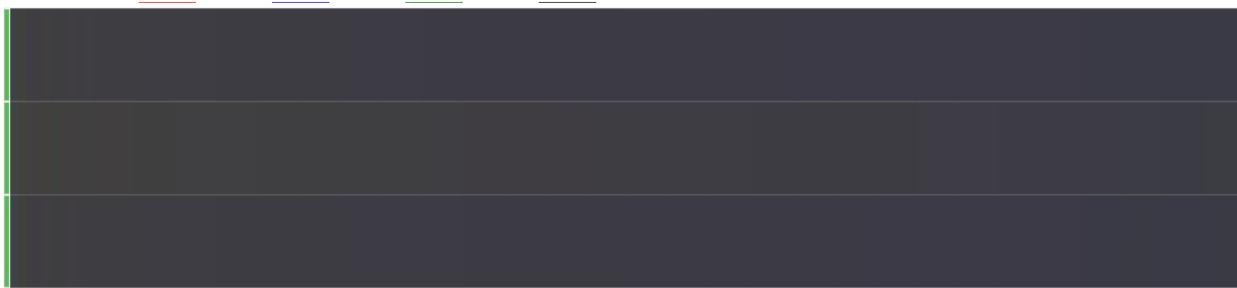
Position: 145-149 bp

%T: 23%

%C: 26%

%A: 24%

%G: 27%



Зависимости нуклеотидного состава от позиции в прочтении нет.

Per Sequence GC Content

3

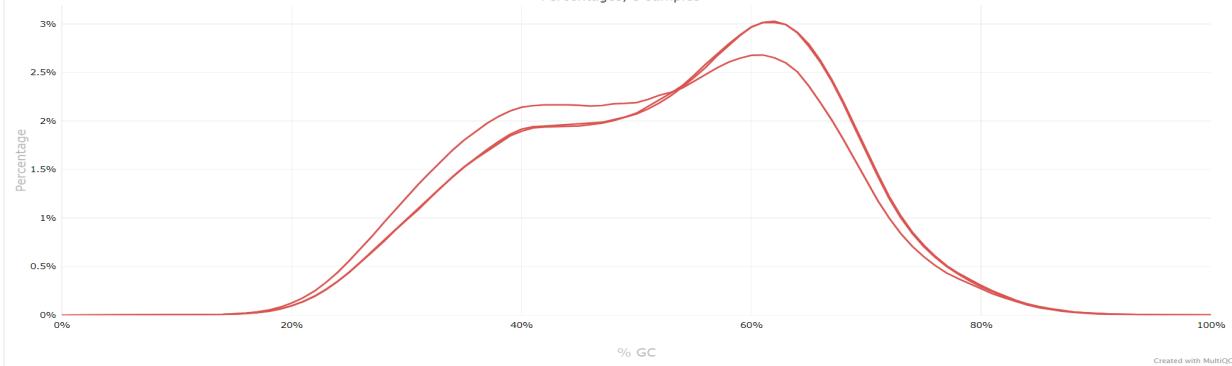
Help

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

Percentages Counts

Export Plot

FastQC: Per Sequence GC Content
Percentages, 3 samples



Created with MultiQC

GC состав, полученный с помощью FastQC. Тоже отличается от нормального распределения из-за того, что у нас экзомное секвенирование.

Per Base N Content

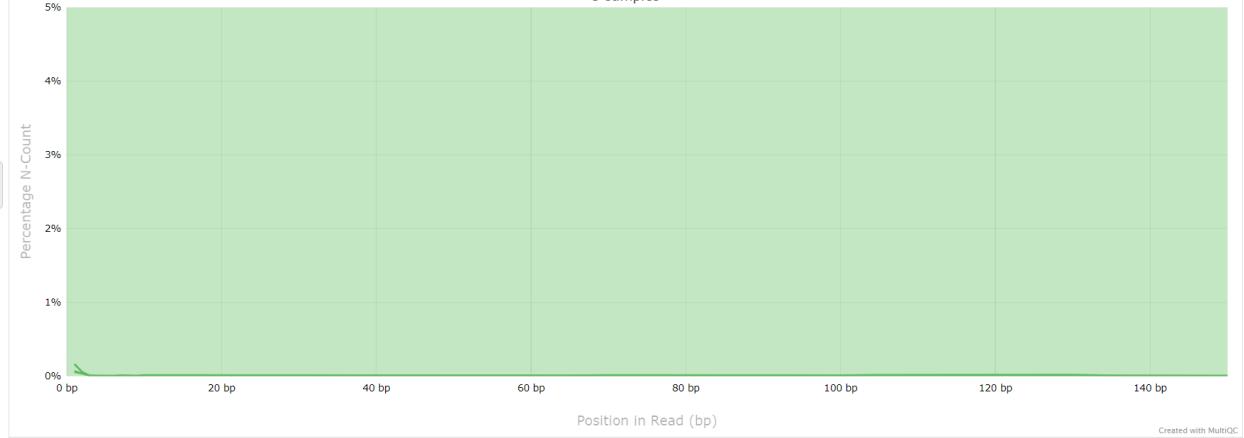
3

The percentage of base calls at each position for which an `N` was called.

Help

Export Plot

FastQC: Per Base N Content
3 samples



Created with MultiQC

Нет неизвестных нуклеотидов.

Sequence Duplication Levels

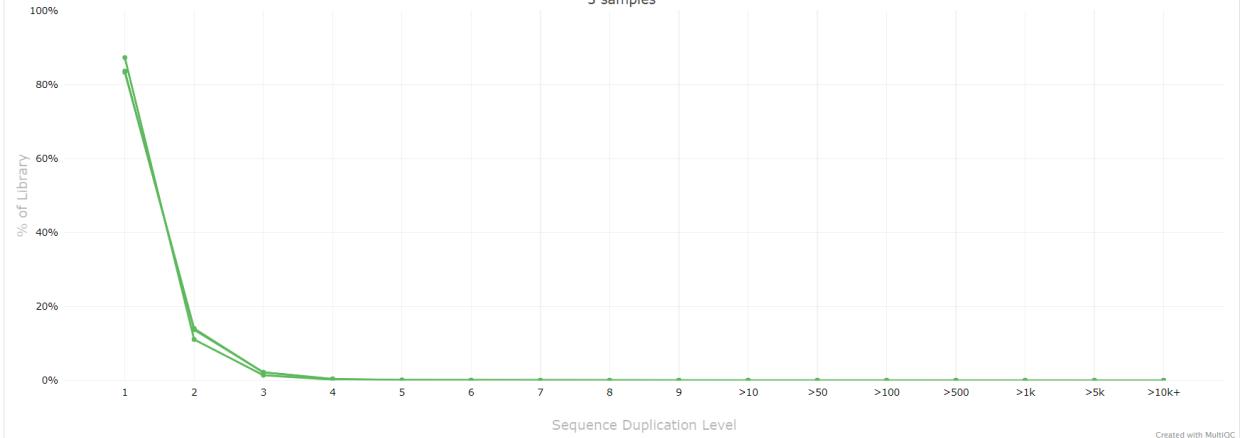
3

The relative level of duplication found for every sequence.

Help

Export Plot

FastQC: Sequence Duplication Levels
3 samples



Created with MultiQC

Уровень дубликации.

Overrepresented sequences by sample

3

Help

The total amount of overrepresented sequences found in each library.

3 samples had less than 1% of reads made up of overrepresented sequences

Export as CSV

Top overrepresented sequences

Top overrepresented sequences across all samples. The table shows 20 most overrepresented sequences across all samples, ranked by the number of samples they occur in.

Copy table Violin plot

Showing 0/0 rows.

Export as CSV

Overrepresented sequence

Нет перепредставленных последовательностей.

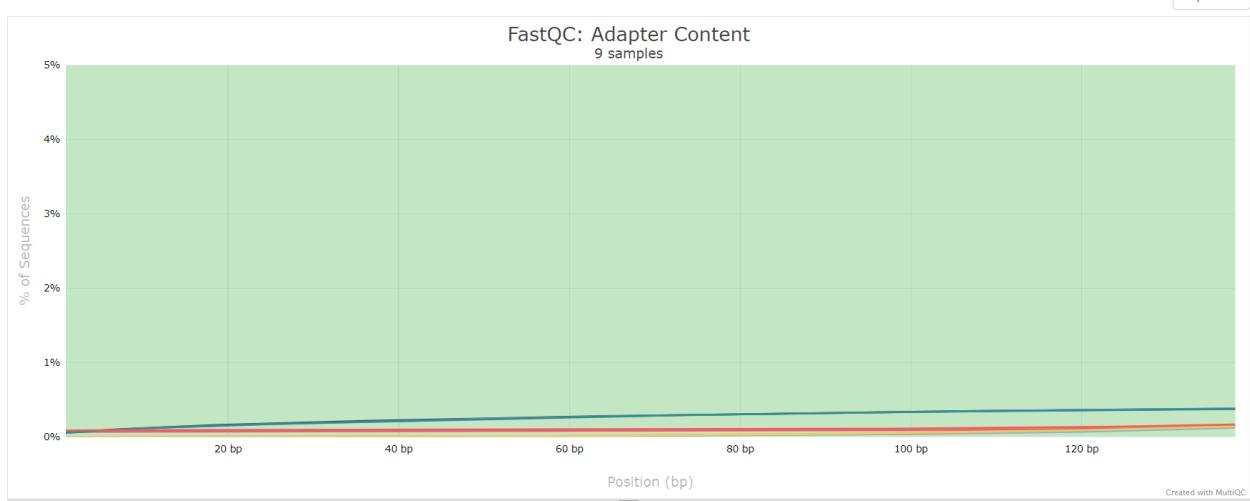
Adapter Content

3

Help

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

Export Plot



Адаптеры – количество низкое.

Status Checks

Help

Status for each FastQC section showing whether results seem entirely normal (green), slightly abnormal (orange) or very unusual (red).

Min: 0,0

Max: 1,0

Export Plot

FastQC: Status Checks
11 samples

HG00103.final_1 -

HG03280.final_1 -

NA18966.final_1 -

Basic Statistics Per Base Sequence Quality Per Tile 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

Created with MultiQC

Сводная статистика. Единственная проблема – GC состав, но у нас экзомное секвенирование, поэтому данные можно использовать для анализа.

Проведена проверка на родство образцов с помощью plink, получены три файла:

chr_all.wgs.genome

chr_all.wgs.log

chr_all.wgs.nosex

Таблица из файла chr_all.wgs.genome:

IID1	FID2	IID2	RT	EZ	Z0	Z1	Z2	PI_HAT	PHE	DST	PPC	RATIO
HG00103.final_1.fasta	HG03280.final_1.fasta	HG03280.final_1.fasta	UN	NA	1.0000	0.0000	0.0000	0.0000	-1	0.608039	0.0000	0.5460
HG00103.final_1.fasta	NA18966.final_1.fasta	NA18966.final_1.fasta	UN	NA	1.0000	0.0000	0.0000	0.0000	-1	0.688757	0.0000	0.6859
HG03280.final_1.fasta	NA18966.final_1.fasta	NA18966.final_1.fasta	UN	NA	1.0000	0.0000	0.0000	0.0000	-1	0.597150	0.0000	0.4367

Судя по значениям PI_HAT (у всех сочетаний = 0), образцы ДНК получены от людей, не являющихся друг другу родственниками.

2. Аннотация вариантов.

Для аннотации вариантов был использован Annovar и базы данных: refGene, avsnp150, clinvar_20220320, gnomad30_genome, gnomad211_exome, dbnsfp35c, dbscsnpv11. Получены следующие файлы:

chr all final renamed.avinput

Файл	Правка	Формат	Вид	Справка
chr_all_final_renamed.avinput	Блокнот			
chr1	69270	A	G	0.75 190.87 2
chr1	69511	A	G	0.8333 1069.92 15
chr1	69897	T	C	0.8333 502.67 5
chr1	942451	T	C	1 3810.73 32
chr1	944101	G	C	0.1667 307.29 17
chr1	946247	G	A	0.5 777.95 14
chr1	952421	A	G	1 2663.73 22
chr1	953259	T	C	1 523.99 5
chr1	953279	T	C	1 693.26 5
chr1	957171	C	T	0.1667 282.29 12
chr1	961945	G	C	0.8333 3359.93 40
chr1	962358	G	C	0.1667 708.29 32
chr1	965017	C	T	0.1667 459.29 21
chr1	965125	G	C	0.1667 191.3 11
chr1	970892	A	C	0.1667 732.29 35
chr1	973858	G	C	1 3816.73 32
chr1	973862	G	A	0.1667 507.29 32
chr1	973929	T	C	0.1667 602.29 34
chr1	973946	T	C	0.1667 306.29 31
chr1	976215	A	G	0.8333 1451.98 18
chr1	976536	C	T	0.3333 986.14 21
chr1	978953	G	C	0.6667 1787.96 26
chr1	979472	G	C	0.6667 4161.96 55
chr1	979496	T	C	1 5778.73 54
chr1	979560	T	C	0.6667 3898.96 50
chr1	979847	A	G	1 4813.73 40
chr1	980388	A	G	0.1667 800.29 37
chr1	981169	G	A	1 2908.73 21
chr1	999842	B	C	0.6667 1770.96 23
chr1	1014228	G	A	0.3333 1309.14 29
chr1	1014274	A	G	1 3508.73 24
chr1	1020217	G	T	0.3333 524.23 7
chr1	1041505	A	G	0.1667 274.29 23
chr1	1046551	A	G	0.8333 3364.93 37
chr1	1047741	T	C	1 5000.73 15

chr all final renamed.hg38 multianno.txt

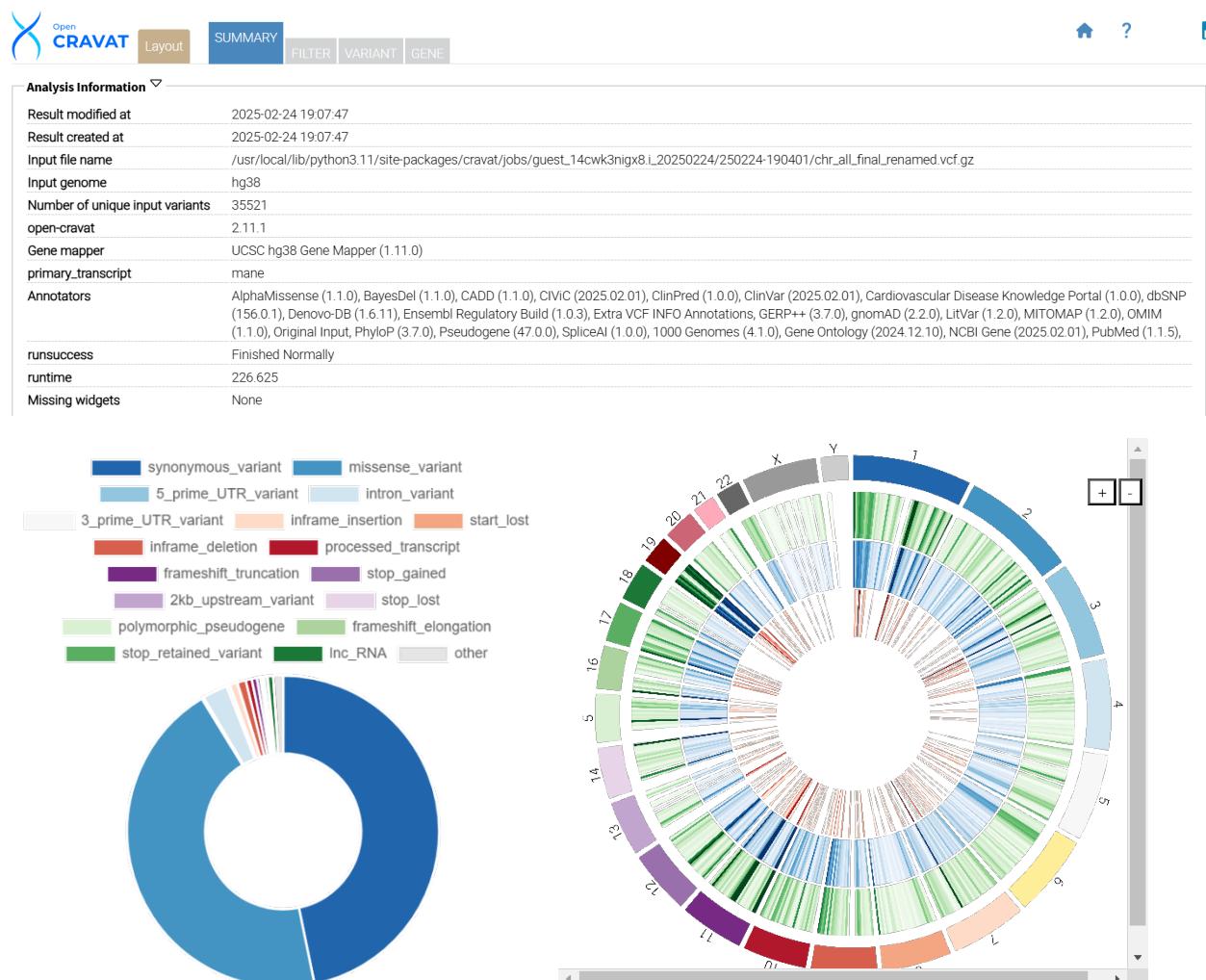
chr all final renamed hg38 multianno vcf

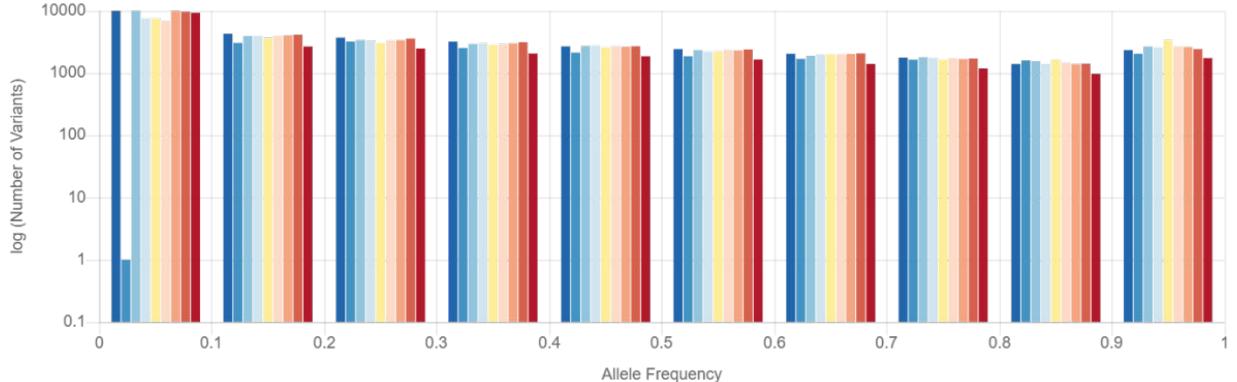
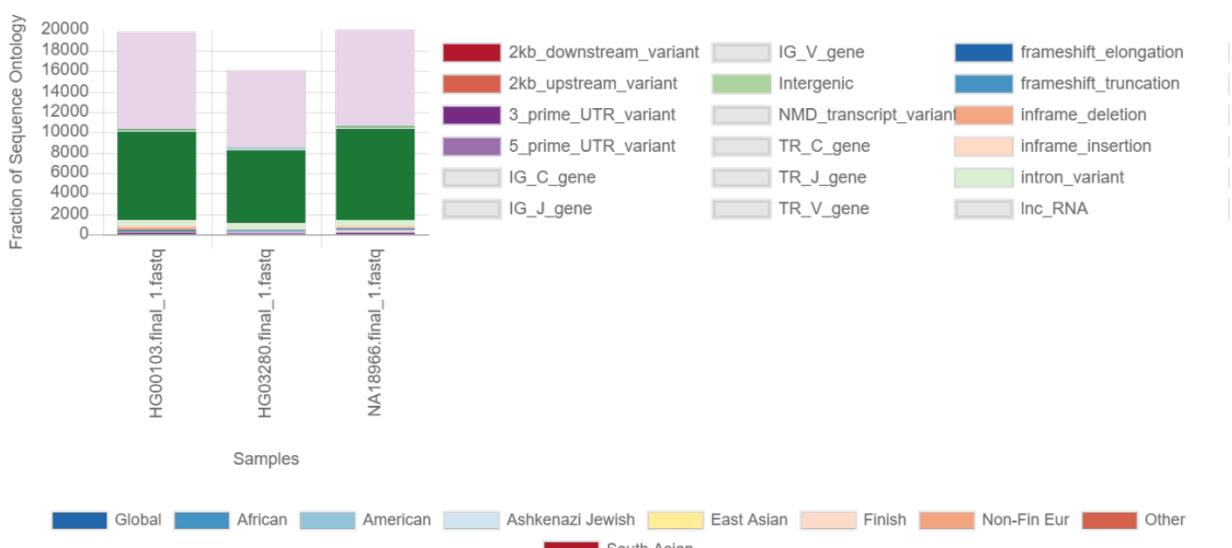
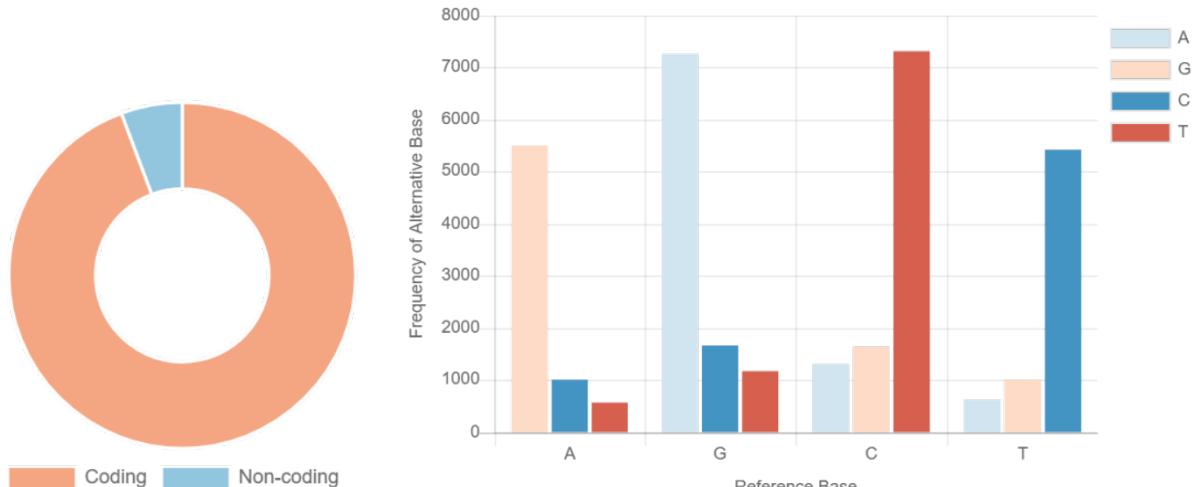
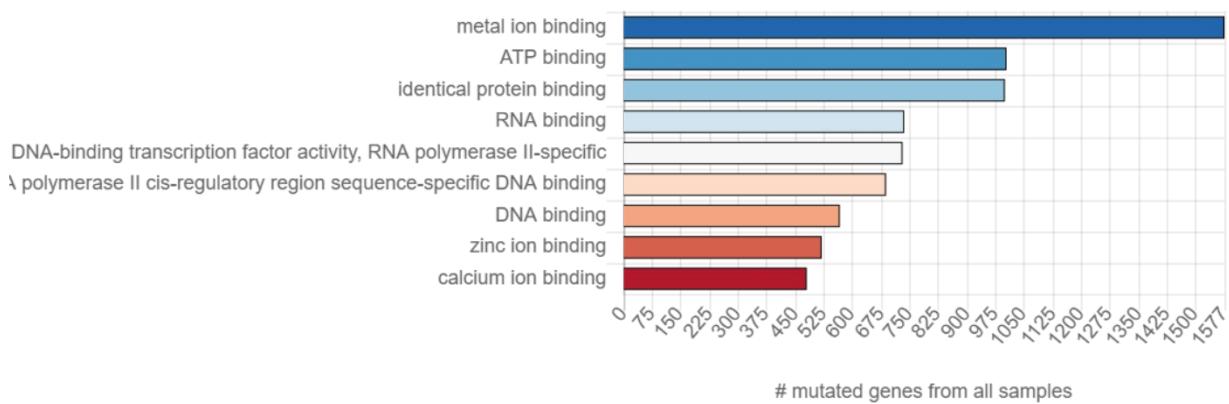
```

chr_all_final_renamed.hg38_multianno.vcf - блокнот
Файл Правка Формат Вид Справка
##INFO<=ID=GERP++,RS_rankscore,Number=.,Type=String,Description="GERP++_RS_rankscore annotation provided by ANNOVAR">
##INFO<=ID=phyloP100way_vertebrate,Number=.,Type=String,Description="phyloP100way_vertebrate annotation provided by ANNOVAR">
##INFO<=ID=phyloP100way_vertebrate_rankscore,Number=.,Type=String,Description="phyloP100way_vertebrate_rankscore annotation provided by ANNOVAR">
##INFO<=ID=phyloP20way_mammalian,Number=.,Type=String,Description="phyloP20way_mammalian annotation provided by ANNOVAR">
##INFO<=ID=phastCons100way_vertebrate,Number=.,Type=String,Description="phastCons100way_vertebrate annotation provided by ANNOVAR">
##INFO<=ID=phastCons100way_vertebrate_rankscore,Number=.,Type=String,Description="phastCons100way_vertebrate_rankscore annotation provided by ANNOVAR">
##INFO<=ID=phastCons20way_mammalian,Number=.,Type=String,Description="phastCons20way_mammalian annotation provided by ANNOVAR">
##INFO<=ID=phastCons20way_mammalian_rankscore,Number=.,Type=String,Description="phastCons20way_mammalian_rankscore annotation provided by ANNOVAR">
##INFO<=ID=SiPhy_29way_logOdds,Number=.,Type=String,Description="SiPhy_29way_logOdds annotation provided by ANNOVAR">
##INFO<=ID=SiPhy_29way_logOdds_rankscore,Number=.,Type=String,Description="SiPhy_29way_logOdds_rankscore annotation provided by ANNOVAR">
##INFO<=ID=Interpro_domain,Number=.,Type=String,Description="Interpro_domain annotation provided by ANNOVAR">
##INFO<=ID=GTEX_V6p_gene,Number=.,Type=String,Description="GTEX_V6p_gene annotation provided by ANNOVAR">
##INFO<=ID=GTEX_V6p_tissue,Number=.,Type=String,Description="GTEX_V6p_tissue annotation provided by ANNOVAR">
##INFO<=ID=dbcsNVN_ADA_SCORE,Number=.,Type=String,Description="dbcsNVN_ADA_SCORE annotation provided by ANNOVAR">
##INFO<=ID=dbcsNVN_RF_SCORE,Number=.,Type=String,Description="dbcsNVN_RF_SCORE annotation provided by ANNOVAR">
##INFO<=ID=ALLELE_END,Number=0,Type=Flag,Description="Flag the end of ANNOVAR annotation for one alternative allele">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG00103.final_1.fastaq NA18966.final_1.fastaq
chr1 69720 . A G 199.87 . AC=3;AF=0.75;AN=4;BaseRankSum=-3.93;DP=24;ExcessHet=0;FS=0;MLEAC=3;MLEAF=0.75;MQ=35.89;MQRankSum=-0.
onTaster_converted_rankscore=.;MutationTaster_pred=.;MutationAssessor_score=.;MutationAssessor_score_rankscore=.;MutationAssessor_pred=.;FATHMM_score=.;FATHMM_Ph_29way_logOdds_rankscore=.;Interpro_domain=.;GTEX_V6p_gene=.;GTEX_V6p_tissue=.;dbcsNVN_ADA_SCORE=.;dbcsNVN_RF_SCORE=.;ALLELE_END GT:AD:DP:QV:PL ./..1
chr1 69511 . A G 1099.92 . AC=5;AF=0.833;AN=6;BaseRankSum=-0.073;DP=70;ExcessHet=0;FS=8.638;MLEAC=5;MLEAF=0.833;MQ=47.43;MQRank ionTaster_score=1.000;MutationTaster_converted_rankscore=0.189;MutationTaster_pred=P;MutationAssessor_score=0.855;MutationAssessor_score_rankscore=0.015;Mut e=0.000;phastCons100way_vertebrate_rankscore=0.063;phastCons20way_mammalian=0.765;phastCons20way_mammalian_rankscore=0.326;SiPhy_29way_logOdds=4.198;SiPhy_29 chr1 69897 . T C 502.67 . AC=5;AF=0.833;AN=6;BaseRankSum=-3.87;DP=34;ExcessHet=0;FS=0;MLEAC=5;MLEAF=0.833;MQ=32.86;MQRankSum=-0.;MutationTaster_converted_rankscore=.;MutationTaster_pred=.;MutationAssessor_score=.;MutationAssessor_score_rankscore=.;MutationAssessor_pred=.;FATHMM_score dds=.;SiPhy_29way_logOdds_rankscore=.;Interpro_domain=.;GTEX_V6p_gene=.;dbcsNVN_ADA_SCORE=.;dbcsNVN_RF_SCORE=.;ALLELE_END GT:AD:DP:QV:PL ./..1
chr1 944251 . T C 3810.73 . AC=6;AF=1;AN=6;DP=108;ExcessHet=0;FS=0;MLEAC=6;MLEAF=1;MQ=60;QD=25.36;SOR=0.916;ANNOVAR_DATE=2017-07-17;LRT_pred=N;MutationTaster_score=1.000;MutationTaster_converted_rankscore=0.205;MutationTaster_pred=P;MutationAssessor_score=2-1;MutationAssessor_score_rankscore=r tetrabate=0.656;phastCons100way_vertebrate_rankscore=0.280;phastCons20way_mammalian=0.866;phastCons20way_mammalian_rankscore=0.362;SiPhy_29way_logOdds=7.519;S chr1 944101 . G C 307.29 . AC=1;AF=0.167;AN=6;BaseRankSum=-3.233;DP=52;ExcessHet=0;FS=0;MLEAC=1;MLEAF=0.167;MQ=60;MQRankSum=0;Q FT_score=0.002;SIFT_converted_rankscore=0.721;SIFT_pred=D;LRT_score=0.000;LRT_converted_rankscore=0.843;LRT_pred=D;MutationAssessor_score=1.000;MutationTaster score=0.474;phyloP20way_mammalian=0.138;phyloP20way_mammalian_rankscore=0.222;phastCons100way_vertebrate=1.000;phastCons100way_vertebrate_rankscore=0.7 chr1 946247 . G A 777.95 . AC=3;AF=0.5;AN=6;BaseRankSum=-2.594;DP=58;ExcessHet=0;FS=0;MLEAC=3;MLEAF=0.5;MQ=60;MQRankSum=0;QD=24

```

Для более удобного просмотра вариантов была проведена также аннотация в Open Cravat:





3. Описание самого патогенного варианта для каждого образца.

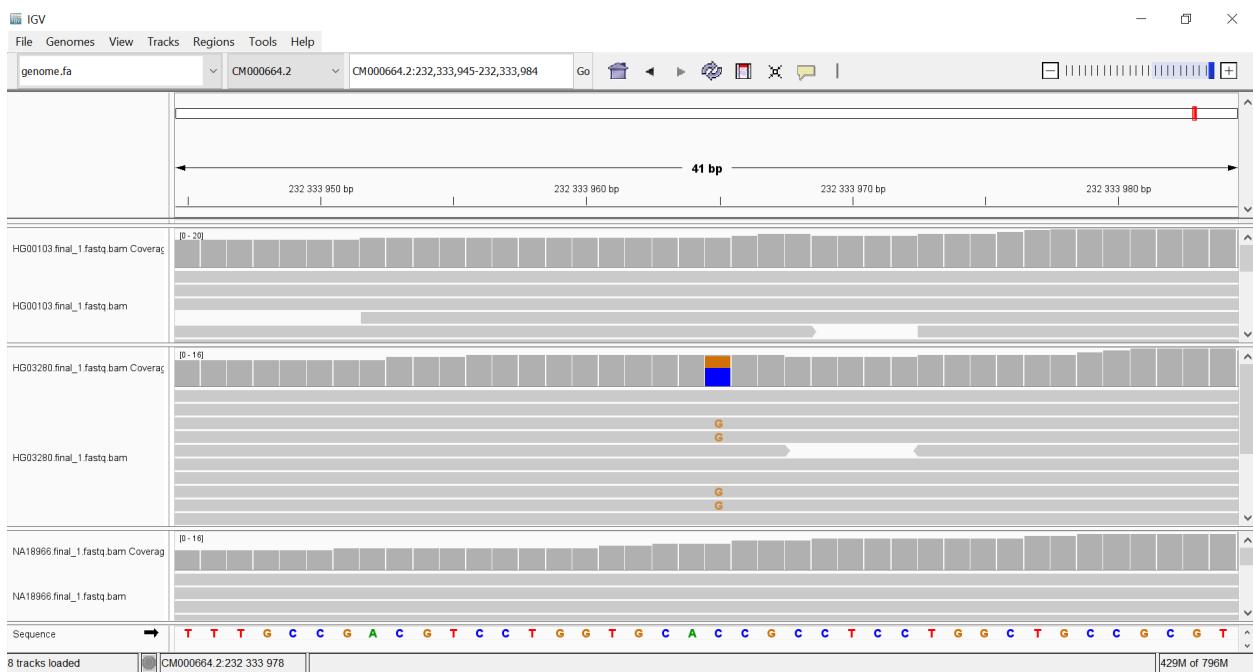
HG00103

Ген	Положение (GRCh38.p14)	Генотип	Экзон	Положение в кДНК	Замена АК	Транскрипт	Частота аллеля	Глубина прочтения
HFE	chr6: 26090951C>G	C/G	2	c.187C>G	p.(H63D)	NM_000410.4	0.1082	19
Всего прочтений		6651775		Всего выявлено вариантов		4492144		
Средняя длина прочтений		150		Вариантов после фильтрации по базовым критериям патогенности и оценки по клиническим критериям		1		
Прочитано нуклеотидов		997766250		Доля региона с покрытием меньше 50x		99.97		
Среднее покрытие		0.3014		Доля региона с покрытием меньше 10x		98.97		



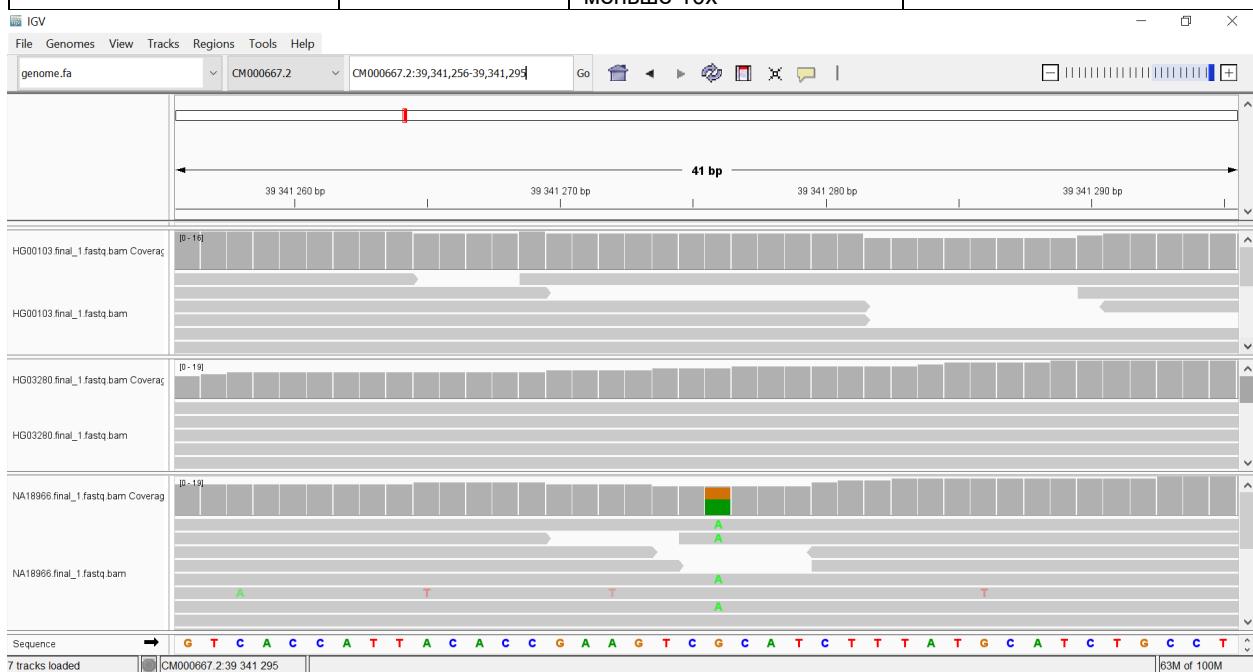
HG03280

Ген	Положение (GRCh38.p14)	Генотип	Экзон	Положение в кДНК	Замена АК	Транскрипт	Частота аллеля	Глубина прочтения
DIS3L2	chr2: 232333965C>G	C/G	17	c.2136C>G	p.(H712Q)	NM_152383.5	4.047e-6	12
Всего прочтений		4439988		Всего выявлено вариантов		3108755		
Средняя длина прочтений		150		Вариантов после фильтрации по базовым критериям патогенности и оценки по клиническим критериям		1		
Прочитано нуклеотидов		665998200		Доля региона с покрытием меньше 50x		99.97		
Среднее покрытие		0.2011		Доля региона с покрытием меньше 10x		99.33		



NA18966

Ген	Положение (GRCh38.p14)	Генотип	Экзон	Положение в кДНК	Замена АК	Транскрипт	Частота аллеля	Глубина прочтения
C9	chr5: 39341276G>A	G/A	4	c.346C>T	p.(R116X)	NM_001737.5	0.003	13
Всего прочтений		6708479		Всего выявлено вариантов		4432719		
Средняя длина прочтений		150		Вариантов после фильтрации по базовым критериям патогенности и оценки по клиническим критериям		1		
Прочитано нуклеотидов		1 006 271 850		Доля региона с покрытием меньше 50x		99.96		
Среднее покрытие		0.3039		Доля региона с покрытием меньше 10x		98.98		



4. Установление Y и MT гаплогрупп.

Для анализа Y хромосомы был использован LineageTracker classify, получены три файла:

classify_Y.ClassificationLog.log classify_Y.hapresult.hg classify_Y.lineageresult.txt

В .log файле предупреждение:

There are no mutation matched for sample HG00103.final_1.fastq

There are no mutation matched for sample HG03280.final_1.fastq

В файле .txt:

SampleID	Haplogroup	KeyHaplogroup	Mutations	LineageTrack
HG00103.final_1.fastq	.	.	.	
HG03280.final_1.fastq	.	.	.	
NA18966.final_1.fastq	D1	.	M174/Page30	D1(1/101)->D->DE(1/28)->CT->BT->A1b->A1->A0T->A00T->A000T->Y-Adam

Из-за того, что секвенирование экзомное, не нашлось совпадающих мутаций, чтобы определить гаплогруппу для образцов HG00103 и HG03280. Образец NA18966 определен как гаплогруппа D1 по одной замене из 101 - M174/Page30. Судя по сайту <https://www.yfull.com/tree/>, эта гаплогруппа распространена в Китае и Японии:





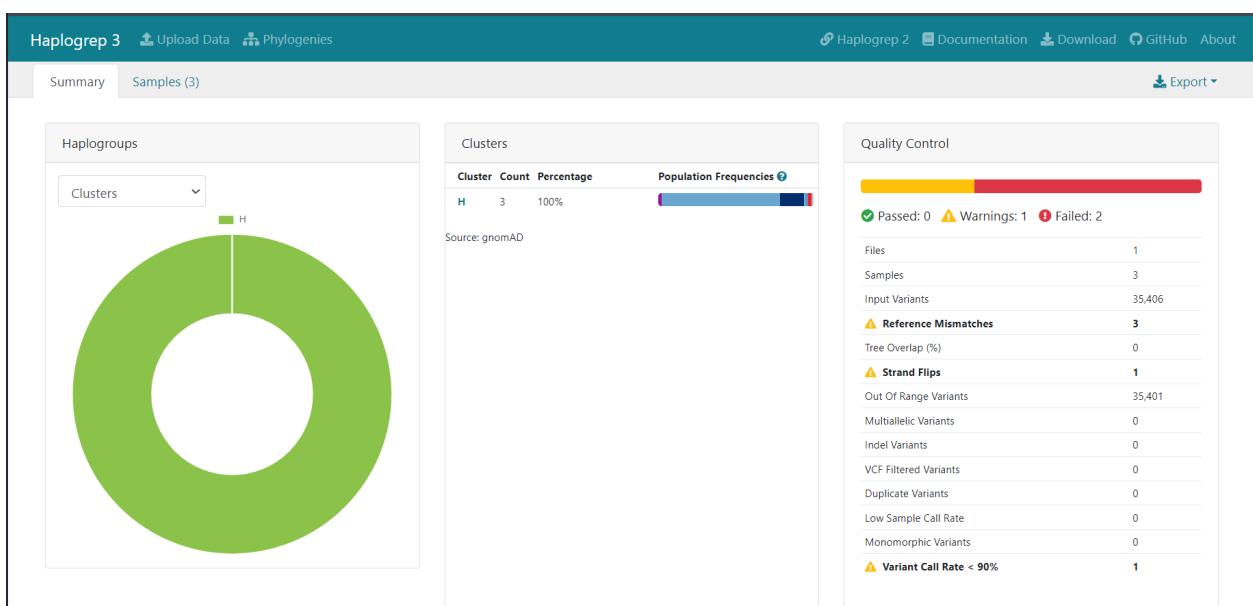
Y хромосома в файле присутствует, но в значительно меньшем количестве, чем нужно для определения гаплогруппы.

МТ анализ проведен с помощью haplogrep classify, получен файл haplogroups.txt:

"SampleID"	"Haplogroup"	"Rank"	"Quality"	"Range"
"HG00103.final_1.fastq"	"H2a2a1"	"1"	"0.5000"	"1-16569"
"HG03280.final_1.fastq"	"H2a2a1"	"1"	"0.5000"	"1-16569"
"NA18966.final_1.fastq"	"H2a2a1"	"1"	"0.5000"	"1-16569"

Гаплогруппа не определена ни у одного из образцов с достаточной точностью, судя по столбцу "Quality": уверенность на уровне 0,5.

На сайте <https://haplogrep.i-med.ac.at/> получены те же результаты:



Haplogrep 3 Haplogrep 2 GitHub About

Summary Samples (3)

Samples:	All (3)	Mutations:	first 8 mutations	Search:		
Sample	Haplogroup	Quality	Ns	Coverage	Range	Mutations
HG00103.final_1.fastq	H2a2a1	<div style="width: 100px; height: 10px; background-color: green; margin-bottom: 5px;"></div> <div style="width: 100px; height: 10px; background-color: yellow; border: 1px solid black; margin-bottom: 5px;"></div>	0	16569	1 ranges	
HG03280.final_1.fastq	H2a2a1	<div style="width: 100px; height: 10px; background-color: green; margin-bottom: 5px;"></div> <div style="width: 100px; height: 10px; background-color: yellow; border: 1px solid black; margin-bottom: 5px;"></div>	0	16569	1 ranges	16026T
NA18966.final_1.fastq	H2a2a1	<div style="width: 100px; height: 10px; background-color: green; margin-bottom: 5px;"></div> <div style="width: 100px; height: 10px; background-color: yellow; border: 1px solid black; margin-bottom: 5px;"></div>	0	16569	1 ranges	

Showing 1 to 3 of 3 entries Previous **1** Next

Classification Time: 1 sec

The screenshot shows the Haplogrep 3 dashboard interface. The top navigation bar includes links for 'Upload Data', 'Phylogenies', 'Haplogrep 2', 'Documentation', 'Download', 'GitHub', and 'About'. Below the navigation, there are tabs for 'Summary' (selected) and 'Samples (3)'. On the right, there is an 'Export' button.

Haplotype Data: A donut chart titled 'Haplotype Data' showing the distribution of Haplotype H2a2a1. A dropdown menu labeled 'Haplotype' is open, showing the selected category.

Clusters: A table titled 'Clusters' with columns 'Cluster', 'Count', and 'Percentage'. It shows one cluster named 'H' with a count of 3 and 100% percentage. A 'Population Frequencies' bar chart is shown below, with a tooltip for 'European (non-Finnish)' at 76.0%. The source is listed as 'gnomAD'.

Quality Control: A section with a progress bar and a summary table. The progress bar is mostly red. The summary table includes:

Category	Value
Warnings	1
Failed	2
Files	1
Samples	3
Input Variants	35,406
Reference Mismatches	3
Tree Overlap (%)	0
Strand Flips	1
Out Of Range Variants	35,401
Multiallelic Variants	0
Indel Variants	0
VCF Filtered Variants	0
Duplicate Variants	0
Low Sample Call Rate	0
Monomorphic Variants	0
Variant Call Rate < 90%	1

Для всех образцов определилась одна и та же МТ гаплогруппа - H2a2a1, которая больше всего распространена у европейцев.

Судя по всему, митохондриального генома, содержащегося в данных, недостаточно для точной оценки гаплогруппы образцов.