Задание 1.
Загрузите файлы с идентификаторами SRR10426968, SRR10426969 и SRR10426970 (найдите их самостоятельно в любой архивной базе данных), проведите контроль качества, сконструируйте выравнивание на референс (любой Homo sapiens), проведите дальнейшую предподготовку и коллинг вариантов (проще всего – с помощью VarScan, но данная программа некорректно генерирует VCF-файлы; нужно будет сделать так, чтобы выходной файл был похож на входной формат для VEP*).
Отправьте в качестве результата свой пайплайн (скрипт) и получившийся vcf-файл.

fastqc SRR10426968.fastq
fastqc SRR10426969.fastq
fastqc SRR10426970.fastq

cutadapt --quality-cutoff=20 -o SRR10426968_trimmed.fastq SRR10426968.fastq
cutadapt --quality-cutoff=20 -o SRR10426969_trimmed.fastq SRR10426969.fastq
cutadapt --quality-cutoff=20 -o SRR10426970_trimmed.fastq SRR10426970.fastq

bwa mem -t 24 GRCh38_latest_genomic.fna SRR10426968_trimmed.fastq

samtools view -bS SRR10426968.sam > SRR10426968.bam
samtools view -bS SRR10426969.sam > SRR10426969.bam
samtools view -bS SRR10426970.sam > SRR10426970.bam

java -jar picard.jar SortSam INPUT=SRR10426968.bam OUTPUT=SRR10426968_sorted.bam SORT_ORDER=coordinate
java -jar picard.jar SortSam INPUT=SRR10426969.bam OUTPUT=SRR10426969_sorted.bam SORT_ORDER=coordinate
java -jar picard.jar SortSam INPUT=SRR10426970.bam OUTPUT=SRR10426970_sorted.bam SORT_ORDER=coordinate

java -jar picard.jar MarkDuplicates INPUT=SRR10426968_sorted.bam OUTPUT=SRR10426968_marked_duplicates.bam METRICS_FILE=SRR10426968_marked_dup_metrics.txt
java -jar picard.jar MarkDuplicates INPUT=SRR10426969_sorted.bam OUTPUT=SRR10426969_marked_duplicates.bam METRICS_FILE=SRR10426969_marked_dup_metrics.txt
java -jar picard.jar MarkDuplicates INPUT=SRR10426970_sorted.bam OUTPUT=SRR10426970_marked_duplicates.bam METRICS_FILE=SRR10426970_marked_dup_metrics.txt

gatk AddOrReplaceReadGroups -I SRR10426968.bam -O SRR10426968_rg.bam -RGID 1 -RGLB lib1 -RGPL ILLUMINA -RGPU unit1 -RGSM test_sample1
gatk AddOrReplaceReadGroups -I SRR10426969.bam -O SRR10426969_rg.bam -RGID 2 -RGLB lib2 -RGPL ILLUMINA -RGPU unit2 -RGSM test_sample2
gatk AddOrReplaceReadGroups -I SRR10426970.bam -O SRR10426970_rg.bam -RGID 3 -RGLB lib3 -RGPL ILLUMINA -RGPU unit33 -RGSM test_sample3

gatk SortSam -I SRR10426968_rg.bam -O SRR10426968_rg_sorted.bam -SO coordinate
gatk SortSam -I SRR10426969_rg.bam -O SRR10426969_rg_sorted.bam -SO coordinate
gatk SortSam -I SRR10426970_rg.bam -O SRR10426970_rg_sorted.bam -SO coordinate

samtools index SRR10426968_rg_sorted.bam & samtools index
SRR10426969_rg_sorted.bam & samtools index SRR10426970_rg_sorted.bam

amtools flagstat SRR10426968_sor
ted.bam
samtools flagstat SRR10426970_sorted.bam
samtools flagstat SRR10426969_sorted.bam
1105460 + 0 in total (QC-passed reads + QC-failed reads)
1105460 + 0 primary
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
1081884 + 0 mapped (97.87% : N/A)
1081884 + 0 primary mapped (97.87% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
1152718 + 0 in total (QC-passed reads + QC-failed reads)
1152718 + 0 primary
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
1127654 + 0 mapped (97.83% : N/A)
1127654 + 0 primary mapped (97.83% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
1272004 + 0 in total (QC-passed reads + QC-failed reads)
1272004 + 0 primary
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
1244739 + 0 mapped (97.86% : N/A)
1244739 + 0 primary mapped (97.86% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2

0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

BaseRecalibrator

gatk BaseRecalibrator |
-R GRCh38_latest_genomic.fna |
-I SRR10426968_sorted.bam |
--known-sites GRCh38_latest_clinvar.vcf |
-O SRR10426968_recal.table
gatk BaseRecalibrator |
-R /GRCh38_latest_genomic.fna |
-I SRR10426969_sorted.bam |
--known-sites GRCh38_latest_clinvar.vcf |
-O SRR10426969_recal.table
gatk BaseRecalibrator |
-R GRCh38_latest_genomic.fna |
-I SRR10426970_sorted.bam |
--known-sites GRCh38_latest_clinvar.vcf |
-O SRR10426970_recal.table

BQCR

gatk ApplyBQSR |
-R GRCh38_latest_genomic.fna |
-I SRR10426968_sorted.bam |
-bqsr SRR10426968_recal.table |
-O SRR10426968_bqsr.bam
gatk ApplyBQSR |
-R GRCh38_latest_genomic.fna |
-I SRR10426969_sorted.bam |
-bqsr SRR10426969_recal.table |
-O SRR10426969_bqsr.bam
gatk ApplyBQSR |
-R GRCh38_latest_genomic.fna |
-I SRR10426970_sorted.bam |
-bqsr SRR10426970_recal.table |
-O SRR10426970_bqsr.bam

samtools mpileup -B -f nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw7/genome/GCF_000001405.40_GRCh38.p14_genomic.fna SRR10426968_bqsr.bam SRR10426969_bqsr.bam SRR10426970_bqsr.bam > 2/output.pileup

java -jar /home/nastyslav/miniconda3/envs/varscan/share/varscan-2.4.6-0/VarScan.jar mpileup2snp |
nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw7/trimmed/2/output.pileup |
--min-coverage 4 |
--min-reads2 1 |

--min-var-freq 0.1 |
--p-value 0.05 |
--output-vcf 1 > 2/variants_sensitive2.vcf

*Variant Effect Predictor: https://www.ensembl.org/Tools/VEP
Познакомьтесь с данным инструментом: если сложно сразу работать с ним, попробуйте загрузить в него свои результаты или пробные варианты, вроде такого:
1 65568 . A C . . .
2 265023 . C T . . .
3 319780 . GA G . . .
Подумайте, почему на один вариант после анализа в таблице приходится не одна строка, а несколько (отвечать не надо)? Этот инструмент (VEP) относится к следующему занятию, но разбирать его подробно не будем, поэтому чем больше Вы его изучите, тем проще будете понимать предстоящий материал.

Аннотация .vcf в https://www.ensembl.org/Tools/VEP

Проведена аннотация вариантов в формате VCF с помощью инструмента VEP от Ensembl. После фильтрации по клинической значимости (патогенные и потенциально патогенные) и влиянию на белок (высокое и умеренное) были выявлены следующие варианты:

- ***Миссенс-мутация в гене ASL (ENSG00000126522) с заменой A/G;***
- ***Мутация в сайте сплайсинга приемника в гене BRCA1 (ENSG00000012048) с заменой C/T;***
- ***Миссенс-мутация в гене APOE (ENSG00000130203) с заменой T/C.***