| Команда | Результат | Комментарий |
| --- | --- | --- |
| bwa index virus\_ref\_gene.fasta | virus\_ref\_gene.fasta.amb  virus\_ref\_gene.fasta.ann  virus\_ref\_gene.fasta.bwt  virus\_ref\_gene.fasta.pac  virus\_ref\_gene.fasta.sa | Индексируем референсный ген |
| bwa mem virus\_ref\_gene.fasta virus\_SRA.fastq | samtools view -S -b - | samtools sort - -o virus\_SRA\_sorted.bam | virus\_SRA\_sorted.bam | Выровнять риды на референсный ген, перевести результат в .bam формат и отсортировать |
| samtools flagstat virus\_SRA\_sorted.bam | 99.94% mapped | checking that alignment is successful |
| samtools index virus\_SRA\_sorted.bam | virus\_SRA\_sorted.bam.bai | Indexing alignment |
| samtools mpileup -d 100000 -f virus\_ref\_gene.fasta virus\_SRA\_sorted.bam > my.mpileup | my.mpileup | Create an intermediate file required for varscan. -d is “per file depth” and 100000 is just some random number for now. Need to find out what this parameter actually is and what number should be passed to it |
| java -jar ./VarScan.v2.4.4.jar mpileup2snp my.mpileup --min-var-freq 0.95 --variants --output-vcf 1 > VarScan\_results\_conservative\_snp.vcf | VarScan\_results\_conservative\_snp.vcf | Search for snp variants. Here we search only for highly conservative variants: those which are present in >=95% reads. 5 snps found 0 indel |
| java -jar ./VarScan.v2.4.4.jar mpileup2snp my.mpileup --min-var-freq 0.001 --variants --output-vcf 1 > VarScan\_results\_rare\_snp.vcf | VarScan\_results\_rare\_snp.vcf | search for rare variants. Found 21 SNP, 2 indel |
| bwa mem virus\_ref\_gene.fasta control\_1.fastq | samtools view -S -b - | samtools sort - -o control\_1\_sorted.bam | control\_1\_sorted.bam | Aligning the first control samples to the reference sequence |
| samtools index control\_1\_sorted.bam | control\_1\_sorted.bam.bai |  |
| samtools mpileup -d 100000 -f virus\_ref\_gene.fasta control\_1\_sorted.bam > control\_1\_mpileup | control\_1\_mpileup | Creating intermediate file for varscan |
| java -jar ./VarScan.v2.4.4.jar mpileup2snp control\_1\_mpileup --min-var-freq 0.01 --variants --output-vcf 1 > control\_1\_snp.vcf | control\_1\_snp.vcf | Calling rare variants in the first control sample |
| … the same procedure for control\_2.fasta and control\_3.fasta |  |  |

Given an aligning, for each position in the reference gene we associate four values: frequencies of A, C, G and T in the reads aligned on this position. We focus on the second highest frequency, as it may indicate a rare variant. Here are data on second highest frequencies in the control samples and the real sample

| Source/Statistic | Mean | Max | Max - Mean | Max/Mean |
| --- | --- | --- | --- | --- |
| Control 1 | 0.0005929303625901071 | 0.0065676255925473095 | 0.005974695229957202 | 11.07655469667272 |
| Control 2 | 0.0006007021535627968 | 0.006229471182556449 | 0.005628769028993652 | 10.370316046994539 |
| Control 3 | 0.0006003635225677369 | 0.006992548740370984 | 0.006392185217803247 | 11.64719120586151 |
| Real sample | 0.00039844029990385236 | 0.009383632038352886 | 0.008985191738449035 | 23.550910991225663 |

Only positions with the second highest frequency greater than 0.007 are considered to be putative rare variants, as smaller frequencies appear in control samples. Only two positions are of interest:

| Position/Statistic | Second highest frequency (SHF) | SHF - Mean | SHF/Mean |
| --- | --- | --- | --- |
| 307 | 0.009383632038352886 | 0.008985191738449035 | 23.550910991225663 |
| 1458 | 0.008348622042063483 | 0.007950181742159631 | 20.953257097934344 |

All other positions have SHF less than 0.00318 and are not considered as rare variants.