Article

Abstract

Mechanisms of antibiotic resistance on the one hand are crucial for bacterias and on the other hand can be critically dangerous for people who are getting these drugs. In order to find which mutations can lead to resistance we made this research. Here are provided assembly and search for SNP’s which can cause antibiotic resistance for E.coli strain K-12 substrain MG1655, laboratory workhorse using reference assembled and annotated genome and genome with SNP’s. Detected SNP’s showed that two of three mismatches help bacteria to adjust metabolism and get resistance. There are possible efflux pumps from the gene responsible for sensor histidine kinase EnvZ, Multidrug efflux pump RND permease AcrB and peptidoglycan DD-transpeptidase FtsI enabling penicillin binding.

Introduction

Howdays the problem of antibiotic resistance is a really huge challenge to solve for humankind. High frequency of bacterial mutations and absence of paired chromosomes provide them a framework to avoid effects of antibiotics and they use it with pleasure as MRSA [1] even when they get just SNP’s [2]. That’s why there is a great need in detecting SNP's using bioinformatic tools because they provide speed and high effectivity in comparison to routine wet lab methods [3]. Resistance mechanisms can involve various alterations, such as modifications to the target site, changes in the antibiotic itself, shifts in metabolic pathways, efflux pumps, decreased permeability, and cell membrane modifications [4]. Our purpose was to make reference genome assembly from raw sequence data of antibiotic resistance strain and compare it with the reference genome E.coli strain K-12 substrain MG1655, laboratory workhorse. It may show different modifications that provide resistance for E. coli through SNP’s.

Methods

Sequencing Data and Reference Genome

We initiated our study with raw sequencing data obtained from an Illumina sequencing run, targeting an Escherichia coli strain exhibiting resistance to ampicillin. The dataset comprised 455,876 reads, formatted in FASTQ files. A non-resistant E. coli strain (K-12 substrain MG1655) served as the reference sequence for comparative analyses. The reference genome was retrieved from the NCBI genome database [7] .

Quality Control Analysis

In order to evaluate the quality metrics of the DNA sequence data and identify potential issues that could impact our analyses, we utilized FastQC(version 0.12.1) [8], a tool for quality control assessment of raw sequence data. Post-trimming, FastQC was employed once again to inspect the quality of the trimmed sequences, facilitating a direct comparison between pre- and posttrimming data quality (Fig 1, Results section)

Read Trimming

For the preprocessing of sequence data, we applied Trimmomatic(version 0.39) [9] to trim lowquality bases from the reads. The trimming parameters were set as follows: HEADCROP:20 to remove the first 20 bases from every read, TRAILING:20 to cut 20 bases from the end of each read, SLIDINGWINDOW:10:20 to apply a sliding window trimming, and MINLEN:20 to exclude reads shorter than 20 bases.

Alignment and Variant Analysis

Alignment of the trimmed reads against the reference genome was performed utilizing the Burrows-Wheeler Aligner (BWA)[10] . We then processed the SAM files into BAM format with Samtools[11] and further sorted and indexed these files to optimize analyses and visualization. Variants within each sample were identified using VarScan2[12] , enabling the categorization of variants into missense, nonsense, or synonymous alterations based on their potential effects on the genome.

Visualization of Aligned Sequences

For a comprehensive examination of aligned sequences against the reference genome, we employed the Integrative Genomics Viewer (IGV) [13] , which facilitated the visualization of genetic variants and overall alignment quality.

Results

Our initial evaluation of the forward and reverse raw data, comprising 455,876 reads, revealed low per-base sequence quality towards the ends of reads and an uneven distribution of base composition (A, T, G, C) at the start of reads. These observations suggest potential biases arising during library preparation or sequencing processes. Subsequent processing with Trimmomatic significantly improved the data quality by eliminating suboptimal reads. The post-trimming statistics revealed that of the initial 455,876 read pairs, 431,982 pairs (97.89%) remained with sequence length (20-80 bp). A mere 128 read pairs (0.03%) were completely discarded due to insufficient quality (Figures 1 and 2) . Further quality control analysis conducted via Samtools after alignment confirmed the high quality of our processed dataset, with all reads passing the QC standards (Figure 3). A detailed breakdown of alignment statistics showed that, of the total reads, 863,964 were classified as primary alignments without secondary alignments or duplicates. Notably, 88 reads aligned to multiple locations, hinting at potential repetitive elements or regions within the genome. Of the processed reads, 860,382 (99.58%) successfully mapped to the reference genome, with 857,350 mapped in proper pairs, indicating correct orientation and spacing in line with the sequencing protocol. The analysis underscored that all sequences stemmed from paired-end sequencing, evenly distributed among Read 1 and Read 2. Within this subset, 1,576 reads were identified as singletons, with only one member of the pair aligning to the reference genome. No instances were observed where reads had their mate mapped to a different chromosome, including those reads with a high mapping quality (mapQ>=5).

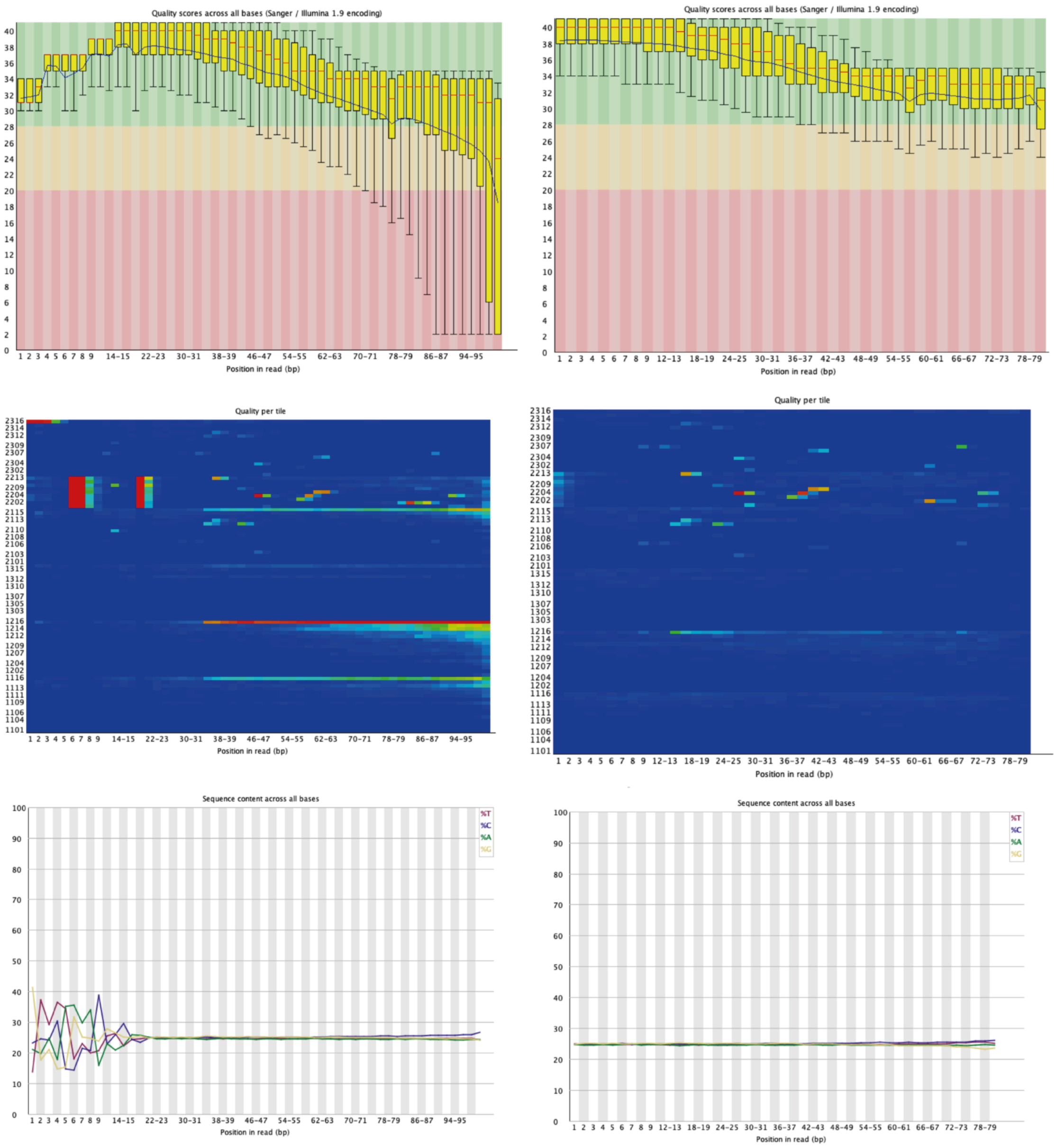


Figure 1. Difference between row(left) and trimmed(right) FastQC data on amp_res_1_fastqc sequence

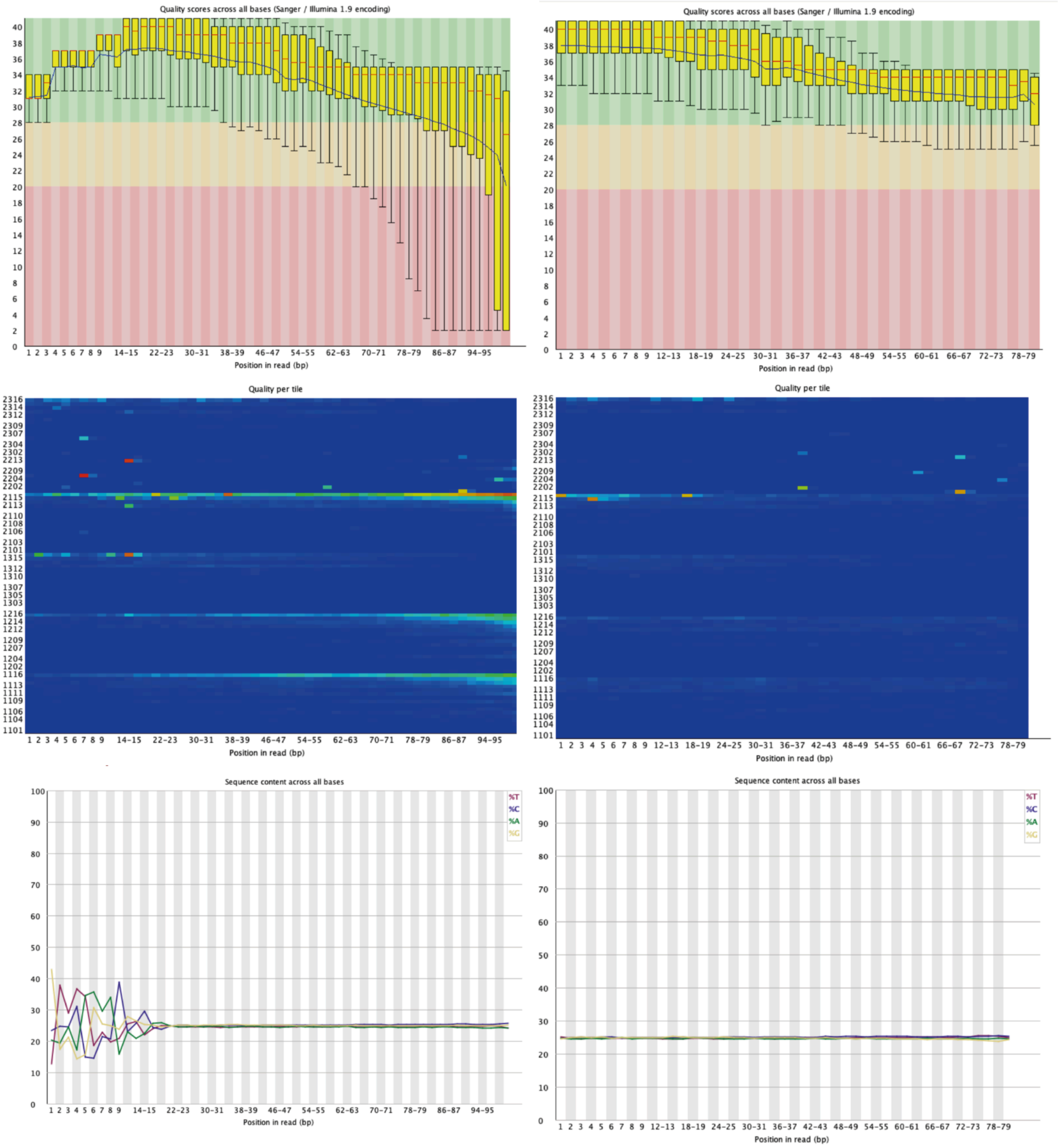
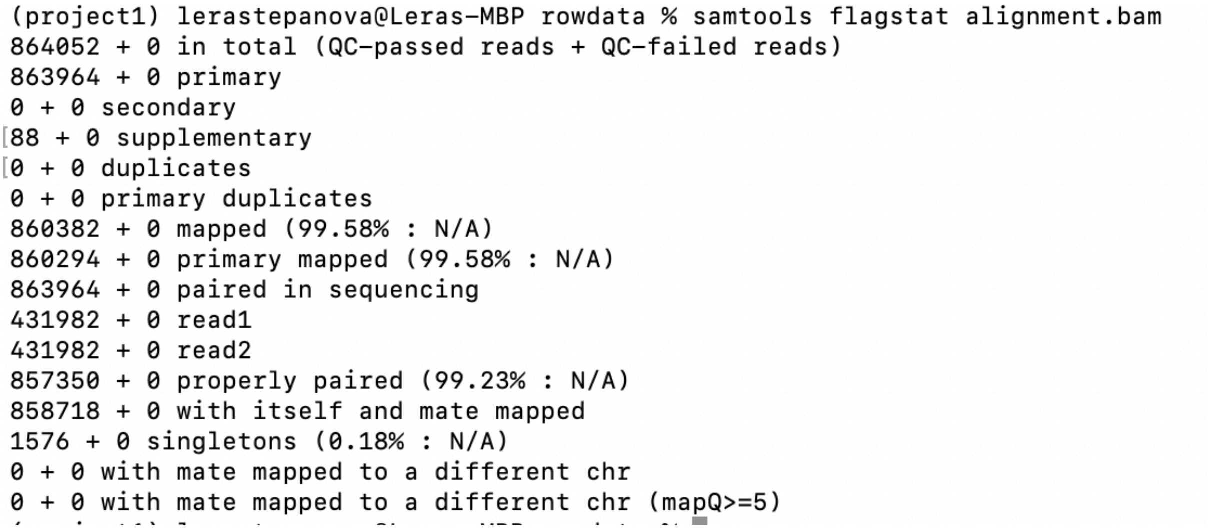


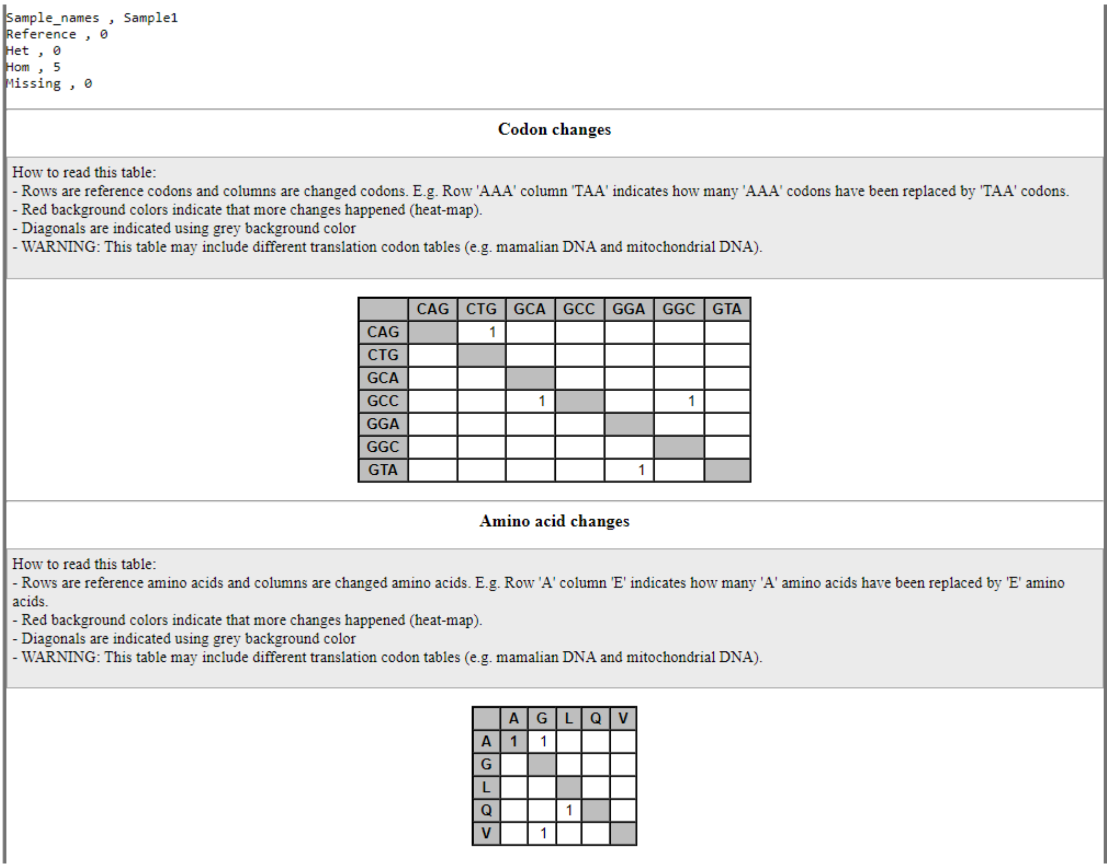
Figure 2. Difference between row(left) and trimmed(right) FastQC data on amp_res_2_fastqc sequence

The variant analysis conducted with VarScan via the mpileup2snp function was aimed at identifying single nucleotide polymorphisms (SNPs) and indels from the generated mpileup file, utilizing specific filtering criteria to ensure the reporting of high-confidence variants. VarScan identified 7 variant positions within the pileup file, of which 6 were single nucleotide polymorphisms (SNPs) and 1 an indel (insertion or deletion). One variant was excluded based on the strand-filter criterion, leaving 5 variant positions to be reported. These comprised exclusively SNPs, with indels being effectively filtered out or not meeting the criteria for reporting (Figure 3).



```
(project1) lerastepanova@Leras-MBP rowdata % samtools flagstat alignment.bam
864052 + 0 in total (QC-passed reads + QC-failed reads)
863964 + 0 primary
0 + 0 secondary
88 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
860382 + 0 mapped (99.58% : N/A)
860294 + 0 primary mapped (99.58% : N/A)
863964 + 0 paired in sequencing
431982 + 0 read1
431982 + 0 read2
857350 + 0 properly paired (99.23% : N/A)
858718 + 0 with itself and mate mapped
1576 + 0 singletons (0.18% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figure 3. Samtools analysis results



```
Sample_names , Sample1
Reference , 0
Het , 0
Hom , 5
Missing , 0
```

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).

	CAG	CTG	GCA	GCC	GGA	GGC	GTA
CAG		1					
CTG							
GCA							
GCC			1			1	
GGA							
GGC							
GTA					1		

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	G	L	Q	V
A	1	1			
G					
L					
Q			1		
V		1			

Figure 4. snpEFF analysis results.

SnpEFF report (Fig. 4) posses us to evaluate number of synonymous mutations, triplets and aminoacid variants.

After visualization of our data we found five SNP’s, two of which can be responsible for the mechanism of antibiotic resistance (Table 1).

Table 1. Single nucleotide polymorphisms presented in the genome.

Gene	Reference/ Aminoacid	Assembly/ Aminoacid	Function and involvement
FtsI	C/Ala	G/Gly	Peptidoglycan DD-transpeptidase
EnvZ	T/Val	G/Gly	Sensor histidine kinase
AcrB	A/Gln	T/Leu	Multidrug efflux pump RND permease AcrB
rybA	A/Phe	G/Ser	Small iRNA
rsgA	C/Ala	A/Ala	Ribosome small subunit-dependent GTPase A.

Discussion

Processed data allows us to make some predictions about capability to cause antibiotic resistance.

FtsI

Missense mutation. Protein coding. This gene responsible for Peptidoglycan DDtranspeptidase production and involved in penicillin binding and peptidoglycan glycosyltransferase activity which can be important in protection. Mismatch here causes change in the protein sequence from polar alanine to non-polar Glycine without indels. These facts make it promising. Looking at this situation it seems obvious that treatment is supposed to imply absence of the penicillin group in prescription [13].

EnvZ

Missense mutation. Protein coding. Gene has replacing mutation which switches valine to glycine. This gene is responsible for sensor histidine kinase, can regulate efflux pumps, and it also can provide antibiotic resistance. The best way to overcome this type of protection is to use antibiotics which violate protein synthesis or cell wall [14].

rybA

Missense mutation. Non-protein coding. Responsible for Small iRNA which can regulate expression, so it could theoretically cause higher activity of efflux pump due to overexpression. This problem could be solved by protein synthesis impairment or folic acid inhibition with antibiotics such as azithromycin or some sulfonomyde. But this particular gene represents an intragenic variant and does not appear to be the cause of resistance [15].

AcrB

Missense mutation glutamine to leucine. Protein coding. Multidrug efflux pump RND permease. And looks like a really good candidate. So patients can get MDR inhibitors to decrease the activity of pumps such as isoflavones. SNP’s on our point of view does not like a useful for the cells [16].

RsgA

Synonymous mutation. Protein coding. Enables GDP binding, enables GTP binding, enables GTPase activity, enables RNA binding, enables guanosine, tetraphosphate binding, enables hydrolase activity, enables metal ion binding, nucleotide binding, enables rRNA binding, enables rRNA binding. RsgA mutation causes just a synonymous mutation and rybA is not a protein coding gene. This SNP on our point of view can't make a contribution to cell viability [17].

References

1. Vestergaard M., Frees D., Ingmer H. Antibiotic resistance and the MRSA problem //Microbiology spectrum. – 2019. – Т. 7. – №. 2. – P. 10.1128/microbiolspec. gpp3-0057-2018.
2. Figueroa J. et al. Analysis of single nucleotide polymorphisms (SNPs) associated with antibiotic resistance genes in Chilean Piscirickettsia salmonis strains //Journal of fish diseases. – 2019. – Т. 42. – №. 12. – P. 1645-1655.
3. Ndagi U. et al. Antibiotic resistance: bioinformatics-based understanding as a functional strategy for drug design //RSC advances. – 2020. – Т. 10. – №. 31. – P. 18451-18468.
4. Munita J. M., Arias C. A. Mechanisms of antibiotic resistance //Virulence mechanisms of bacterial pathogens. – 2016. – P. 481-511.
5. Stavri M., Piddock L. J. V., Gibbons S. Bacterial efflux pump inhibitors from natural sources //Journal of antimicrobial chemotherapy. – 2007. – Т. 59. – №. 6. – P. 1247-1260.
6. E. coli K-12 substrain MG1655 Reference Genome.
7. FastQC: A Quality Control Tool for High Throughput Sequence Data. FastQC Website.
8. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120. Trimmomatic Paper.
9. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25(14), 1754-1760. BWA
10. Li, H., Handsaker, B., Wysoker, A., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079. Samtools.
11. Koboldt, D. C., Zhang, Q., Larson, D. E., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research, 22(3), 568-576. VarScan2.
12. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., et al. (2011). Integrative genomics viewer. Nature Biotechnology, 29, 24–26. IGV.
13. National Center for Biotechnology Information. Gene 944799 [Electronic resource]. URL: https://www.ncbi.nlm.nih.gov/gene/944799 (accessed: 27 October 2023).
14. National Center for Biotechnology Information. Gene 945108 [Electronic resource]. URL: https://www.ncbi.nlm.nih.gov/gene/945108 (accessed: 27 October 2023).
15. National Center for Biotechnology Information. Gene rybA [Electronic resource]. URL: https://www.ncbi.nlm.nih.gov/gene/?term=rybA (accessed: 27 October 2023).
16. National Center for Biotechnology Information. Gene 947272 [Electronic resource]. URL: https://www.ncbi.nlm.nih.gov/gene/947272 (accessed: 27 October 2023).

17. National Center for Biotechnology Information. Gene 948674 [Electronic resource]. URL: https://www.ncbi.nlm.nih.gov/gene/948674 (accessed: 27 October 2023).