DE NOVO ASSEMBLY AND WHOLE GENOME ANALYSIS OF THE *E.COLI* PATHOGENIC STRAIN IN THE CONTEXT OF 2011 HEMOLYTIC UREMIC SYNDROME OUTBREAK

ANNA MALYKHINA
Bioinformatics Institute
anja12@yandex.ru

OKSANA SIDORENKO
Bioinformatics Institute
oxiksid@gmail.com

MIKHAIL RAYKO
Bioinformatics Institute

November 28, 2022

ABSTRACT

2011 Germany *E.coli* O104:H4 outbreak affected more than 4,075 people in 16 countries, caused 53 deaths and inflicted economic costs for several countries.

In this work, we performed a *de novo* assembly and whole genome analysis of the *E.coli* pathogenic strain to identify the factors that contributed to the rapid spread and lethality of the disease.

It was found that the strain acquired pathogenicity due to horizontal genetic transfer: an *stx*2 (Shiga toxin)-encoding prophage is responsible for the toxigenic properties, plasmid-encoded CTX-M-15, TEM-1 along with a number of other genes of antibiotic resistance and one missence mutation in DNA gyrase subunit A gene explain its spread and resistance to therapy.

Keywords *Escherichia coli* · *E.coli* · NGS · *de novo* genome assembly · Shiga toxin · foodborne infections · Antibiotic resistance · 2011 German outbreak

1 INTRODUCTION

In April 2011, hundreds of people in Germany were hospitalized with hemolytic uremic syndrome (HUS), a deadly blood disease that often starts as food poisoning with bloody diarrhea and can lead to kidney failure.

A total of 3816 cases (including 54 deaths) were reported in Germany, 845 of which (22%) involved the hemolytic–uremic syndrome. The outbreak was centered in northern Germany and peaked around May 21 to 22. The incidence map is shown on Figure 1. Most of the patients in whom the hemolytic–uremic syndrome developed were adults (88%; median age, 42 years), and women were overrepresented (68%). The estimated median incubation period was 8 days, with a median of 5 days from the onset of diarrhea to the development of the hemolytic–uremic syndrome. Among 59 patients prospectively followed at Hamburg University Medical Center, the hemolytic–uremic syndrome developed in 12 (20%), with no significant differences according to sex or reported initial symptoms and signs. The outbreak strain was typed as an enteroaggregative Shiga-toxin–producing *E.coli* O104:H4, producing extended-spectrum beta-lactamase [1].

At the beginning of the spread of the disease, doctors' suspicions fell on the strain of *Escherichia coli* O157:H7. This strain causes tens of thousands of hospitalizations each year and often leads to HUS.

However, samples from patients with unknown disease back then did not pass biochemical tests for known strains of *E.coli* that cause HUS (including O157:H7). In addition, the symptoms of the new disease (for example, blood in

the stool) did not match the symptoms of pathologies caused by known strains of $Escherichia\ coli$. By this point, it was clear that humanity was facing a previously unknown pathogen (we now know it was a new O104:H4 strain) that somehow gained an additional virulence factors. Bioinformatics approaches were needed to solve the problem of obtaining these factors.

To investigate the evolutionary origin and pathogenic potential of the strain that caused the outbreak, the researchers launched a crowdsourced research program. They released sequencing data from an isolate of a girl from Hamburg, named sample TY2482. It allowed the discovery of the closest relatives of the new strain, which was a necessary step to further establish the sources of its acquisition of pathogenicity factors. There is a widely recognized mechanism for adaptation in bacteria and archaea called horizontal gene transfer (HGT), which is the sharing of genetic material between organisms that are not in a parent-offspring relationship. It is related to a variety of exchangeable genetic elements, like plasmids, bacteriophages, transposons and pathogenicity islands, which influences creation of new dangerous bacterial strains as microbial antibiotic resistance and pathogenicity are often associated with HGT. [3][4] As it turned out, it took place in our case as well.

The aim of our study was to establish the genetic cause of the outbreak and to characterize the properties of the $E.coli$ O104:H4 strain, as researchers did in 2011 for a new pathogen. When researching a new organism or strain of a bacterium, especially if it is pathogenic and deadly, as it was in our case, understanding the mechanisms of its virulence is extremely important for the control of the disease distribution and the development of approaches to its prevention, diagnosis and treatment. In this case, it is important to find and analyze all the significant genomic rearrangements (signs of which were manifested in the differences in the results of biochemical tests and symptoms of the disease from known strains, as we said above), which may conceal answers to questions that medicine and science poses and which cannot be revealed by aligning on reference. Only $de\ novo$ genome assembly can answer all these questions.

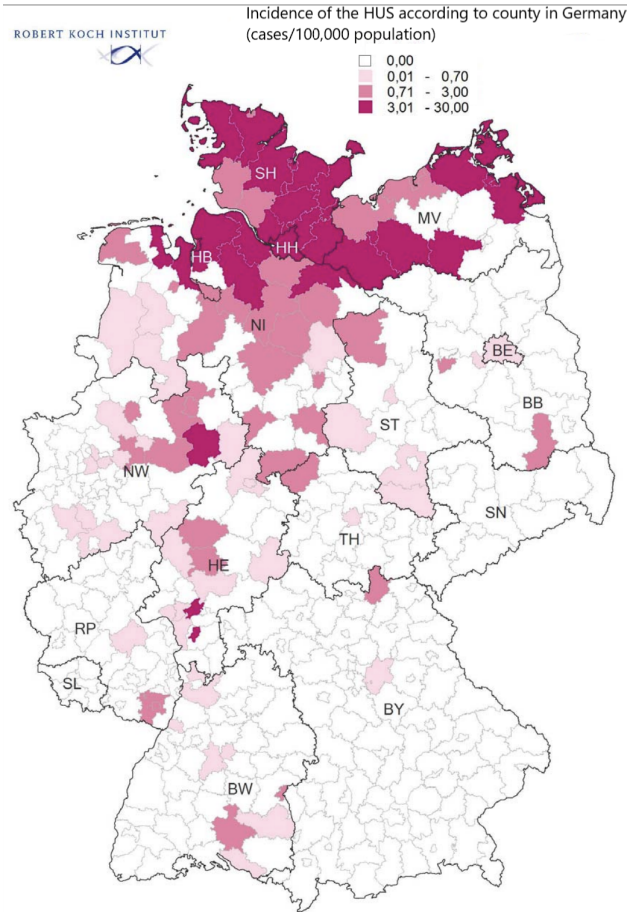

Figure 1: Incidence of HUS in the outbreak by district in which the infection probably took place (district of residence, or in the case of a travel history, district of residence at the time of infection). [2]

2 METHODS

2.1 SEQUENCE SOURCE

For this project, three libraries from the TY2482 sample were used with the following insert sizes and orientation:

- SRR292678 - paired end, insert size 470 bp
- SRR292862 – mate pair, insert size 2 kb
- SRR292770 – mate pair, insert size 6 kb

Raw Illumina sequencing reads were downloaded from the following links:

```
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R1_001.fastq.gz
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R2_001.fastq.gz
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R1_001.fastq.gz
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R2_001.fastq.gz
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R1_001.fastq.gz
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R2_001.fastq.gz
```

2.2 DATA ANALYSIS

For reads quality analysis FastQC software was used (ver.0.11.9) [5]. K-mer profile and genome size were estimated by Jellyfish program (ver.2.3.0) with the hash table size parameter 5000000 and k-mer size as 31 [6]. $E.coli$ genome was assembled by SPAdes (ver.3.13.1) with default parameters [7]. Assembling quality was checked by QUAST (ver. 5.2.0) with default settings [8]. Assembled genome was annotated by Prokka (ver.1.14.5) with –centre BI argument for sequencing ID [9]. Barrnap (ver. 0.9) was used as rRNA genes prediction tool [10]. After that, BLAST [11] search was performed for the genome in the RefSeq database with 16S rRNA that is most similar to the 16S rRNA in sample. The time period was specified (1900/01/01:2011/01/01[PDAT]) in order to be in character with the date of the sample. Genomic browser Mauve (ver. 2.4.0) was used for alignment of the sample and reference genomes [12]. Moreover, detailed search for toxicity and antibiotics resistance genes were also performed on Mauve. Antibiotics resistance analysis was done with web-version of ResFinder (ver.4.1) using both Chromosomal point mutations and Acquired antimicrobial resistance genes flags [13]. Visualization was done on SWISS-MODEL software [14].

3 RESULTS

3.1 $DE\ NOVO$ GENOME ASSEMBLY OF THE $E.COLI$ UNKNOWN PATHOGENIC STRAIN

All reads in the libraries of the TY2482 sample were of high quality according FastQC (see Supplementary). Genome size estimated by Jellyfish program was approximately 5 143 895 bp. By this method we estimated the shape of the k-mer distribution and the size of the genome prior to the beginning of the genome assembly. These parameters turned out to be very similar to the real ones, so we could start assembling. It should be noted that according to Jellyfish histogram some reads contain sequencing mistakes [Figure 2], which were corrected after genome assembly with SPAdes (rare k-mers are considered errors).

Genome was assembled in two variants: using one library (SRR292678) and using all three libraries (SRR292678, SRR292862, SRR292770). Assembling variants were compared in QUAST [Table 1] which showed significant improvement of all main scaffold assembly parameters in three-library variant.

	One-library assembly	Three-library assembly
N50	105 346	114 227
N90	21 421	24 915
L50	15	14
L90	53	50
Genome length	5 252 701 bp	5 253 564 bp

Table 1: Comparison of two assembly variants. Main scaffold parameters assessed by QUAST.

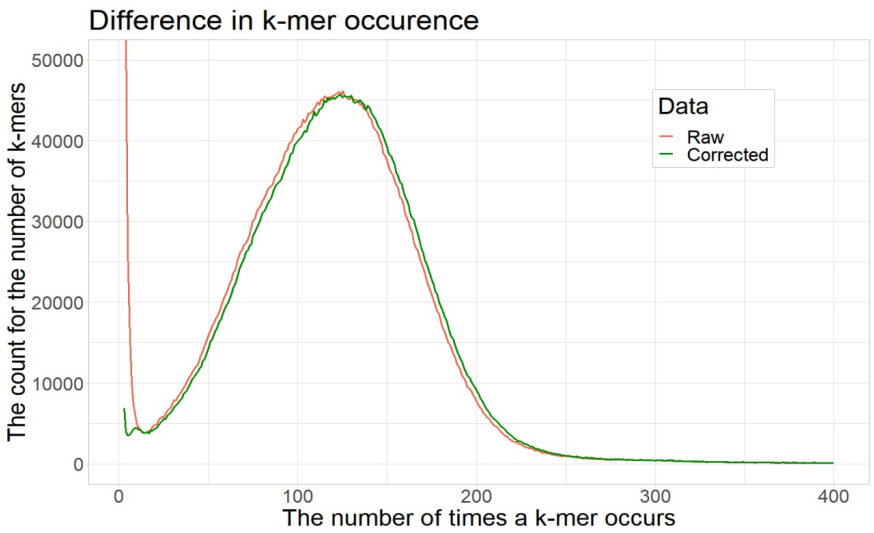

Figure 2: Difference in k-mer occurrence. In raw data (red line) there are a lot of reads appeared for few times which represents sequencing mistakes. These mistakes were corrected (green line) by SPAdes leading to significant decrease of the line in the left part of the graph.

Indeed, mate pair reads of 2 kb and 6 kb size helped to resolve some repeats resulting in decreased contigs number included in scaffolds , especially of small size [Figure 3]. Thus, for further genome analysis three-library assembly was used. The genome was annotated with Prokka.

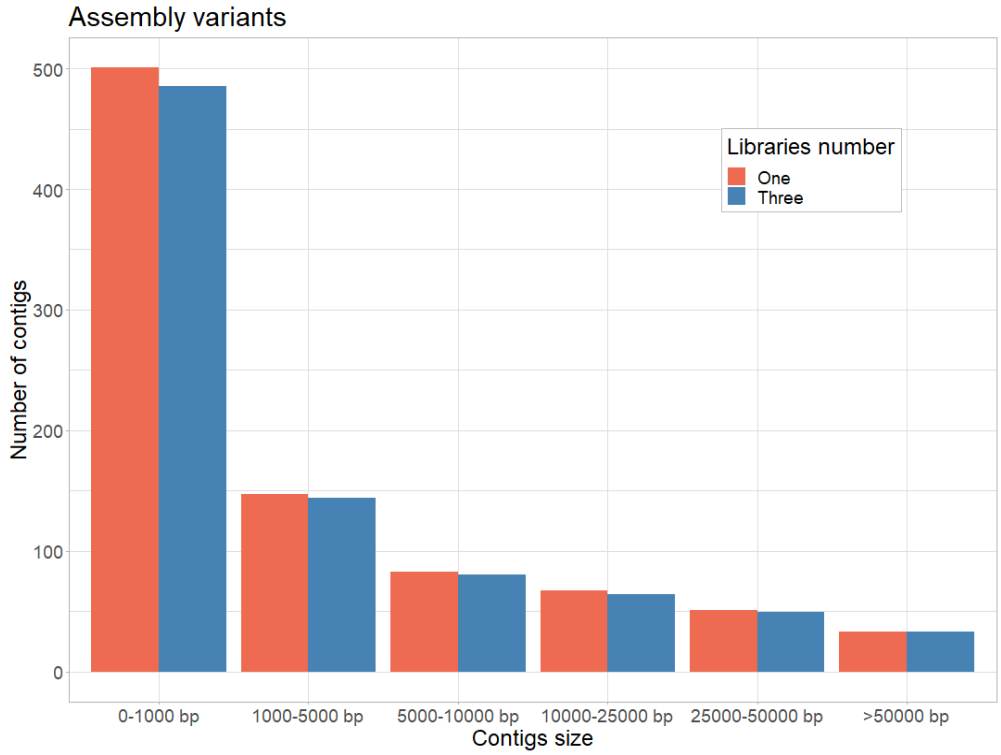

Figure 3: Influence of applying additional libraries on number of contigs in assembly (QUAST scaffold data).

3.2 GENOME ANALYSIS

The first step in analysis of *de novo* assembled genome was to find the genome that is the most similar to this pathogenic strain. For that purpose, we decided to use an important and evolutionarily conserved gene for comparison, namely 16S ribosomal RNA. Corresponding sequence in novel genome was located by rRNA genes prediction tool Barrnap (2 genes were found of 1538 bp and 406 bp accordingly) and subsequently applied to BLAST search. The best match to our strain is *Escherichia coli* strain 55989 (NC_011748.1). The genome of this strain was downloaded and aligned with our novel genome in Mauve browser [Figure 4].

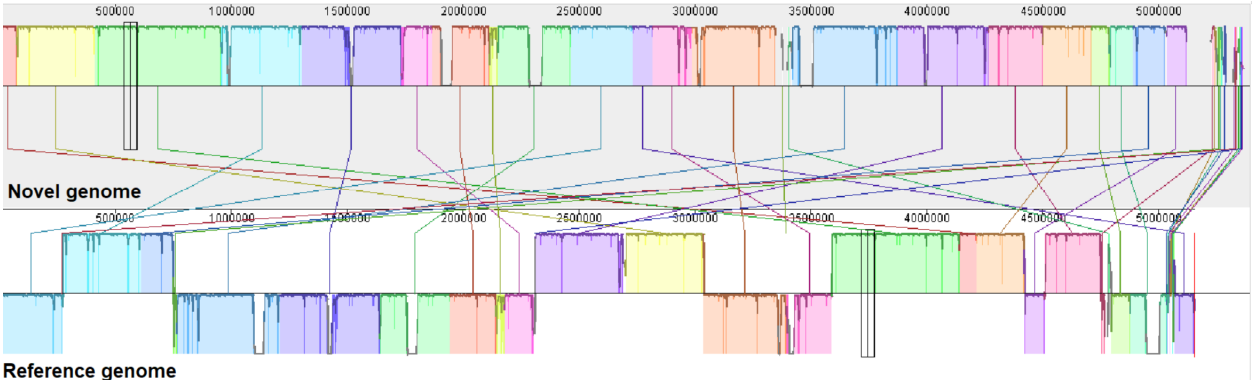

Figure 4: Alignment of novel and reference genomes in Mauve browser. Colors represent LCBs.

Mauve visualizes an alignment as the series of conserved segments called Locally Collinear Blocks (LCBs), which are similar to synteny blocks. Insertions and deletions in LCBs correspond to insertions and deletions in a bacterial chromosome. Separate unaligned regions that have no flanking regions from reference DNA correspond to extrachromosomal elements such as plasmids. Despite visual similarities with reference genome, novel genome clearly underwent some major genomic re-arrangement. Another finding that could be spotted is that novel genome is longer than reference genome which implies several insertions of unknown origin. Our interest lied in unaligned regions because they could contain information of how pathogenic $E.coli$ acquired additional virulence factors. In particular, we checked whether novel genome encodes Shiga toxins that cause internal bleeding. As a matter of fact, we found that two genes for Shiga toxin are present in novel genome: $stxA$ and $stxB$. The genes are located closely together (positions for $stxA$: 3483605-3483874; for $stxB$: 3483886-3484845) surrounded by many phage genes (for instance, phage Rha protein, phage endopeptidase Rz, phage antirepressor protein, phage lysozime R). We could presume that Shiga toxin genes were brought into $E.coli$ by phage. To identify genes responsible for antibiotic resistance, we applied ResFinder to our novel genome and found 10 genes [Table 2] while in reference (*Escherichia coli* strain 55989 (NC_011748.1)) genome there was only 1 ($tetB$ gene – for resistance to Doxycycline, Tetracycline, Minocycline). Obviously, pathogenic strain picked up not only virulence factors but also multi-drug resistance which leaded to serious disease outbreak.

Gene	Function/Product	Antibiotics (Class/Drugs)
bla_1	Beta-lactamase CTX-M-1 precursor	**Beta-lactam** Amoxicillin, ampicillin, aztreonam, cefepime, cefotaxime, ceftazidime, ceftriaxone, piperacillin, ticarcillin
bla_2	Beta-lactamase TEM precursor	**Beta-lactam** Amoxicillin, ampicillin, cephalothin, piperacillin, ticarcillin
$sul1$	Dihydropteroate synthase type-2, Sulfonamide resistance protein	**Folate pathway antagonist** Sulfamethoxazole
$sul2$	Dihydropteroate synthase type-2, Sulfonamide resistance protein	**Folate pathway antagonist** Sulfamethoxazole
$dfrA7$	Dihydrofolate reductase	**Folate pathway antagonist** Trimethoprim
$APH(3") - l$	Aminoglycoside 3"-phosphotransferase	**Aminoglycoside** Streptomycin
$APH(6)$ - $LC/APH(6) - ld$	Aminoglycoside 6-phosphotransferase	**Aminoglycoside** Streptomycin
$qacE$	Small multidrug resistance (SMR) efflux transporter	**Quaternary ammonium compound** Benzylkonium chloride, ethidium bromide, chlorhexidine, cetylpyridinium chloride
$tet(A)$	Tetracycline resistance, MFS efflux pump	**Tetracycline** Doxycycline, tetracycline
$gyrA$	DNA gyrase subunit A	**Quinolone** Nalidixic acid, ciprofloxacin

Table 2: Genes responsible for antibiotic resistance in pathogenic $E.coli$ strain (ResFinder)

4 DISCUSSION

Our analysis showed the $Escherichia\ coli$ strain 55989 (NC_011748.1) was the closest known relative of outbreak strain $E.coli$ O104:H4 (sample TY2482). Within the framework of this project, a full-fledged phylogenetic analysis of the strains was not carried out, however, there is reason to believe that the progenitor of the considered pathogenic strain is very similar to strain 55989.

It was shown that the pathogenic properties of strain O104:H4 (sample TY2482) include both "attack factors" (Shiga toxin-2 genes) [Figure 5] and "protection factors" (antibiotic resistance genes, including two β-lactamases: TEM-1 (broad-spectrum β-lactamase, ESBL) and CTX-M-15), which together make the strain extremely dangerous and complicates the differential diagnosis and treatment. We believe that almost all of the listed pathogenetic factors were acquired as a result of HGT. The acquisition of Shiga toxin-2 genes supposed to be of a phage nature: they follow immediately one after another in the genome sequence and are surrounded by numerous phage protein genes (at least 10 flanking genes at each end). These genes could be derived from any Shiga-like toxin-producing types of $E.coli$ (e.g. $Escherichia\ coli$ O157:H7 already mentioned or from $Shigella$ which is genetically closely related to $E.coli$).

As for the acquisition of antibiotic resistance, it can be assumed that it is associated with the getting a plasmid carrying the listed genes, since it was noted that all of them were surrounded by genes indicating the presence in the plasmid, such as $Incl1$ plasmid conjugative transfer proteins genes, transposases genes, mobile element proteins genes, genes of Integron integrase $Intl\ etc.$ Moreover, it has been observed that most antibiotic resistance genes are very close to each other (within 15000 bp): $sul2$, $APH(3") - l$, $APH(6) - LC/APH(6) - ld$, $dfrA7$, $qacE$, $sul1$, $tet(A)$, forming a

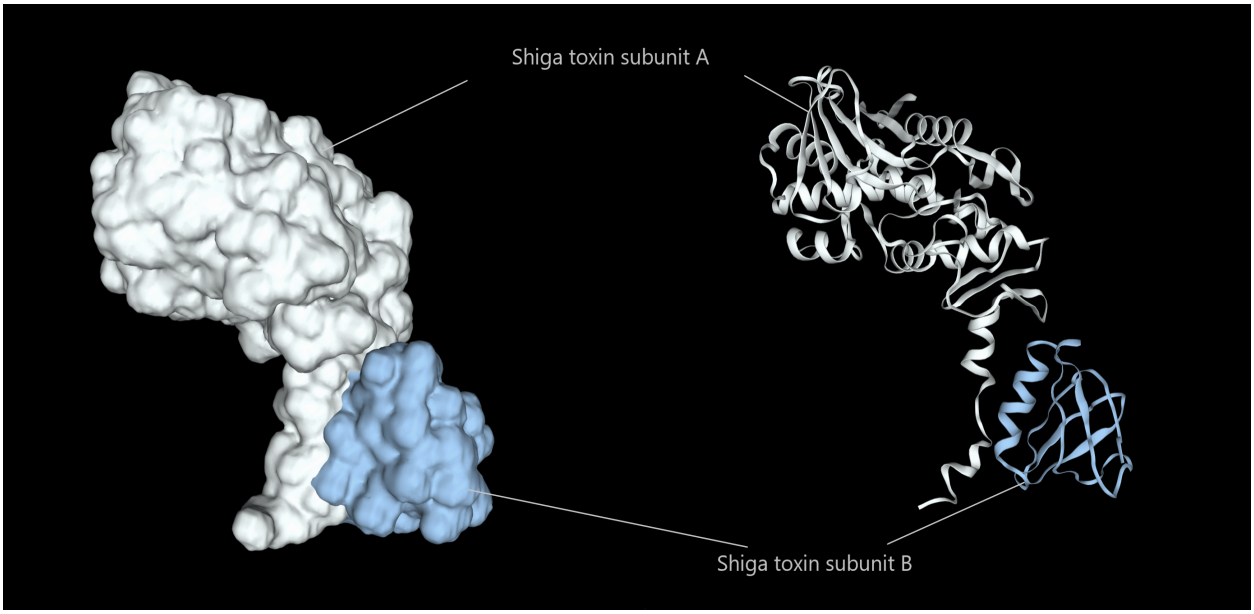

Figure 5: Crystal structure of Shiga-toxin surface (on the left) and its Ribbon diagram (on the right). Software: SWISS-MODEL[14]

multidrug-resistant cluster. Finding such a large cluster on a plasmid is worrisome, as plasmids imply a rapid spread of resistance among bacteria.

No signs of the appearance of the $gyrA$ gene responsible for resistance to nalidixic acid and ciprofloxacin due to HGT were found. But the result of processing the assembled genome of the TY2482 sample with the ResFinder software showed the presence of a point missense mutation TCG -> GCG, which caused the replacement of amino acids (S -> A). Serine and alanine are chemically different enough (serine is a hydrophilic and polar, while alanine is hydrophobic and nonpolar) amino acids to theoretically cause a change in the structure of the DNA gyrase subunit A molecule and hence - DNA gyrase avoidance of inhibition by quinolones.

As it is shown in the Figure 6, Ser83 can form a hydrogen bond with Asp87 which is absent in the case of Ala83.

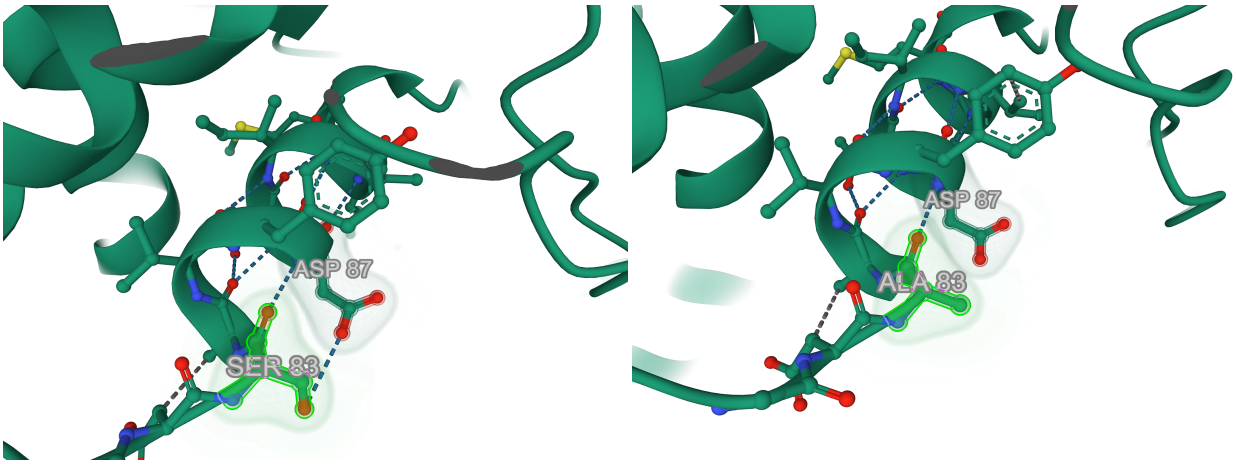

Figure 6: Model of changes in chemical bonds in the case of a serine-to-alanine substitution mutation in the 83 position of DNA gyrase subunit A molecule. Software: SWISS-MODEL[14], RCSB PDB[15]

Without considering this mutation, carbapenems, since they have a structure that makes them highly resistant to beta-lactamases, nalidixic acid and ciprofloxacin could be recommended for the therapy, but since the mutation presumably provides resistance to the latter, only carbapenems or modern macrolides, rifampicin or rifaximin, which act only in the intestine could be theoretically applicable, subject to mandatory prior thorough diagnosis. Without a reasonable

approach to the use of antibiotic therapy, careful selection of antibiotics with antibiotic susceptibility testing and dosage control, one can only aggravate the situation both for a particular patient and, more generally, become involved in the development of antibiotic resistance in bacteria.

In general, the history of the $E.coli$ O104:H4 strain in Germany in 2011 showed how important open research and the collaboration of epidemiologists, doctors, bioinformaticians and other specialists are in order to prevent new outbreaks of known diseases, timely monitor dangerous changes in the genomes of pathogens, and also find methods of diagnosis, prevention, treatment and control of an already manifested disease.

5 SUPPLEMENTARY

The lab journal with the detailed pipeline, settings and details can be found at the link `https://docs.google.com/document/d/13wa__wNz2SZg0tD-n_0av0ojIxQIDe8Hr9WYruQ0iMk/edit?usp=sharing`.

References

[1] Christina Frank, Ph.D., Dirk Werber, D.V.M., Jakob P. Cramer, M.D., Mona Askar, M.D., Mirko Faber, M.D., Matthias an der Heiden, Ph.D., Helen Bernard, M.D., Angelika Fruth, Ph.D., Rita Prager, Ph.D., Anke Spode, M.D., Maria Wadl, D.V.M., Alexander Zoufaly, M.D., et al. Epidemic Profile of Shiga-Toxin–Producing $Escherichia\ coli$ O104:H4 Outbreak in Germany. The New England Journal of Medicine. Massachusetts Medical Society.N Engl J Med 2011; 365:1771-1780. doi: 10.1056/NEJMoa1106483

[2] https://www.rki.de/DE/Content/InfAZ/E/EHEC/EHEC_O104/EHEC_Sachstandsbericht.pdf?__blob=publicationFile

[3] Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. Nat Rev Genet. 2015 Aug;16(8):472-82. doi: 10.1038/nrg3962. PMID: 26184597.

[4] Bloch SK, Felczykowska A, Nejman-Faleńczyk B. $Escherichia\ coli$ O104:H4 outbreak–have we learnt a lesson from it? Acta Biochim Pol. 2012;59(4):483-8. Epub 2012 Dec 13. PMID: 23240107.

[5] Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2022 Nov 24]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[6] gmarcais/Jellyfish: A fast multi-threaded k-mer counter [Internet]. [cited 2022 Nov 24]. Available from: https://github.com/gmarcais/Jellyfish

[7] SPAdes – Center for Algorithmic Biotechnology [Internet]. [cited 2022 Nov 24]. Available from: http://cab.spbu.ru/software/spades/

[8] QUAST – Center for Algorithmic Biotechnology [Internet]. [cited 2022 Nov 24]. Available from: https://cab.spbu.ru/software/quast/

[9] Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9.

[10] tseemann/barrnap: Bacterial ribosomal RNA predictor [Internet]. [cited 2022 Nov 24]. Available from: https://github.com/tseemann/barrnap

[11] BLAST: Basic Local Alignment Search Tool [Internet]. [cited 2022 Nov 24]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi

[12] the Darling lab | computational (meta)genomics [Internet]. [cited 2022 Nov 24]. Available from: https://darlinglab.org/mauve/download.html

[13] ResFinder 4.1 [Internet]. [cited 2022 Nov 24]. Available from: https://cge.food.dtu.dk/services/ResFinder/

[14] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018 Jul 2;46(W1):W296-W303. doi: 10.1093/nar/gky427. PMID: 29788355; PMCID: PMC6030848. https://swissmodel.expasy.org/

[15] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242. https://www.rcsb.org/