**De novo assembly and annotation application in discovery of novel pathogenic strain of E. Coli causing Hemolytic Uremic Syndrome (HUS)**

*Evgenia Khokhlova, Eugenia Ivanova (Strebulaeva)*

**Abstract**

In our study we investigate the unknown *E.coli* strain causing hemolytic uremic syndrome (HUS). For our analysis we used an Illumina reads from the patient sample for de novo assembly and genome annotation of this strain to investigate the pathogenic and antibiotic resistance potential of this unknown strain and figure out what was the cause of the *E.coli* pathogenicity. Finally, we describe new features of the novel strain and speculate about the origin of these features.

**Introduction**

Hemolytic uremic syndrome is defined by the characteristic triad of microangiopathic hemolytic anemia, thrombocytopenia, and acute renal failure. HUS is often caused by enterohemorrhagic *Escherichia coli* (EHEC) serogroup strains, mainly *Escherichia coli* O157:H7. Enterohemorrhagic *E. coli* is an emerging food- and water-borne pathogen family whose Shiga-like toxins induce painful hemorrhagic colitis with potentially lethal complications. These bacteria are not invasive, so bacteremia is rare, but they secrete ribosome inactivating Shiga-like toxins (Stx1 and Stx2 with variants) which are responsible for much of the organ damage, and Stx2 is more frequently associated with severe disease. Shiga toxin (Stx-) mediated injury to vascular endothelial cells in the kidney, brain, and other organs underlies the pathogenesis of HUS caused by EHEC. These potent cytotoxins are released in the gut by bacteria, enter the bloodstream, and cause endothelial injury through the binding to the globotriaosylceramide (Gb3) receptor on the plasma membrane of target cells. Gb3 is a sphingolipid receptor expressed on endothelial cells, podocytes, and proximal tubular cells in humans. Stx binding to Gb3 leads to Stx internalization by receptor-mediated endocytosis and its retrograde transport to the endoplasmic reticulum. This triggers a cascade of signaling events, involving NF-*κ*B activation, which induces apoptosis and the binding of leukocytes to endothelial cells [1].

Although most strains of *E. coli* are harmless for people, some strains have the potential to cause severe diseases. It is happening because of horizontal transfer of virulence genes from other species. Horizontal transfer affects larger parts of the genome and has a greater influence on the evolution of bacteria, especially on the gain of pathogenic properties and antibiotic resistance. Horizontal gene transfer may occur via three main mechanisms: transformation, transduction or conjugation. Transformation involves uptake of short fragments of naked DNA by naturally transformable bacteria. Transduction involves transfer of DNA from one bacterium into another via bacteriophages. Conjugation is a genetic recombination in which there is a transfer of DNA from a living donor bacterium to a living recipient bacterium by cell-to-cell contact. [2]. In all of these cases bacterial genomes acquire new elements which didn't exist previously. That’s why sometimes it’s important to assemble genome de novo because mapping to reference can’t reflect the new insertion in the genome.

**Methods**

**Description of the data and quality control**

The data were obtained from the NCBI Sequence Read Archive from the TY2482 sample with the following insert sizes and orientation:

SRR292678 - paired end, insert size 470 bp ([forward reads](https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R1_001.fastq.gz), [reverse reads](https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R2_001.fastq.gz), 400 Mb each)  
 SRR292862 – mate pair, insert size 2 kb, ([forward reads](https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R1_001.fastq.gz), [reverse reads](https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R2_001.fastq.gz), 200 Mb each)  
 SRR292770 – mate pair, insert size 6 kb, ([forward reads](https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R1_001.fastq.gz), [reverse reads](https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R2_001.fastq.gz), 200 Mb each)

To assess the quality of the data a fastqc program was used. The overall quality of the reads was good.

**K-mer profile and genome size estimation**

For assembly we used a de Bruijn graph strategy, which breaks the reads up into k-mers to facilitate assembly of correctly connected contigs. So, we need to assess k-mer distribution. We used a kmer counting program “Jellyfish” with the following command parameters (*jellyfish count -m 31 -C -s 10M*) that count the frequency of all possible k-mers of a given length in our data. Then we made a histogram file with the command *jellyfish histo* and visualized it in RStudio. To estimate the genome size, we used thefollowing formulas:

*N = (M\*L)/(L-K+1)*

*Genome\_size = T/N*

*\*(N: Depth of coverage, M: Kmer peak, K: Kmer-size, L: avg read length T: Total bases)*

## **Assembling *E. coli* X genome**

## For genome assembly we used the SPAdes program with the pair-end mode. The quality of the resulting assembly was assessed with QUAST program. Then we run SPAdes providing all three libraries: SRR292678 as a paired end, SRR292862 and SRR292770 as a mate pair to assess how the quality of our assembly changes when we add mate-pair libraries.

**Genome Annotation**

To annotate the assembled genome we runProkka programm on the “scaffolds.fasta” file from the SPAdes output with default parameters.

**Finding the closest relative of *E. coli* X**

Next, we compared each contig in our assembly against the 16S ribosomal RNA in the RefSeq database using BLAST. To locate 16S rRNA in the assembled *E. coli* X genome we used rRNA genes prediction tool Barrnap. Then we search for the similar genome in the RefSeq database with our 16S rRNA that we just found in the previous step. The closest genome relative to the *E. coli X* was used as a reference genome.

To understand the genetic cause of HUS and tracing the source of toxin genes in *E. coli* X we compared the *E. coli* X with the reference genome with the program Mauve, which visualizes an alignment as a series of conserved segments called Locally Collinear Blocks (LCBs), which are similar to synteny blocks.

**Antibiotic resistance detection**

To search for genes responsible for antibiotic resistance, we used the ResFinder tool, which specifically searches a database of genes implicated in antibiotic resistance, identifying similarities between the sequenced genome and this database using local alignment.

**Results**

**K-mer profile and genome size estimation**

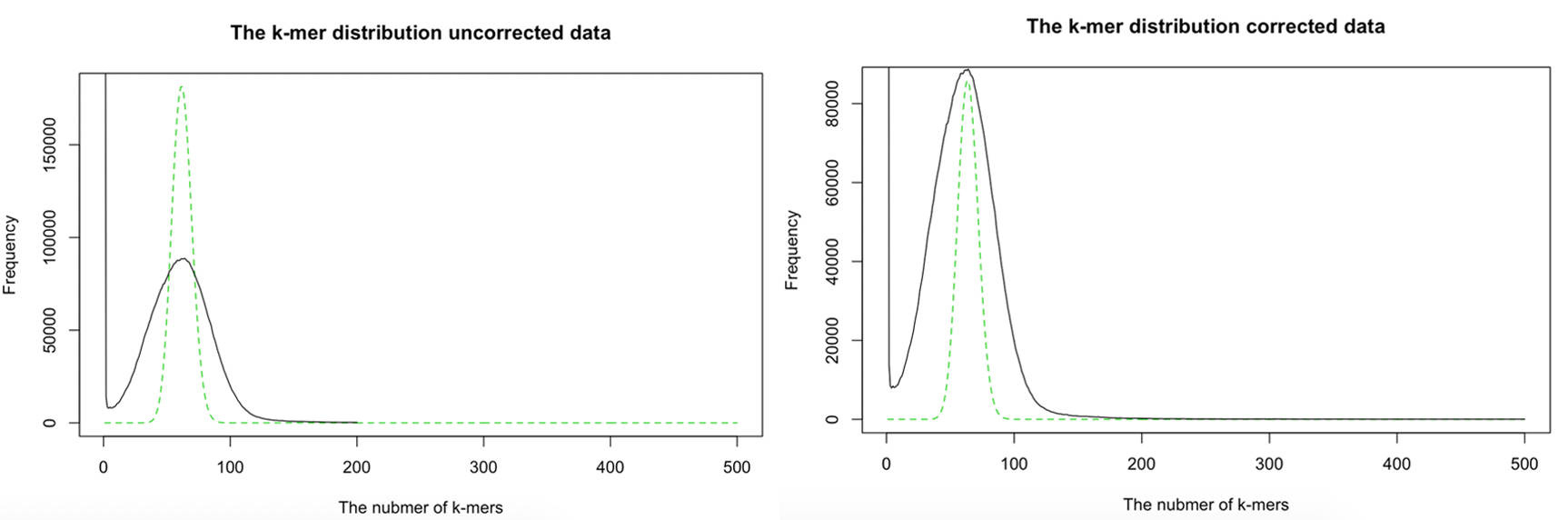
****

Figure 1 - Distribution of all possible k-mers of given length before and after correction by SPAdes.

N = *(M\*L)/(L-K+1) = (62\*90)/(90-31+1)=5580/60=93*

*Genome\_size = T/N = 5499346/93 = 59132,7*

*N\_corrected =*  *(M\*L)/(L-K+1) = (64\*90)/(90-31+1)=5760/60=96*

*Genome\_size\_corrected = T/N = 5498906/96 = 57280,3  
\*(N: Depth of coverage, M: Kmer peak, K: Kmer-size, L: avg read length T: Total bases)*  
The distribution of k-mers became more narrow and more similar to the ideal curve of scaled poisson distribution estimated by single copy region size, which indicates that quality of the reades was visibly improved by SPAdes inner correction.

**Genome assembling**

Assembly statistics displayed in table 1. It can be seen that running SPAdes with three libraries - one containing paired-end reads and two mate-pair reads - improves the quality of assembly. The length of N50 fragments increased threefold and the number of contigs decreased significantly.

Table 1 - Assembly statistics for one and three libraries.

|  |  |  |
| --- | --- | --- |
| Metrics | One paired-end library | Three libraries |
| N50 | 105346 | 335515 |
| Number of contigs | 519 | 369 |

**Genome annotation**

Annotation statistics are shown in table 2.

Table 2 - Summary for the annotation.

|  |  |
| --- | --- |
| Feature | Number |
| tRNAs | 80 |
| rRNAs | 0 |
| CRISPRs | 1 |
| CDS | 5064 |
| Unique gene codes | 2923 |

**The closest relative of *Escherichia coli* X**

To find relative strains we needed to extract 16s rRNA sequence from the assembly. With the Barrnap tool we found positions of all six copies of 16s rRNA genes in the assembly and extracted one of them using Python script. Information about 16s rRNA genes displayed in table 3. According to search in BLAST by 16s rRNA gene sequence the closest relative strain was Escherichia coli 55989, sequence ID [NC\_011748.1](https://www.ncbi.nlm.nih.gov/nucleotide/NC_011748.1?report=genbank&log$=nuclalign&blast_rank=1&RID=WJ7GBF1J013).

Table 3 - 16s rRNA genes description.

|  |  |  |
| --- | --- | --- |
| Number | Position | Length |
| 1 | NODE\_1 326359 - 327896 | 1537 |
| 2 | NODE\_1 595966 - 597503 | 1537 |
| 3 | NODE\_5 43835 - 45372 | 1537 |
| 4 | NODE\_5 85462 - 86999 | 1537 |
| 5 | NODE\_6 111955 - 113492 | 1537 |
| 6 | NODE\_71 314 - 719, partial | 405 |

**Causes of pathogenicity**

The results of alignment in Mauve showed shiga-like toxin genes presence. Characteristics of these genes are listed in table 4.

Table 4 - Shiga-like toxin genes in the E. coli X genome.

|  |  |  |
| --- | --- | --- |
| Gene | Position | Length |
| stxB | 3483605-3483874 | 269 |
| stxA | 3483886-3484845 | 959 |

**Antibiotic resistance estimation**

The results of antibiotic resistance estimation are given in table 5.

Table 5 - Comparison of antibiotic resistance of E. coli 55989 and E. coli X and genetic foundation.

|  |  |  |  |
| --- | --- | --- | --- |
| Antibiotic\resistance | E. coli 55989 | E. coli X | Resistance gene |
| unknown macrolide | resistance | resistance | mdf(A) |
| unknown fluoroquinolone | resistance | resistance | mdf(A) |
| unknown aminoglycoside | resistance | resistance | mdf(A) |
| doxycycline | resistance | resistance | tet(B) |
| unknown phenicol | resistance | resistance | mdf(A) |
| unknown tetracycline | resistance | resistance | mdf(A) |
| unknown rifamycin | resistance | resistance | mdf(A) |
| tetracycline | resistance | resistance | tet(B) |
| minocycline | resistance | resistance | tet(B) |
| ampicillin | no resistance | resistance | blaCTX-M-15 |
| sulfamethoxazole | no resistance | resistance | sul1, sul2 |
| trimethoprim | no resistance | resistance | dfrA7 |
| ceftriaxone | no resistance | resistance | blaCTX-M-15 |
| cephalothin | no resistance | resistance | blaTEM-1B |
| piperacillin | no resistance | resistance | blaCTX-M-15 |
| amoxicillin | no resistance | resistance | blaCTX-M-15 |
| cefepime | no resistance | resistance | blaCTX-M-15 |
| ceftazidime | no resistance | resistance | blaCTX-M-15 |
| ticarcillin | no resistance | resistance | blaCTX-M-15 |
| cefotaxime | no resistance | resistance | blaCTX-M-15 |
| aztreonam | no resistance | resistance | blaCTX-M-15 |
| streptomycin | no resistance | resistance | aph(6)-Id |

**Discussion**

Inspection of the assembly alignment revealed presence of shiga-like toxin (stx, subunits A and B) that was absent in the reference genome. Stx apparently originates from the bacteriophage lambda family and is found in *Shigella dysenteriae* and some *Escherichia coli* serotypes [3]. Subunit A inactivates 60s ribosomal subunit and thus inhibits protein synthesis. Subunit B is responsible for binding to cell surface receptors and assembling of holotoxin with subunit A [4]. In the genome assembly we found some evidence that *E. coli* X obtained shiga-like toxin via horizontal gene transfer. The toxin genes were surrounded by prophage genes - phage lysin, phage antirepressor protein, phage exonuclease etc., so the toxin seemed to be integrated with prophage DNA. This toxin causes hemorrhagic colitis and hemolytic-uremic syndrome which indicates that acquisition of stx genes was the crucial difference between reference enteroaggregative E. coli strain and the strain X caused outbreak of foodborne illness in Germany.

According to ResFinder reference strain of *E. coli* and strain under the study possessed a gene of resistance to tetracycline - tet - and we discovered it in the annotation of *E. coli* X. The tet gene (<https://www.uniprot.org/uniprot/P02981>) encodes a transmembrane protein that performs drug efflux and hence decreases antibiotic concentration in the cell. In contrast to the reference strain, X strain had two bla genes (bla\_1 and bla\_2, <https://www.uniprot.org/uniprot/P28585>, <https://www.uniprot.org/uniprot/P62593>). The bla genes encode beta-lactamase. This enzyme catalyzes transformation of beta-lactam antibiotics causing a loss of their activity. Presence of gene repA\_1 nearby the bla and tet genes - replication initiation protein - and other genes related to plasmids (such as ssbF, plasmid-derived single-stranded DNA-binding protein, and parM\_2, plasmid segregation protein) indicates that *E. coli* X received plasmids with genes coding beta-lactamase and efflux protein. In addition to mentioned ones several other antibiotic resistance genes were found on the chromosome of *E. coli* X and the reference strain. By the ResFinder and Mauve multiple drug resistance gene mdfA was discovered, but it is unclear which level of resistance it confers. Product of mdfA is an efflux pump which contributes to resistance to a broad spectrum of drugs [5]. We also found many multiple drug resistance genes with the same mechanism in the E. coli X annotation - emr, mdt, fsr, drrA, etc. - and multiple antibiotic resistance associated genes marA and marR. Mar genes supposed to affect antibiotic transport [6]. ResFinder showed additional antibiotic resistance genes for *E. coli* X that were not found in reference strain: sul1, sul2, aph and dfr. Those genes also could have been received via HGT since 2002 when the reference strain was discovered. Summarizing all this information we can propose two major mechanisms of resistance that *E. coli* X strain could use: inactivation of antibiotics by beta-lactamases and reducing concentration of the drug by product of the tet gene and less so by other products of genes involved in transport regulation.

**Conclusion**

Given results of our study point out that *Escherichia coli* X strain became pathogenic presumably due to transduction - activation of prophage with shiga-like toxin in pathogenic *E. coli* or *Shigella dysenteriae* could lead to *E. coli* X infection by this bacteriophage. Also, elevation of antibiotic resistance level related to acquisition of bla genes apparently occurred in consequence of bacterial conjugation with strain carried resistance plasmids. The increasing resistance requires changes in therapeutic strategies. Macrolides instead of beta-lactam antibiotics can be used to treat patients infected by novel *Escherichia coli* strain.

**References**

1. Boyer, O.; Niaudet, P. Hemolytic Uremic Syndrome: New Developments in Pathogenesis and Treatment. *International Journal of Nephrology* **2011**, *2011*, 1–10, doi:10.4061/2011/908407.

2. Messerer, M.; Fischer, W.; Schubert, S. Investigation of horizontal gene transfer of pathogenicity islands in Escherichia coli using next-generation sequencing. *PLoS ONE* **2017**, *12*, e0179880, doi:10.1371/journal.pone.0179880.

3. Friedman, D.I.; Court, D.L. Bacteriophage lambda: alive and well and still doing its thing. *Current Opinion in Microbiology* **2001**, *4*, 201–207, doi:10.1016/S1369-5274(00)00189-2.

4. Donohue-Rolfe, A.; Acheson, D.W.K.; Keusch, G.T. Shiga Toxin: Purification, Structure, and Function. *Clinical Infectious Diseases* **1991**, *13*, S293–S297, doi:10.1093/clinids/13.Supplement\_4.S293.

5. Lewinson, O.; Adler, J.; Poelarends, G.J.; Mazurkiewicz, P.; Driessen, A.J.M.; Bibi, E. The Escherichia coli multidrug transporter MdfA catalyzes both electrogenic and electroneutral transport reactions. *Proceedings of the National Academy of Sciences* **2003**, *100*, 1667–1672, doi:10.1073/pnas.0435544100.

6. Cohen, S.P.; Hächler, H.; Levy, S.B. Genetic and functional analysis of the multiple antibiotic resistance (mar) locus in Escherichia coli. *Journal of Bacteriology* **1993**, *175*, 1484–1492, doi:10.1128/JB.175.5.1484-1492.1993.