Student report

***E.coli* outbreak investigation**

Authors: Ilyina Margarita1 , Grabarnick Ilya 1 , Julia Solomennikova1

1ITMO University, St. Petersburg, Russian Federation

**Abstract**

For the research of a bacterial infection outbreak in Germany, clinically manifesting as HUS (hemolytic-uremic syndrome), it was necessary to obtain complete data on the putative infectious agent, the unknown strain of *E. coli* which had acquired the pathogenetic quality and is resistant to some of the common antibiotics.

For the investigation and data analyze, such methods of bioinformatics as de novo genome assembly using the de Bruijn graph strategy and beforehand k-mer profiling, similarity-based annotation and gene prediction using the sequence alignment algorithm, taxonomic identification by the 16S rRNA gene region and the search for other functional DNA sections using local alignment tool were used.

Sequences of a biomaterial isolate sample from one of the German patients hospitalized with such disease were obtained by Illumina sequencing technology, the resulting raw data libraries of 470 bp, 2 kb and 6 kb size were used to assemble the genome of the supposed infectious *E. coli X* strain. The most complex resulting assembly of all three libraries has a length of 5311537 bp, 90 contigs and statistical characteristics of N50 and N90 are 2815616 and 180369 respectively. In the resulting sequence, the conservative 16S rRNA genes regions were identified, which were needed to search for the complete genome of the known closest relative that E. coli 55989 strain was found to be. It was found as well that determining the virulence of the strain features are the Shiga-toxin genes, Shiga toxin II subunit A (StxA) and subunit B (StxB). Their appearance is due to the interaction between a bacterium with a prophage carrying pathogenicity, which is confirmed by the presence of the protein encoded within prophage CP-933V and number of other phage related genes in an ensemble with shiga-toxin genes. Also, 9 antibiotic resistance genes coding specific enzymes that destroy antibiotic molecules were found.

As a result of this study, data on the genetic structure and microbiological features of the pathogen were obtained, which was necessary to develop timely countermeasures and stop the epidemic.

**Introduction**

Almost every year mankind or individual countries are faced with outbreaks of viral and bacterial diseases, such as influenza (Uwishema et al., 2021)⁠, ebola (Aruna et al., 2019)⁠, plague (Rust Jr et al., 1971)⁠, pneumonia (Zhou et al., 2020)⁠⁠

In April 2011 there was an outbreak of hemolytic uremic syndrome (HUS). The characteristic triad of hemolytic uremic syndrome (HUS) is nonimmune hemolytic anemia, thrombocytopenia, and renal failure (Loos et al., 2017)⁠.

Some of the patients had chronic kidney disease (CKD), some still had major neurological symptoms at the latest follow-up. Proteinuria (in 15%–30%) and hypertension (in 5%–15%) are frequently detected in patients after HUS, which might indicate a higher risk of renal or cardiovascular disease later in life (Loos et al., 2017)⁠.

After 1 year, kidney function was affected with proteinuria [26.7%; 95% confidence interval (CI) 13.8-39.6], increased serum creatinine (4.4%, CI 0.0-10.4), increased cystatin C (46.7%, CI 32.1-61.3) and reduced (<90 mL/min) estimated glomerular filtration rate (46.7%, CI 32.1-61.3). Nine of the 36 patients without previous hypertension developed de novo hypertension (25%, CI 10.9-39.1) (Derad et al., 2016)⁠.

E.coli acquire specific virulence factors from bacteriophages, plasmids, and DNA horizontal transfer transposon (Kaper et al., 2004)⁠.

This new type of Escherichia coli was previously unknown. We need to collect raw data on this pathotype, compare it with its closest relative, and determine from where the new virulence factor originated.

**Methods**

1. *Exploring the dataset with FastQC*

We used [FastQC](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to learn more about our new E.coli strain reads: length, quantity, quality. We uploaded 6 files from open source: pair-end reads (2) and mate-pair reads (4).

1. *K-mer profile with Jellyfish*

We used [jellyfish](https://github.com/gmarcais/Jellyfish) to build a k-mers profile. The entered parameters are labeled as:

$ jellyfish count -m 31 - C -o output -c 3 -s 10000000 PE\_forw.fastq PE\_rev.fastq

With these settings we marked the size of the k-mers as 31 bp; the option C allows us to read forward and backward assembly in separate files; the output to the output file; -c 3 means Length of counting field in bits, we had to put a value greater than two because we don't care about unique reads; the -s parameter stands for hash size, which we set to be slightly larger than the genome of E. coli.

We then converted the resulting file using the jellyfish histo option into a table, which we used to plot the frequency of occurrence versus multiplicity of the reads.

$ jellyfish histo -o output.histo output

The plot was called by that command:

> spec1\_25 <- read.table("output.histo")

> plot(spec1\_25[5:400,],type="l")

> sum(as.numeric(spec1\_25[16:1550,1]\*spec1\_25[16:1550,2]))

> sum(as.numeric(spec1\_25[16:1550,1]\*spec1\_25[16:1550,2]))/125

[1] 5143609

singleC <- sum(as.numeric(spec1\_25[16:1550,1]\*spec1\_25[16:1550,2]))/125

> plot(1:400,dpois(1:400, 51)\*singleC, type = "l", col=3, lty=2)

> lines(spec1\_25[1:400,],type="l")

1. *Assembling E. coli X genome from paired reads with SPAdes*

We used [SPAdes](https://docs.google.com/spreadsheets/u/0/d/1deSlcGApT3rN1_3e0NWV-xwZlyMlyXD198nBHhhSROs/edit) to assemble the genome of a bacterium whose reeds we had with command:

$ python quast.py ~/dz3/spades\_results/contigs.fasta ~/dz3/spades\_results/scaffolds.fasta

and got 4 files to analyze.

1. *Impact of reads with large insert size with Quast*

[Quast](https://github.com/ablab/quast) was used to evaluate changes in the quality of the [SPAdes](https://github.com/ablab/spades) assembly by adding mate-pair libraries. Thus, we will be able to keep short and long repeats in our assembly.

1. *Genome Annotation with Prokka*

To evaluate gene functions, we annotated the bacterial genome with [Prokka](https://github.com/tseemann/prokka). It suggests where the gene is located in the genome, comparing it with the included database. From the output file, we can isolate the genes of interest and compare them with the databases already submitted.

1. *Finding the closest relative of E. coli X with Barrnap tool and BLAST*

At the first step we extract 16S rRNA sequences with the [Barrnap](https://github.com/tseemann/barrnap) tool, responsible for finding rRNA genes in genome files. We deleted 5S and 23S rRNA from the file by hand and uploaded it to NCBI [BLAST](https://blast.ncbi.nlm.nih.gov/Blast.cgi) online program - Basic Local Alignment Search Tool.

1. *Genetic cause of HUS*

Instead of using Mauve aligner, we examined the annotation results manually. Firstly, we found shiga toxin-related genes in annotated sequence: **StxA and StxB (Shiga ToxinClassification, Structure, and Function**) and examined its nearest genes.

1. *Tracing the source of toxin genes in E. coli X*

In order to identify the origin of these toxin genes, we performed annotation using [RAST](http://rast.nmpdr.org/). We uploaded the resulting .gbk file and obtained annotation with additional information about the functions of genes.

1. *Antibiotic resistance detection*

To find the genes responsible for antibiotic resistance, we used [ResFinder](https://cge.cbs.dtu.dk/services/ResFinder/). We uploaded the “scaffolds.fasta” file from the SPAdes output.

**Results**

We examined all SRR data (SRR292678, SRR292862, SRR292770). The supplement materials available in the following link:

1. SRR292678:

forward read: [SRR292678sub\_S1\_L001\_R1\_001\_fastqc.html](https://drive.google.com/file/d/1loH8CdgqMt84dVTkn3JNnzPn2vqy7fw3/view?usp=sharing) (5499346 reads)

reverse read: [SRR292678sub\_S1\_L001\_R2\_001\_fastqc.html](https://drive.google.com/file/d/1rOpEeegCd2aMWJDpo4aAkPIAEbIITxM6/view?usp=sharing) (5499346 reads)

1. SRR292862

forward read: [SRR292862\_S2\_L001\_R1\_001\_fastqc.html](https://drive.google.com/file/d/1_L44rt1g536p9c0jOWLGMQsNHdRywoZ3/view?usp=sharing) (5102041 reads)

reverse read: [SRR292862\_S2\_L001\_R2\_001\_fastqc.html](https://drive.google.com/file/d/1lC6Dsx3MSq__Tx3AV3c9jHVVy0e-QJ9Q/view?usp=sharing) (5102041 reads)

1. SRR292770

reverse read: [SRR292770\_S1\_L001\_R1\_001\_fastqc.html](https://drive.google.com/file/d/1F2YZkHsvQUgvCUwAiV5FkPUDKvurffje/view?usp=sharing) ( 5102041 reads)

reverse read: [SRR292770\_S1\_L001\_R2\_001\_fastqc.html](https://drive.google.com/file/d/1W6Fg8b5Jf4jxgWiR2mQ1byHY43fnvYtu/view?usp=sharing) ( 5102041 reads)

The data is of high quality according to FastQC. There is a normal distribution of GC content, where the central peak corresponds to the overall GC content. A low level of duplication in all samples indicates a very high level of coverage of the sequence.

For paired-end libraries (SRR292678) the data are as follows: total number of k-mer in the distribution: 610671524, genome size: 5311537, the size of single copy region: 5003577, proportion: 0.9420207. The results of running [Jellyfish](http://www.cbcb.umd.edu/software/jellyfish/) are available in the supplemental materials 1-3.

According to the [QUAST](https://cab.spbu.ru/software/quast/) report for paired-read sequence (supplement material 4), the number of contigs are 221, GC content is 50.53%, N50 and N90 are 111860 and 18506 respectively.

Obviously, the quality of the assembly while using multiple libraries with different insert sizes increased substantially. For the paired library, the number of contigs are 221, while for the multiple libraries the number of contigs are 90. The reason for that is because mate pair sequencing involves generating long-insert paired-end DNA libraries. Combining data from mate pair sequencing with those from short-insert paired-end reads provides increased information for maximizing sequencing coverage across a genome (Tritt et al., 2012). ⁠⁠

After assembling multiple libraries with different insert sizes (paired and mate-paired), the data according to [QUAST](https://cab.spbu.ru/software/quast/) is as follows: the number of contigs are 90, GC content is 50.57%, N50 and N90 are 2815616 and 180369 respectively.

According to genome annotation, the sequence has 80 tRNAs, 0 rRNAs, 1 CRISPRs, 5064 CDS, 2923 unique gene codes.

After using BLAST, run on the 16S rRNA gene matrix, we found that the closest relative for our bacteria is Escherichia coli 55989 strain, complete sequence (Accession number: NC\_011748). After performing a genome-wide comparison with the reference genome and analyzing the regions where our and reference strains differ from each other, we found a region encodes a new virulence factor or a new gene responsible for antibiotic resistance (StxA and StxB) (Melton-Celsa, 2014). ⁠

By examining the scaffolds.gbk file, we found several genes related to shiga-toxin such as stxA and stxB. Below, the genes express a hypothetical protein according to Prokka annotation.

Using [RAST](http://rast.nmpdr.org/) annotation service it became clear that genes that have above mentioned products associated with: Shiga toxin II subunit B located in NODE\_115\_length\_2156\_cov\_117.079486 and Shiga toxin II subunit A located in the same locus.

Most importantly, the genes close to above mentioned have a such products:

unknown protein encoded within prophage CP-933V, UPF0380 proteins YafZ and homologs, which Phage or Prophage Related, Phage tail, component T Phage DNA adenine methylase (EC 2.1.1.72).

According to [ResFinder](https://cge.cbs.dtu.dk/services/ResFinder/) output in the supplement materials (5-7), our strain is resistance to some antibiotics from beta-lactam class such as cefepime, ampicillin, cefotaxime, ceftazidime, piperacillin, amoxicillin, doxycycline, ceftazidime, cefotaxime, aztreonam; tetracycline, trimethoprim, sulfamethoxazole, streptomycin.

Hence, genetic plasticity of bacterial pathogens causing specific reactions, mutational adaptations, acquisition of genetic material or changes in gene expression leads to resistance to the listed antibiotics available in clinical practice.

**Discussion**

As a result of horizontal gene transfer, this E.coli strain acquired antibiotic resistance genes from phages during transduction.

Our studies showed that the reference genome belongs to the enteroaggregative E. coli strain (EAEC). Enteroaggregative strains of E. coli adhere to mucosal cells and are a new cause of gastroenteritis (Ahmed et al., 2012)⁠

Our strain of E.coli has developed resistance to antibiotics through evolution and intense selective pressure. As bacterial genomes undergo rearrangement, different regions have been rearranged.

The presence of stx phage genes indicates that the current isolates may be the result of a transfer of a new phage element into a current strain. Since shiga-related genes are surrounded by phage-related regions, it is obvious that these genes were acquired from phages.

Hence, accurate characterization of such isolates will be increasingly important. Bioinformatic approaches can be vital for the right treatment in clinical practice, particularly for choosing the class of antibiotics.

Therefore, understanding the biochemical and genetic basis of resistance is an important part of developing strategies to contain the emergence and spread of resistance and to develop innovative therapeutic approaches against multidrug-resistant organisms. The urgent need certainly includes the development of new antibiotics, but further bioinformatics research will also be needed to help contain bacterial outbreaks before they can become epidemics.

**References**

1. Ahmed, S. A., Awosika, J., Baldwin, C., Bishop-Lilly, K. A., Biswas, B., Broomall, S., Chain, P. S. G., Chertkov, O., Chokoshvili, O., Coyne, S., Davenport, K., Detter, J. C., Dorman, W., Erkkila, T. H., Folster, J. P., Frey, K. G., George, M., Gleasner, C., Henry, M., … Gibbons, H. S. (2012). Genomic Comparison of Escherichia coli O104:H4 Isolates from 2009 and 2011 Reveals Plasmid, and Prophage Heterogeneity, Including Shiga Toxin Encoding Phage stx2. *PLoS ONE*, *7*(11). https://doi.org/10.1371/journal.pone.0048228
2. Aruna, A., Mbala, P., Minikulu, L., Mukadi, D., Bulemfu, D., Edidi, F., Bulabula, J., Tshapenda, G., Nsio, J., & Kitenge, R. (2019). Ebola virus disease outbreak—democratic republic of the Congo, August 2018–November 2019. *Morbidity and Mortality Weekly Report*, *68*(50), 1162.
3. Derad, I., Obermann, B., Katalinic, A., Eisemann, N., Knobloch, J. K.-M., Sayk, F., Wellhöner, P., Lehnert, H., Solbach, W., & Süfke, S. (2016). Hypertension and mild chronic kidney disease persist following severe haemolytic uraemic syndrome caused by Shiga toxin-producing Escherichia coli O104: H4 in adults. *Nephrology Dialysis Transplantation*, *31*(1), 95–103.
4. Kaper, J. B., Nataro, J. P., & Mobley, H. L. (2004). Nature reviews. Microbiology. *Nat Rev Microbiol*, *2*(2), 123–140.
5. Loos, S., Aulbert, W., Hoppe, B., Ahlenstiel-Grunow, T., Kranz, B., Wahl, C., Staude, H., Humberg, A., Benz, K., & Krause, M. (2017). Intermediate follow-up of pediatric patients with hemolytic uremic syndrome during the 2011 outbreak caused by E. coli O104: H4. *Clinical Infectious Diseases*, *64*(12), 1637–1643.
6. Melton-Celsa, A. R. (2014). Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiology Spectrum*, *2*(4), 1–19. https://doi.org/10.1128/microbiolspec.ehec-0024-2013
7. Rust Jr, J. H., Cavanaugh, D. C., O’Shita, R., & Marshall Jr, J. D. (1971). The role of domestic animals in the epidemiology of plague. I. Experimental infection of dogs and cats. *Journal of Infectious Diseases*, *124*(5), 522–526.
8. Tritt, A., Eisen, J. A., Facciotti, M. T., & Darling, A. E. (2012). *An integrated pipeline for de novo assembly of microbial genomes*.
9. Uwishema, O., Adriano, L. F., Chalhoub, E., Onyeaka, H., Mhanna, M., David, S. C., Nasrallah, Y., Ribeiro, L. L. P. A., & Berjaoui, C. (2021). Bird flu outbreak amidst COVID‐19 pandemic in South Africa: efforts and challenges at hand. *Journal of Medical Virology*, *93*(10), 5676–5679.
10. Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., & Huang, C. L. (2020). Regulation of inflammatory cytokines and inhibition of T and B lymphocytes X. *Wang, XS Zheng, K. Zhao, QJ Chen, F. Deng, LL Liu, B. Yan, FX Zhan, YY Wang, GF Xiao, ZL Shi, A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin, Nature*, *579*, 270–273.

**Supplement materials:**

1. <https://drive.google.com/file/d/1EYehLDEAJjAvPaM6uf1naSJnZTwAIFo5/view?usp=sharing>
2. <https://drive.google.com/file/d/1BMAdVm3gZP1z4pRJVmxqmw18xb9EC3mz/view?usp=sharing>
3. <https://drive.google.com/file/d/1l7OYScm8H-7h5G2eSQReNm2eiBpFLzAH/view?usp=sharing>
4. <https://drive.google.com/file/d/14cHm-SFIXbV9uj6w2Yv64gKLblO2iDLJ/view?usp=sharing>
5. <https://drive.google.com/file/d/1YjwYpVNnhmli2Opgd8ILCsOiAPFBr_fT/view?usp=sharing>
6. <https://drive.google.com/file/d/1tykbPvfVcOkHZuPyD7U-HzEvS8JkFkGI/view?usp=sharing>
7. <https://drive.google.com/file/d/13i6PIJW9prxziApg49JvPvzUU8Gr0zl2/view?usp=sharing>