De novo genome assembly unravels the evolutionary origin of a pathogenic *E. coli* strain responsible for a hemolytic uremic syndrome outbreak

Michael Greenberbg, Polina Vaganova
Bioinformatics Institute, Saint Petersburg, Russia

Abstract. Pathogenic *E. coli* strains pose a significant global health threat, in some cases leading to large outbreaks. Horizontal gene transfer is a key factor in the emergence of such strains. This study utilizes *de novo* genome assembly to trace the origins of a pathogenic *E. coli* strain notorious for causing a massive outbreak associated with hemolytic uremic syndrome (HUS) in 2011. Our findings suggest that this strain evolved from an enteroaggregative strain, acquiring Shiga toxin genes through phage transduction and incorporating β-lactamase genes into its chromosome via mobile elements.

Keywords: *E. coli* pathogenic strains, hemolytic uremic syndrome (HUS), horizontal gene transfer, genome assembly, phage transduction, mobile elements, antibiotic resistance.

1 Introduction

Escherichia coli is a type of Gram-negative bacterium that is a common part of the natural gut flora. However specific pathogenic strains have the potential to lead to a wide range of infections. For example, Shiga toxin (*Stx*)-producing *Escherichia coli* (STEC) are notorious foodborne pathogens that cause several human diseases, such as diarrhea, hemorrhagic colitis and hemolytic uremic syndrome (HUS).[1] The main cause of this disease is the horizontal transfer of *Stx* toxin gene from one bacterium to another.[2] Phages perform one of the options for such gene transmission through insert their DNA into the host bacterium, providing the host genome with new functions and, as a result, may convert harmless strains into the virulent ones.[3]

Besides virulence factors, antibiotic resistance genes, encoded by mobile genetic elements such as plasmids can be transferred among *E. coli*, causing antibiotic resistance to different agents and lead to worse outcomes in case of virulent strain invasion.

In some cases, such mechanisms lead to global risks. A large outbreak of diarrhea and the hemolytic–uremic syndrome caused by an unusual of Shiga-toxin–producing *Escherichia coli* strain (called *E. coli* X) began in Germany in May 2011, leading to numerous numerous deaths. To investigate the evolutionary origins and pathogenic potential of the outbreak strain, researchers started its genetic characterization.[2,4]

The standard genetic analysis is the alignment of reads from sample of interest to the known reference genome, however it works bad with entirely novel strains. Another approach is the *de novo* assembly which allows the reconstruction of the complete genome sequence of the pathogenic bacteria, including novel regions which may not be present in the reference genome.[5]

This article follows the steps of scientists studying *E. coli* X.[4,6–8] It explains why they preferred de novo assembly over aligning to references and how it sheds light on broader aspects of pathogen genomics, revealing hidden details of their evolution and what makes them dangerous.

2 Materials and methods

For this study we used three Illumina sequencing reads libraries from the TY2482 sample with the following insert sizes and orientation: paired end, insert size 470 bp (forward reads, reverse reads, 400 Mb), mate pair, insert size 2 kb, (forward reads, reverse reads, 200 Mb each), mate pair, insert size 6 kb, (forward reads, reverse reads, 200 Mb each). Read quality was assessed using FastQC tool (version 0.12.1).[9]

For the reads assembly we used assembler SPAdes (version 3.15.2).[10] For the assessment of the quality of the resulting assembly we used QUAST (version 5.0.1).[11]

To perform functional gene annotation we used RAST online server.[12] To locate 16S rRNA in the assembled genome we used the rRNA genes prediction tool Barrnap (version 0.9). Then we used BLAST algorithm to search for the genome in the RefSeq database[13] with 16S rRNA that is most similar to the *E. coli* X 16S rRNA. To restrict our search to only those genomes that were present in the GenBank database at the beginning of 2011, set the time range using parameter PDAT in the "Entrez Query" field. All other parameters were specified as default.

For analysis of the genome-wide alignment of *E. coli* X and the reference genome we used Mauve software (version 2.4.0).

For antibiotic resistance genes search we used ResFinder (version 4.4.1).[14]

3 Results

In our research we used 3 libraries of froward and reverse reads of *E. coli* X strain: SRR292678 - paired-end reads (insert size 470kb) and two sets of mate pair reads: SRR292862 and SRR292770 (insert size 2kb and 6kb, respectively). Info about forward/reverse reeds from each library is given in table 1.

Table 1: *Summary of forward/reverse reads from each library*

Library	Total bases (Mbp)	Number of reads	Read length (bp)
SRR292678	494.9	5499346	90
SRR292862	250	5102041	49
SRR292770	250	5102041	49

We performed two types of genome assembly: only with pair-end reads and with all 3 libraries consolidated. Based on number of contigs and N50, L50 statistics, the assembly from consolidated libraries overall yielded fewer contigs and larger contig sizes (table 2, QUAST

reports can be found in supplementary materials), indicating its higher quality. All further steps were done with consolidated assembly.

Table 2: *Assembly statistics for scaffolds*

Assembly type	**N50**	**L50**	**Number of contigs($\geq$ 500 bp)**
Paired-end library	105346	15	504
3 libraries consolidated	1046832	2	460

Selected assembly was then annotated. Separate identification of 16S RNA sequences was carried out using Barrnap. In total there were identified 7 whole 16S RNA genes (1537 bp each) and one partial sequence (405 bp). Fasta file with these 16S genes was used for identification of closest relative strain of *E. coli* in RefSeq database up to 2011/01/01. We found that that *E. coli* 55989 (enteroaggregative strain) seem to be the closest.

The genome alignment between *E. coli* 55989 and the annotated genome of the researched strain was performed using Locally Collinear Blocks visualization to track the origin of the shiga toxin in strain X. Our analysis revealed that the shiga toxin genes, absent in the 55989 strain, were located near the CP-933V phage genes in strain X.

Our analysis of antibiotic resistance in both strains revealed that strain X exhibits resistance to β-lactam antibiotics (amoxicillin, ampicillin, cefepime, cefotaxime, ceftazidime, piperacillin, aztreonam, ticarcillin, ceftriaxone) as well as some other antibiotic types when compared to the 55989 strain (table 3).

Table 3: *Antibiotic Resistance Comparison between 55989 strain and X strain of E. coli*

Antibiotic	***E. coli* 55989 strain resistance**	***E. coli* X strain resistance**
tetracycline	+	+
doxycycline	+	+
minocycline	+	-
streptomycin	-	+
β-lactam antibiotics	-	+
sulfamethoxazole	-	+
trimethoprim	-	+

The resistance to β-lactam antibiotics in X strain was found to be due to the presence of β-lactamase class A genes (blaCTX-M-15 and blaTEM-1B). Using Locally Collinear Blocks visualization we found that these genes are associated with mobile elements proteins.

4 Discussion

Our findings suggest that *E. coli* X strain evolved from 55989 strain, that is known for it enteroaggregative capabilities, but lacks Shiga toxin and thus do not cause HUS.[15] Our data implies that *E. coli* X strain acquired Shiga toxin from CP-933V phage, which is know to

transfer these toxin genes in *E. coli*.[16] Additionally β-lactamase class A genes (blaCTX-M-15 and blaTEM-1B) were acquired, which led to increased stability against antibiotic treatment. β-lactamase is an enzyme responsible for bacterial multi-resistance against β-lactam antibiotics due to its ability to break down the β-lactam ring through hydrolysis. Bla genes are often transferred via plasmids,[17] however in our case, these genes were placed into bacterial chromosome with mobile elements. Plasmids often place a selection burden on host cells due to the negative fitness cost incurred by the replication and translation of plasmid genes.[18] Consequently, maintaining plasmids in host cells requires continuous antibiotic selective pressure. However, events such as plasmid mobilization to the chromosome eliminate the necessity for ongoing antibiotic presence to sustain them,[19] promoting more stable vertical gene transfer. This could have been one of the factors contributing to the rapid spread of *E. coli* X strain.

Possible recommendations for patients infected with the X strain may initially include necessary physiological therapy. Given the X strain's resistance to several antibiotic classes, exploring the use of novel types of antibiotics may be a potential treatment approach. Some antibiotics like polymyxins or tigecycline are already used.[20]

References

1 J. M. Hunt, "Shiga toxin–producing escherichia coli (stec)," *Clinics in Laboratory Medicine* **30**, 21–45 (2010).

2 D. A. Rasko, D. R. Webster, J. W. Sahl, *et al.*, "Origins of thee. colistrain causing an outbreak of hemolytic–uremic syndrome in germany," *New England Journal of Medicine* **365**, 709–717 (2011).

3 Y. Deng, H. Xu, Y. Su, *et al.*, "Horizontal gene transfer contributes to virulence and antibiotic resistance of vibrio harveyi 345 based on complete genome sequence analysis," *BMC Genomics* **20** (2019).

4 M.-K. Cheung, L. Li, W. Nong, *et al.*, "2011 german escherichia coli o104:h4 outbreak: whole-genome phylogeny without alignment," *BMC Research Notes* **4** (2011).

5 X. Liao, M. Li, Y. Zou, *et al.*, "Current challenges and solutions of de novo assembly," *Quantitative Biology* **7**, 90–109 (2019).

6 D. Li, F. Xi, M. Zhao, *et al.*, "Genomic data from escherichia coli o104:h4 isolate ty-2482," (2011).

7 A. Mellmann, D. Harmsen, C. A. Cummings, *et al.*, "Prospective genomic characterization of the german enterohemorrhagic escherichia coli o104:h4 outbreak by rapid next generation sequencing technology," *PLoS ONE* **6**, e22751 (2011).

8 H. Rohde, J. Qin, Y. Cui, *et al.*, "Open-source genomic analysis of shiga-toxin–producinge. colio104:h4," *New England Journal of Medicine* **365**, 718–724 (2011).

9 S. Andrews, "Fastqc. a quality control tool for high throughput sequence data," (2010).

10 A. Bankevich, S. Nurk, D. Antipov, *et al.*, "Spades: A new genome assembly algorithm and its applications to single-cell sequencing," *Journal of Computational Biology* **19**(5), 455–477 (2012). PMID: 22506599.

11 A. Gurevich, V. Saveliev, N. Vyahhi, *et al.*, "Quast: quality assessment tool for genome assemblies," *Bioinformatics* **29**, 1072–1075 (2013).

12 R. K. Aziz, D. Bartels, A. A. Best, *et al.*, "The rast server: Rapid annotations using subsystems technology," *BMC Genomics* **9** (2008).

13 N. A. O'Leary, M. W. Wright, J. R. Brister, *et al.*, "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation," *Nucleic Acids Research* **44**, D733–D745 (2015).

14 A. F. Florensa, R. S. Kaas, P. T. L. C. Clausen, *et al.*, "Resfinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes," *Microbial Genomics* **8** (2022).

15 M. Touchon, C. Hoede, O. Tenaillon, *et al.*, "Organised genome dynamics in the escherichia coli species results in highly diverse adaptive paths," *PLoS Genetics* **5**, e1000344 (2009).

16 L. D. Teel, A. R. Melton-Celsa, C. K. Schmitt, *et al.*, "One of two copies of the gene for the activatable shiga toxin type 2d in escherichia coli o91:h21 strain b2f1 is associated with an inducible bacteriophage," *Infection and Immunity* **70**, 4282–4291 (2002).

17 C. Branger, A. Ledda, T. Billard-Pomares, *et al.*, "Extended-spectrum -lactamase-encoding genes are spreading on a wide range of escherichia coli plasmids existing prior to the use of third-generation cephalosporins," *Microbial Genomics* **4** (2018).

18 A. San Millan and R. C. MacLean, "Fitness costs of plasmids: a limit to plasmid transmission," *Microbiology Spectrum* **5** (2017).

19 N. Hülter, J. Ilhan, T. Wein, *et al.*, "An evolutionary perspective on plasmid lifestyle modes," *Current Opinion in Microbiology* **38**, 74–80 (2017).

20 M. M. Walker, J. A. Roberts, B. A. Rogers, *et al.*, "Current and emerging treatment options for multidrug resistant escherichia coli urosepsis: A review," *Antibiotics* **11**, 1821 (2022).

5 Supplementary materials

Data availability

1. SRR292678 - paired end library: forward reverse
2. SRR292862 – mate pair(2kb) library: forward reverse
3. SRR292770 – mate pair (6kb) library: forward reverse

FastQC and QUAST results: google disc