E.coli 2011 outbreak investigation

Irina Bogdanova, Valeriya Rubinova
Bioinformatics Institute

Abstract

In 2011, an outbreak of the disease caused by E. Coli occurred in Germany, accompanied by hemolytic uremic syndrome (HUS). It was not clear which bacterial strain caused this disease, so in this work we investigated it: the precursor strain, distinctive features (genes and phenotypic traits). We worked with Illumina reads from the TY2482 sample, which were generated at Beijing Genome Institute. De novo assembly of the genome of the bacterium was carried out, a bacterial strain was found that was as similar as possible to the studied strain *E.Coli X*. Further, the genomes of these two strains were compared, and it was found that genes of Shiga-toxins (*stxA, stxB*), which are the cause of the disease, were obtained by bacteriophage-mediated horizontal gene transfer. Also, the new strain was distinguished by an antibiotic resistance to the following groups of antibiotics: beta-lactam, aminoglycoside, tetracycline.

Introduction

Shiga Toxin-producing Escherichia coli (E. coli) is frequently found in the intestines of humans and warm-blooded animals. Most strains of E. coli are harmless. However, some strains, such as enterohemorrhagic E. coli (STEC), can cause severe foodborne illness. This bacterium is transmitted to humans primarily through the consumption of contaminated foods such as raw or undercooked ground meats, raw milk, and contaminated raw vegetables and sprouts [1].

STEC produces toxins known as Shiga toxins, so named because of their similarity to the toxins produced by Shigella dysenteriae. The most important public health STEC serotype is E. coli O157:H7; however, sporadic cases and outbreaks are often caused by other serotypes [1, 2].

One of the most common mechanisms for the formation of new, including pathogenic, bacterial strains is horizontal gene transfer (HGT) - the sharing of DNA that can spread neutral or beneficial genes, pathogenic factors [3]. This horizontal gene transfer is particularly prevalent in prokaryotes, where it is one of the main mechanisms contributing to

genetic variation and thus evolution. HGT can occur by transformation, transduction, and conjugation, as well as other mechanisms [3].

Although the mutated strain is very similar to the strain that is its predecessor, it is desirable to use de novo assembly of its genome in some cases. It is a good practice to assemble a genome de novo first, then find the most similar bacterial strain and compare them, rather than aligning the obtained reads to the reference. By finding mismatches and examining the genes around this mismatch, we can understand the nature of the occurrence of a given strain, where its special factors come from [4].

Methods

Following the footsteps of the explorers who determined the origin and the source of pathogenicity of the 2011 outbreak strain, we carried out the genome assembling of the *E. coli* strain of interest using three libraries from the TY2482 sample that was isolated from the girl in Hamburg. Illumina reads (Illumina HiSeq 2000), generated at Beijing Genome Institute, were obtained from the Short Read Archive[1]. Libraries characteristics are shown in Table 1.

Table 1.Illumina reads characteristics

Name	**SRA accession number**	**Library type**	**Insert size**
SRR292678	SRX079804	paired end	470 bp
SRR292862	SRX079806	mate pair	2 kb
SRR292770	SRX079805	mate pair	6 kb

The raw data quality was assessed using the FastQC tool with default parameters [5]. K-mer profile and genome size estimation was carried out using Jellyfish count [6] with the following parameters: -m 31 -C -s 5000000 -F 2.

The *E. coli* X strain genome was then assembled using the SPAdes assembler [7, 8]. Quality of the resulting assembly was assessed with QUAST [9].

[1] https://www.ncbi.nlm.nih.gov/sra

Then, the genome was annotated using prokka [10] with following parameters: --cpu 2 --force --compliant --centre XXX.

To find the closest relative of our *E. coli* X strain, the rRNA genes prediction tool, barrnap, was implemented with the '-k bac' parameter [11]. 16S rRNA sequence found was then blasted using NCBI BLAST[2] with Entrez query 1900/01/01:2011/01/01[PDAT] parameter. The Fasta file of the closest relative defined (Escherichia coli 55989; NCBI Reference Sequence: NC_011748.1) was then downloaded for subsequent search of pathogenicity source *E. coli* X. To achieve this aim, the Mauve tool was implemented [12].

For antibiotic resistance search ResFinder was used[3].

Results

The quality assessment was performed using the FastQC tool before the K-mer profile and genome size estimation. The unusual results in the "Kmer Content" quality metric were obtained for both forward and reverse reads of SRR292862 library [Supplementary materials, Figure 1]. This module issues a warning if any k-mer is imbalanced with a binomial p-value < 10^{-5}. In order to save more reads for later genome assembly, no trim step was carried out.

For K-mer profile and genome size estimation via Jellyfish only the paired-end library (SRR292678) was used. A hisogram file, containing the list of k-mer 'depth' and the count for the number of k-mers in the data that fit into each category, was then visualized. K-mer plotting step was repeated for corrected reads after genome assembly with SPAdes as this tool works in two-step mode – error correction and assembly [Supplementary Materials, Figure 2]. As can be seen from the Figure we get a clearer peak of kmers.

Genome size was estimated with the following formulas:

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

$$\text{Genome Size} = T/N$$

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

The calculated genome size was 5279372.

After genome assembling via SPAdes with one paired-end and all three libraries used, the quality assessment of resulting assemblies was carried out with QUAST. The results were then compared (Table 2).

[2] http://blast.ncbi.nlm.nih.gov/
[3] https://cge.food.dtu.dk/services/ResFinder/

Table 2. Part of QUAST reports for one paired-end (1) and three (2) libraries assemblies. All statistics are based on contigs of size ≥ 500 bp

Assembly	**contigs (1)**	**scaffolds (1)**	**contigs (2)**	**scaffolds (2)**
contigs	210	221	105	90
Largest contig	300763	300763	698474	2815616
Total length	5295721	5304595	5350156	5391554
GC (%)	50.56	50.53	50.59	50.57
N50	111860	111860	335515	2815616
N90	18506	18506	79998	180369
auN	131921.2	131704.5	319603.4	1633387.0
L50	14	14	6	1
L90	53	53	20	7
N's per 100 kbp	0.00	33.74	0.00	627.52

As can be seen from the table, the impact of reads with large insert sizes is significant. It is well known that using multiple libraries with different insert sizes may improve an assembly, since the library with a small insert size can resolve short repeats, while the library with a larger insert size can resolve longer repeats. The improvement in quality when using different libraries that we see in results obtained may indicate the presence of long repeats that are resolved with libraries with a large insertion size.

After genome annotation via Prokka tool, the main task was to find the most similar pathogenic strain of E. coli. The 16S rRNA coding gene was chosen as a conservative gene, based on the sequence of which the search for the nearest strain was carried out. Following this step, a search for the nearest strain using NCBI BLAST allowed us to identify the closest relative of the E. coli X that is Escherichia coli 55989 (NCBI Reference Sequence: NC_011748.1).

To find the genetic cause of HUS and its source, the Mauve tool was used. It was found that there are Shiga toxin-related genes in E. coli X strain, coding stxB (genome position: 3483605-3483874) and stxA (genome position: 34838886-3484845) Shiga toxin subunits, which ended up in the genome of our X strain due to the bacteriophage, the genes of which surround the region with the Shiga toxin coding genes found.

The search for genes responsible for antibiotic resistance with ResFinder allowed us to identify a number of antibiotics to which the X strain may be resistant. The results are presented in Table 3

Table 3. ResFinder partial results

Resistance gene	**Identity**	**Coverage**	**Phenotype**	**Accession no.**
blaTEM-1B	100.00	100.0	Beta-lactam resistance Alternate name; RblaTEM-1	AY458016
blaCTX-M-15	100.00	100.0	Beta-lactam resistance Alternate name; UOE-1	AY044436
tet(A)	100.00	100.0	Tetracycline resistance	AJ517790
qacE	100.00	84.6846846846846	Disinfectant resistance	X68232
dfrA7	100.00	100.0	Trimethoprim resistance	AB161450
sul1	100.00	91.9082125603864	Sulphonamide resistance	AY522923
sul1	100.00	90.5952380952381	Sulphonamide resistance	U12338
sul1	100.00	86.2811791383219	Sulphonamide resistance	DQ914960
sul2	100.00	100.0	Sulphonamide resistance	HQ840942
sul1	100.00	90.5952380952381	Sulphonamide resistance	AY115475
aph(6)-Id	100.00	100.0	Aminoglycoside resistance Alternate name; aph(6)-Id	M28829
aph(3")-Ib	100.00	100.0	Aminoglycoside resistance Alternate name; aph(3")-Ib	AF321551

Discussion

Based on the described results, it can be concluded that the genes obtained by the bacterium using horizontal gene transfer (stxA, stxB) are associated with Shiga-toxin. Thus, the *E.Coli X* strain became pathogenic, the produced shiga-toxins of which caused hemolytic-uremic syndrome in patients.

Antibiotic resistance was also analyzed in the test sample and in the reference. Study of the test sample showed prevalence of resistance against different types of antibiotics including beta-lactam (amoxicillin, ampicillin, aztreonam, cefepime, cefotaxime, ceftazidime, ceftriaxone, piperacillin, ticarcillin), aminoglycoside (streptomycin), tetracycline (doxycycline, tetracycline) [13]. The reference sample showed sensitivity to the antibiotic tetracycline (doxycycline, tetracycline, minocycline) [14]. Therefore, we can conclude that the E.Coli X strain has acquired resistance to several groups of antibiotics at once.

Antibiotic resistance occurs in one of several ways: natural resistance in certain types of bacteria through random mutations and/or exposure to the antibiotic; genetic mutation; or by acquiring resistance by some bacterial species from others (through horizontal gene transfer). In addition, antibiotic resistance of microorganisms can be created artificially by genetic transformation. For example, the introduction of artificial genes into the genome of a microorganism [15]. The strain acquired antibiotic resistance by bacteriophage-mediated horizontal gene transfer.

Therefore it is necessary to consider alternative treatment for the affected patients that includes supportive care (BSC), with particular attention to fluid replacement, renal support, and treatment of neurological manifestations of the disease [16]. Several other treatments for STEC HUS, including oral administration of Shiga toxin-binding agents, corticosteroids, antiplatelet agents, and heparin, have been suggested but have not proven beneficial [17].

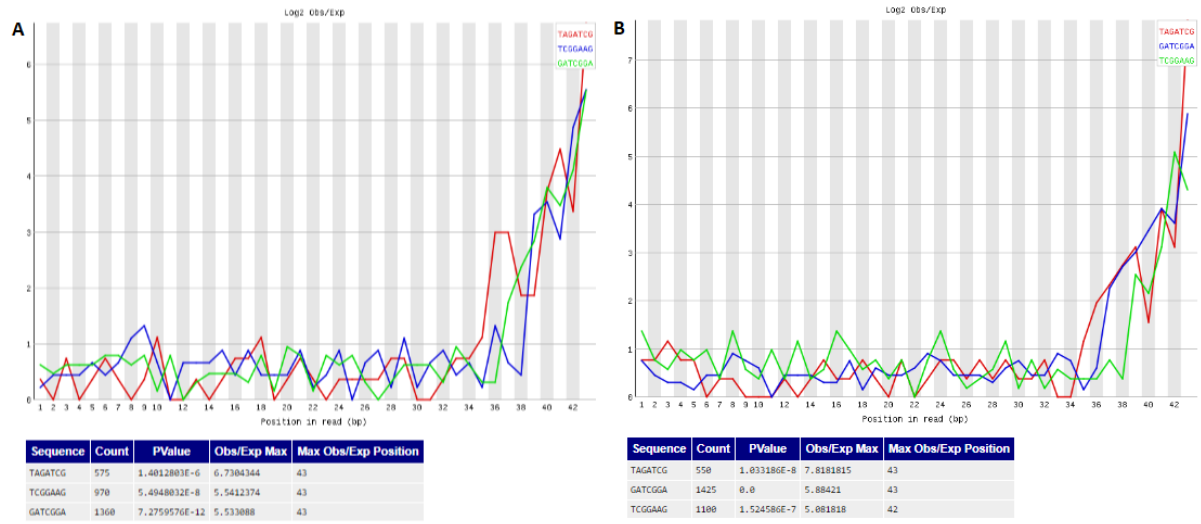
Citations

1. Smith J. L., Fratamico P. M., Gunther IV N. W. Shiga toxin-producing Escherichia coli //Advances in applied microbiology. – 2014. – T. 86. – C. 145-197.
2. Tarr P. I., Gordon C. A., Chandler W. L. Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome //The lancet. – 2005. – T. 365. – №. 9464. – C. 1073-1086.
3. Hall RJ, Whelan FJ, McInerney JO, Ou Y, Domingo-Sananes MR. Horizontal Gene Transfer as a Source of Conflict and

Cooperation in Prokaryotes. Front Microbiol. 2020 Jul 17;11:1569. doi: 10.3389/fmicb.2020.01569. PMID: 32849327; PMCID: PMC7396663.

4. Lischer, H.E.L., Shimizu, K.K. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 18, 474 (2017). https://doi.org/10.1186/s12859-017-1911-6
5. Andrews, S. (2010). *FastQC* (v0.11.9) [Software]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
6. Marcais, G., & Kingsford, C. (2011). *Jellyfish* (2.3.0) [Software]. https://github.com/gmarcais/Jellyfish
7. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455-77. doi: 10.1089/cmb.2012.0021. Epub 2012 Apr 16. PMID: 22506599; PMCID: PMC3342519.
8. *SPAdes* (SPAdes 3.15.5). (2011). [Software]. https://cab.spbu.ru/software/spades/
9. Gurevich, A., Saveliev, V., & Vyahhi, N. (2013). *QUAST* (5.2.0) [Software]. https://cab.spbu.ru/software/quast/
10. Seemann, T. (2011). *Prokka: rapid prokaryotic genome annotation* (1.14.6) [Software]. https://github.com/tseemann/prokka
11. Seemann, T. (n.d.). *Barrnap* (0.9) [Software]. https://github.com/tseemann/barrnap/
12. Darling, A. (n.d.). *Mauve* (Version 20150226) [Software]. https://darlinglab.org/mauve/download.html
13. https://cge.food.dtu.dk//cgi-bin/webface.fcgi?jobid=6389F15B000072544260FE66
14. https://cge.food.dtu.dk//cgi-bin/webface.fcgi?jobid=6389F90300002222F51269F5
15. Cesur S., Demiroz A.P.Antibiotics and the Mechanisms of Resistance to Antibiotics // Medical Journal of Islamic World Academy of Sciences. 2013. Vol. 21, 4.
16. Kavanagh D, Raman S, Sheerin NS. Management of hemolytic uremic syndrome. F1000Prime Rep. 2014 Dec 1;6:119. doi: 10.12703/P6-119. PMID: 25580273; PMCID: PMC4278190.
17. Michael M, Elliott EJ, Ridley GF, Hodson EM, Craig JC. Interventions for haemolytic uraemic syndrome and thrombotic thrombocytopenic purpura. Cochrane Database Syst Rev. 2009:CD003595. doi: 10.1002/14651858.CD003595.pub2.

Supplementary materials



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TAGATCG	575	1.4012803E-6	6.7304344	43
TCGGAAG	970	5.4948032E-8	5.5412374	43
GATCGGA	1360	7.2759576E-12	5.533088	43

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TAGATCG	550	1.033186E-8	7.8181815	43
GATCGGA	1425	0.0	5.88421	43
TCGGAAG	1100	1.524586E-7	5.081818	42

Figure 1. "K-mer Content" FastQC result for SRR292862. A - forward reads, B - reverse reads.

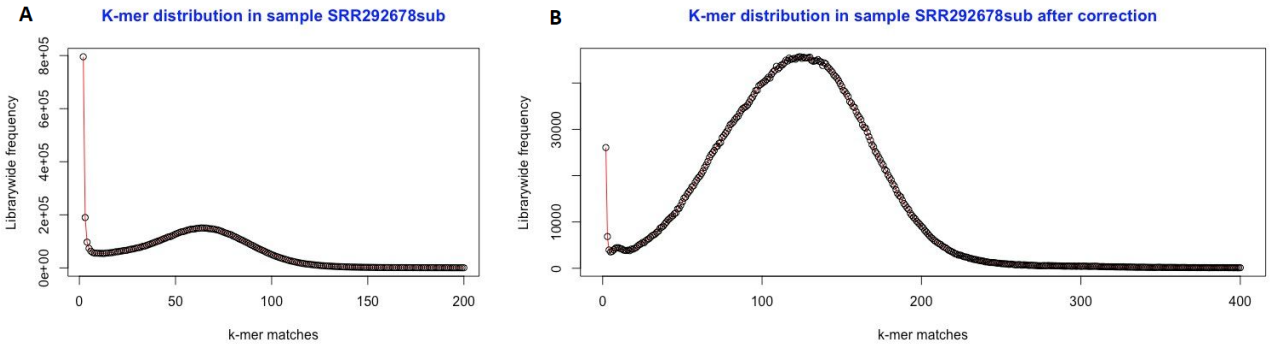


Figure 2. K-mer distribution in SRR292678. A - before SPAdes correction, B - after SPAdes correction.