E.coli outbreak investigation

Evdokimova Anastasiia, Bioinformatics Institute

Abstract

This study focuses on the mysterious strain that we will call *E. coli* X, that has caused hundreds of cases hemolytic uremic syndrome (HUS), a deadly blood disease that often starts as food poisoning with bloody diarrhea and can lead to kidney failure. Using bioinformatics methods, we tried to find out what is the cause of the high pathogenicity of this strain and to give recommendations for the treatment of patients.

Introduction

E. coli is a well-known bacterium, and has hundreds of strains, most of which are harmless. These strains are part of the normal intestinal flora of humans and animals. E. coli benefits the host, for example by synthesizing vitamin K and by preventing the development of pathogenic microorganisms in the intestines. But there are also pathogenic strains that can cause various diseases: gastroenteritis, inflammation of the genitourinary system, and meningitis in newborns. In rare cases, virulent strains also cause hemolytic-uremic syndrome, peritonitis, mastitis, sepsis, and gram-negative pneumonia.

Pathogenicity is determined by virulence factors, various compounds that it produces to facilitate colonizing the host and evading or inhibiting the host's immune system. Bacteria can acquire pathogenic factors (genes encoding toxins and antibiotic resistance genes) as a result of horizontal gene transfer, or they can get them along with the prophage.

In this work, we will try to deal with the massive food poisoning that swept Europe in 2011. They are thought to be caused by a new and particularly aggressive strain of E. coli. To do this, we will assemble a genome of new strain from libraries, obtained by sequencing DNA isolated from an isolate from a girl with food poisoning. So we can understand what genes led to the pathogenicity of this strain.

Methods

For this research, we provide three libraries from the TY2482 sample (sequencing data of the isolate from the girl in Hamburg) with the following insert sizes and orientation:

- SRR292678 - paired end, insert size 470 bp (forward reads, reverse reads)
- SRR292862 – mate pair, insert size 2 kb, (forward reads, reverse reads)
- SRR292770 – mate pair, insert size 6 kb, (forward reads, reverse reads)

We used different libraries, because there is an advantage in using multiple libraries with different insert sizes, since the library with a small insert size can resolve short repeats, whereas the library with a larger insert size can resolve longer repeats. In

particular, we are interested in how the quality of our assembly changes when we add mate-pair libraries.

The quality of raw reads was assessed using the program FastQC (version 0.11.9) [2]. For assembling E.coli X genome we used assembler SPAdes (version 3.14.1) [3]. Firstly, we tried to assemble a single library of sequencing (paired end) reads from *E. coli* X: run SPAdes in the paired-end mode, providing paired reads of *E. coli* X. from the library SRR292678 (forward and reverse). For the checking of the quality of the resulting assembly we used the console version of QUAST (version 5.2.0) [4].
To compare the quality of single-library and three-library assemblies, we repeated the last two steps by consolidating three libraries. We run SPAdes providing all three libraries: SRR292678 as a paired ends, SRR292862 and SRR292770 as a mate pairs and run QUAST again.
Next, we want to find the closest relative of the sequenced strain. We will use the "comparative genomics" approach to gene annotation, which is based on the (not always ideal) assumption that similar genes in different organisms perform similar functions.
For gene prediction and annotation we used Prokka (version 1.12) [5].
To find the known genome that is the most similar to the pathogenic strain we could compare each contig in our assembly against the entire RefSeq database using BLAST, but this could take several hours depending on server workload. A more efficient approach is to select one important and evolutionarily conserved gene for comparison with all other sequenced genomes. The gene that we used is 16S ribosomal RNA.
To find the 16s rRNA we used rRNA genes prediction tool Barrnap [6]. We got several matches here, because rRNA genes in bacteria are typically organized in ribosomal operons, and bacteria often possess several copies of this operon.
To search for the genome in t he RefSeq database with 16S rRNA that is most similar to the 16S rRNA that we found we used BLAST [7] (with parameter PDAT 1900/01/01:2011/01/01[PDAT] to restrict our search to only those genomes that were present in the GenBank database at the beginning of 2011). The comparison tells us that the sample under study is indeed a distinct strain. Further we will use the closest strain as reference.
To find out which genes can cause HUS we compare the *E. coli* X with the reference genome we used Mauve (version 2.4.0) [8].
To search for genes responsible for antibiotic resistance of new strain, we used ResFinder [9]. The result of quality control of raw reads from test sample data and control data you can see in Supplementary Information (SI)

Results

As we mentioned before, we provide three libraries from the TY2482 sample (sequencing data of the isolate from the girl in Hamburg) with the following insert sizes and orientation: SRR292678 (paired end, insert size 470 bp), SRR292862 (mate pair, insert size 2 kb) and SRR292770 (mate pair, insert size 6 kb). The result of quality control of raw reads

from test sample data and control data you can see in Supplementary Information (SI). We have no problem with reads.
After that we run assembling with SPAdes twice: in paired-end mode, providing paired reads of *E. coli* X. from the library SRR292678, and providing all three libraries: SRR292678 as a paired ends, SRR292862 and SRR292770 as a mate pairs. Results of assessing the quality of the resulting assembly you can see in Table 1 (for more information go to SI).

	Total length	N50
one library (contigs)	5295721	111860
one library (scaffolds)	5304595	111860
all libraries (contigs)	5403327	335515
all libraries (scaffolds)	5391554	2815616

Table 1. N50 and total length of assembly for two assembly: based on one library (SRR292678 - paired end, insert size 470 bp) and three libraries (SRR292862 - mate pair, insert size 2 kb, and SRR292770 - mate pair, insert size 6 kb). All statistics are based on contigs of size >= 500 bp.

We see that the quality of our assembly increases significantly when we are using three libraries. Therefore, for further analysis, we will use this particular assembly.

The summary of assembly and annotation (Prokka) you can see in Table 2.

assembly summary		annotation summary	
Assembly length	5390599	tRNAs	80
number of contigs	247	rRNAs	0
contig N50	335515	CRISPRs	1
		CDS	5064
		Unique gene codes	2923

Table 2. Assembly and annotation summary for assembly based on three libraries with different insert size. Data from Prokka output.

Using Barrnap we found seven 16S rRNA genes. Six of them have length = 1527 b.p. and one - 405 b.p. (this one has note=aligned only 25 percent of the 16S ribosomal RNA).
Next we used BLAST to search for the genome in the RefSeq database with 16S rRNA that is most similar to the 16S rRNA that we just found. We find out that the closest relative of our strain *E. coli* X is *E.coli* 55989.

Using Mauve, we aligned *E. coli* X on *E.coli* 55989 and found shiga toxin-related genes. These results you can see in Table 3.

GENE	START	STOP	LENGTH	GENE	START	STOP	LENGTH
vapB_1	193926	194192	266	***higA_2***	3324586	3325005	419
pasI	350604	350930	326	***apxIB***	3386125	3388227	2102
ratA	350920	351396	476	***stxB***	3483605	3483874	269
prlF	451073	451408	335	***stxA***	3483886	3484845	959
yhaV	451408	451872	464	***yoeB***	3880747	3881001	254
tabA_1	538027	538491	464	***yefM***	3880998	3881249	251
higB-2_1	688247	688564	317	***parD1***	3956315	3956566	251
tabA_2	931670	932137	467	***parE1***	3956568	3956864	296
dinj	965777	966037	260	***pezT***	4373822	4374748	926
hicB	1323001	1323438	437	***tabA_3***	4841382	4841834	452
cbtA	1732584	1732958	374	***chpS***	4869476	4869727	251
doc	2298176	2298661	485	***higA-2_2***	5081129	5081419	290
higA-2_1	2298712	2299029	317	***higB-2_2***	5081420	5081731	311
hipB	2398235	2398501	266	***tcpE***	5174508	5175593	1085
higA_1	2820799	2821215	416	***tcpT***	5175606	5177159	1553
cptB	3114437	3114703	266	***ccdB***	5218091	5218396	305
cptA	3114684	3115091	407	***ccdA***	5218398	5218616	218
mazE	3250811	3251059	248	***vapB_2***	5219211	5219441	230

Table 3. Shiga toxin-related genes found in *E. coli* X by Mauve.

To find the optimal treatment, it is necessary to check the antibiotic resistance of the strain *E. coli* X and compare the results with the reference strain *E.coli* 55989. The strain *E.coli* 55989 is resistant only to tetracycline, while while the studied strain is resistant to cefepime, ceftazidime, cefotaxime, ampicillin (class beta-lactam), sulfamethoxazole, trimethoprim (class folate pathway antagonist) and tetracycline (more information in SI).

Discussion

We found out that massive food poisoning was caused by a new strain of *E. coli*, rRNA analysis courts, a close relative of the strain *E.coli* 55989, previously found in the Central African Republic and causing diarrhea. Its extremely high pathogenicity is explained by the fact that its genome contains about 40 genes encoding the so-called shiga toxin. As is known, these toxins cause bleeding by breaking down the lining of the colon, and they can lead to HUS if they reach the kidneys, which corresponds to the observed symptoms. Shiga toxins attack highly specific receptors on the surface of human cells, and so species that do not have this receptor, such as cows, may harbor toxigenic bacteria without any ill effects.

Some genes nearly shige toxin genes (like these genes themselves) are characteristic of other bacteria. So, we believe that this could have happened as a result of horizontal gene transfer (HGT). There are several mechanisms for horizontal gene transfer: transformation (introduction of foreign genetic material, rather possible in laboratories), transduction (the process in which bacterial DNA is moved from one bacterium to another by a bacteriophage), bacterial conjugation (a process that involves the transfer of DNA via a plasmid from a donor cell to a recombinant recipient cell during cell-to-cell contact).

The situation is complicated by the fact that the observed strain has resistance to a number of antibiotics (unlike its close relative), which significantly complicates treatment. A bacterium can acquire resistance to new antibiotics as a result of its own mutation or HGT. Considering the above about the genes of shiga toxins, in our case the second option is most likely.

How to treat a person infected with this strain? You need to be very careful when using antibiotics: they can worsen an *E. coli* infection. When the bacteria die, they release the toxin in massive amounts. But it is still possible, using highly effective antibiotics, for example carbapenems [10]. These drugs do not trigger a major toxin release.

It is much easier to prevent the spread of *E. coli*. The methods are simple and very effective: hand-washing with soap, washing and hygienically preparing food, and properly heating/cooking food, so the bacteria are destroyed.

Citations

1. Data:
   SRR292678 - paired end, insert size 470 bp (forward reads, reverse reads)

   SRR292862 – mate pair, insert size 2 kb, (forward reads, reverse reads)

   SRR292770 – mate pair, insert size 6 kb, (forward reads, reverse reads)

2. FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
3. SPAdes: https://cab.spbu.ru/software/spades/
4. QUAST: http://cab.cc.spbu.ru/quast/

5. Prokka: http://www.vicbioinformatics.com/software.prokka.shtml
6. Barrnap: http://www.vicbioinformatics.com/software.barrnap.shtml
7. BLAST: http://blast.ncbi.nlm.nih.gov/
8. Mauve: http://darlinglab.org/mauve/download.html
9. ResFinder: https://cge.cbs.dtu.dk/services/ResFinder/
10. Papp-Wallace KM, Endimiani A, Taracila MA, Bonomo RA (2011). "Carbapenems: past, present, and future". *Antimicrob. Agents Chemother.* **55**(11): 4943–60. doi:10.1128/AAC.00296-11

Supplementary Information

FastQC reports:
https://drive.google.com/file/d/1XLxHLo7vTOKdF7rOYjAMYMnElb1Hr_bw/view?usp=sharing

QUAST reports:
https://drive.google.com/file/d/13wA4uhnT1AIugDT7wGjLKLDmm3GVsd52/view?usp=sharing

Antibiotic resistance detection:
https://cge.cbs.dtu.dk//cgi-bin/webface.fcgi?jobid=5FC89AD3000030BA43F5ECA9
(*E.coli* 55989)
https://cge.cbs.dtu.dk//cgi-bin/webface.fcgi?jobid=5FC899F300002F3844903330
(*E.coli* X)