E.coli outbreak investigation

Dudkovskaya A.

Abstract

We investigate a reason of Europe outbreak in 2011 of hemolytic uremic syndrome (HUS), a disease that often starts as food poisoning with bloody diarrhea and can lead to kidney failure. We know that the cause was the new strain of *E. coli*, so we assemble the genome of unknown strain *de novo* to find why people had HUS. We annotated the assembly, identified the closest relative of unknown E. coli strain and found that investigated strain has *stxA* and *stxB* genes. These genes produce Shiga-toxin, that could be the cause of HUS. Also we found out that this strain is resistant to many clases of antibiotics, including beta-lactam, tetracycline, folate pathway antagonist, aminoglycoside, quaternary ammonium compound, so it can be the cause of outbreak.

1 Introduction

E. coli is a bacterium commonly found in the lower intestines of warm-blooded animals. Most strains of *Escherichia coli* are harmless to humans, but some can cause severe food poisoning, in special cases they cause severe illness - hemolytic uremic syndrome - acute renal failure, hemolytic anemia and thrombocytopenia.

One of the purpose of hemolytic uremic syndrome caused by strains of *E. coli* is Shiga toxin, a toxin whose gene is encoded in the DNA of some *E. coli* strains. All of the Shiga toxins have a single 30 kDa A-subunit and a pentamer of non-covalently attached identical 7 kDa B-subunits [2]. The enzymatic activity is in the A subunit, and the binding properties of the cell recognition receptor are in the B subunit. The Stx receptor, Gb3, has been detected in homogenates of total kidney from human and animals [3]. Binding to Gb3 on eukaryotic target cells, the Shiga toxins receptor complex is located inside endosomes. In the endoplasmic reticulum the Shiga toxin A subunit is cleaved by a furin-like protease, releasing the enzymatically active A1 subunit into the cytoplasm, where it acts on ribosomes. Shiga toxins enzymatically inactivate eukaryotic ribosomes by removing one adenine base from 28S rRNA in the large (60S) ribosomal subunit, so the cell cannot do the translation [4]. So due to the location of the Shiga receptor, the toxin causes acute renal failure, hemolytic anemia and thrombocytopenia.

Shiga toxin converting phages can lyse their hosts transfer shiga toxin genes from cell to cell via transduction, representing a significant mechanism of horizontal gene transfer, and is most likely the cause of shiga toxin genes dissemination among a multitude of Shiga toxin-producing *E. coli* serotypes. [5]

In 2011 thousands of people were infected by unknown strain of *E. coli* that was the cause of hemolytic uremic syndrome [1]. To understand what was the cause of that outbreak we use bioinformatic tools. Due to we don't know how that strain leads to the disease, we assemble the whole genome to find the genes that are the reason. We do *de novo* assembling because if another genome doesn't have the searched genes, we can't find them if we do the alignment.

2 Methods

Data acquisition

We have three libraries from the TY2482 sample: SRR292678 [6] - paired end, SRR292862 [7] – mate pair, SRR292770 [8] – mate pair.

All raw data quality was checked with `fastqc` tool [9].

K-mer profile and genome size estimation

We use the `jellyfish`[10] tool to learn about k-mer profile. We use it on the first paired-ended library and also use '-C' and '-s 10000000' options and 31 as size of k-mer.

To find an approximate genome size we use a formula: genome size = $\frac{T \cdot (L-K+1)}{(M \cdot L)}$, where M is K-mer peak, K is K-mer-size, L is average read length and T is count of total bases.

Assembling *E. coli* genomes

For assembling *E. coli* genome we use `SPAdes` [11]. For assess the quality of the resulting assembly we use `QUAST` [12]. First time we use only paired-end library, in the second time we use all three libraries.

Genome Annotation

For genome annotation we use `Prokka` [13].

Finding the closest relative of *E. coli*

We use `BLAST` [14] to find the closest relative to oue strain of *E. coli*. 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 use is 16S ribosomal RNA. We find the location of this gene using `Barrnap` [15].

Then we use the output fasta file in `BLAST`. We select "Nucleotide blast", "Reference Genome Database" in the "Database" field, and *E. coli* in the "Organism" field. To restrict our search to only those genomes that were present in the GenBank database at the beginning of 2011, we set the time range using parameter PDAT in the "Entrez Query" field: `1900/01/01:2011/01/01[PDAT]`. We find the closest genome and then use it like a reference genome.

Finding the genetic cause of HUS and antibiotic resistance.

We use a program called `Mauve`, [16], which visualizes an alignment as a series of conserved segments called Locally Collinear Blocks (LCBs).

For finding the antibiotic resistance cause, we use `ResFinder` [17]. We use our scaffold `SPAdes` output to find genes implicated in antibiotic resistance.

When we got the list of resistant antibiotics, by using `Muave` we looked for genes that can provide the resistance and the neighbourhood for understanding the origin of these genes.

3 Results

Readings general information

Number of reads, average read length and count of total bases are in the Table 1 below.

K-mer profile and genome size estimation

The visualization of k-mer distribution is belong. There is a peak at 46189 position at point 125.

So we can count the approximate genome size: $\frac{2 \cdot 494941140 \cdot (90-31+1)}{(125 \cdot 90)}$ 5279372

Table 1: Number of reads and average read lengths in target sample and control samples.

File type	Number of reads	Average read length, bp	Total bases
SRR292678 forw	5499346	90	494941140
SRR292678 rev	5499346	90	494941140
SRR292862 forw	5102041	49	250000009
SRR292862 rev	5102041	49	250000009
SRR292770 forw	5102041	49	250000009
SRR292770 rev	5102041	49	250000009

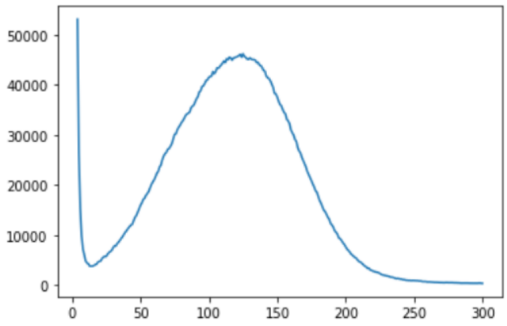

Figure 1: Distribution if 31 k-mers.

Assembling *E. coli* genome from paired reads and mated reads.

We use `SPAdes` to assemble our genome, then run `QUAST` on `contigs.fasta` file. The results of `QUAST` you can see below in Table 2.

Table 2: Assembling statistics

Libraries	Contigs number	Contigs N50	Contigs L50	Scaffolds contigs number	Scaffolds N50	Scaffolds L50
Pair-end	205	105346	15	213	105346	15
Pair-end + mate pairs	178	151014	11	120	1048041	2

Finding the closest relative of E. coli X.

After using `Barrnap` we find 5 16srRNA genes. The information about them you can find in Table 3.

After using `BLAST` we find one the most suitable reference genome - *Escherichia coli 55989 [NC_011748.1]* [18].

Finding the genetic cause of HUS and antibiotic resistance.

By using `Mauve` we find that the genome of our strain of *E. coli* has Shiga-toxin genes: *stxB*, has length 269, from 5301809 to 5302078, and *stxA*, has length 959, from 5302090 to 5303049. To find the origin of these genes, we search near genes too, and find that they encode phage proteins (for example phage DNA adenine methylase).

Antibiotic resistance detection

We use `ResFinder` to detect the antibiotic resistant genes. The result is the Table 4

Table 3: 16srRNA genes that was found by `Barrnap`.

Gene, strand	First position	End position	Lenght
16S_rRNA::NODE_14 forw (-)	45	1583	1538
16S_rRNA::NODE_41 rev (+)	26351	27889	1538
16S_rRNA::NODE_65 forw (-)	3511	5049	1538
16S_rRNA::NODE_67 rev (-)	3419	4957	1538
16S_rRNA::NODE_197 forw (+)	33	439	406

Table 4: Antibiotics to which the strain is resistant.

Class of antibiotics	Name of antibiotics	Genetic background
beta-lactam	cefepime, ampicillin, cefotaxime, ceftazidime, piperacillin, amoxicillin, ticarcillin, aztreonam	blaTEM-1B (blaTEM-1B_AY458016), blaCTX-M-15 (blaCTX-M-15_AY044436)
aminoglycoside	streptomycin,	aph(6)-Id (aph(6)-Id_M28829), aph(3")-Ib (aph(3")-Ib_AF321551)
folate pathway antagonist	sulfamethoxazole, trimethoprim	sul1 (sul1_U12338), sul2 (sul2_HQ840942), sul1 (sul1_AY115475), sul1 (sul1_AY522923), dfrA7 (dfrA7_AB161450)
tetracycline	tetracycline, doxycycline	tet(A) (tet(A)_AJ517790)
quaternary ammonium compound	cetylpyridinium chloride, ethidium bromide, benzylkonium chloride, chlorhexidine	qacE (qacE_X68232)

Then we searched β-lactamase in `Mauve` and found that it is encodes in the part of genome that is absent in the reference strain. We have looked at this region and found genes related to plasmids.

4 Discussion

In this project we assemble *de novo* the genome of *E. coli* that was the cause of outbreak in German and Europe in 2011. We did two assemblies, used one pair-end library and three libraries, one of which was the library from the first assembling and two others were mate-pair libraries. We used `Quast` to check our assemblies. The N50 metric in the first assembly is 105346, in the second is 151014, so we can conclude that the second one is better. So we use it through our project further.

We find the closest relative genome to our strain *E. coli* 55989 strain. Then we align them and find out genes that could provide Hemolytic uremic syndrome in infected patients. We found that the unknown strain has Shiga-toxin genes, *stxA* and *stxB*, and they are of phage origin. So the investigated strain could become pathogenic by phage genome integration.

Futher we explore our strain for antibiotic resistance. We find out that the strain is resistant to wide variety of types of antibiotics. Also we find genes of the bla-family that encode β-lactamase. It was surrounded by plasmid-associated genes, so we conclude that the outbreak strain of E. coli strain obtained antibiotic resistance by plasmid horizontal gene transfer from another bacteria.

Thus, we investigate the strain that has *Stx* genes, that produce Shiga-toxin which provide the Hemolytic uremic syndrome. We can suppose that the origin of these genes is some phage. The strain also has resistance to many type of antibiotics, include beta-lactam, tetracycline, folate pathway antagonist, aminoglycoside, quaternary ammonium compound. So we suggest to treat patients that were infected with the this E.coli strain with another type, like oxazolidinone (linezolid) or polymyxin (colistin).

Labjournal.

You can see the labjournal here: https://docs.google.com/document/d/1owONLdLfRaP_zpfDqCXZ5inpshwwo-VIs_SwQxIRQaE/edit?usp=sharing

References

[1] Beutin L, Martin A. Outbreak of Shiga toxin-producing Escherichia coli (STEC) O104:H4 infection in Germany causes a paradigm shift with regard to human pathogenicity of STEC strains. J Food Prot. 2012 Feb;75(2):408-18. doi: 10.4315/0362-028X.JFP-11-452. PMID: 22289607.

[2] Fraser M.E., Fujinaga M., Cherney M.M., Melton-Celsa A.R., Twiddy E.M., O'Brien A.D., James M.N. Structure of shiga toxin type 2 (Stx2) from Escherichia coli O157:H7. J. Biol. Chem. 2004;279:27511–27517.

[3] Obrig TG. Escherichia coli Shiga Toxin Mechanisms of Action in Renal Disease. Toxins (Basel). 2010 Dec 2;2(12):2769-2794. doi: 10.3390/toxins2122769. PMID: 21297888; PMCID: PMC3032420.

[4] Obrig T.G., Moran T.P., Brown J.E. The mode of action of Shiga toxin on peptide elongation of eukaryotic protein synthesis. Biochem. J. 1987;244:287–294.

[5] McAllister LJ Bent SJ Petty NK et al. Genomic comparison of Two enterohemorrhagic Escherichia coli isolates from a historic hemolytic-uremic syndrome Outbreak in Australia Infect Immun 2016 84 775 81

[6] Forward reads: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R1_001.fastq.gz, reverse read: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R2_001.fastq.gz.

[7] Forward reads: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R1_001.fastq.gz, reverse read: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R2_001.fastq.gz.

[8] Forward reads: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R1_001.fastq.gz, reverse read: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R2_001.fastq.gz.

[9] Wingett SW, Andrews S. (2018) FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res. 24*, 1338.

[10] Guillaume Marcais and Carl Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics (2011) 27(6): 764-770 doi:10.1093/bioinformatics/btr011

[11] Andrey Prjibelski, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, Anton Korobeynikov, Using SPAdes De Novo Assembler, Current Protocols in Bioinformatics Volume: 70 Issue 1 (2020) ISSN: 1934-3396 Online ISSN: 1934-340X

[12] Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics, 2013

[13] Seemann T. Prokka: rapid prokaryotic genome annotation Bioinformatics 2014 Jul 15;30(14):2068-9. PMID:24642063

[14] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

[15] Seemann T, barrnap 0.9 : rapid ribosomal RNA prediction, https://github.com/tseemann/barrnap

[16] Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004 Jul;14(7):1394-403. doi: 10.1101/gr.2289704. PMID: 15231754; PMCID: PMC442156.

[17] Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. ResFinder - an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microb Genom. 2022 Jan;8(1):000748. doi: 10.1099/mgen.0.000748. PMID: 35072601; PMCID: PMC8914360.

[18] https://www.ncbi.nlm.nih.gov/nucleotide/NC_011748.1?report=genbank&log$=nuclalign&blast_rank=1&RID=RPV31UBC016