BIOINFORMATICS INSTITUTE

Search of *Escherichia coli* virulence factors using *de novo* assembly

Nikita Vaulin[1, 2,†]

[1]Bioinformatics Institute, Saint-Petersburg, Russia
[2]Skolkovo Institute of Science and Technology, Moscow, Russia

[†]vaulin@ro.ru

Abstract

In this work, we have shown that the strain of *Escherichia coli* that has become pathogenic has simultaneously acquired genes encoding Shiga toxins and a beta-lactam antibiotic resistance cassette. We show that the Shiga toxin genes were most likely acquired during infection with a phage that had previously infected the Shigella bacterium. The acquisition of the antibiotic resistance cassette most likely occurred during conjugative plasmid transfer. Also, both of these processes may be related to transposons. Thus, we can definitely say that the emergence of patenogenesis is related to horizontal gene transfer and mobile genetic elements. We suggest other treatment options such as other antibiotics, monoclonal antibodies and Shiga toxin inhibitors.

Key words: pathogenic bacteria; HUS; antibiotic resistance; *Escherichia coli*; *de novo* genome assembly

Introduction

Escherichia coli are the most well-studied bacteria, that are commonly found in gastrointestinal tract of vertebrates. Usually they do not harm host organism and play a role in a work of gastrointestinal tract [1]. However, in contrast to the described strains, called commensal, there are also several pathogenic *E. coli*. This pathogenic strains are able to cause different diseases, namely intestinal and diarrheal diseases, urinary tract infections and sepsis and meningitis [2]. Thus, despite all the benefits and widespread prevalence of *E. coli*, its pathogenesis can lead to terrible consequences.

Commensal bacteria can become pathogenic in different ways. Mobile genetic elements, for instance plasmids, transposons, bacteriophages and pathogenicity islands (PAIs) can encode virulence factors. The process of loss and gain of this mobile genetic elements, called horizontal gene transfer, may lead to the spread of virulence factors among different strains of bacteria. Moreover, horizontal gene transfer coupled with *de novo* mutations of commensal *E. coli* can lead to the emergence of new pathogenic *E. coli* strains. It this case, it is considered to use bioinformatics methods to search virulence factors, due to absence of specific biochemical test to this particular strain [1].

One of the bioinformatics techniques, that can be used to find virulence factors is the genome *de novo* assembly. This makes it possible not to focus on the reference genome, as in the case of alignment, and avoid missing of the biological effect of mobile genetic elements, carrying virulence factors. Thus, possible large fragments of DNA responsible for the acquisition of pathogenicity, that may not be aligned with the reference genome, but can be detected after *de novo* assembly.

In this work, we focus on enterohaemorrhagic *E. coli*, which are pathogens associated with food and water contamination. People infected with this bacteria suffer from enterohaemorrhagic colitis, possible complications leading to the haemolytic uraemic syndrome (HUS), which is a potentially fatal disease [3, 4]. We are faced with the task of determining the strain of the *E. coli* that led to the outbreak of HUS in Germany in 2011 using bioinformatics methods and search of resistance to antibiotics, which were generally used in treatment.

Materials and methods

Dataset

In this work data from Illumina HiSeq 2000 deep sequencing of a *Escherichia coli* O104:H4 sample TY-2482 were used. Data was submitted by Beijing Genome Institute (BGI) into Short Read Archive (SRA) and available on following identifiers: SRR292678 (paired end library), SRR292862 (first mate pair library), SRR292770 (second mate pair library).

Raw reads were analysed with FastQC v.0.11.9 [5]. For SRR292678 sequence data total number of reads (*TNR*) and average length of reads (*L*) were obtained. All FastQC reports are presented in project repository at GitHub (see Supplemental materials).

K-mer profile and genome size estimation

In order to assess data quality k-mer profile construction was performed with Jellyfish v. [6]. The `jellyfish count` and `jellyfish hist` functions were used to count the frequency of all possible k-

Compiled on: December 2, 2022.
Draft manuscript prepared by the author.

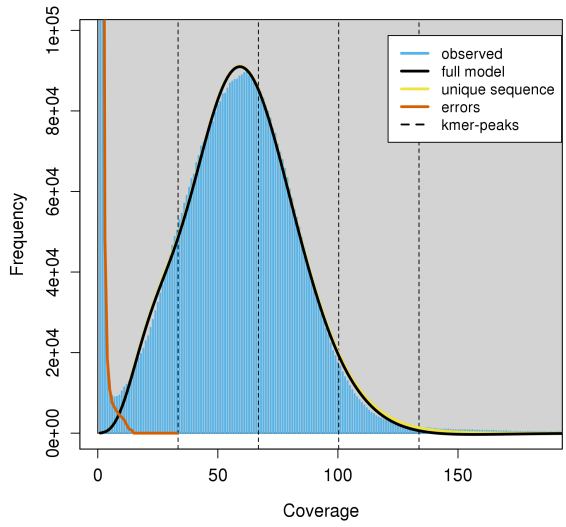

Figure 1. Illumina paired-end library 31-mers distribution.

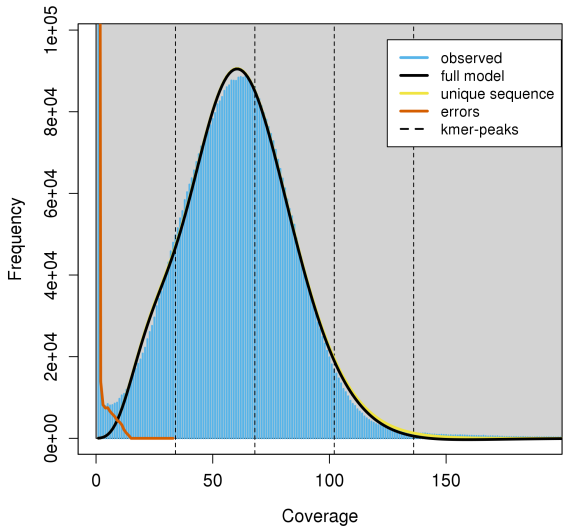

Figure 2. Corrected Illumina paired-end library 31-mers distribution.

mers of the given length ($K = 31$) in SRR292678 sequence sample. The k-mers distribution were visualised with GenomeScope browser [7].

The k-mers distribution were used to estimate the genome size. Genome size (*EGS*) was estimated according to the formula:

$$EGS = \frac{T}{N},$$

where T is total number of bases ($T = L \cdot TNR$) and N is depth of coverage, estimated using following formula:

$$N = \frac{M \cdot L}{L - K + 1}$$

M is k-mer peak (See Fig. 1). L and T were obtained as a result of FastQC analysis.

Genome assembly

Two genome assembly was performed via SPAdes tool v. [8] using *(i)* SRR292678 only and (ii) all three libraries to get higher assembly quality: SRR292678 (flags `pe1-1` and `pe1-1`), SRR292862 (flags `mp1-1` and `mp1-2`) and SRR292770 (flags `mp2-1` and `mp2-2`). Due to fact that SPAdes performs error correction, additional k-mer profiling was performed as described in section K-mer profile and genome size estimation. Assembly quality metrics, such as L50, N50 and number of miss-assemblies, were evaluated with QUAST v. 5.0.0 [9].

Genome annotation and search of closest *E. coli* strain

To perform gene prediction and annotation Prokka v was used with parameters `-genus Escherichia -species coli -gram neg -centre` [10]. The closest strain to that one under the study was found basing on 16S rRNA sequence with help of nucleotide BLAST (blastn) search against RefSeq database [11, 12]. The 16S rRNA were located with Barrnap tool v [13]. The blastn search was used with the following restrictions on time range to find only entries, available at the beginning of 2011: `1900/01/01:2011/01/01[PDAT]` (all other parameter were set by default).

The bacterial reference strain found was used as reference for search of pathogenicity and antibiotic resistance factors. FASTA file of reference strain was obtained from the NCBI nucleotide database.

Search of pathogenicity and antibiotic resistance factors

Obtained as a result of genome annotation .gbk file was used to perform a genome-wide comparison of pathogenic strain with the reference one. Comparison was performed using Mauve v. 20150226build [14]. Search of responsible for antibiotic resistance genes was performed with help of ResFinder [15]. The genes, located close to those one related to the pathogenicity or antibiotic resistance, was found in the Prokka annotation .gbk file using *bash* script, presented in the Supplemental materials section. The genes products list was submitted to the BLAST blastp search with default parameters. The blastp results were downloaded as a .txt alignment file and the very proteins were extracted using another *bash* script (see as well).

Results

Data quality assessment

In this study here were used 3 sequencing Illumina libraries: a paired end one and two mate pair libraries. Some libraries metrics are presented in the Table 1. The k-mers distribution, obtained with jellyfish tool and plotted with GenomeScope is presented on Fig. 1. As SPAdes assembler provides reads correction as the first step of assembly, we also inspected k-mer profile of corrected reads (Fig. 2).

Additionally, the genome size was estimated basing on k-mers distribution. Firstly, the formula presented in the section was used, resulted in genome size equal to 5 321 948 base pairs (for the k-mer length 31, read length 90, and 31-mers peak 62). Secondly, GenomeScope also provided its own genome size estimation equal to 4 824 834. This quantities are slightly different, but they indicate the overall expected genome size.

Genome assembly and annotation

The genome was assembled with SPAdes tool basing either on a single paired-end library and on all three libraries. The assembly quality metrics obtained with QUAST are presented in the Table 2.

Gene annotation procedure, carried out with Prokka, resulted in 5106 CDS, 19 rRNAs, 69 tRNAs, 1 tmRNA and 1 repeat region. Additionally, rRNAs were predicted with Barrnap tool. Barrnap resulted 19 rRNAs: 8 16S, 5 23S and 6 5S rRNAs. The predicted 16S rRNAs have the length of 1538 nucleotides.

Library		Read length	Insert size	№ of reads
SRR292678	PE	90	470 bp	10 998 692
SRR292862	MP	49	2 kb	10 204 082
SRR292770	MP	49	6 kb	10 204 082

Table 1. Illumina libraries statistics; *PE* - paired end, *MP* - mate pair

Libraries	N50	L50	№ N's / 100 kbp	№ of scaffolds
PE	105 346	15	32.23	502
PE and MP	1 046 849	2	1149.12	445

Table 2. Assembly quality statistics

E.coli strain 55989 is a closest relative

In order to determine the closest *E. coli* strain we used rRNA genes to do a blastn search in the RefSeq database. Based on rRNAs, *E.coli* strain 55989 was found as a closest relative.

This strains is an enteroaggregative *E. coli* (EAEC) strain, harboring the pAA plasmid, which contains aggregative adherence fimbria (AAF) genes allowing bacteria to stick to cells in the intestine [16].

This particular *E. coli* strain obtained additional antibiotics resistance and pathogenecity factors

Using Mauve we inspected an alignment of assembled genome and the reference one. Firstly, we found two genes related to the shiga toxin: *stxA* and *stxB*. These genes are not present in the reference genome.

Secondly, we used ResFinder in order to determine antibiotic resistance genes. ResFinder resulted in two *bla* genes: *blaCTX-M-15* and *blaTEM-1B*, that are not present in reference genome and provide resistance to ampicillin, amoxicillin, ceftriaxone and some other beta-lactam antibiotics. Interestingly, RefFinder does not predict resistance to the penicillin and some other beta-lactam antibiotics even these *bla* genes are exist.

Shiga toxin and *bla* genes emergence might be due to mobile genetic elements

We've carried out screening of genes, located near to the *bla* and *stx* ones. Since most of the ORFs were unannotated, we've done blastp search for their products.

For the Shiga toxin related genes we found genes, that encode proteins related to the *transposones* (IS66 family element, Orf2 protein; ISAs1 family transposase, transposase family protein) and to the *phages* (tail fiber domain-containing protein, receptor recognizing protein Gp38, phage tail assembly protein T).

For the *bla* genes we found genes, that encode proteins related to the *transposones* (ParA family plasmid-partitioning AAA ATPase, ParB-like protein, ParB-like protein, Tn3-like element Tn3 family transposase, TnpA) and *plasmid conjugation* (PilK, PilI type IV pilus biogenesis protein, ProQ/FINO family protein, pilus assembly protein PilP, pe IV pilus biogenesis proteins PilL and PilM, conjugal transfer proteins TraA and TraC).

Discussion

Genome assembly quality

As can be seen from the QUAST results, supplementing paired-end libraries with mate-pair ones of different insert sizes can dramatically improve the assembly. Here the N50 metric is increased 10-times and L50 became really small (Table 2). At the same time number of unidentified positions became larger. At the first glance, it can be assumed as a bad sign. In the contrary, this can indicate, that addition of mate-pair libraries helps to better locate scaffolds between each other. In this case we obtain additional information about what regions of the genome were unconsidered in assembly.

The additional SPAdes step of reads correction, seemingly, does not really fixes the situation of k-mers with too low coverage. The lest one peak became thinner and lower, but not so much enough. Maybe, in the others versions of SPAdes this process of correction works better.

The emergency of pathogenecity

Here we identified, that the strain being studied became pathogenic due to the appearance of Shiga toxin related genes: *stxA* and *stxB*. Near to the Shiga toxins genes two groups of genes have been identified: the genes related to the transposones and related to the phages. From the phage side, we observed several genes responsible for the phage structure (such as phage tail assembly protein) and for the phage infecting process (receptor recognizing protein Gp38 uses outer membrane protein ompA as a target). From the transposone site, here we see some transposones related proteins and the ISAs1 transposases itself. *Nota bene*, that transposones-related genes are located both upstream and downstream from the *stx* genes region. Thus we can conclude that most probably, the *stx* genes cassette appearance occured with the transposone insertion. On the other hand, such gene set is not presented in the reference strain, hence it cannot appear from nowhere. It can be possible, that transposone DNA fragment was obtained by cell *via* some DNA uptake from the environment. Since there are some phage-related genes nearby, it can be possible that transposone cassette was brought with the phage insertion. This hypothesis is also supported by the fact that a phage lysis protein from the bacterium *Shigella sonnei* was found in the region nearby. This allows us to construct a picture of such a scenario that the *E. coli* under the study was infected by a phage that had previously infected *Shigella sonnei*, from which it had inadvertently taken the Shiga toxins during excision.

The mechanism of pathogenecity

The Shiga toxin family includes Shiga toxin from Shigella dysenteriae serotype 1 and the Shiga toxins that are produced by enterohaemorrhagic *Escherichia coli* (EHEC) strain [17]. Firstly Shiga toxins were characterized by the Japanese microbiologist Kiyoshi Shiga after the dysentery outbreak in Japan in 1897 [17].

An enzymatically active monomeric StxA subunit is noncovalently associated with a StxB pentamer which is responsible for binding to cell surface receptors [17]. StxA and the StxB fragments are secreted into the bacterial periplasm, where they assemble noncovalently into the a holotoxin [17].

StxB binds to the neutral glycosphingolipid globotriaosylceramide which is present on the surface of cells leading to subsequent internalization of the toxin [17]. StxA possesses a highly specific RNA N-glycosidase activity that cleaves an adenine base at position 4,324 on the α-sarcin loop located on domain VI of 28S ribosomal RNA (Fig. 3) [17]. Although Shiga toxins are extremely potent ribosome-modifying enzymes, it should be noted that their action is not limited to the inhibition of protein synthesis. They have several cellular effects, including the induction of cytokine expression by macrophages, which in turn may increase the susceptibility of certain cells to toxins [17].

The emergency of antibiotic resistance

We also determined that this particular strain carries two additional beta-lactam antibiotic resistance genes. Since we mentioned above that the bacterium is resistant to, for example, ampicillin, but not

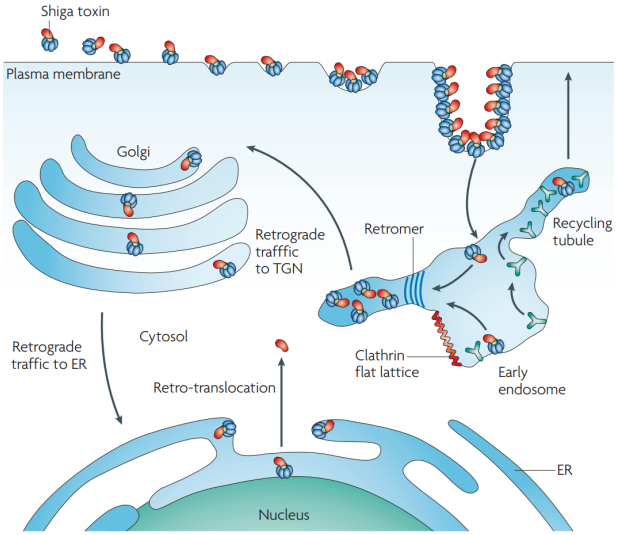

Figure 3. The scheme of Shiga toxin acting within the cell [17].

to penicillin, this suggests that there is a specific mechanism. We found that the antibiotics to which the bacterium is resistant (ampicillin, ceftriaxone, ceftrazidime) have additional nitro groups in their structure. Nevertheless, the specific determination of the mechanism of partial resistance may be a subject for further research.

Here we found that there are two groups of genes next to antibiotic resistance genes. First, there are a large number of transposon-related genes. Secondly, many genes related to plasmid conjugation are found. Thus, several mechanisms can be assumed. It is possible that the antibiotic resistance cassette was introduced by a transposon. Maybe even in the same phage infection process. At the same time, it could have occurred in a conjugative transfer process, such as an F-plasmid. It may be worth examining the sequencing data more closely for the presence of plasmid fragments.

The possible ways of treatment

As additional antibiotic resistance has emerged, new ways of treating patients must be found. First, screening for other possible antibiotics is worthwhile. For example, lincomycin or tiamulin may be helpful here. At the same time, treatment scenarios with monoclonal antibodies against Shiga toxins or Shiga toxin inhibitors can be considered [17].

References

1. Gambushe SM, Zishiri OT, Zowalaty MEE. Review of *Escherichia coli* O157:H7 Prevalence, Pathogenicity, Heavy Metal and Antimicrobial Resistance, African Perspective. Infection and Drug Resistance 2022 Aug;Volume 15:4645–4673. https://doi.org/10.2147/idr.s365269.
2. Kaper JB. Pathogenic *Escherichia coli*. International Journal of Medical Microbiology 2005 Oct;295(6-7):355–356. https://doi.org/10.1016/j.ijmm.2005.06.008.
3. Croxen MA, Finlay BB. Molecular mechanisms of *Escherichia coli* pathogenicity. Nature Reviews Microbiology 2009 Dec;8(1):26–38. https://doi.org/10.1038/nrmicro2265.
4. Michael M, Bagga A, Sartain SE, Smith RJH. Haemolytic uraemic syndrome. The Lancet 2022 Nov;400(10364):1722–1740. https://doi.org/10.1016/s0140-6736(22)01202-8.
5. FastQC tool;. Accessed: 2022-10-27. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
6. Kingsford C MG. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011;27(6):2769–2794.
7. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 2017;33(14):2202–2204.
8. et al BA. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 2012;.
9. et al GA. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013;.
10. T S. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30(14):2068–9.
11. et al ASF. Basic local alignment search tool. Journal of Molecular Biology 1990;215(3):403–410.
12. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research 2015 Nov;44(D1):D733–D745. https://doi.org/10.1093/nar/gkv1189.
13. T S, Barrnap v0.9; 2018. https://github.com/tseemann/barrnap/blob/master/README.md.
14. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research 2004;14(7):1394–1403.
15. et al BV. ResFinder 4.0 for predictions of phenotypes from genotypes . J Antimicrob Chemother 2020;75(12):3491–3500.
16. Mossoro C, Glaziou P, Yassibanda S, Lan NTP, Bekondi C, Minssart P, et al. Chronic diarrhea, hemorrhagic colitis, and hemolytic-uremic syndrome associated with HEp-2 adherent Escherichia coli in adults infected with human immunodeficiency virus in Bangui, Central African Republic. Journal of Clinical Microbiology 2002;40(8):3086–3088.
17. Johannes L, Römer W. Shiga toxins—from cell biology to biomedical applications. Nature Reviews Microbiology 2010;8(2):105–116.

Supplemental materials

GitHub repository of the project: Pathogenicity_project_BI_2022

Extracting related genes from the .gbk annotation files:

```
cat ecoliX_.gbk | grep -P '(CDS|gene|translation| [A-Z]+"*$)' | grep 'stx' -C 100 | tr -d "\n" | tr -d " " |
grep -oP '(\d+..\d+).+?translation="[A-Z]+."' | tr -d ")" | sed "s/^/> /g" | sed 's/\/.\+\?=/\n/g' | tr -d '"'
```

Extracting hits from BLAST results:

```
cat ecoliX_.gbk | grep -P '(CDS|gene|translation| [A-Z]+"*$)' | grep 'stx' -C 100 | tr -d "\n" | tr -d " " |
grep -oP '(\d+..\d+).+?translation="[A-Z]+."' | tr -d ")" | sed "s/^/> /g" | sed 's/\/.\+\?=/\n/g' | tr -d '"'
```