Revealing the reason for the pathogenic outbreak of *E.coli* strain by the genome assembly

Abstract

Hemolytic Uremic Syndrome (HUS) is often triggered by infection with enterohemorrhagic *Escherichia coli* (EHEC) strains, particularly *Escherichia coli* O157:H7, producing Shiga-like toxins. Horizontal gene transfer (HGT) is a potential mechanism by which non-pathogenic *E.coli* strains acquire pathogenic traits. This study explores the genomic evolution of *E.coli* strain TY2482, associated with HUS, employing *de novo* genome assembly and comparative genomics. Three ion torrent libraries with varying insert sizes were sequenced, and SPAdes assembly produced two sets of contigs. Genomic analysis revealed a 5.3 Mb genome with an inserted segment carrying Shiga toxin-related genes, most likely acquired through HGT from *Escherichia phage P13374*. Antibiotic resistance analysis revealed multiple antibiotic resistance, which most probably occurred due to plasmids transfer. Overall, this project enhances understanding of the genomic basis for *E.coli* strain causing HUS and antibiotic resistance and provides insights into potential treatment strategies.

Introduction

Hemolytic uremic syndrome (HUS) is a condition that can occur when small blood vessels become damaged and inflamed. Most often, infection with certain strains of *E. co*li is the cause, namely by enterohemorrhagic *Escherichia coli* (EHEC) serogroup strains, mainly *Escherichia coli* O157:H7 [1]. Enterohemorrhagic *E.coli* (EHEC) are not invasive, so bacteremia is rare, but they secrete ribosome inactivating Shiga-like toxins (Stx1 and Stx2 with variants) which are responsible for much of the organ damage [2, 3], and Stx2 is more frequently associated with severe disease.

Those *E.coli* strains that initially do not contribute to HUS can also suddenly start to produce Shiga toxin and cause HUS. One of the possible reasons for such a transformation is a horizontal gene transfer (HGT) [4]. HGT is a process of transmition of genetic material between organisms other than vertical transmission from parent to offspring. There are 3 main mechanisms of HGT in nature: transformation, transduction, and conjugation. In the HGF transformation mode extracellular naked DNA is taken up by cells that have developed genetic competence. HGT transduction occurs when the bacteriophage carries a portion of the bacterial genetic information from a bacterium to another bacterium. In the process of conjugation, a complete DNA sequence, e.g., a plasmid, is transmitted between bacterial cells via direct cell-to-cell connection by conjugating pilus [4, 5].

HGT usually can lead to dramatic genome structural variations and changes. In order to be able to detect large genome insertions, deletions, inversions and translocations de novo genome assembly is usually used. Compared to aligning reads to a reference genome, de novo assembly helps in discovering novel genes that may not be presented in the reference and detect gene content difference.

The aim of this project was to understand the mechanism of how *E.coli* strain acquired pathogenic features to become the causative of HUS and describe its distinctive genetic traits responsible for pathogenicity.

Materials & Methods

The sequencing data of an isolate TY2482 [6] from SRA database [7] was used for genome assembly. Three libraries are available: one paired (SRR292678) end and two mate-pairs (SRR292862, SRR292770) with different insert size. *FastQC v0.11.9* was used to control the quality of the reads [8].

Jellyfish v2.3.0 was used for k-mer profile and genome size estimation [9]. Genome size evaluations were carried out for the raw reads and for the corrected ones. To count k-mers, jellyfish count (-C -m 31 -s 500M) and jellyfish hist were run on raw data and SPAdes-corrected reads from paired end library (SRR292678). The genome size was estimated by the ratio T/N, where T is the total number of bases in reads, T= L*number_of_reads, N - the depth of coverage, N = (M*L)/(L-K+1), M - the average peak in k-mer histogram, L - the average reads length, K - k-mer size.

The genome assembly and reads correction are provided by *SPAdes v3.13.1* [10]. Since the library with a small insert size can resolve short repeats, whereas the library with a larger insert size can resolve longer repeats, two assemblies were provided: for one paired end library (SRR292678) and for all three libraries. Assembly statistics for contigs quality were obtained using the online version of Quast [11].

Further work was carried out with the assembly based on three libraries. Genes were annotated with Prokka v1.14.6 (--centre XXX) [12]. The search for a close relative of an unknown strain of E.coli X is carried out on the basis of 16s rRNA sequences predicted by Barrnap v0.9 [13]. Predicted 16s rRNA sequences was uploaded in BLAST (refseq_genomes Database for organism *E.coli* (taxid:562) and Entrez Query: 1900/01/01:2011/01/01[PDAT] for searching to only those genomes that were present in the GenBank database at the beginning of 2011) [14]. To determine the differences between the genome of an unknown strain and its close relative, Mauve v2.4.0 was used [15]. ResFinder was used to detect resistance genes [16]. The pipeline is shown in figure 1.

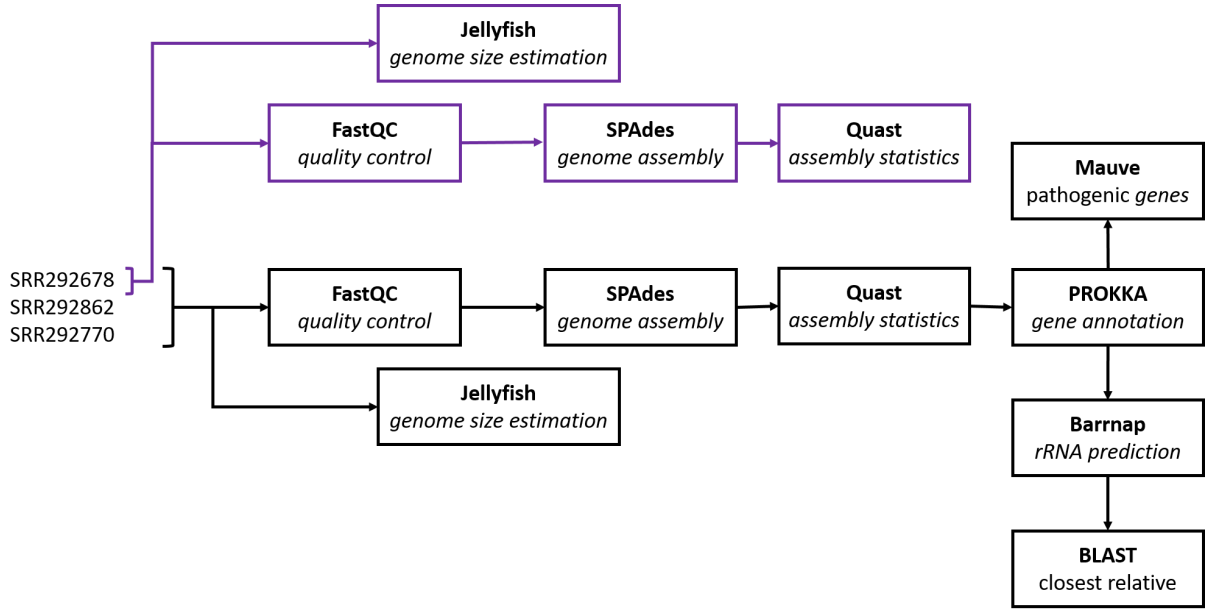

Fig. 1 The pipeline for *E.coli* strain investigation. The main part of analysis is colored in black, the additional one is purple

Results

There are available three ion torrent libraries from the TY2482 with different inset size:

- SRR292678 - paired end, insert size 470 bp, 5499346 reads
- SRR292862 – mate pair, insert size 2 kb, 5102041 reads
- SRR292770 – mate pair, insert size 6 kb, 5102041 reads (for each forward and reverse)

Estimation of the genome size using the library of paired ends (SRR292678) resulted in approximately 5.28 MB, using corrected reads for 3 libraries gave the value 5.32 Mb. As a result of the correction, the score has become higher, but not by much. The peak of the distribution for the second estimate has shifted to the region of lower values (Figure 2).

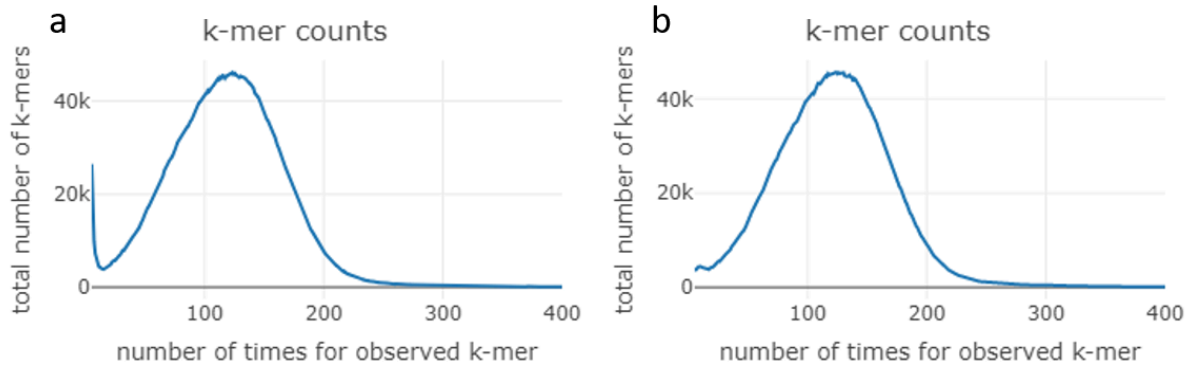

Fig. 2 K-mer histograms for (a) raw SRR292678 library and (b) corrected SRR292678 library.

SPAdes assembly was provided for the paired end library and for all three libraries. The assembly from one library contains more configs with lower N50 than for three libraries one (Table 1). So it could be concluded that this assembly is better than the first one. Thus, using multiple libraries with different insert sizes allows to get better assembly than using only one library.

	One-library assembly		**Three-libraries assembly**	
	contigs	**scaffolds**	**contigs**	**scaffolds**
Contigs (>=5000 bp)	81	82	33	16
Total length (>=5000 bp)	5076685	5081904	5202939	5258076
N50	111860	11860	335515	2815616

Table 1. Basic assembly statistics for one and three libraries.

After Prokka gene annotation and prediction, the following statistics was obtained:

80	tRNAs
0	rRNAs
1	CRISPRs
5064	CDS
2923	Unique gene codes

The use of the Barrnap tool allowed to locate seven 16S rRNA genes with the length of 1538 bp each except one (406 bp) in *E.coli* X genome assembly (Table 1. Supp.). After running BLAST with determined *E.coli* X 16S rRNA gene sequences, the closed relative to *E.coli* X strain was found, which is *E.coli* 55989 strain (Sequence ID: NC_011748.1). In comparison with the reference strain, *E.coli* X strain contains an insertion including *stxA* (959 bp, coords:3483886,3484845; locus_tag PROKKA_03324) and *stxB* (269 bp;

coords:3483605,3483874; locus_tag PROKKA_03323) genes which produce Shiga toxin - the main cause of HUS syndrome. According to RAST annotation, *stxA* and *stxB* are surrounded by phage PPwrbA genes; the first insert's gene is a phage PPwrbA integrase.

Discussion

An insertion containing shiga-toxin-related genes (*stxA*, *stxB*) was found in the genome of the studied strain of *E.coli.* Shiga-toxin is a known toxin of several strains of *E.coli.* These toxins cause bleeding by destroying the lining of the colon, and they can lead to HUS if they get into the kidneys [https://doi.org/10.3389/fmicb.2020.01472]. In addition to the toxin genes, the insertion contains *Escherichia phage P13374* genes. Bacteriophages can provide horizontal genetic transfer (HGT) and transfer genes between two organisms. In such a way bacteriophages transform non-pathogenic strains.

According to ResFinder antibiotic resistance analysis, the E.coli X is resistant to a broad spectrum of antibiotics (Table 2), while the reference E.coli 55989 strain is vulnerable to all tested antibiotics. By aligning with BLAST a series of genes flanking the beta-lactam antibiotic resistance genes bla1 and bla2, it was found that these genes most probably came into E. coli X with plasmid pKC90-L DNA or *E.coli* plasmid pB5-L DNA through HGT conjugation. The same mechanism is most likely to be responsible for resistance to other antibiotics: for example, according to ResFinder aph(3")-Ib and aph(6)-Id genes responsible for streptomycin resistance were transferred from Shigella flexneri plasmid pSTR1 and Escherichia coli plasmid RSF1010 correspondingly. In case of streptomycin, the resistance mechanism is related to strA and strB genes, which encode enzymes that modify streptomycin by phosphorylation which prevents it from binding to bacterial ribosomes.

As for alternative ways of patient treatment, the antibiotics which do not belong to resistance groups mentioned in Table 2. can be suggested for patients therapy. For example, ciprofloxacin, which belongs to quinolones, or any kind of aminoglycoside antibiotics (fortimicin, butiromycin) might be effective.

	Resistance.gene	Phenotype	Accession.no.
1	sul2	Sulphonamide resistance	HQ840942
2	sul1	Sulphonamide resistance	AY115475
3	sul1	Sulphonamide resistance	U12338
4	sul1	Sulphonamide resistance	AY522923
5	tet(A)	Tetracycline resistance	AJ517790
6	dfrA7	Trimethoprim resistance	AB161450
7	blaTEM-1B	Beta-lactam resistance Alternate name; RblaTEM-1	AY458016
8	blaCTX-M-15	Beta-lactam resistance Alternate name; UOE-1	AY044436
9	aph(3")-Ib	Aminoglycoside resistance Alternate name; aph(3")-Ib	AF321551
10	aph(6)-Id	Aminoglycoside resistance Alternate name; aph(6)-Id	M28829
11	qacE	Disinfectant resistance	X68232

Table 2. Genes responsible for antibiotic resistance in *E.coli X* strain.

Citations

1. Riley L.W., Remis R.S., Helgerson S.D., McGee H.B., Wells J.G., Davis B.R., Hebert R.J., Olcott E.S., Johnson L.M., Hargrett N.T., et al. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. N. Engl. J. Med. 1983;308:681–685.
2. Schmidt H., Geitz C., Tarr P.I., Frosch M., Karch H. Non-O157:H7 pathogenic Shiga toxin-producing *Escherichia coli*: Phenotypic and genetic profiling of virulence traits and evidence for clonality. J. Infect. Dis. 1999;179:115–123.
3. Mayer CL, Leibowitz CS, Kurosawa S, Stearns-Kurosawa DJ. Shiga toxins and the pathophysiology of hemolytic uremic syndrome in humans and animals. Toxins (Basel). 2012 Nov 8;4(11):1261-87. doi: 10.3390/toxins4111261.
4. Emamalipour M, Seidi K, Zununi Vahed S, Jahanban-Esfahlan A, Jaymand M, Majdi H, Amoozgar Z, Chitkushev LT, Javaheri T, Jahanban-Esfahlan R, Zare P. Horizontal Gene Transfer: From Evolutionary Flexibility to Disease Progression. Front Cell Dev Biol. 2020 May 19;8:229.
5. Hasegawa H, Suzuki E, Maeda S. Horizontal Plasmid Transfer by Transformation in *Escherichia coli*: Environmental Factors and Possible Mechanisms. Front Microbiol. 2018 Oct 4;9:2365.
6. https://www.ncbi.nlm.nih.gov/sra/?term=TY2482
7. https://www.ncbi.nlm.nih.gov/sra
8. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
9. 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
10. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes de novo assembler. Current Protocols in Bioinformatics, 70, e102. doi: 10.1002/cpbi.102
11. http://cab.cc.spbu.ru/quast/
12. Seemann T. Prokka: rapid prokaryotic genome annotation Bioinformatics 2014 Jul 15;30(14):2068-9. PMID:24642063
13. Seemann T. 2018. Barrnap 0.9: rapid ribosomal RNA prediction
14. Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, David J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research, Volume 25, Issue 17, 1 September 1997, Pages 3389–3402.
15. https://darlinglab.org/mauve/download.html
16. https://cge.food.dtu.dk/services/ResFinder/

Supplementary materials

location	length
>16S_rRNA::NODE_16_length_113537_cov_78.5276_ID_565907:111954-113492(+)	1538
>16S_rRNA::NODE_1_length_698474_cov_70.3598_ID_565849:45-1583(-)	1538
>16S_rRNA::NODE_24_length_45417_cov_81.3463_ID_565899:43834-45372(+)	1538
>16S_rRNA::NODE_25_length_38063_cov_95.809_ID_565903:36480-38018(+)	1538
>16S_rRNA::NODE_5_length_353792_cov_75.7002_ID_565879:42578-44116(-)	1538
>16S_rRNA::NODE_7_length_269063_cov_84.1067_ID_565845:45-1583(-)	1538
>16S_rRNA::NODE_110_length_440_cov_1.36364_ID_566047:33-439(+)	406

Table 1. Supp. 16S rRNA genes found in *E.coli X* strain with Barrnap tool.