Sprouting diarrhoea with bloody consequences

Polina Bogaichuk, Polina Guseva

December 6, 2022

Abstract

Once in a while out of the blue humanity faces new severe resistant pathogenic bacteria. As usually it contains too many changes in comparison to an ideal reference, *de novo* genome assembly is required to decipher virulent characteristics. Herein such a procedure is performed for *Escherichia coli* strain X from sprouts causing hemolytic uremic syndrome (HUS) outbreak in Germany 2011. Among other features annotated, Shiga toxins are responsible for extensive blood loss and multiple drug-resistance genes could be a reason for treatment difficulties. These adaptations are likely obtained by horizontal gene transfer: via phages and plasmids respectively. As an alternative, macrolides are advised.

Introduction

Most strains of *Escherichia coli* are harmless to humans. Yet the acquisition of virulence factors via horizontal gene transfer [1] by bacteriophages or plasmid exchange often leads to pathogenicity and antibiotic resistance [2]. For example, some strains can cause severe diseases, including harsh hemorrhagic diarrhoea and hematuria [3]. Pathogenic factors that often damage in such a way could be Shiga toxins. By shutting down protein synthesis [4], Shiga toxins thin blood vessel walls [5]. When the pressure becomes too high it tears capillaries up and blood is coming to the intestinal tract [6] or out of the kidneys causing hemolytic uremic syndrome (HUS) [7].

Genome assembly is an important component of many genetic studies. There are two ways to assemble a genomic sequence: mapping and assembly or *de novo* assembly [8]. The latest one has advantages when it is necessary to study the genome of a highly variable organism. *De novo* assembly of an *E. coli* strain can help establish the pathogenic factors of the hemolytic uremic syndrome and their sources. Assembling the genome of *E. coli* strain X can help establish the pathogenesis of the infection and thus improve the treatment of severe complications.

In our study, we look at 1) the *de novo* assembly of the genome of the *E. coli* strain X that caused the German HUS outbreak in 2011 [9] using three paired-end libraries with different insertion sizes to increase validity, 2) the origins of virulence and antibiotic resistance, 3) and how they are acquired.

Methods

The raw data is obtained by Illumina HiSeq reads from the TY2482 sample of *Escherichia coli* X HUS-causing the outbreak in Germany 2011. Three libraries were used: SRR292678 (paired-end, forward [10], reverse [11]), SRR292862 (mate-pair, forward [12], reverse [13]), SRR292770 (mate-pair, forward [14], reverse [15]). As for the reference to antibiotic resistance and toxins, the closest genome of *E. coli* strain released earlier than 2011 is chosen based on 16S rRNA.

To extract new features of our sample, reads should be *de novo* assembled into a genome and annotated with reference to the nearest relative or similar genes. First, to estimate approximate genome size, k-mers are counted (`jellyfish count -m 31 -s 6M -C`) and visualize (`jellyfish histo`) by Jellyfish v2.2.8-3build1 [16]. After that, using the SPAdes v3.15.5 [17] reads are compiled into contigs and scaffolds (`spades.py --pe1-1 forward --pe1-2 reverse`) in the paired-end mode and also with all three libraries (`spades.py --pe1-1 forward1 --pe1-2 reverse1 --mp1-1 forward2 --mp1-2 reverse2 --mp2-1 forward3 --mp2-2 reverse3`). The same process of genome size evaluation by Jellyfish is repeated for obtained error-corrected reads from the SPAdes output. The quality of both assemblies is checked (`quast.py`) with QUAST v5.2.0 [18]. For genome annotation and feature prediction, Prokka v1.14.6 [19] sorts through similar genes and based on that assumes characteristics of the tested strain (`prokka --centre X`). Whereas, barrnap v0.9 [20] specifically finds the 16S rRNA gene (`barrnap`) for the closest relative search. With that sequence, the previously mentioned reference is discovered by BLAST setting no later than 2011. If the opposite is not stated, the default settings are used.

With most genes annotated and the reference point located, new characteristics could be compared to explain a potential cause of the HUS outbreak. For that, modifications are aligned progressively against the reference sequence in Mauve [21]. To make an assumption of traits' origin, the surrounding genes are run through protein BLAST no later than 2011. Furthermore, our strain and the reference are analysed for antimicrobial configurations using ResFinder v4.1 [22] to identify acquired multi-drug resistance. The same procedure of origin search is applied to founded resistance genes again by Mauve.

Table 1: K-mers and reads attributes

	K (k-mer size)	**L** (average read length)	**M** (k-mer peak)	**T, bp** (total base)	**N** (depth of coverage)	**Genome size, bp***
SRR292678	31	90	125	989 882 280	187.5	5 279 372
corrected	31	90	124	989 803 080	186	5 321 948

* genome size is calculated as T / N = T / ((M * L) / (L - K + 1))

Results

Based on the FastQC report [23] quality of all raw data is way more than significant with slight deviations of GC content. The total amount of reads is ~31.4M as a sum of SRR292678 (5 499 346 x2, L=90), SRR292862 (5 102 041 x2, L=49), and SRR292770 (5 102 041 x2, L=49) which covers ~750M bp. They have insert sizes of lengths 470, 2000, and 6000 nucleotides respectively.

De novo assembly with all three libraries shows improved performance in comparison to only the pair-ended one (tab.2). The corrected reads obtained after the SPAdes assembly have an estimated genome size of 5 321 948 bp (tab.1) based on k-mers frequency (fig.3). 16S ribosomal RNA (5276_ID_565907: 111954-113492(+)) is matched with the closest organism as *E. coli* strain 55989 (NCBI Reference Sequence NC_011748.1). There are 296 contigs annotated with 10 245 coding sequences in total (fig.4).

Using Mauve, we found that the genome of the new *E. coli* strain X differs from the related genome by insertion with the genes *stxB* and *stxA*. These genes are associated with Shiga toxins with phage genes nearby (fig.1). We downloaded the data into ResFinder and determined that the new strain of *E. coli* is resistant to several antimicrobials including beta-lactams (tab.3, tab.4). Among them, absent in the reference *blaCTX-M-15* and *blaTEM-1B* genes are surrounded by some plasmid genes (fig.2).

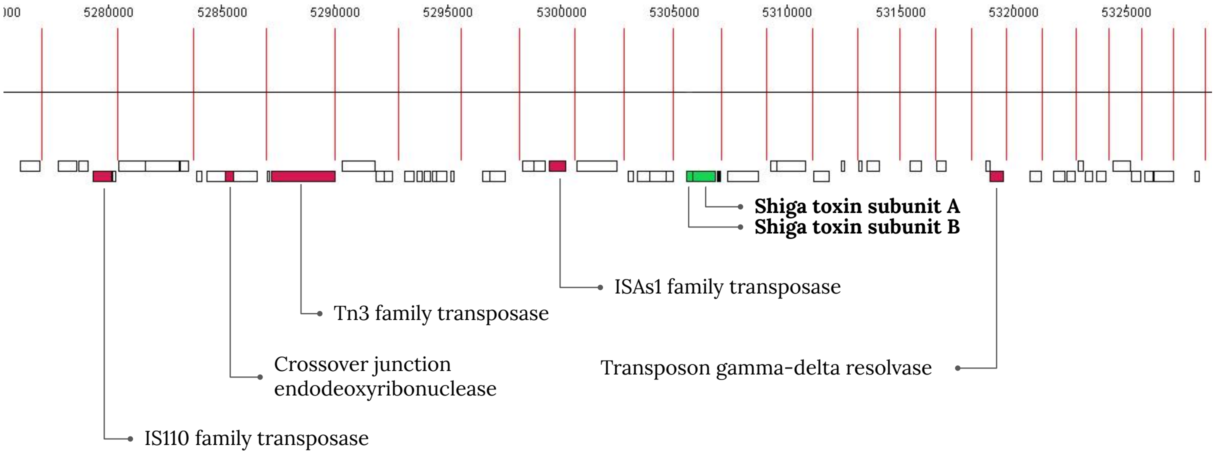

Figure 1: Shiga toxins: their gene location and surrounding

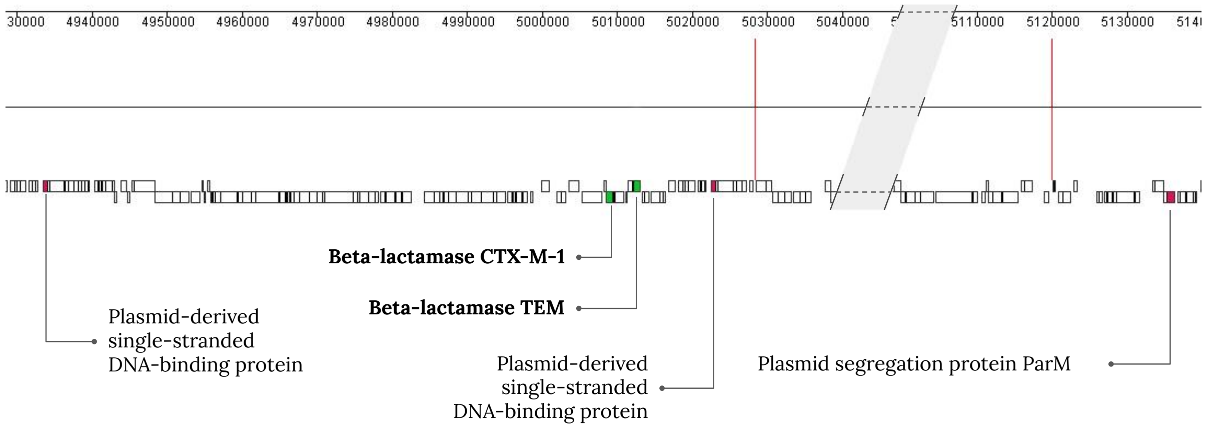

Figure 2: Beta-lactam resistance genes: their location and surrounding

Discussion

In our study, we used genome assembly from three libraries (one pair-end and two mate-pair) with different insertion sizes. That method significantly improves the quality of the genome assembly (tab.2) by resolving short repeats [24]. In addition, this correction led to a change in the distribution of k-mers (fig.3), which is associated with the filtering of some reads with errors out. Yet it does not significantly change the estimated genome size (tab.1).

Complex pathogenic traits that improve the adaptability of our *E. coli* X strain are probably acquired not by evolution from scratch but by receiving via mobile elements from the environment [25]. For example, Shiga toxins are found in the presence of prophage genes (fig.1) and have no aligned similarity within the reference. Therefore, most likely they are obtained by prophage incorporation of stx bacteriophages [26]. As for antibiotic resistance, there are dozens of different genes interacting with several classes of antimicrobial compounds (tab.3). That leads to extensive drug resistance. At least beta-lactamase (enzymes degrading beta-lactams) genes are obtained by horizontal gene transfer within plasmids as their surrounding contains plasmid-associated genes (fig.2). To discover the origin of others additional research is required. Yet it is highly likely that they are acquired by the same mechanism since the reference does not contain such genes. The possible exception is tetracycline because the reference has *tet(B)* instead of *tet(A)*. All of these features are presumably supported by evolutionary selection as they help host *E. coli* to survive and be distributed more effectively than the earlier less damaging version of our strain [27].

The reviewed strain of *E. coli* is resistant to multiple compounds and classes of drugs including beta-lactams (tab.4). Therefore, it is not wise to use such last-resort treatment in *E. coli* infection as carbapenems [28]. Hence our recommendation will be to use azithromycin as another one of the most common antibiotics [29]. Unfortunately, there is no guarantee that bacteria will not adapt to macrolides too.

Table 2: Assembly statistics: main characteristics

	single-library		three-library	
	contigs	scaffolds	contigs	scaffolds
N50	105346	105346	151014	1048022
number of contigs	522	504	543	458

References

[1] M. Petridis, M. Bagdasarian, M. Waldor, and E. Walker, “Horizontal transfer of Shiga toxin and antibiotic resistance genes among Escherichia coli strains in house fly (Diptera: Muscidae) gut,” *Journal of medical entomology*, vol. 43, no. 2, pp. 288–295, 2014. doi:10.1093/jmedent/43.2.288.

[2] S. M. Soucy, J. Huang, and J. P. Gogarten, “Horizontal gene transfer: building the web of life,” *Nature Reviews Genetics*, vol. 16, no. 8, pp. 472–482, 2015. doi:10.1038/nrg3962.

[3] S. Makvana and L. R. Krilov, “Escherichia coli infections.,” *Pediatrics in review*, vol. 36, no. 4, pp. 167–70, 2015. doi:10.1542/pir.36-4-167.

[4] H. Nakao and T. Takeda, “Escherichia coli Shiga toxin.,” *Journal of natural toxins*, vol. 9, no. 3, pp. 299–313, 2000.

[5] E. V. O’Loughlin and R. M. Robins-Browne, “Effect of Shiga toxin and Shiga-like toxins on eukaryotic cells,” *Microbes and infection*, vol. 3, no. 6, pp. 493–507, 2001. doi:10.1016/S1286-4579(01)01405-8.

[6] L. W. Riley, R. S. Remis, S. D. Helgerson, H. B. McGee, J. G. Wells, B. R. Davis, R. J. Hebert, E. S. Olcott, L. M. Johnson, N. T. Hargrett, *et al.*, “Hemorrhagic colitis associated with a rare Escherichia coli serotype,” *New England journal of medicine*, vol. 308, no. 12, pp. 681–685, 1983. doi:10.1056/NEJM198303243081203.

[7] M. A. Karmali, M. Petric, C. Lim, P. C. Fleming, G. S. Arbus, and H. Lior, “The association between idiopathic hemolytic uremic syndrome and infection by verotoxin-producing Escherichia coli,” *Journal of Infectious Diseases*, vol. 151, no. 5, pp. 775–782, 1985. doi:10.1093/infdis/151.5.775.

[8] P. C. Ng and E. F. Kirkness, “Whole genome sequencing,” *Genetic variation*, pp. 215–226, 2010. doi:10.1007/978-1-60327-367-1$_1$2.

[9] C. Frank, D. Werber, J. P. Cramer, M. Askar, M. Faber, M. an der Heiden, H. Bernard, A. Fruth, R. Prager, A. Spode, *et al.*, “Epidemic profile of Shiga-toxin–producing Escherichia coli O104: H4 outbreak in Germany,” *New England Journal of Medicine*, vol. 365, no. 19, pp. 1771–1780, 2011. doi:10.1056/NEJMoa1106483.

[10] Beijing Genome Institute (BGI), "SRR292678 forward," NCBI Sequence Read Archive (SRA), 6 2011.

[11] Beijing Genome Institute (BGI), "SRR292678 reverse," NCBI Sequence Read Archive (SRA), 6 2011.

[12] Beijing Genome Institute (BGI), "SRR292862 forward," NCBI Sequence Read Archive (SRA), 6 2011.

[13] Beijing Genome Institute (BGI), "SRR292862 reverse," NCBI Sequence Read Archive (SRA), 6 2011.

[14] Beijing Genome Institute (BGI), "SRR292770 forward," NCBI Sequence Read Archive (SRA), 6 2011.

[15] Beijing Genome Institute (BGI), "SRR292770 reverse," NCBI Sequence Read Archive (SRA), 6 2011.

[16] G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics*, vol. 27, no. 6, pp. 764–770, 2011. doi:10.1093/bioinformatics/btr011.

[17] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, *et al.*, "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing," *Journal of computational biology*, vol. 19, no. 5, pp. 455–477, 2012. doi:10.1089/cmb.2012.0021.

[18] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "QUAST: quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, 2013. doi:10.1093/bioinformatics/btt086.

[19] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, 2014. doi:10.1093/bioinformatics/btu153.

[20] T. Seemann, "Barrnap 0.9: rapid ribosomal RNA prediction." GitHub: `https://github.com/tseemann/barrnap`, 2018.

[21] A. E. Darling, A. Tritt, J. A. Eisen, and M. T. Facciotti, "Mauve assembly metrics," *Bioinformatics*, vol. 27, no. 19, pp. 2756–2757, 2011. doi:10.1093/bioinformatics/btr451.

[22] V. Bortolaia, R. S. Kaas, E. Ruppe, M. C. Roberts, S. Schwarz, V. Cattoir, A. Philippon, R. L. Allesoe, A. R. Rebelo, A. F. Florensa, *et al.*, "ResFinder 4.0 for predictions of phenotypes from genotypes," *Journal of Antimicrobial Chemotherapy*, vol. 75, no. 12, pp. 3491–3500, 2020. doi:10.1093/jac/dkaa345.

[23] S. Andrews, "FASTQC. A quality control tool for high throughput sequence data," 2010. url.

[24] J. Wetzel, C. Kingsford, and M. Pop, "Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies," *BMC bioinformatics*, vol. 12, no. 1, pp. 1–14, 2011. doi:10.1186/1471-2105-12-95.

[25] R. Jain, M. C. Rivera, J. E. Moore, and J. A. Lake, "Horizontal gene transfer accelerates genome innovation and evolution," *Molecular biology and evolution*, vol. 20, no. 10, pp. 1598–1602, 2003. doi:10.1093/molbev/msg154.

[26] H. Schmidt, "Shiga-toxin-converting bacteriophages," *Research in microbiology*, vol. 152, no. 8, pp. 687–695, 2001. doi:10.1016/S0923-2508(01)01249-9.

[27] C. Mossoro, P. Glaziou, S. Yassibanda, N. T. P. Lan, C. Bekondi, P. Minssart, C. Bernier, C. Le Bouguénec, and Y. Germani, "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*, vol. 40, no. 8, pp. 3086–3088, 2002. doi:10.1128/JCM.40.8.3086-3088.2002.

[28] P. Bajaj, N. S. Singh, and J. S. Virdi, "Escherichia coli β-lactamases: what really matters," *Frontiers in microbiology*, vol. 7, p. 417, 2016. doi:10.3389/fmicb.2016.00417.

[29] L. M. King, M. C. Lovegrove, N. Shehab, S. Tsay, D. S. Budnitz, A. I. Geller, J. N. Lind, R. M. Roberts, L. A. Hicks, and S. Kabbani, "Trends in US outpatient antibiotic prescriptions during the coronavirus disease 2019 pandemic," *Clinical Infectious Diseases*, vol. 73, no. 3, pp. e652–e660, 2021. doi:10.1093/cid/ciaa1896.

Supplementary materials

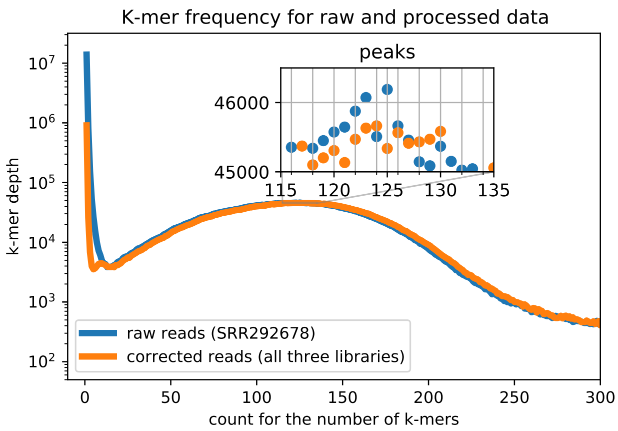

Figure 3: K-mer distribution

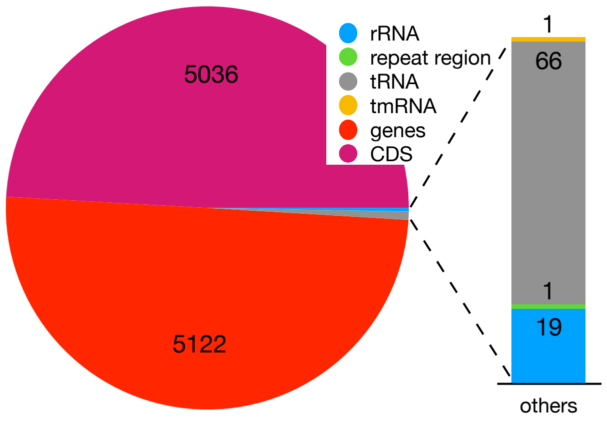

Figure 4: Annotation statistics

Table 3: Genes associated with antimicrobial resistance

Resistance gene	**Length Alignment/Gene**	**Position in contig**	**Phenotype resistance**	**Accession no.**
blaCTX-M-15[*]	876/876	81599-82474	Beta-lactam	AY044436
blaTEM-1B[**]	861/861	85296-86156	Beta-lactam	AY458016
aph(6)-Id	837/837	2262-3098	Aminoglycoside	M28829
aph(3”)-Ib	804/804	3098-3901	Aminoglycoside	AF321551
dfrA7	474/474	11442-11915	Trimethoprim	AB161450
tet(A)	1200/1200	956-2155	Tetracycline	AJ517790
sul1	761/840	10111-10871	Sulphonamide	AY115475
sul1	761/828	10111-10871	Sulphonamide	AY522923
sul2	816/816	3962-4777	Sulphonamide	HQ840942
sul1	761/840	10111-10871	Sulphonamide	U12338
sul1	761/882	10111-10871	Sulphonamide	DQ914960
qacE	282/333	10931-11212	Disinfectant	X68232

[*] alternative name: *UOE-1*

[**] alternative name: *RblaTEM-1*

Table 4: Antimicrobial resistance

Antimicrobial compound	**Class**	***E. coli* X**	***E. coli* 55989**	**Genetic background**
Amoxicillin	Beta-lactam	Resistant	No resistance	*blaCTX-M-15*, *blaTEM-1B*
Ampicillin	Beta-lactam	Resistant	No resistance	*blaCTX-M-15*, *blaTEM-1B*
Aztreonam	Beta-lactam	Resistant	No resistance	*blaCTX-M-15*
Benzylkonium chloride	Quaternary ammonium compound	Resistant	No resistance	*qacE*
Cefepime	Beta-lactam	Resistant	No resistance	*blaCTX-M-15*
Cefotaxime	Beta-lactam	Resistant	No resistance	*blaCTX-M-15*
Ceftazidime	Beta-lactam	Resistant	No resistance	*blaCTX-M-15*
Ceftriaxone	Beta-lactam	Resistant	No resistance	*blaCTX-M-15*
Cephalothin	Beta-lactam	Resistant	No resistance	*blaTEM-1B*
Cetylpyridinium chloride	Quaternary ammonium compound	Resistant	No resistance	*qacE*
Chlorhexidine	Quaternary ammonium compound	Resistant	No resistance	*qacE*
Doxycycline	Tetracycline	Resistant	Resistant	*tet(A)*[*]
Ethidium bromide	Quaternary ammonium compound	Resistant	No resistance	*qacE*
Minocycline	Tetracycline	No resistance	Resistant	[*]
Piperacillin	Beta-lactam	Resistant	No resistance	*blaCTX-M-15*, *blaTEM-1B*
Streptomycin	Aminoglycoside	Resistant	No resistance	*aph(6)-Id*, *aph(3”)-Ib*
Sulfamethoxazole	Folate pathway antagonist	Resistant	No resistance	*sul1* (all), *sul2*
Tetracycline	Tetracycline	Resistant	Resistant	*tet(A)*[*]
Ticarcillin	Beta-lactam	Resistant	No resistance	*blaCTX-M-15*, *blaTEM-1B*
Trimethoprim	Folate pathway antagonist	Resistant	No resistance	*dfrA7*

[*] for the reference that is *tet(B)*