HGT augments *E. Coli* strain virulence

Vsevolod Zvezdin, Nikita Vyatkin

December 2022

1 Abstract

In this study, we figured out the chain of events that led a strain of *E. Coli* to obtain pathogenic properties along with antibiotic resistance. Using several libraries of Illumina paired-end reads, we produced a scaffold-level genome assembly and found these new harmful bacteria to derive from a non-pathogenic strain 55989. After having done genome-to-genome alignment we were able to identify newly acquired genes as well as bacteriophage DNA in the same region of the genome. We suspect this *E. Coli* variant arose as a result of horizontal gene transfer that had a beneficial effect on the bacteria and was supported in a selection process.

2 Introduction

E. Coli inhabits the lower digestive tract of various animals including humans. Being a part of the normal microflora, it is involved in such processes as vitamin B12 and K synthesis and consuming oxygen to maintain anaerobic conditions for other symbiotic bacteria in the gut. [1] However, this widely distributed species has got a lot of strains and serotypes, many of which bear pathogenic properties and might endanger the life of their host. Such strains are characterized by the presence of virulence factors, various molecules that enhance their ability to infect the host or inhabit them. These molecules might, as well, endanger the life of the infected person. [2]

One example of a condition brought by them is a hemolytic uremic syndrome. People experiencing it see their blood cells disrupted which leads to hemorrhage and associated symptoms (kidney failure etc.). This is typically caused by the presence of Shiga-like toxins, proteins produced by some *E. Coli* strains, similar to the toxins of *S. dysenteriae*, dysentery agent. In humans, these proteins are able to bind the receptors located mainly on the endothelial cells' surface. This allows bacteria to penetrate the blood vessels surrounding the intestines, which results in bloody diarrhea and later other symptoms. [2]

A common cause of acquiring virulence factors is horizontal gene transfer. Bacteria are known to exchange fragments of their DNA, the plasmids, with each other, procure DNA pieces from the environment or receive it from bacteriophages, viruses that inhabit bacteria. Brought-in DNA can have next to any possible properties, and if they are useful for the cell, there are chances that these sequences will stay in the population. It might result in rapid changes in bacterial phenotypes, notably in their virulence and toxicity, giving rise to new deadly strains. [3]

Thus, it is important to address these events and detect the origins of pathogenic strains to learn what made them such in the first place to have a notion of a possible treatment. Another problem taking place often is that genes encoding virulence factors co-occur with others that benefit bacteria directly, namely those related to antibiotic resistance. [3] It can happen thus that people infected with these pathogens will not respond to standard therapy, and it might cost the time required to save patients' lives. As such, precise tracking of pathogen lineage is needed, and often it can best be done through de novo sequencing followed by the search for the closest relative strain - the approach we selected in our study.

3 Methods

For this project, we used paired-end Illumina reads of a pathogenic *E. Coli* strain (links can be accessed in the Lab Notebook). Three libraries with insert sizes of 470 bp, 2k bp, and 6k bp were taken for genome assembly to resolve repeats of different lengths in the DNA. We applied FastQC to reads to check their quality before moving on with the assembly. [4] At this stage, we also drew the k-mers distribution with the Jellyfish tool to estimate the size of the genome using its -count key to count k-mer number (for 31-mers) and then the -histo key. [5]

The genome was assembled de novo from the SRR292678 library reads using SPAdes tool. [6] An additional run was performed that included mate-pair reads to provide information for scaffolds building. The resulting assembly quality was assessed with QUAST analyzer which was run on both contigs and scaffolds. [7]

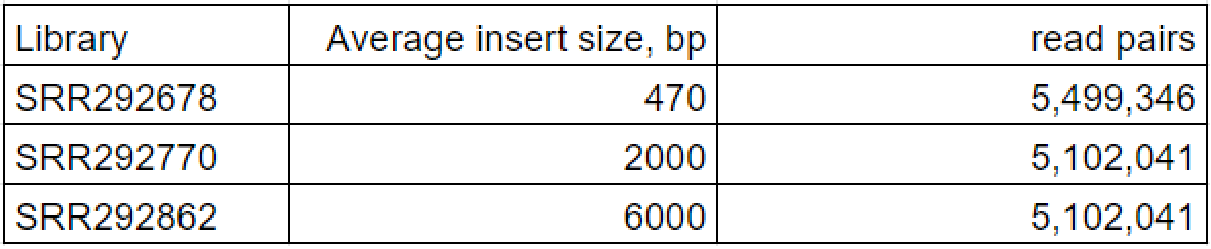
Afterward, we used PROKKA for rapid genome annotation. [8] The conservative 16S rRNA gene sequence was extracted from the annotation data with the Barrnap tool. [9] We then searched for similar sequences in NCBI refseq_genomes database (among *E. Coli* records) with Nucleotide BLAST to determine the most likely ancestral *E. Coli* strain, limiting ourselves to entries that were known before the year 2011. We chose one that held the highest similarity (100%) to the studied pathogen and used it for all subsequent analyses. [10]

We then compared our scaffold-level assembly to the completed genome of the 55989 strain, loading both into the Mauve genome aligner. [11] We looked for any additions in a new strain that contributed to its obtained pathogenic properties. In addition, we aligned the ResFinder database against our assembly (the ResFinder contains sequences that are linked to antibiotic resistance). [12] The respective genes were found and analyzed in the Mauve aligner as well.

4 Results

The number of reads is presented in Table 1.

Table 1. The number of reads in sequencing libraries.



Library	Average insert size, bp	read pairs
SRR292678	470	5,499,346
SRR292770	2000	5,102,041
SRR292862	6000	5,102,041

Reads in the libraries met all the quality standards, as was assessed with the FastQC. The total number of 31-mers in reads was equal to 643044370 according to JellyFish -count. The occurrence number peaked at 125 (see distribution in figure 1), therefore the estimated genome length was thus taken as 5144355.

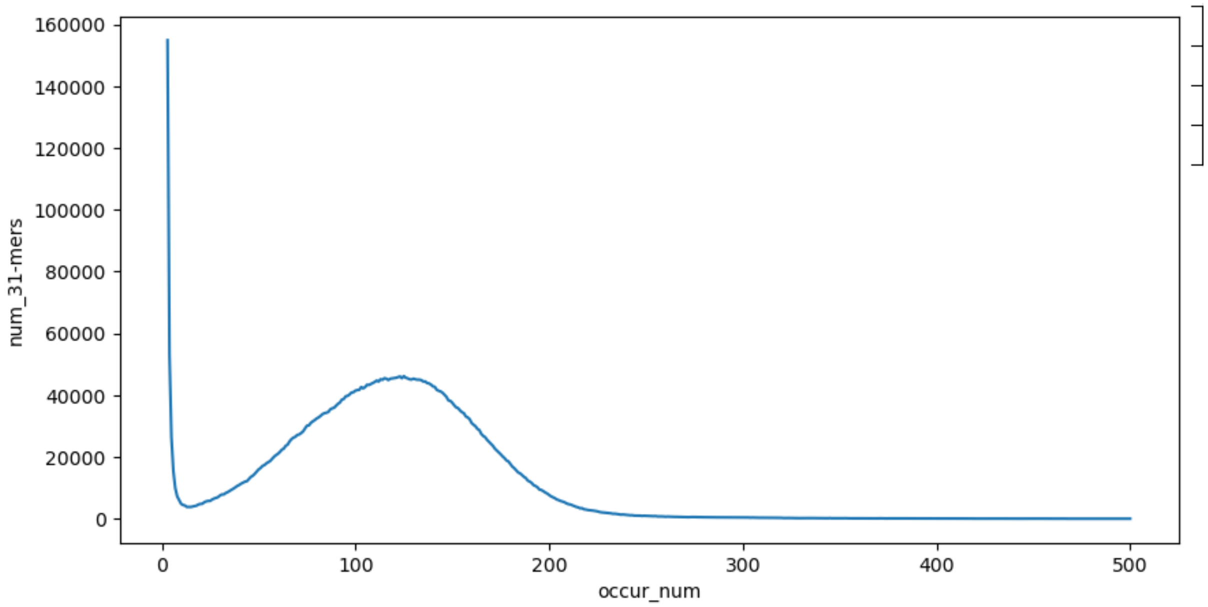
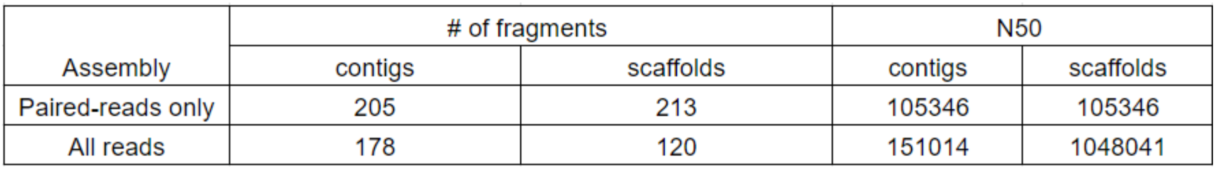


Figure 1. Histogram showing the distribution of occurrence number of various 31-mers

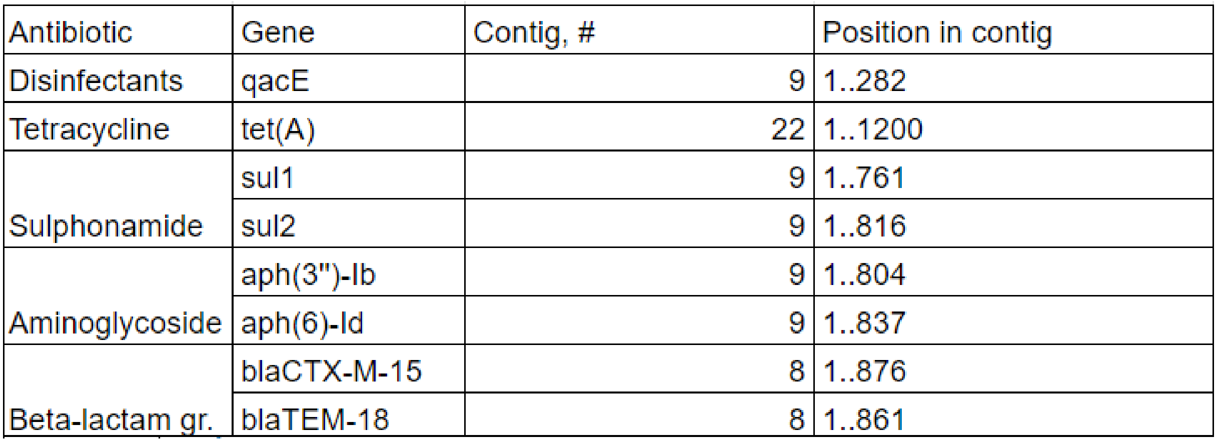
Using the insert-size data of mate-pair reads allowed us to combine contigs to build the scaffolds and improve the quality of our assembly. This became possible due to long repeats in bacterial DNA resolved by larger inserts that were able to cover all their length. As can be seen in table 2, half of the genome can be covered by significantly longer fragments, and their total amount is twice as less.

Table 2. The assembly parameters in QUAST output.



Assembly	# of fragments		N50	
	contigs	scaffolds	contigs	scaffolds
Paired-reads only	205	213	105346	105346
All reads	178	120	151014	1048041

After performing annotation we were able to identify seven copies of the 16S rRNA gene with Barrnap, all of them having an identical length of 1537 nucleotides. On using this gene



Antibiotic	Gene	Contig, #	Position in contig
Disinfectants	qacE	9	1..282
Tetracycline	tet(A)	22	1..1200
Sulphonamide	sul1	9	1..761
	sul2	9	1..816
Aminoglycoside	aph(3")-Ib	9	1..804
	aph(6)-Id	9	1..837
Beta-lactam gr.	blaCTX-M-15	8	1..876
	blaTEM-18	8	1..861

sequence in nucleotide BLAST, we found it to be 100% identical to the variant in strain 55989 (accession number NC_011748.1).

Muave annotation allowed us to find the new additions that the hemorrhagic strain possesses, neither of which were aligned to a reference 55859 genome: Shiga toxin subunit A, 89 amino acids Shiga toxin subunit B, 317 amino acids Both of these units were aligned not far from each other along with genes that are leftovers of bacteriophages.

Also having looked in our assembly for any genes possibly related to antibiotics in ResFinder, we received the next results (Table 3.)

Table 3. Hits in the alignment of ResFinder database genes against the assembled strain.

Bla genes presented a particular interest as they are associated with resistance against the beta-lactam group, the most commonly used antibiotics. After similarly analyzing them in Muave we observed features that are typically found within plasmids, such as transposon resolvase CDS. Besides these two, situated in the 8th contig, there were also protective genes in contigs 9 and 22. As all contigs are sorted by their size in descending order, it can be assumed that these smaller contigs are plasmids as well that provide some resistance.

5 Discussion

The hallmarks of the horizontal gene transfer we found in *E. Coli* genome show us that one or several such events indeed have taken place. Most probably, Shiga toxin genes could be introduced by bacteriophage bearing their sequences. It is known that *Shigella* is, in fact, a subgenus of *Escherichia*, and it is trivial for these two groups of bacteria to coinhabit the same host or environment. [13] For example, a lot of cattle can also serve as a host to *E. Coli* or *Shigella* but they do not have the receptors that bind Shiga toxins. [13] As such, they are not endangered by the pathogens, instead becoming reservoirs for them to multiply in. Under these conditions, there could be a place for gene exchange in a variety of ways.

Unlike her parental strain, *E. Coli* also seems to have obtained protection against beta-lactam antibiotics, borrowing a plasmid from some other bacteria. These two events might be independent, yet there also exists such a phenomenon as joint fixation. One mutation (in this case probably an acquired plasmid DNA) that greatly enhances the survival also ”pulls ahead” all the other mutations that occurred. When the subpopulation outperforms everyone else, their other mutations are fixed as well in the end. It can not be clearly deducted from our analysis but apparently, both these changes were acquired not far in time from each other and it might be a hint of such development. [14]

This strain is able to cleave the beta-lactam ring using beta-lactamase enzyme encoded *bla* genes. Currently, the standard treatment includes the adjuvant, clavulanic acid, which is able to overcome such resistance by acting as a competitive substrate. These particular groups of lactamases (CTX-M and TEM) seem to be susceptible to clavulanic acid inhibitory action. [15]

Our lab notebook and files associated with the project can be found on GitHub:
https://github.com/Nik9kin/workshopsBI/blob/main/Project3

References

[1] Zachary D Blount. “The unexhausted potential of E. coli”. In: *eLife* 4 (2015), e05826. DOI: 10.7554/eLife.05826.

[2] Marina S Palermo, Ramón A Exeni, and Gabriela C Fernández. “Hemolytic uremic syndrome: pathogenesis and update of interventions”. In: *Expert Review of Anti-infective Therapy* 7.6 (2009). PMID: 19681698, pp. 697–707. DOI: 10.1586/eri.09.49. eprint: https://doi.org/10.1586/eri.09.49. URL: https://doi.org/10.1586/eri.09.49.

[3] Melissa Emamalipour et al. “Horizontal Gene Transfer: From Evolutionary Flexibility to Disease Progression”. In: *Frontiers in Cell and Developmental Biology* 8 (2020). ISSN: 2296-634X. DOI: 10.3389/fcell.2020.00229. URL: https://www.frontiersin.org/articles/10.3389/fcell.2020.00229.

[4] *FastQC*. June 2015. URL: https://qubeshub.org/resources/fastqc.

[5] Guillaume Marçais and Carl Kingsford. “A fast, lock-free approach for efficient parallel counting of occurrences of k-mers”. In: *Bioinformatics* 27.6 (Jan. 2011), pp. 764–770. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr011. eprint: https://academic.oup.com/bioinformatics/article-pdf/27/6/764/16902460/btr011.pdf. URL: https://doi.org/10.1093/bioinformatics/btr011.

[6] Andrey Prjibelski et al. “Using SPAdes De Novo Assembler”. In: *Current Protocols in Bioinformatics* 70.1 (2020), e102. DOI: https://doi.org/10.1002/cpbi.102. eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpbi.102. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.102.

[7] Alexey Gurevich et al. “QUAST: quality assessment tool for genome assemblies”. In: *Bioinformatics* 29.8 (Feb. 2013), pp. 1072–1075. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt086. eprint: https://academic.oup.com/bioinformatics/article-pdf/29/8/1072/17106244/btt086.pdf. URL: https://doi.org/10.1093/bioinformatics/btt086.

[8] Torsten Seemann. “Prokka: rapid prokaryotic genome annotation”. In: *Bioinformatics* 30.14 (Mar. 2014), pp. 2068–2069. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu153. eprint: https://academic.oup.com/bioinformatics/article-pdf/30/14/2068/7250406/btu153.pdf. URL: https://doi.org/10.1093/bioinformatics/btu153.

[9] Torsten Seemann. *Barrnap*. github. https://github.com/tseemann/barrnap. 2013.

[10] Stephen F. Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (1990), pp. 403–410. ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2. URL: https://www.sciencedirect.com/science/article/pii/S0022283605803602.

[11] Aaron E. Darling et al. “Mauve: multiple alignment of conserved genomic sequence with rearrangements.” In: *Genome research* 14 7 (2004), pp. 1394–403.

[12] Valeria Bortolaia et al. “ResFinder 4.0 for predictions of phenotypes from genotypes”. In: *Journal of Antimicrobial Chemotherapy* 75.12 (Aug. 2020), pp. 3491–3500. ISSN: 0305-7453. DOI: 10.1093/jac/dkaa345. eprint: https://academic.oup.com/jac/article-pdf/75/12/3491/34291500/dkaa345.pdf. URL: https://doi.org/10.1093/jac/dkaa345.

[13] Lothar Beutin et al. “Identification of Human-Pathogenic Strains of Shiga Toxin-Producing ¡i¿Escherichia coli¡/i¿ from Food by a Combination of Serotyping and Molecular Typing of Shiga Toxin Genes”. In: *Applied and Environmental Microbiology* 73.15 (2007), pp. 4769–4775. DOI: 10.1128/AEM.00873-07. eprint: https://journals.asm.org/doi/pdf/10.1128/AEM.00873-07. URL: https://journals.asm.org/doi/abs/10.1128/AEM.00873-07.

[14] Marius Moeller, Benjamin Werner, and Weini Huang. “Accumulating waves of random mutations before fixation”. In: *bioRxiv* (2022). DOI: 10.1101/2022.06.04.494827. eprint: https://www.biorxiv.org/content/early/2022/06/05/2022.06.04.494827.full.pdf. URL: https://www.biorxiv.org/content/early/2022/06/05/2022.06.04.494827.

[15] Catherine L. Tooke et al. “-Lactamases and -Lactamase Inhibitors in the 21st Century”. In: *Journal of Molecular Biology* 431.18 (2019). The molecular basis of antibiotic action and resistance, pp. 3472–3500. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2019.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0022283619301822.