BIOINFORMATICS INSTITUTE

Bio-informatics Practice

November 25, 2023

Unraveling the Origins: Investigating the Cause of a Gastrointestinal Infection Outbreak and Characterizing the Properties of a Novel Pathogenic Strain

by Artem Vasilev and Tatiana Lisitsa

Abstract

Horizontal gene transfer (HGT) is one of the key mechanisms of bacterial evolution, facilitating the exchange of genetic material between organisms. In this work, we investigate what cause the *E. coli* outbreak, and describe properties a novel pathogenic strain of *E. coli X* using next-generation sequencing and *de novo* assembly. Our comprehensive analysis revealed high similarity between *E. coli X* and *E. coli* serotype O104:H4, highlighting the importance of HGT in shaping bacterial genomes. The identification of genes encoding Shiga toxin (*stxA* and *stxB*) flanked by transposase genes and genes encoding phage proteins allowed us to elucidate the process of acquisition of virulence factors that possibly account for the pathogenicity of the strain. In addition, our study revealed the antibiotic resistance profile of *E. coli X*.

Keywords: *E. coli*, HGT, HUS, Shiga toxin

Introduction

Horizontal gene transfer (HGT) is the exchange of genetic material between organisms that are not in a parent-offspring relationship. HGT is a characteristic feature of bacterial evolution, and evidence of it can be found in the majority of bacterial genomes. While HGT can significantly alter bacterial genomes, not all transfer events may be biologically significant. Nevertheless, adaptive transfer and positive selection can lead to notable properties in bacteria, such as novel mechanisms of pathogenicity and antibiotic resistance [1, 2].

Escherichia coli (*E. coli*) is a bacterium commonly found in the lower intestines of warm-blooded organisms. Most strains of *E. coli* are harmless. However, there are pathogenic strains, such as enteroaggregative *E. coli* (EAEC) serotype O104:H4 or enterohemorrhagic *E. coli* (EHEC). Shigatoxin-producing *E. coli* (STEC) can cause life-threatening illness, including hemolytic uremic syndrome (HUS), especially in young children and the elderly [3, 4].

When aligning sequences to a reference genome, the genetic material acquired by bacteria as a result of horizontal gene transfer may go unnoticed, and important features of a particular strain may be missed. Therefore, *de novo* assembly is necessary. The aim of this study is to investigate a novel strain of pathogenic *E. coli* using NGS-based approach, followed by data processing involving genome *de novo* assembly.

Materials and methods

We used three libraries from the TY2482 sample deposited in the Short Read Archive (SRA) for public access: SRR292678 - paired end, insert size 470 bp; SRR292862 – mate pair, insert size 2 kb; and SRR292770 – mate pair, insert size 6 kb. Links for forward and reverse reads can be found in lab notes on the GitHub page. Then we ran FastQC [5] on all 6 files to find out the number of reads and MultiQC [6] in order to summarize the obtained results. To count k-mers, we used the Jellyfish program (k-mer sizes of 31; -s 500M for paired end and -s 260M for mate pairs) [7]. Genome of *E. coli X* strain was assembled with SPAdes assembler in the in the paired-end and combined modes [8]. The resulting assembly was assessed with N50 statistic using QUAST [9]. Visualisation was performed with python script. Prokka used for gene prediction and annotation [10]. To find the known genome that is the most similar to the pathogenic strain we located 16S rRNA genes in the assembled *E. coli X* genome with Barrnap tool [11] and performed BLAST [12] in the RefSeq database (to simulate the situation in 2011 we changed parameter PDAT in the "Entrez Query" field: 1900/01/01:2011/01/01[PDAT]). After finding the closest relative we found a region where *E. coli X* encodes a new virulence factors and a new genes responsible for antibiotic resistance using Mauve program [13]. To search for genes responsible for antibiotic resistance, we used ResFinder web-site. The versions of the programs used can be also found on the GitHub page in the environment.yaml file.

Results

From the FastQC report, we determined that the raw data files of paired-end reads contain a total of 5,499,346 reads, with an average length of 90 base pairs. All reads exhibit good quality.

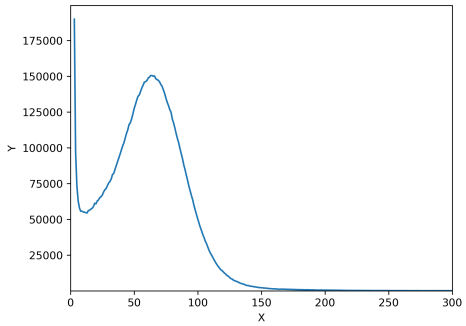
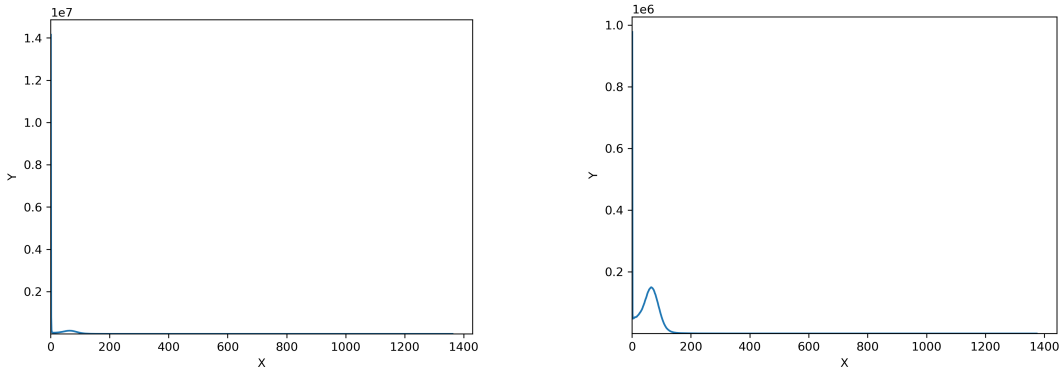


Figure 1: k-mer distribution

Knowing the k-mer distribution, we can estimate the genome size:

- M (k-mer peak) = 64
- K (k-mer size) = 31
- L (avg read length) = 90
- T (total bases) = 5499346 * 90
- N (depth of coverage) = (M*L)/(L-K+1)
- genome size = T/N = 5,155,637



(a) before correction

(b) after correction

Figure 2: k-mer distribution before and after correction

We repeated the k-mer profile plotting step for the data corrected by SPAdes during the assembly and compared it with the profile for uncorrected reads. There were far fewer single k-mers after correction. The difference in k-mer profiles reflects the impact of sequencing error correction on data quality and, consequently, on subsequent genome assembly. A more accurate k-mer profile derived from corrected reads contributes to more accurate and reliable genome assembly.

As we can see in Table 1, the mate-pair reads provide additional information about the relative positions and orientations of DNA fragments, especially in regions with repeats or structural

Table 1: Main assembly metrics

	N50 scaffolds	number of contigs
single-library	105346	212
three-libraries	1048022	117
single-library provided	111860	221
three-libraries provided	2815616	90

variations. This additional information helps the assembly algorithm to correctly scaffold and order contigs, resulting in a more accurate and contiguous assembly.

We observed that *E. coli X* strain exhibits the highest similarity to enteroaggregative *E. coli* serotype O104:H4 (Escherichia coli 55989, NCBI Reference Sequence: NC_011748.1). Strain 55989 was isolated in the Central African Republic from a stool sample obtained from a male with persistent diarrhea. However, strain 55989 did not demonstrate enterohemorrhagic properties or the ability to induce HUS. In contrast, *E. coli X* strain possesses these properties due to the presence of genes stxA and stxB encoding Shiga toxins, acquired most likely through HGT. The Shiga toxin genes are flanked by transposase genes and genes encoding phage proteins. We also discovered that *E. coli X* strain carries antibiotic resistance genes for various classes of antibiotics (Table 2), including tetracyclines (tetracycline, doxycycline), unprotected beta-lactams (amoxicillin, ampicillin, cefepime, cefotaxime, ceftazidime, piperacillin, aztreonam, ticarcillin, ceftriaxone, cephalothin), aminoglycosides (streptomycin), and folate pathway antagonists (sulfamethoxazole, trimethoprim), while the strain 55989 was only resistant to tetracyclines.

Table 2: Antibiotic resistance

Antibiotics	Resistance genes
aminoglycoside	*neo*
beta-lactam	*ampC*, *bla1*, *bla2*
tetracycline	*tetA*, *tetR*

Discussion

Based on our studies, we can hypothesize that *E. coli* acquired pathogenic properties and antibiotic resistance during HGT. The transposase genes and genes encoding phage proteins flanking the Shiga toxin genes indicate that this genome section is a mobile element. Additionally, resistance to aminoglycosides may be acquired through HGT. The *neo* gene encodes the enzyme Aminoglycoside 3'-phosphotransferase, which phosphorylates and inactivates many aminoglycoside antibiotics. The genes *ampC*, *bla1*, and *bla2* encode beta-lactamases, proteins that cleave the beta-lactam ring of penicillin antibiotics. The *tetA* gene encodes a transporter protein that enhances the tetracyclin efflux from the cell, and the *tetR* gene is a transcriptional repressor required for the inducible expression of the tetracycline "efflux pump" tetA. Protected penicillins (such as amoxiclav), macrolides, and fluoroquinolones can be employed to treat such infections.

In conclusion, we would like to highlight the importance of such studies in the sphere of molecular epidemiology for monitoring infections, discovering novel mechanisms of antibiotic resistance, and developing new antimicrobial agents.

Supplementary materials

GitHub

References

[1] Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. "Horizontal gene transfer: building the web of life". In: *Nature Reviews Genetics* 16.8 (2015), pp. 472–482.

[2] Brian J Arnold, I-Ting Huang, and William P Hanage. "Horizontal gene transfer and adaptive evolution in bacteria". In: *Nature Reviews Microbiology* 20.4 (2022), pp. 206–218.

[3] Fernando Navarro-Garcia. "Escherichia coli O104: H4 pathogenesis: an enteroaggregative E. coli/Shiga toxin-producing E. coli explosive cocktail of high virulence". In: *Enterohemorrhagic Escherichia coli and Other Shiga Toxin-Producing E. coli* (2015), pp. 503–529.

[4] WHO. *E. coli*. `https://www.who.int/news-room/fact-sheets/detail/e-coli`. Accessed: November 25, 2023.

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

[6] Philip Ewels et al. "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19 (June 2016), pp. 3047–3048. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btw354`. eprint: `https://academic.oup.com/bioinformatics/article-pdf/32/19/3047/49021359/bioinformatics\_32\_19\_3047.pdf`. URL: `https://doi.org/10.1093/bioinformatics/btw354`.

[7] Guillaume Marçais and Carl Kingsford. "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers". In: *Bioinformatics* 27.6 (2011), pp. 764–770.

[8] Anton Bankevich et al. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". In: *Journal of computational biology* 19.5 (2012), pp. 455–477.

[9] Alexey Gurevich et al. "QUAST: quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8 (2013), pp. 1072–1075.

[10] Torsten Seemann. "Prokka: rapid prokaryotic genome annotation". In: *Bioinformatics* 30.14 (2014), pp. 2068–2069.

[11] Torsten Seemann. *Barrnap: bacterial ribosomal RNA predictor*. 2018.

[12] Stephen F Altschul et al. "Basic local alignment search tool". In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[13] Aaron CE Darling et al. "Mauve: multiple alignment of conserved genomic sequence with rearrangements". In: *Genome research* 14.7 (2004), pp. 1394–1403.