Bioinformatics
doi.10.1093/bioinformatics/xxxxxx

Genome analysis

E. coli outbreak investigation. De novo assembly and annotation of bacterial genomes.

Bogdanova Lilia[1,3], Vlasenko Daniil[2,3]

[1]Sechenov University, Specialist in Bioengineering and Bioinformatics, Moscow, Russia, E-mail: liliia.bgdnv@gmail.com
[2]St. Petersburg State University, School of Mathematics and Mechanics, St. Petersburg, Russia, E-mail: vlasnko.daniil.vl@gmail.com
[3]Bioinformatics Institute.

Abstract

In this paper, we studied how *E.coli* could lead to an epidemic in Germany in 2011 that caused dozens of deaths. For this purpose we have taken the steps of assembling the genome of the *E.coli X* strain, comparing it to another *E.coli* strain, determining which genes our strain has acquired and how it became pathogenic. During the work we found out that strain *E.coli X* acquired the genes stxB and stxA from the phage, which led to the ability to secrete shiga toxin. We also found the bla_1 and bla_2 genes encoding beta-lactamase in the CTX-M family and TEM family, respectively.
Supplementary information: https://github.com/Daniil-Vlasenko/IBBioinformaticsWorkshop/blob/main/Project%203

1 Introduction

The acquisition of the stxB and stxA genes by E. coli made the shiga toxin-producing *E.coli* strain (STEC) [1]. Shiga toxins (Stx) of STEC are typically encoded in the genome of lambdoid bacteriophages. These bacteriophages spend the majority of their life integrated as prophages in certain places of the bacterial chromosome. The Stx genes are co-transcribed during spontaneous induction as well as induction by chemical or physical stimuli. The mature bacteriophage particles are built in the cytoplasm, and when the host cell has been lysed, they are discharged into the environment together with Stx. Several prophages, either cryptic or functional, may also be present in the genome of pathogenic STEC bacteria in addition to Stx phages. In addition to the genes required for phage life, these prophages may also contain foreign genes, some of which are connected to virulence. The primary pathogenicity component of STEC is thought to be the development of one or more Stx [2].

The majority of STEC infections that result in severe consequence like hemolytic uremic syndrome (HUS) are brought on by pathogens that carry the locus of enterocyte effacement (LEE) [3]. Stx phages are very mobile elements that facilitate the horizontal gene transfer of stx genes and as a result help novel pathotypes to develop. The strain of E.coli that was responsible for the severe enterohemorrhagic E.coli outbreak that occurred in Germany in 2011 has the highest genome sequence similarity to enteroaggregative *E.coli*, and it lacks the locus of enterocyte effacement (LEE) pathogenicity island that is present in most isolates of enterohemorrhagic *E.coli* [4]. The exchanging of genetic information between species who do not have a parent-offspring connection is known as horizontal gene transfer (HGT). HGT is a well-known adaptation process in bacteria and archaea. HGT is frequently linked to microbial pathogenicity and antibiotic resistance [5].

In our study, we used genome assembly from three paired-end read libraries with different insertion sizes. In cases where the genome contains a large number of repeats of different lengths, it is preferable to use genome assembly rather than alignment to the reference genome, because the reads can align to several parts of the reference genome at once. So libraries with small insert sizes are better suited for resolution of short repeats, libraries with large insert sizes are better suited for resolution of long repeats.

2 Methods

We took forward and reverse pair end reads of length 470bp [6, 7] and forward and reverse mate pare reads of length 2kb [8, 9] and 6kb [10, 11]. We estimated genome size with Jellyfish [12] and this formula:

$$N = \frac{M - L}{L - K + 1}, \quad SizeOfGenome = \frac{T}{N},$$

where $M = 125$ is peak of 31-mers, $K = 31$ is length of k-mers, $L = 32.57$ is average length of k-mers, $N = 35.96$ is average depth, T is total number of bases.

We used SPAdes [13] and QUAST [14] for assembling genome and estimating results of assembling, then we used Prokka [15] for annotation genome and prediction of genes.

We found the closest relative of E. coli X by selecting one important and evolutionarily conserved gene for comparison with all other sequenced genomes — 16S ribosomal RNA gene. We used Barrnap [16] for localisation 16S ribosomal RNA genes, then we searched most relative genome with NCBI Blast [17]. To restrict our search to only those genomes

Table 1. Results of the search for causes of new pathogenic properties

gene	start position	end position
stxB	3483605	3483874
srxA	3483886	3484845
Phage DNA adenine methylase	3485228	3485380
Phage antitermination protein Q	3485629	3486063
Phage head completion protein	3480580	3480825

Table 2. Results of the search for causes of β-lactomase antibiotic resistance

gene	start position	end position
bla_1	5195566	5196441
bla_2	5199263	5200123
Plasmid conjugative transfer protein TraA	5185482	5185667
Plasmid conjugative transfer protein TraC	5183570	5184253
Plasmid SOS inhibition protein PsiB	5211132	5211569

that were present in the GenBank database at the beginning of 2011, we set the time range using parameter PDAT: 1900/01/01:2011/01/01.

We found out with ResFinder [18] that new strain resistant to β-lactamase antibiotics. We used Mauve [19] to search for the causes of new pathogenic properties of the strain and the causes of resistance to β-lactamase antibiotics.

3 Results

Our libraries contain 5499346, 5499346, 5102041, 5102041, 5102041 and 5102041 respectively. Estimate size of genome is 4980915.

Assembling the genome with short reads resulted in N50 equal to 105346 and the number of contigs ($\geq$ 500bp) equal to 205, and assembling the genome with all libraries of reads resulted in N50 equal to 151014 and the number of contigs ($\geq$ 500bp) equal to 170. Note that assembling with reads of different length improved results of single-library assembling.

Barrnap found 8 16S ribosomal RNA genes (See more details in the supplementary information). NCBI Blast found with the first of the 16S ribosomal RNA genes that our strain more likely relieves to Escherichia coli 55989 (NC_011748.1).

Mauve found stxB and srxA genes that are associated with shiga toxin and phage genes next to them. The results are presented in Table 1.

With ResFinder we found out that new strain of E. coli is resistant to β-lactomase antibiotics. Mauve found the bla_1, bla_2 genes that are related to beta lactamase. The region with bla_1 and bla_2 did not align to the reference and contained plasmid's genes. The results are presented in Table 2.

4 Discussion

4.1 Shiga toxin

The study found that the E. coli strain became pathogenic because of the stxB (348360-3483874) and stxA (3483886-348484845) genes encoding A and B subunits which formed stx [20]. A non covalently joined A subunit that forms a pentamer with five identical B subunits. The toxin's A component damages the eukaryotic ribosome and prevents target cells from synthesizing proteins. The B pentamer's job is to attach to the Gb3 globotriaosylceramide receptor, which is largely present on endothelial cells. The Stxs migrate retrogradely throughout the cell, thus the toxin's A component doesn't enter the cytosol until it has passed from the endosome to the Golgi to the endoplasmic reticulum [21].

The emergence of genes is caused by the lysogenic cycle. During the lysogenic cycle, the bacteriophage bound to a receptor on the surface of *E.coli*, the phage DNA was introduced into the cytoplasm, then recombination occurred into the bacterial chromosome, resulting in the integration of the phage DNA into the chromosome and transforming *E.coli* into STEC.

Stx-phage induction and, in turn, the expression of Stx by the bacterial hosts are both influenced by the environmental factors present in the human body. Toxins can enter the bloodstream and penetrate intestinal epithelial cells by mechanisms such as bacterial invasion, destruction of epithelial cells, or transcytosis, where they bind to the surface of carrier cells like neutrophils or blood monocytes with low affinity. Leukocytes or cellular microvesicles connected to leukocytes may transfer the toxins to susceptible endothelial cells once they are in the microvasculature serving the target organs through interactions with a membrane Gb3, which has a high affinity for the toxins [2] [21].

4.2 Antibiotic resistance

The E. coli strain we studied has resistance to 17 antibiotics, among which the largest group is β-lactam antibiotics resistance to which is due to the acquisition of a specific enzyme, β-lactamase, which is encoded by genes called bla. The main factor of gram-negative bacteria's resistance to β-lactam antibiotics is the synthesis of β-lactamases. These enzymes break the amide link in the β-lactam ring, rendering bacterial resistance to β-lactam antibiotics impossible [22]. Bla genes are found in the genome belonging to the TEM and CTX-M families. The extended-spectrum β-lactamases (ESBLs) known as CTX-M, TEM and SHV enzymes, or plasmid-mediated cefotaximases, are a fast expanding family with important therapeutic implications [23].

The environment of these genes shows that these genes were derived from the plasmid of the InkL group and conjugatively transferable. Mobile genetic elements (plasmids and transposons) transmit mostly across bacteria by conjugative transfer as indicated by the PilI, PilK, PilM, PilN, PilO, PilP, PilR and PilS proteins. A mating bridge is facilitated by plasmid-encoded conjugative machinery, which encourages the localized union of two cells [24]. The close association of the donor and recipient cells' cell surfaces is a crucial requirement for conjugative transfer. Complex extracellular filaments known as sex pili are responsible for establishing this physical contact in gram-negative bacteria [25].

References

[1]Lee MS, Koo S, Jeong DG, Tesh VL. Shiga Toxins as Multi-Functional Proteins: Induction of Host Cellular Stress Responses, Role in Pathogenesis and Therapeutic Applications. Toxins (Basel). 2016 Mar 17;8(3):77. doi: 10.3390/toxins8030077. PMID: 26999205; PMCID: PMC4810222.

[2]Rodriguez-Rubio, L., Haarmann, N., Schwidder, M., Muniesa, M., & Schmidt, H. (2021). Bacteriophages of Shiga Toxin-Producing Escherichia coli and Their Contribution to Pathogenicity. Pathogens (Basel, Switzerland), 10(4), 404. https://doi.org//10.3390/pathogens10040404

[3]Newton, H. J., Sloan, J., Bulach, D. M., Seemann, T., Allison, C. C., Tauschek, M., Robins-Browne, R. M., Paton, J. C., Whittam, T. S., Paton, A. W., & Hartland, E. L. (2009). Shiga toxin-producing Escherichia coli strains are negative for locus of enterocyte effacement. Emerging infectious diseases, 15(3), 372–380. https://doi.org//10.3201/eid1503.080631

[4]Berger, P., Kouzel, I.U., Berger, M. et al. Carriage of Shiga toxin phage profoundly affects Escherichia coli gene expression and carbon source utilization. BMC Genomics 20, 504 (2019). https://doi.org//10.1186/s12864-019-5892-x

[5]Soucy, S., Huang, J. & Gogarten, J. Horizontal gene transfer: building the web of life. Nat Rev Genet 16, 472–482 (2015). https://doi.org//10.1038//nrg3962

[6]Forward paired end reads; Available from: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R1_001.fastq.gz.

[7]Reverse paired end reads; Available from: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R2_001.fastq.gz.

[8]Forward mate pair reads; Available from: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R1_001.fastq.gz.

[9]Reverse mate pair reads; Available from: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R2_001.fastq.gz.

[10]Forward mate pair reads; Available from: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R1_001.fastq.gz.

[11]Reverse mate pair reads; Available from: https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R2_001.fastq.gz.

[12]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 (first published online January 7, 2011) doi:10.1093/bioinformatics/btr011

[13]Andrey Prjibelski, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, Anton Korobeynikov. Using SPAdes De Novo Assembler. Current Protocols in Bioinformatics, 2020.

[14]Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics (2013) 29 (8): 1072-1075. doi: 10.1093/bioinformatics/btt086 First published online: February 19, 2013.

[15]Seemann T, "Prokka: Rapid Prokaryotic Genome Annotation", Bioinformatics, 2014 Jul 15;30(14):2068-9.

[16]Seemann T, barrnap 0.9 : rapid ribosomal RNA prediction, https://github.com/tseemann/barrnap

[17]Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097

[18]Valeria Bortolaia, Rolf S Kaas, Etienne Ruppe, Marilyn C Roberts, Stefan Schwarz, Vincent Cattoir, Alain Philippon, Rosa L Allesoe, Ana Rita Rebelo, Alfred Ferrer Florensa, Linda Fagelhauer, Trinad Chakraborty, Bernd Neumann, Guido Werner, Jennifer K Bender, Kerstin Stingl, Minh Nguyen, Jasmine Coppens, Basil Britto Xavier, Surbhi Malhotra-Kumar, Henrik Westh, Mette Pinholt, Muna F Anjum, Nicholas A Duggett, Isabelle Kempf, Suvi Nykäsenoja, Satu Olkkola, Kinga Wieczorek, Ana Amaro, Lurdes Clemente, Joël Mossong, Serge Losch, Catherine Ragimbeau, Ole Lund, Frank M Aarestrup, ResFinder 4.0 for predictions of phenotypes from genotypes, Journal of Antimicrobial Chemotherapy, Volume 75, Issue 12, December 2020, Pages 3491–3500, https://doi.org/10.1093/jac/dkaa345

[19]Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004 Jul;14(7):1394-403. doi: 10.1101/gr.2289704. PMID: 15231754; PMCID: PMC442156.

[20]Habib NF, Jackson MP. Roles of a ribosome-binding site and mRNA secondary structure in differential expression of Shiga toxin genes. J Bacteriol. 1993 Feb;175(3):597-603. doi: 10.1128/jb.175.3.597-603.1993. PMID: 7678590; PMCID: PMC196194.

[21]Melton-Celsa A. R. (2014). Shiga Toxin (Stx) Classification, Structure, and Function. Microbiology spectrum, 2(4), 10.1128/microbiolspec.EHEC-0024-2013. https://doi.org/10.1128/microbiolspec.EHEC-0024-2013

[22](Bonnet R. Growing group of extended-spectrum beta-lactamases: the CTX-M enzymes. Antimicrob Agents Chemother. 2004;48(1):1-14. doi:10.1128/AAC.48.1.1-14.2004

[23]Zhao, W. H., & Hu, Z. Q. (2013). Epidemiology and genetics of CTX-M extended-spectrum β-lactamases in Gram-negative bacteria. Critical reviews in microbiology, 39(1), 79–101. https://doi.org//10.3109//1040841X.2012.691460

[24]Malgorzata Zatyka, Christopher M. Thomas, Control of genes for conjugative transfer of plasmids and other mobile elements, FEMS Microbiology Reviews, Volume 21, Issue 4, February 1998, Pages 291–319, https://doi.org/10.1111/j.1574-6976.1998.tb00355.x

[25]Grohmann, E., Muth, G., & Espinosa, M. (2003). Conjugative plasmid transfer in gram-positive bacteria. Microbiology and molecular biology reviews : MMBR, 67(2), 277–301. https://doi.org//10.1128//MMBR.67.2.277-301.2003