Fall BIOINF

Search for genetic causes of the 2011 *E.coli* outbreak

Abstract

Bacterial genomes differ from strain to strain because of the mechanisms of horizontal gene transfer. In addition, bacteria borrow an arsenal of pathogenicity and resistance from each other at present. In this work, we show the importance of high quality *de novo* assembly of the bacterial genome in order to identify the strain that caused the outbreak of an infectious disease. We also try to explain the mechanism of obtained multidrug resistance.

Introduction

Escherichia coli is a bacteria species that normally can be found in the intestines of a healthy human. But, although most of *E.coli* strains a harmless, some strains can be highly pathogenic. Sometimes new pathogenic E.coli strains emerge as a result of the integration of newly acquired pathogenic genes into the bacterial genome. These new pathogenic genes can come as plasmids transferred from other bacteria or as a part of the transferred viral prophage. One of the known pathogenic *E.coli* strains that produce Shiga toxins often causes diarrhea associated with hemolytic-uremic syndrome (HUS). These types of strains are often called enterohemorrhagic *E.coli* [1].

When a new strain of pathogenic bacteria emerges it is necessary to understand what particular protein is causing symptoms and therefore it is substantial to obtain the genome of the pathogenic strain. There are two main approaches to assembling genomes: *de novo* assembly and assembling by mapping reads to a reference genome. The choice of the approach depends on the organism and it is considered preferable to use the first method with organisms that have highly variable genomes to detect all pathogenic factors.

Materials and methods

Raw data analysis Reads from the TY2482 sample were generated by Illumina HiSeq 2000 at Beijing Genome Institute and downloaded from Short Read Archive. This sample contains reads of E.coli that caused the outbreak of hemolytic-uremic syndrome in Germany in 2011. For *de novo* genome assembly three libraries were used: SRR292678 (paired-end), SRR292862 (mate pair) and SRR292770 (mate pair) [2]. All raw reads were evaluated with the FastQC program [3]. To estimate genome size k-mers were counted (k-mer size = 31) and visualized with jellyfish program [4].

De novo assembly E.coli genome was assembled twice: using only one library (paired-end mode) and using all three libraries to compare the quality of assemblies. These assemblies were created using SPAdes v3.15.5 [5]. The qualities of both assemblies were checked with QUAST v5.2.0 [6]. The quality of the second assembly that was made with three libraries was better and was used in all the following steps.

Gene prediction and annotation Genome annotation and feature prediction were performed in Prokka v1.14.6 (with *centre* and *compliant* flags) [7]. To identify the most similar strain we then located 16S rRNA of our strain using Barrnap [8]. The acquired sequence of 16S rRNA was searched in NCBI databases using BLAST and the closest relative strain was identified.

Search of pathogenic genes To explain the cause of the HUS outbreak we performed a comparison of the reference genome and the obtained genome using Mauve program [9]. To understand the genetic background for antibiotic resistance we searched in the database of genes implicated in antibiotic resistance using ResFinder v.4.1 [10]

Results

Three libraries that came from the TY2482 sample were used: SRR292678, with insert size 470 bp; SRR292862, with insert size 2 kb and SRR292770, with insert size 6 kb. The total amount of reads in all three libraries is 31.4 M. According to the FastQC report the quality of all libraries is high. The estimated genome size based on k-mer distribution created by Jellyfish is 5.1 Mb. We created two *de novo* genome assembles using one library (SRR292678, pair-end) and all tree libraries. The quality of the second assembly based on the Quast report was overall better (Table 1).

Table 1 Assembly statisistics (Quast) for one-librarys and three-library assemblies

Library	Assembly	N50	Number of contigs
One-library	scaffolds	111860	372
One-library	contigs	111860	386
Three-library	scaffolds	2815616	327
Three-library	contigs	335515	369

In annotated genome we found 1542 bp 16S rRNA gene and the following search of this sequence in NCBI databases showed

the closest related to pathogenic *E.coli* strain called *Escherichia coli* 55989 (NCBI reference Sequence: NC 011748.1). Other rRNA genes which was found by barrnap were 5S rRNA and 23S rRNA but they had not annotated yet.

Comparison of genomes of unknown *E.coli* and *E.coli* 55989 with Mauve software revealed two Shiga toxin-related genes: *stxA* and *stxB*. Further search for genes that can cause antibiotic resistance showed that the unknown E.coli strain is resistant to a whole spectrum of antibiotics and desinfectants, but predominantly to the beta-lactam type of antibiotics (Table 2).

Table 2 Predicted resistance phenotype via ResFinder v4.1

Antibiotics or desinfectants	Class	Phenotype	Resistance genes
Streptomycin	Aminoglyc.	Resistant	aph(6)-Id, aph(3")-Ib
Cefepime, Ceftriaxone, Aztreonam, eftazidime, efotaxime	beta-lactam	Resistant	blaCTX-M-15
Ampicillin, Piperacillin, Amoxicillin, Ticarcillin	beta-lactam	Resistant	blaTEM-1B, blaCTX-M-15
Cephalothin	beta-lactam	Resistant	blaTEM-1B
Sulfamethox.	Folate pathway antagonist	Resistant	sul1, sul2
Ethidium bromide, Chlorhexidine, Benzylkonium chloride, Cetylpyridinium chloride	Quaternary ammonium compound	Resistant	qacE
Tetracycline, Doxycycline	Tetracycline	Resistant	tet(A)

Discussion

Escherichia coli has the most flexible genome and all strains can be divided into non-pathogenic, commensal, enteric pathogenic and extraintestinal pathogenic strains. Such functional diversity is determined by plasmids, bacteriophages, and genomic islands with pathogenicity factors encoded in them. These parts of the genome are responsible for the adaptation of *E. coli*, as they often undergo rearrangements, excision and transfer, and further acquisition of additional DNA. It is known that various types of shigatoxins (Stx) are usually encoded by bacteriophages [11]. However, it is impossible to say unambiguously about the mechanism of acquisition of these pathogenicity factors, since pathogenicity islands (PAIs) and plasmids are also involved in this process [12]. In (Table 3) are shown the mechanisms and localization of resistance determinants according to the CARD database [13]. In sad cases of multidrug-resistant Escherichia coli, therapy is most effectively carried out using lytic phages.

Table 3 Genetic determinants of antimicrobial resistance

Gene	Genome part	Mechanism
aph(6)-Id	plasmids, integrative conjugative elements and chromosomal genomic islands	antibiotic inactivation
aph(3")-Ib	plasmids, transposons, integrative conjugative element	antibiotic inactivation
blaCTX-M-15, blaTEM-1B	plasmids	antibiotic inactivation
sul1	integrons	antibiotic target replacement
sul2	small plasmids	antibiotic target replacement
qacE, tet(A)	plasmids	antibiotic efflux

References

[1] David A. Rasko et al. "Origins of the E. coli Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany". In: *New England Journal of Medicine* 365.8 (2011). PMID: 21793740, pp. 709–717. DOI: 10.1056/NEJMoa1106920.

[2] Beijing Genome Institute (BGI). *SRX079804: sequencing data*. URL: https://www.ncbi.nlm.nih.gov/sra/?term=SRR292678.

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

[4] 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.

[5] 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.

[6] 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.

[7] Torsten Seemann. "Prokka: rapid prokaryotic genome annotation". In: *Bioinformatics* 30.14 (Mar. 2014), pp. 2068–2069. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu153.

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

[9] Aaron C.E. Darling et al. "Mauve: Multiple alignment of conserved genomic sequence with rearrangements". In: *Genome Research* 14.7 (2004), pp. 1394–1403. DOI: 10.1101/gr.2289704.

[10] Valeria Bortolaia et al. “ResFinder 4.0 for predictions of phenotypes from genotypes”. In: *Journal of Antimicrobial Chemotherapy* 75 (Aug. 2020). DOI: 10.1093/jac/dkaa345.

[11] Sylvia Herold, Helge Karch, and Herbert Schmidt. “Shiga Toxin-Encoding Bacteriophages–Genomes in Motion”. In: *International journal of medical microbiology : IJMM* 294 (Oct. 2004), pp. 115–21. DOI: 10.1016/j.ijmm.2004.06.023.

[12] Ulrich Dobrindt et al. “Genomic islands in pathogenic and environmental microorganisms”. In: *Nat Rev Microbiol* 7 (Jan. 2004), pp. 50–60.

[13] Brian Alcock et al. “CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database”. In: *Nucleic acids research* 48 (Oct. 2019). DOI: 10.1093/nar/gkz935.