Practice

BIOINFORMATICS INSTITUTE

Investigation of the 2011 *E. coli* outbreak reveals the rise of the new highly pathogenic strain gained Shiga-toxin genes

Anna Chechenina

Abstract

There are many known strains of the *Escherichia coli* that could be present in the human microbiome. However, only some of them could cause serious health problems. One of these highly pathogenic strains appeared in Germany in 2011 when the bacterium had infected thousands of people and killed more than 50 within a few months.
Supplementary materials can be found via link: https://github.com/checheanya/BI_git/tree/main/HW3.

Keywords: E. coli; hemolytic uremic syndrome; de novo assembly

Introduction

The *E. coli* outbreak in 2011 revealed a new highly pathogenic strain, that caused hemolytic uremic syndrome (HUS). First cases were detected in Germany but then the disease spread around other countries in central Europe. Turned out that the source of the pathogenic bacteria was bean sprouts, which are cultivated in a very humid environment that favors the bacteria's growth and development.

The microbiological analysis by sorbitol assay showed that the sample strain (lately named O104:H4) did not belong to any known *E. coli* strains. Usually, harmless strains of the bacteria can easily metabolize sorbitol, while pathogenic strains are not able to do it. However, the unknown strain showed a reaction typical for the non-pathogenic types of bacteria. The main challenge for the research of the newly appeared strain was to identify mutations and analyze the genome and plasmid difference with the known *E. coli* in order to find an effective treatment and to create a procedure to identify the presence of this strain in the sample to determine this strain.

It is known that the pathogenic properties of different bacteria might arise due to some toxins synthesized. It was shown that the O104:H4 strain is able to produce the Shiga toxin (Scheutz *et al.* 2011; Bielaszewska *et al.* 2011). There might be several sources of the pathogenic agents n the bacteria. First of all, they may arise during the mutations and permutations in the bacteria genome. Apart from it, the more transferable source is the plasmid sequence, which could carry many different genes. Plasmid genes are not always transcribed in bacteria and sometimes do not affect the phenotype thus making the plasmid harder to detect. Moreover, due to their small size plasmids are harder to detect using the whole genome sequencing protocols. Yet other sources of new genes and variations are phages, which could exist not only as viral agents but also in the prophage form inserted in the bacterial genome. The bacteriophages lifecycle includes the stage of packing the phage genome in the capsid which will be later transmitted to another bacteria. Depending on the bacteriophage type, they might hijack some plasmids, RNAs and proteins or genes from the bacteria genome. Those genes could be then inserted in the recipient cell and give some evolutionary advantage in sertain conditions.

Some of the tested O104:H4 strain isolates were positive for the *aggR* gene, which is typical for the group of enteroaggregative *Escherichia coli*. Besides this, the samples showed the presence of genes related to biofilm formation and haemagglutination with human erythrocytes Scheutz *et al.* (2011). In another research, scholars showed that unique molecular features of the strain include genes *rfbO104*, *fliCH4*, *stx2*, and *terD*. Gene related to the Shiga toxin synthesis by stx_2 and adhesins (*eae*, *iha*, lpf_{O26}, and lpf_{O113}) were described as well Bielaszewska *et al.* (2011). Taken together, it was strong evidence that the newly appeared strain was a combination of bacteria of two groups: Shiga-toxin-producing *E. coli* and enteroaggregative *E. coli*.

Thus, in our study, we performed the genome assembly of the whole genome sequencing data obtained from the Short Reads Archive for the TY2482 sample of the *E. coli* outbreak strain. Then we performed the phylogenetic analysis to determine the most closely related strains and suggest possible changes.

Methods

The raw reads libraries for the TY2482 sample were obtained from the Short Reads Archive added by the Beijing Genome Institute and mentioned in the paper Rohde *et al.* (2011). In our study, we used three libraries: SRR292678 (paired-end, insert size 470 bp), SRR292862 (mate pair, insert size 2 kb), and SRR292770 (mate pair, insert size 6 kb). The number of reads in the library was 5499346, 5102041, and 5102041, respectively. We performed the primary analysis of the reads using the FastQS tool Andrews (2010).

Then in order to build a K-mer profile and estimate the genome size we used a Jellyfish tool Marçais and Kingsford (2011) on the paired-end library (SRR292678) with k-mer size parameter of 31. For the gene size estimation we used the following formula: Genome size $= T/N$, $N = (M * L)/(L - K + 1)$, where N is a depth of coverage, m - a K-mer peak, k - Kmer-size, L - average read length, and T - total bases.

For reads correction and assembly we used SPAdes assembler

Prjibelski *et al.* (2020). We performed two runs, the first one only using a library with paired-end reads and the second one using three libraries with paired-end reads and mate reads. We used multiple libraries with different insert sizes to improve the quality of our final assembly. The idea is that the library with a small insert size can resolve short repeats, whereas the library with a larger insert size can resolve longer repeats. To check how the quality of our assembly changes when we add mate-pair libraries, we used QUAST tool Gurevich *et al.* (2013), which calculated the main parameters for assembly contigs and scaffolds obtained with SPAdes for both cases. You can find a full comparison in the Supplementary, Table 2.

After the quality check, we perform annotation using Prokka Seemann (2014). This tool identifies the coordinates of putative genes within contigs and then uses BLAST for similarity-based annotation using all proteins from sequenced bacterial genomes in the RefSeq database. We performed a run specified for the *E. coli*, which means that in our case the annotation was restricted only to this species.

To find the closest relative for further comparison with examined strain we searched for the 16S rRNA coding genes in the assembly using Barrnap tool Seemann (2013). We found 23 genes coding 16S rRNA, possibly, because bacteria usually carry multiple copies of this operon. Then we used nucleotide blast restricted to date before 2011 to find out the closest relative.

For the alignment of our assembly to the reference genome and visualization, we used Mauve program Darling *et al.* (2004).

Results

Sequence quality exploration

The k-mers distribution could be seen in Figure 1. The estimated size of the genome is $5175209 = 5,18$Mb with the peak at 62, which corresponds to the normal *E. coli* genome size. If we regard the single copy region as a frequency between 16 and 100, the size of the single copy region will be roughly equal to 4640551, which is an $89,6\%$ of the genome.

As was mentioned before, we performed two assemblies using only SRR292678 library as a paired ends and using all three libraries: SRR292678 as paired ends, SRR292862 and SRR292770 as mate pairs. The number of contigs of all length increased in the second assembly compared to the first as well as the size of the largest contig (from 300763 to 698474). The N50 metrics also improved from 11860 for paired-ends to 335515 for mate pairs. Thus we can say that the quality of the assembly significantly improved because we added mate paired libraries with a big insertion length that gives an understanding of sequences on bigger distances and thus can resolve more complicated repeats.

Reference genome and the closest relative

To find the closest relative for further comparison with examined strain we searched for the 16S rRNA coding genes in the assembly. We found 23 genes coding 16S rRNA with length of $1300 - 1500$bp, possibly, because bacteria usually carry multiple copies of this operon.

Using the nucleotide blast search on found 16S rRNA, restricted to the *Escherichia coli* sequences known by 2011, we have found out that the closest relative to our set of sequences with 100% identity is strain *Escherichia coli* 55989 (NC_011748.1). According to the previous research Mossoro *et al.* (2002) enteroaggregative *E. coli* (EAEC) strain, harboring the pAA plasmid, which contains aggregative adherence fimbria (AAF) genes allowing bacteria to stick to cells in the intestine. This strain was

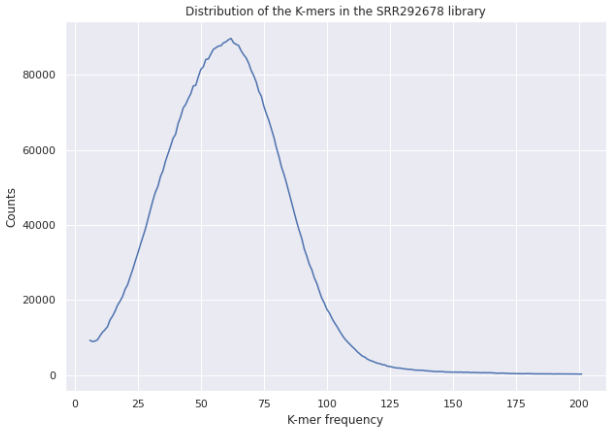

Figure 1 The k-mer frequencies distribution in the SRR292678 library.

isolated in the Central African Republic from a stool sample obtained from a man with persistent diarrhea Mossoro *et al.* (2002). However, as we know from the reports of the 2011 outbreak, in the stool of patients with the pathogenic strain was detected blood. This implies that the pathogenic strain gained some other virulence factors in addition to the reference sequence.

Sources of increased pathogenicity

To find out what is responsible for the increased toxicity, we performed a genome-wide comparison with the reference genome and will analyze the regions where these strains differ from each other. Since we know from the former papers that one of the most possible pathogenic factors for this group of bacteria that cause same effect was Shiga toxin, we specifically looked for genes associated with Shiga toxin. Shiga toxin is coded by two genes *stxA* and *stxB* that correspond to A and B subunits. In our assembly there are these two genes present in the locus, which is absent in the reference sequence: *stxB* on the position 3483605-3483874 (269bp) and *stxA* on the position 3483886-3484845 (959bp).

In order to determine the source of this segment we explored its gene neighborhood. There we have found many phage genes upstream: Phage head completion protein, Phage holin/antiholin component S, Phage antirepressor protein, Phage endopeptidase Rz, Phage Rha protein, Phage terminase (large subunit T4-like headful), Phage tail fiber protein, Phage recombination protein NinG; and downstream: Phage DNA adenine methylase, Phage antitermination protein Q, Gufsy-2 prophage protein, Phage associated DNA primase, Phage transcriptional regulator cro.

Antibiotics-resistance

Since our gene is situated in the prophage region we have to pay more attention to the antibiotics treatment, because it is known that many antibiotics used in clinical practices against similar strains can activate phages Goerke *et al.* (2006).

To look for the antibiotics-resistance genes we used ResFinder website Bortolaia *et al.* (2020) with the search for *E. coli* systems. The investigated *E. coli* strain was resistant to the following antibiotics: ampicillin, sulfamethoxazole, trimethoprim, cetylpyridinium chloride, cephalothin, piperacillin, tetracycline, amoxi-

cillin, ethidium bromide, cefepime, doxycycline, benzylkonium chloride, ceftriaxone, ceftazidime, ticarcillin, chlorhexidine, cefotaxime, aztreonam, streptomycin. However, the reference strain of *E. coli* 55989 was resistant only to the three types of antibiotics: tetracycline, doxycycline, and minocycline.

Using the annotation and alignment to the reference genome we found out that there are two beta-lactamase A genes at position 5195566-5196441 and at position 5199263-5200123 that appeared in the examined strain. In the annotation these genes are surrounded by several mobile elements, so we can assume conclude that these genes moved to the genome during the mobile element's replication.

Discussion

Our search has shown that the closest relative to the unknown outbreak strain is strain *Escherichia coli* 55989 (NC_011748.1). According to the previous research, Mossoro *et al.* (2002) enteroaggregative *E. coli* (EAEC) strain, harboring the pAA plasmid, which contains aggregative adherence fimbria (AAF) genes allowing bacteria to stick to cells in the intestine. This strain was isolated in the Central African Republic from a stool sample obtained from a man with persistent diarrhea Mossoro *et al.* (2002). Apart from it, as we assumed before based on experimental articles, the most crucial additional mutation in this strain that led to the highest toxicity was a gain of the Shiga-toxin.

Taking into account all genes found in the neighborhood of the Shiga-toxin genes, we can suggest that this segment was inserted in the genome by some phage and now exists in the form of a prophage.

Moreover, we showed that the new strain gained many genes of antibiotic resistance. There are several antibiotics in the beta-lactame group. We searched for the beta-lactamase genes (*bla*) in the reference genome and the assembly and found out that there are two beta-lactamase genes inserted, possibly by being carried by the mobile elements. The beta-lactamase gene codes an enzyme that is capable of breaking the beta-lactam ring, deactivating the molecule's antibacterial properties.

As an alternative treatment, there might be several approaches, First of all, another group of antibiotics could be used. Secondly, there are new chemical variations of the same classes of antibiotics already developed to substitute ones to which bacteria are resistant. So doctors might consider using next generation antibiotics. Besides this, there are several promising ways of treatment being developed currently such as phage therapy, however these applications still need investigation.

Literature cited

Andrews S. 2010. FASTQC. A quality control tool for high throughput sequence data.

Bielaszewska M, Mellmann A, Zhang W, Köck R, Fruth A, Bauwens A, Peters G, Karch H. 2011. Characterisation of the escherichia coli strain associated with an outbreak of haemolytic uraemic syndrome in germany, 2011: a microbiological study. The Lancet Infectious Diseases. 11:671–676.

Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AF *et al*. 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy. 75:3491–3500.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Research. 14:1394–1403.

Goerke C, Koller J, Wolz C. 2006. Ciprofloxacin and trimethoprim cause phage induction and virulence modulation in istaphylococcus aureus/i. Antimicrobial Agents and Chemotherapy. 50:171–177.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29:1072–1075.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27:764–770.

Mossoro C, Glaziou P, Yassibanda S, Lan NTP, Bekondi C, Minssart P, Bernier C, Bougue
nec CL, Germani Y. 2002. Chronic diarrhea, hemorrhagic colitis, and hemolytic-uremic syndrome associated with HEp-2 adherent iescherichia coli/i in adults infected with human immunodeficiency virus in bangui, central african republic. Journal of Clinical Microbiology. 40:3086–3088.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes de novo assembler. Current Protocols in Bioinformatics. 70.

Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, Chen W, Pu F, Peng Y, Li J *et al*. 2011. Open-source genomic analysis of shiga-toxin–producingie. coli/io104:h4. New England Journal of Medicine. 365:718–724.

Scheutz F, Nielsen EM, Frimodt-Møller J, Boisen N, Morabito S, Tozzoli R, Nataro JP, Caprioli A. 2011. Characteristics of the enteroaggregative shiga toxin/verotoxin-producing escherichia coli o104:h4 strain causing the outbreak of haemolytic uraemic syndrome in germany, may to june 2011. Eurosurveillance. 16.

Seemann T. 2013. Barrnap. github. https://github.com/tseemann/barrnap.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30:2068–2069.

Supplemetrary

A

Statistics without reference	contigs	scaffolds	scaffolds_broken
# contigs	210	221	223
# contigs (>= 0 bp)	386	372	223
# contigs (>= 1000 bp)	159	158	159
# contigs (>= 5000 bp)	81	82	81
# contigs (>= 10000 bp)	67	67	67
# contigs (>= 25000 bp)	50	50	50
# contigs (>= 50000 bp)	32	32	32
Largest contig	300 763	300 763	300 763
Total length	5 295 721	5 304 595	5 299 555
Total length (>= 0 bp)	5 334 575	5 336 365	5 299 555
Total length (>= 1000 bp)	5 259 101	5 259 608	5 259 101
Total length (>= 5000 bp)	5 076 685	5 081 904	5 076 685
Total length (>= 10000 bp)	4 977 737	4 977 737	4 977 737
Total length (>= 25000 bp)	4 714 504	4 714 504	4 714 504
Total length (>= 50000 bp)	4 035 821	4 035 821	4 035 821
N50	111 860	111 860	111 860
N90	18 506	18 506	18 506
auN	131 921	131 705	131 826
L50	14	14	14
L90	53	53	53
GC (%)	50.56	50.53	50.54
Mismatches			
# N's per 100 kbp	0	33.74	0
# N's	0	1790	0

B

Statistics without reference	contigs	scaffolds	scaffolds_broken
# contigs	105	90	119
# contigs (>= 0 bp)	369	327	119
# contigs (>= 1000 bp)	79	54	79
# contigs (>= 5000 bp)	33	16	33
# contigs (>= 10000 bp)	30	13	30
# contigs (>= 25000 bp)	26	10	26
# contigs (>= 50000 bp)	22	10	22
Largest contig	698 474	2 815 616	698 474
Total length	5 350 156	5 391 554	5 354 477
Total length (>= 0 bp)	5 403 327	5 437 160	5 354 477
Total length (>= 1000 bp)	5 331 230	5 365 719	5 331 494
Total length (>= 5000 bp)	5 202 939	5 258 076	5 203 203
Total length (>= 10000 bp)	5 183 802	5 238 939	5 184 066
Total length (>= 25000 bp)	5 133 691	5 200 270	5 133 955
Total length (>= 50000 bp)	4 975 501	5 200 270	4 975 765
N50	335 515	2 815 616	335 515
N90	79 998	180 369	79 998
auN	319 603	1 633 387	319 372
L50	6	1	6
L90	20	7	20
GC (%)	50.59	50.57	50.58
Mismatches			
# N's per 100 kbp	0	627.52	0.11
# N's	0	33 833	6

Figure 2 Comparison between SPAdes assembly using only paired-ends library (A) and using three paired- and mate-ends libraries (B).