Practice

BIOINFORMATICS INSTITUTE

Investigating the pathogenic *E.Coli* strain using denovo assembling

Ekaterina Grigorieva[1, 2] and Gleb Khegai[1, 2]

[1]Bioinformatics Institute
[2]Moscow Institute of Physics and Technology

Abstract

Investigating bacteria variety is challenging as they can horizontally transfer genetical information. Nevertheless, as sequencing and bioinformatics technologies developed scientists have a spectrum of tools to investigate particular properties of strains. In this project we demonstrated it performing denovo assembling of target E Coli strain, annotating genome, finding the closest relative using rRNA sequence and proposing pathogenicity and antibiotic resistance.

Keywords: Denovo assembling; phylogenetics; antibiotic resistance; toxins

Introduction

In 2011, an outbreak of pathogenic E. coli strain began [1]. This strain caused hemolytic uremic syndrome (HUS) accompanied with bloody diarrhea and renal failure. As far as the bacteria was extracted scientists performed an assembly of this strain and stopped its spread owing to sequencing and bioinformatics technologies.

Inspecting the symptomes Shiga toxin can be considered as one of the possible pathogenicity reasons [2] being associated with bloody diarrhea. Pathogenic bacteria produce Shiga toxin after entering the host organism. The toxin affects the endothelium of single-layer blood vessels. Next toxin is released into the bloodstream and binds to highly specific receptors in the renal glomeruli. Shiga toxins inhibit protein synthesis, thereby destroying the glomerular endothelium, which in turn leads to renal failure. This results in hemolytic uremic syndrome.

There are several mechanisms of novel strain formation that should be taken into account while investigating. For example, transformation allows a bacterium to absorb a plasmid from the external environment. The process of transduction involves the transfer of a part of DNA with the help of viruses. Horizontal gene transferring allows a bacteria to share some properties even between different species. That is why scientists utilize horizontal transfer in research. However, the same processes often prevent researchers from studying the phylogeny of strains and organisms as rapid variability and gene exchange between distant organisms makes building phylogenetic trees much more difficult. Therefore, in some cases it is easier to assemble the genome de novo and use some stable genome such as rRNA regions for phylogenetics.

Thus in this project we performed denovo assembling of target E Coli strain, annotated genome, found the closest relative using rRNA sequence and proposed pathogenicity and antibiotic resistance.

Materials and methods

In this work we used raw data received from the Bioinformatics institute local storage. The links could be found in Supplementary material. The quality of reads was calculated using FastQC [3]. Then we used Jellyfish (v2.0.2) to count k-mers in order to show the quality is enough for appropriate assemble procedure with the following options: 8 threads, k-mer size 31, using a hash with 50 million elements [4]. After that, assembling was performed using SPAdes tool v3.13.0 (which was installed with conda v22.11.0) [5]. Firstly, assembling was performed with default parameters on on paired-end sequencing. Then we include mate pair runs with large inserted size using flags –mp1-1 –mp1-2 –mp2-1 –mp2-2 in our assembly and performed assembling with SPAdes in the same manner to improve performance. This assemble was used in the following analysis. QUAST v5.2.0 was used to analyse obtained contigs quality [6]. Then we used Prokka tool for annotation [7]. After that we used BLAST in order to align our sequence against all RefSeq genomes [8, 9]. For finding 16S rRNA in the assembled *E. coli* genome we used Barrnap [10]. Then we needed to compare the sequence with the reference genome, so we used Mauve [11]. After that we checked the antibiotic resistance using ResFinder [12].

Results

Genome assembly

First, the quality of given data (SRR292678, 5499346 reads; SRR292770, 5102041 reads; SRR292862, 5102041 reads) was great and investigating k-mer profile allowed us to calculate the genome size (10643895 bp). After that we carried the genome assembly, received quality is presented below (Fig. 1, 2). N50 metrics was better for the assebly received from 3 libraries (335515 against 15325 for one library), so we chose this assembly.

	Assembly	contigs
0	# contigs (>= 0 bp)	744.00
1	# contigs (>= 1000 bp)	181.00
2	# contigs (>= 5000 bp)	104.00
3	# contigs (>= 10000 bp)	83.00
4	# contigs (>= 25000 bp)	59.00
5	# contigs (>= 50000 bp)	32.00
6	Total length (>= 0 bp)	5263025.00
7	Total length (>= 1000 bp)	5141208.00
8	Total length (>= 5000 bp)	4953672.00
9	Total length (>= 10000 bp)	4805753.00
10	Total length (>= 25000 bp)	4412553.00
11	Total length (>= 50000 bp)	3457813.00
12	# contigs	245.00
13	Largest contig	199959.00
14	Total length	5187086.00
15	GC (%)	50.51
16	N50	84630.00
17	N90	15325.00
18	auN	92024.20
19	L50	20.00
20	L90	73.00
21	# N's per 100 kbp	0.00

Figure 1 Quality of assembly for one library

	Assembly	contigs_3_seqs
0	# contigs (>= 0 bp)	369.00
1	# contigs (>= 1000 bp)	79.00
2	# contigs (>= 5000 bp)	33.00
3	# contigs (>= 10000 bp)	30.00
4	# contigs (>= 25000 bp)	26.00
5	# contigs (>= 50000 bp)	22.00
6	Total length (>= 0 bp)	5403327.00
7	Total length (>= 1000 bp)	5331230.00
8	Total length (>= 5000 bp)	5202939.00
9	Total length (>= 10000 bp)	5183802.00
10	Total length (>= 25000 bp)	5133691.00
11	Total length (>= 50000 bp)	4975501.00
12	# contigs	105.00
13	Largest contig	698474.00
14	Total length	5350156.00
15	GC (%)	50.59
16	N50	335515.00
17	N90	79998.00
18	auN	319603.40
19	L50	6.00
20	L90	20.00
21	# N's per 100 kbp	0.00

Figure 2 Quality of assembly for three libraries

Investigating the closest strain

After annotating the assembly we processed the alignment against all known complete genomes and got matched 55989 *E.Coli* strain.

Comparison with the reference

After alignment we found two Shiga toxins, which were the part of an misaligned area (Stx A (959 bp) and Stx B (269 bp)).

Exploring the nearest genes

The Fig. 3 presents all genes which we founf near stx A and stx B.

Attachment invasion locus protein precursor
Lysozyme RrrD
putative 9-O-acetyl-N-acetylneuraminic acid
Shiga toxin subunit B precursor
Shiga toxin subunit A precursor
tRNA-Arg
tRNA-Arg
tRNA-Met
putative HTH-type transcriptional regulator
Type-2 restriction enzyme ApII
Killing protein KilR
Transposase from transposon Tn916

Figure 3 List of the genes around Shiga toxins in corresponding order

Investigation for antibiotic resistance

The Fig. 4 presents all resistant and non-resistant antibiotics for our strain and reference respectively.

	Antimicrobial	Class	WGS-predicted phenotype	Match	Genetic background
21	amoxicillin	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436), blaTEM-1B ...
22	ticarcillin	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436), blaTEM-1B ...
32	ceftriaxone	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
34	aztreonam	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
36	cefepime	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
37	ceftazidime	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
39	cephalothin	beta-lactam	Resistant	3	blaTEM-1B (blaTEM-1B_AY458016)
42	cefotaxime	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
44	piperacillin	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436), blaTEM-1B ...
46	ampicillin	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436), blaTEM-1B ...
48	trimethoprim	folate pathway antagonist	Resistant	3	dfrA7 (dfrA7_AB161450)
49	sulfamethoxazole	folate pathway antagonist	Resistant	3	sul1 (sul1_U12338), sul1 (sul1_DQ914960), sul2...
58	benzylkonium chloride	quaternary ammonium compound	Resistant	1	qacE (qacE_X68232)
59	cetylpyridinium chloride	quaternary ammonium compound	Resistant	1	qacE (qacE_X68232)
60	chlorhexidine	quaternary ammonium compound	Resistant	1	qacE (qacE_X68232)
61	ethidium bromide	quaternary ammonium compound	Resistant	1	qacE (qacE_X68232)
86	streptomycin	aminoglycoside	Resistant	3	aph(3'')-Ib (aph(3'')-Ib_AF321551), aph(6)-Id ...

Figure 4 Table with antibiotics of our interest

Discussion

In this work we investigated unknown until 2011 E Coli strain. We found the closest relative strain to our and carried out matching between "reference" and target strains. We found a genome region containing Shiga toxin genes while in closest relative there were no shiga toxins in genome. Zooming out the mauve viewer we found a group of genes inserted into the genome. Next, we analysed genes, that were present in the insertion (Figure 3). There were a lot of "hypotetical genes", so we decided to filter them out. It is interesting, that that there is a transposone gene, which can be the result of phage incorporation into the genome. Thus, we assume, that a phage genes incorporation led to the toxicity of this particular strain in comparison with the closest relative.

Apart from pathogenicity it is interesting to analyse antibiotics resistance of our strain. We analysed antibiotics resistance of target and closest relative strains using ResFinder [12]. Examining the difference between these strains antibiotics of our interest should have following properties 1) target strain has resistance to them; 2) the closest relative has not resistance. Thus, we obtain antibiotics, which our strain gain resistance to in comparison with the "reference" strain. These antibiotics could be found in Figure 4. As we can see, our strain has resistance to beta-lactam.

In conclusion, we analysed the difference between new strain and the closest reference found in database. Toxin genes were found to be incorporated into the genome presumably by phage. Also we found beta-lactam resistance, which is associated with corresponding lactamases.

Supplementary data

Sequencing data used in investigation: SRR292678 (readgroup 1), SRR292678 (readgroup 2), SRR292862 (readgroup 1), SRR292862 (readgroup 2), SRR292770 (readgroup 1), SRR292770 (readgroup 2)

The code and more details are available at https://github.com/tlebchan/BI_project_3.

Acknowledgments

We thank Mike Rayko for supervising and supporting us.

References

1. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS One. 2011;6(7):e22751. doi: 10.1371/journal.pone.0022751. Epub 2011 Jul 20. PMID: 21799941; PMCID: PMC3140518.

2. Melton-Celsa AR. Shiga Toxin (Stx) Classification, Structure, and Function. Microbiol Spectr. 2014 Aug;2(4):EHEC-0024-2013. doi: 10.1128/microbiolspec.EHEC-0024-2013. PMID: 25530917; PMCID: PMC4270005.

3. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

4. https://github.com/gmarcais/Jellyfish

5. https://doi.org/10.1002/cpbi.102

6. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072-5. doi: 10.1093/bioinformatics/btt086. Epub 2013 Feb 19. PMID: 23422339; PMCID: PMC3624806.

7. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014 Jul 15;30(14):2068-9. PMID:24642063

8. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed

9. https://www.ncbi.nlm.nih.gov/refseq/publications/

10. https://github.com/tseemann/barrnap

11. https://github.com/koadman/mauve

12. Bortolaia V, Kaas RF, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AR, Fagelhauer L, Chakraborty T, Neumann B, Werner G, Bender JK, Stingl K, Nguyen M, Coppens J, Xavier BB, Malhotra-Kumar S, Westh H, Pinholt M, Anjum MF, Duggett NA, Kempf I, NykÃ¤senoja S, Olkkola S, Wieczorek K, Amaro A, Clemente L, Mossong J, Losch S, Ragimbeau C, Lund O, Aarestrup FM. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy, 75(12),3491-3500