E.coli outbreak investigation

Supplementary materials

1. **Exploring the dataset**

Quality of each library and number of total sequences were obtained using fastqc:

```
fastqc -o . *.gz
```

or via “for” cycle (as we have only fastq.gz files in our current directory)

```
for f in ${ls .}; do fastqc -o . ${f}; done;
```

File name	Total sequences
SRR292678sub_S1_L001_R1_001.fastq.gz	5499346
SRR292678sub_S1_L001_R2_001.fastq.gz	5499346
SRR292770_S1_L001_R1_001.fastq.gz	5102041
SRR292770_S1_L001_R2_001.fastq.gz	5102041
SRR292862_S2_L001_R1_001.fastq.gz	5102041
SRR292862_S2_L001_R2_001.fastq.gz	5102041

2. **K-mer profile and genome size estimation**

K-mer profile was obtained for pair-end reads (SRR292678)

```
sudo apt-get install jellyfish
```

hash table size (-s parameter) was estimated using formula, as it was recommended in official manual (http://www.cbcb.umd.edu/software/jellyfish/):

(G+k∗n)/0.8 = (4600000 + 31 * 5499346 * 2)/0.8 = 431949315,
G - approximate genome size; k - desirable length of k-mer; n - number of total sequences

```
jellyfish count -m 31 -C -s 431949315 -o 31mer_R1_R2
SRR292678sub_S1_L001_R1_001.fastq SRR292678sub_S1_L001_R2_001.fastq
```

```
jellyfish histo jellyfish_histo/initial_reads/31mer_R1_R2 -o
jellyfish_histo/initial_reads/31merhisto_R1_R2
```

Obtained k-mer distribution was plotted and genome size was calculated using R (**fig. S1 a**) (see **R script for plotting k-mer distribution** section)

3. Assembling *E. coli X* genome from paired reads

Firstly, *E. coli X* genome was assembled from pair-end library using SPAdes (http://cab.spbu.ru/software/spades/):

```
spades.py -1 SRR292678sub_S1_L001_R1_001.fastq -2
SRR292678sub_S1_L001_R2_001.fastq -o SPAdes_results
```

Assembly quality was checked with QUAST (http://cab.spbu.ru/software/quast/) (**fig. S2 a**):

```
quast.py -o QUAST_results SPAdes_results/contigs.fasta
```

3a. Effect of read correction

```
jellyfish count -m 31 -C -s 431949315 -o
jellyfish_histo/corrected_reads/31mer_R1_R2
SPAdes_results/corrected/SRR292678sub_S1_L001_R2_001.00.0_0.cor.fastq
SPAdes_results/corrected/SRR292678sub_S1_L001_R1_001.00.0_0.cor.fastq

jellyfish histo jellyfish_histo/corrected_reads/31mer_R1_R2 -o
jellyfish_histo/corrected_reads/31merhisto_R1_R2
```

Obtained k-mer distribution was plotted and genome size was calculated using R (**fig. S1 b**)

4. Impact of reads with large insert size

E. coli X genome was assembled again from 3 libraries (pair-end library + 2 libraries of mate-pair reads) using SPAdes:

```
spades.py --pe-1 1 SRR292678sub_S1_L001_R1_001.fastq --pe-2 1
SRR292678sub_S1_L001_R2_001.fastq --mp-1 2 SRR292770_S1_L001_R1_001.fastq
--mp-2 2 SRR292770_S1_L001_R2_001.fastq --mp-1 3
SRR292862_S2_L001_R1_001.fastq --mp-2 3 SRR292862_S2_L001_R2_001.fastq -o
SPAdes_results_matepair
```

The same using bash script:

```
#!/bin/bash
path_to_reads=/home/rybina/BI2020prak/Project3
path_to_output=/home/rybina/BI2020prak/Project3/spades_output
spades.py --careful -o $path_to_output --pe1-1
${path_to_reads}/SRR292678sub_S1_L001_R1_001.fastq.gz --pe1-2
${path_to_reads}/SRR292678sub_S1_L001_R2_001.fastq.gz --mp1-1
${path_to_reads}/SRR292770_S1_L001_R1_001.fastq.gz --mp1-2
${path_to_reads}/SRR292770_S1_L001_R2_001.fastq.gz --mp2-1
```

```
${path_to_reads}/SRR292862_S2_L001_R1_001.fastq.gz  --mp2-2
${path_to_reads}/SRR292862_S2_L001_R2_001.fastq.gz
```

Assembly quality was again checked with QUAST (**fig. S2 b**):

```
quast.py -o QUAST_results_matepair
SPAdes_results_matepair/three_libs_spades_out/contigs.fasta
```

5. Genome annotation

As far as the quality of the second assembly (with mate-pair reads) was higher, we used it for further annotation. Annotation was obtained using Prokka software (https://www.vicbioinformatics.com/software.prokka.shtml)

```
prokka --compliant --outdir prokka_results
SPAdes_results_matepair/scaffolds.fasta
```

If all required prerequisites are installed (BLAST 2.9.0+, Prodigal V2.6.3, tbl2asn, ARAGORN v1.2.36, barrnap 0.9, HMMER 3.3, minced 0.4.2, signalp-4.1, infernal-1.1.3, perl v5.10.1):

```
#!/bin/bash
#PBS -d .
#PBS -l walltime=100:00:00,mem=4gb
path_out=/home/rybina/BI2020prak/Project3/prokka_output
path_assembly=/home/rybina/BI2020prak/Project3/spades_output
prokka   --outdir ${path_out} --force --genus Escherichia --species coli
--prefix X --addgenes --locustag X --kingdom Bacteria --gcode 11
--usegenus --rfam --gram neg  --centre BS --compliant
${path_assembly}/scaffolds.fasta
```

6. Finding the closest relative of *E. coli* X

1) Running barrnap to identify 16 S rRNA

Coordinates of 16S rRNA in the assembled *E. coli* X genome were obtained using a barrnap tool (https://github.com/tseemann/barrnap)

```
~/barrnap/bin/barrnap scaffolds.fasta > rrna.gff
```

Nucleotide sequence of 16S rRNA were obtained using bedtools getfasta:

```
bedtools getfasta -fi scaffolds.fasta -bed rrna.gff > rrna.fasta
```

Similar way to save only sequences as a fasta file

```
barrnap -o ./rrna.fasta scaffolds.fasta
```

Get sequences as fasta file and coordinates as gff file :

```
 barrnap -o ./rrna_mine.fna < scaffolds.fasta > ./rrna_mine.gff
```

2) Get location and length of rRNAs:

using Biopython

```
for record in
SeqIO.parse('/home/rybina/BI2020prak/Project3/rrna_mine.fa','fasta'):
    print(record.id.split(':')[0], record.id.split(':')[3],
len(record.seq))
16S_rRNA 353549-355087(+) 1538
16S_rRNA 185421-186959(-) 1538
16S_rRNA 764546-766084(-) 1538
16S_rRNA 313-719(+) 406
23S_rRNA 5-1023(+) 1018
5S_rRNA 18-129(+) 111
5S_rRNA 641221-641332(-) 111
```

or extract specific columns from the gff file

```
cut -d '=' -f2 rrna_mine.gff | cut -d ';' -f1
##gff-version 3
16S_rRNA
5S_rRNA
5S_rRNA
16S_rRNA
23S_rRNA
16S_rRNA
16S_rRNA
```

```
cut -f4,5 rrna_mine.gff
##gff-version 3
764547  766084
19      129
641222  641332
353550  355087
6       1023
185422  186959
314     719
```

3) BLAST (nucleotide) to search for the genome against complete genomes in the RefSeq database

Database: RefSeq Genomes Database (refseq_genomes)
Organism: Escherichia coli

Entrez Query: 1900/01/01:2011/01/01[PDAT]
We used BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to search for the genome in the RefSeq database with the closest 16S rRNA sequence. As a result, *E. coli* 55989 is the closest species to our object, so we used a genome sequence of *E. coli* 55989 as a reference for comparison.

7. Antibiotic resistance detection

We used ResFinder (https://cge.cbs.dtu.dk/services/ResFinder/) to find genes responsible for antibiotic resistance.

8. Tracing the source of toxin and antibiotic resistance genes in E. coli X

In order to find possible insertions of shiga-toxin and antibiotic resistance genes we used a program called mauve (http://darlinglab.org/mauve/user-guide/introduction.html). We also inspected surrounding genes to uncover a source of these genes.

1) Reference genome (in fasta format) was aligned to the studied strain (gbk file yielded by Prokka) using Mauve v2.3.1 (progressiveMauve).

```
# using prepared prokka annotation file:
progressiveMauve --output=progressiveMauve_notmine2.xmfa scaffolds.gbk
55989.fasta

#using our own prokka annotation file:
progressiveMauve --output=progressiveMauve_mine.xmfa X.gbk 55989.fasta

# using rasta annotation file
progressiveMauve --output=progressiveMauve_rast.xmfa three_libs_rast.gbk
55989.fasta
```

2) To get length of each Shiga-toxin related gene we used biopython

```
from Bio import SeqIO
path = '/home/rybina/BI2020prak/Project3/prokka_output/X.ffn'
for record in SeqIO.parse(path , 'fasta'):
    if 'Shiga' in record.description:
        print(len(record.seq), record.description)
```

270 X_05211 Shiga toxin subunit B
960 X_05212 Shiga toxin subunit A

3) Get names, coordinates of bla genes

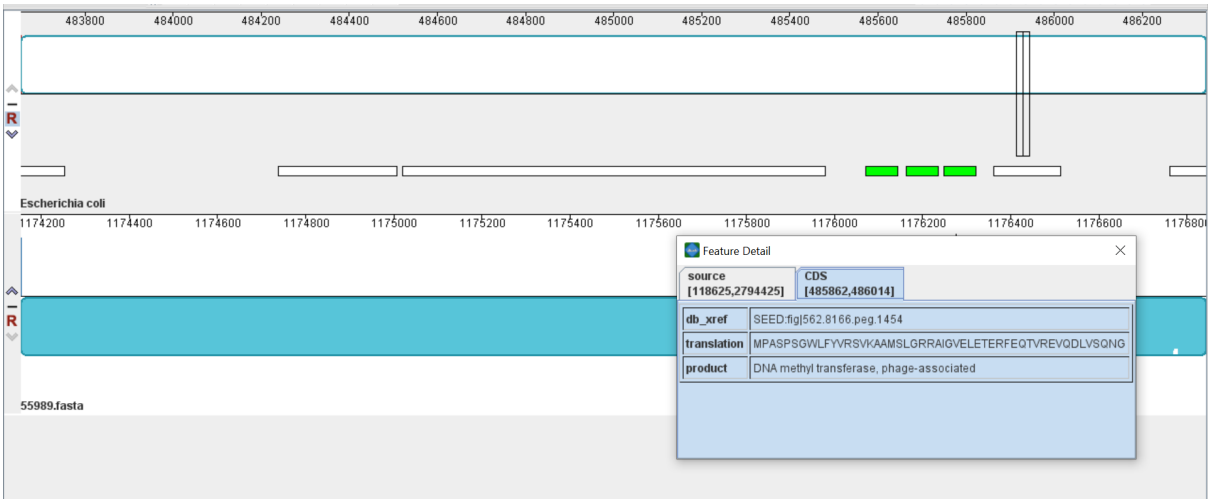
names

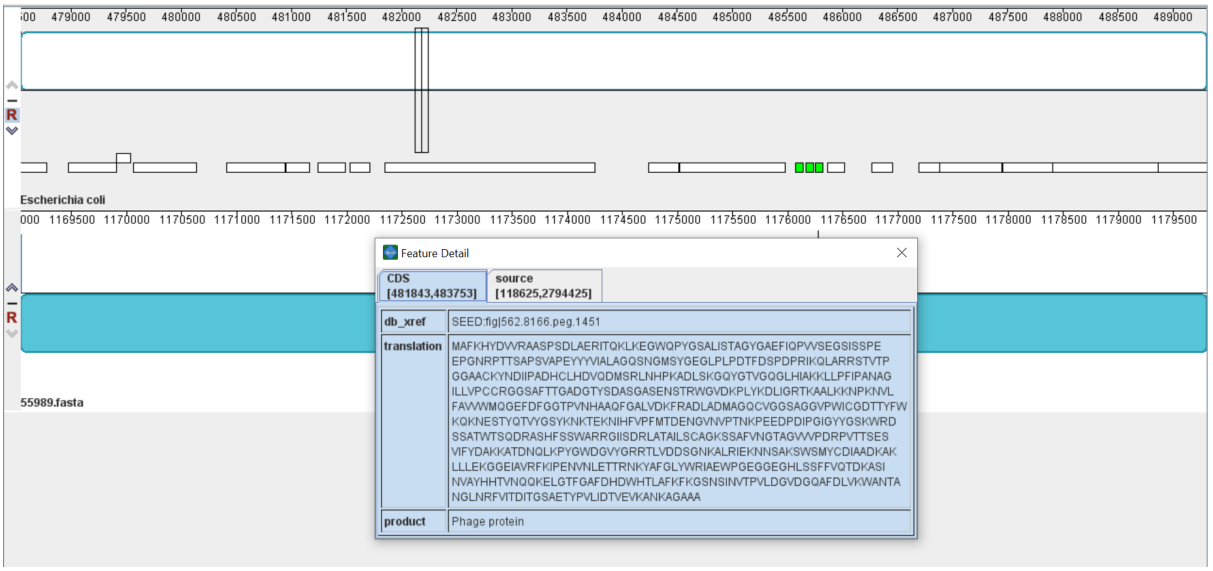
```
$ grep -e 'bla' ./prokka_output/X.gff | grep -e 'CDS' | cut -f9 | awk
-F";" '{print $5,$8}'
gene=bla_1 product=Beta-lactamase CTX-M-1
gene=bla_2 product=Beta-lactamase TEM
```

coordinates

```
$ grep -e 'bla' ./prokka_output/X.gff | grep -e 'CDS' | cut -f4,5
81730   82605
85427   86287
```

4) Visualisation and analysis of genome-wide alignment based on rast annotation - analysis of gene neighborhood





Both identified Shiga-toxin related genes were located between phage- associated genes suggesting that *stxB* and *stxA* genes were acquired via HGT with bacteriophage.

5) Extract amino acid and nucleotide sequences of *bla* gene neighbours using biopython to run blast and reveal their possible function and origin, and therefore, possible origin of *bla* genes

```
from Bio import SeqIO
path = '/home/rybina/BI2020prak/Project3/prokka_output/X.ffn'
list_ids =
['X_04998','X_05001','X_05002','X_05003','X_05004','X_05005','X_05006','X_
05007']
with open('/home/rybina/BI2020prak/Project3/bla_neigbors.fna','w') as
f_in:
    for record in SeqIO.parse(path , 'fasta'):
        for x in list_ids:
            if x in record.id:
                SeqIO.write(record, f_in, 'fasta')
```

```
from Bio import SeqIO
path = '/home/rybina/BI2020prak/Project3/prokka_output/X.faa'
list_ids =
['X_04998','X_05001','X_05002','X_05003','X_05004','X_05005','X_05006','X_
05007']
with open('/home/rybina/BI2020prak/Project3/bla_neigbors.faa','w') as
f_in:
    for record in SeqIO.parse(path , 'fasta'):
        for x in list_ids:
            if x in record.id:
                SeqIO.write(record, f_in, 'fasta')
```

6) Blastp search - tracing the origin of *bla* genes

for X_04998 (location 78542...81091) - Tn3-like element Tn3 family transposase

Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
transposase [Klebsiella pneumoniae]	Klebsiella pneu...	1743	1743	100%	0.0	100.00%	865	AQY76005.1
Tn3-like element Tn3 family transposase [Salmonella enterica subsp. enterica serovar 4,[5],12:i:-]	Salmonella ent...	1742	1742	100%	0.0	100.00%	870	EDI6181308.1
TPA: Tn3-like element Tn3 family transposase [Salmonella enterica subsp. enterica serovar Enteritidis]	Salmonella ent...	1742	1742	100%	0.0	100.00%	868	HAE0622329.1
Tn3 family transposase [Klebsiella quasipneumoniae]	Klebsiella quas...	1742	1742	100%	0.0	100.00%	927	WP_139592497.1
MULTISPECIES: Tn3-like element Tn3 family transposase [Enterobacteriaceae]	Enterobacteria...	1742	1742	100%	0.0	100.00%	869	WP_058351058.1
Tn3-like element Tn3 family transposase [Salmonella enterica]	Salmonella ent...	1742	1742	100%	0.0	100.00%	851	WP_077950247.1
Tn3-like element Tn3 family transposase [Escherichia coli]	Escherichia coli	1742	1742	100%	0.0	100.00%	950	EER5087249.1
Tn3-like element Tn3 family transposase [Salmonella enterica subsp. enterica serovar 4,[5],12:i:-]	Salmonella ent...	1742	1742	100%	0.0	100.00%	853	ECI2026040.1

for X_04999 (location 81411... 81683) - MULTISPECIES: cupin fold metalloprotein, WbuC family

Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
MULTISPECIES: cupin fold metalloprotein, WbuC family [Enterobacteriaceae]	Enterobacteria…	186	186	100%	2e-59	100.00%	90	WP_000243817.1
Tryptophan synthase subunit beta like protein [Klebsiella pneumoniae]	Klebsiella pneu…	186	186	100%	3e-59	98.89%	126	SWY83209.1
WbuC family cupin fold metalloprotein [Escherichia coli]	Escherichia coli	187	187	100%	6e-59	98.89%	158	WP_103809503.1
MULTISPECIES: WbuC family cupin fold metalloprotein [Gammaproteobacteria]	Gammaproteo…	187	187	100%	8e-59	98.89%	158	WP_013023839.1
TPA: cupin fold metalloprotein, WbuC family [Escherichia coli]	Escherichia coli	188	188	100%	8e-59	98.89%	190	HAJ2699363.1
hypothetical protein IOMTU157_1634 [Citrobacter portucalensis]	Citrobacter por…	184	184	100%	9e-59	98.89%	90	BBV16385.1
WbuC family cupin fold metalloprotein [Proteus mirabilis]	Proteus mirabilis	186	186	100%	9e-59	98.89%	158	WP_149127800.1

for X_05001 (location 82758 … 82880) - the best hit is Mobile genetic element

Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
MULTISPECIES: hypothetical protein [Enterobacterales]	Enterobacterales	81.3	81.3	97%	7e-19	100.00%	79	WP_117054877.1
hypothetical protein EA141_27895 [Klebsiella pneumoniae]	Klebsiella pneu…	80.9	80.9	97%	8e-19	100.00%	72	RLZ15192.1
hypothetical protein [uncultured bacterium]	uncultured bact…	79.7	79.7	100%	1e-18	100.00%	45	AMP47490.1
hypothetical protein [Klebsiella pneumoniae]	Klebsiella pneu…	79.3	79.3	100%	1e-18	100.00%	40	WP_000800339.1
hypothetical protein [Escherichia coli]	Escherichia coli	79.3	79.3	100%	1e-18	100.00%	41	AVE22911.1
hypothetical protein ELP95_23235 [Shigella sonnei]	Shigella sonnei	80.1	80.1	97%	1e-18	100.00%	63	RUL49175.1
hypothetical protein [Klebsiella pneumoniae]	Klebsiella pneu…	78.2	78.2	100%	5e-18	97.50%	40	BBE81069.1
hypothetical protein AM428_28785 [Klebsiella pneumoniae]	Klebsiella pneu…	77.8	77.8	97%	5e-18	100.00%	41	APR50793.1
hypothetical protein BAE49_29295 [Klebsiella pneumoniae]	Klebsiella pneu…	77.8	77.8	97%	6e-18	100.00%	39	ODP81153.1
hypothetical protein CSC40_5461 [Klebsiella pneumoniae]	Klebsiella pneu…	77.8	77.8	97%	6e-18	100.00%	41	RCH14301.1
hypothetical protein [Salmonella sp.]	Salmonella sp.	77.8	77.8	100%	7e-18	97.50%	40	QJR97255.1
Mobile element protein [Klebsiella pneumoniae]	Klebsiella pneu…	83.6	83.6	100%	1e-17	100.00%	349	QAR15180.1
hypothetical protein [Klebsiella pneumoniae]	Klebsiella pneu…	77.4	77.4	97%	1e-17	97.44%	63	WP_077254729.1
transposase [Klebsiella pneumoniae subsp. pneumoniae]	Klebsiella pneu…	78.6	78.6	97%	2e-17	97.44%	112	AOZ87095.1
hypothetical protein ACN79_28630 [Escherichia coli]	Escherichia coli	75.9	75.9	100%	4e-17	97.50%	40	OKU30755.1
Putative transposase (identified by ISEscan HMM) [Klebsiella pneumoniae]	Klebsiella pneu…	80.1	80.1	97%	3e-16	97.44%	454	VCZ13553.1

for X_05002 (location 82861..84123) - the best hit is IS1380-like element ISEc9 family transposase

Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
TnpA [Escherichia coli]	Escherichia coli	855	855	100%	0.0	100.00%	421	ADL14066.1
MULTISPECIES: IS1380-like element ISEc9 family transposase [Bacteria]	eubacteria	855	855	100%	0.0	100.00%	420	WP_000608644.1
IS1380-like element ISEc9 family transposase [Salmonella enterica]	Salmonella ent…	855	855	100%	0.0	99.76%	420	EDY2523057.1
IS1380-like element ISEc9 family transposase [Klebsiella pneumoniae]	Klebsiella pneu…	855	855	100%	0.0	99.76%	424	WP_032434420.1
IS1380-like element ISEc9 family transposase [Escherichia coli]	Escherichia coli	855	855	100%	0.0	99.76%	420	WP_119178235.1
IS1380-like element ISEc9 family transposase [Salmonella enterica]	Salmonella ent…	855	855	100%	0.0	99.76%	420	EAA9462488.1
IS1380-like element ISEc9 family transposase [Escherichia coli]	Escherichia coli	855	855	100%	0.0	100.00%	423	WP_113914577.1

for X_05003 (location 84305 , 84523) the best hit is transposase for transposon Tn3

Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
hypothetical protein SM268103_07128 [Salmonella enterica subsp. enterica serovar Typhi]	Salmonella ent…	152	152	100%	4e-46	100.00%	103	AYS80382.1
transposase for transposon Tn3 [Escherichia coli]	Escherichia coli	152	152	100%	4e-46	100.00%	125	ACQ41906.1
hypothetical protein MS6198_B106 [Escherichia coli]	Escherichia coli	152	152	100%	5e-46	100.00%	115	AOM73345.1
MULTISPECIES: DUF4158 domain-containing protein [Enterobacteriaceae]	Enterobacteria…	150	150	100%	5e-46	100.00%	72	WP_001143771.1
hypothetical protein BME44_20695 [Klebsiella pneumoniae]	Klebsiella pneu…	152	152	100%	5e-46	100.00%	125	OVW73836.1
hypothetical protein CMV43_25520 [Escherichia coli]	Escherichia coli	150	150	100%	5e-46	100.00%	73	RJK10274.1
Tn1 transposase [Escherichia coli IS25]	Escherichia col…	150	150	100%	9e-46	98.61%	72	CDK83314.1
transposase [Escherichia coli]	Escherichia coli	150	150	98%	1e-45	100.00%	99	CAA61080.1
DUF4158 domain-containing protein [Escherichia coli]	Escherichia coli	150	150	98%	1e-45	100.00%	80	WP_162879298.1

Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
MULTISPECIES: recombinase family protein [Bacteria]	eubacteria	364	364	100%	7e-127	100.00%	185	WP_001235713.1
recombinase family protein [Aeromonas hydrophila]	Aeromonas hy…	363	363	100%	1e-126	99.46%	185	WP_118881662.1
resolvase, N terminal domain protein [uncultured bacterium]	uncultured bact…	363	363	100%	1e-126	100.00%	199	AMP56037.1
transposon Tn3 resolvase [Haemophilus influenzae 86-028NP]	Haemophilus i…	363	363	100%	1e-126	100.00%	197	AAX87127.2
MULTISPECIES: recombinase family protein [Enterobacteriaceae]	Enterobacteria…	363	363	100%	2e-126	99.46%	185	WP_087835521.1
TPA: recombinase family protein [Escherichia coli]	Escherichia coli	363	363	100%	2e-126	99.46%	185	HAH3863661.1
Tn3 resolvase [Salmonella enterica subsp. enterica serovar Heidelberg str. 92-0138]	Salmonella ent…	363	363	100%	2e-126	100.00%	204	KJU62498.1
recombinase family protein [Salmonella enterica]	Salmonella ent…	363	363	100%	2e-126	99.46%	185	WP_143551345.1
TPA: recombinase family protein [Escherichia coli]	Escherichia coli	363	363	100%	2e-126	99.46%	185	HAO2597947.1
recombinase family protein [Escherichia coli]	Escherichia coli	363	363	100%	2e-126	99.46%	185	WP_097560614.1

Based on the protein homolog search via blastp, we can suggest that functions and possible origin of hypothetical proteins whose genes are neighbouring to beta-lactamase genes are associated with transposon from Tn3 family.

R script for plotting k-mer distribution

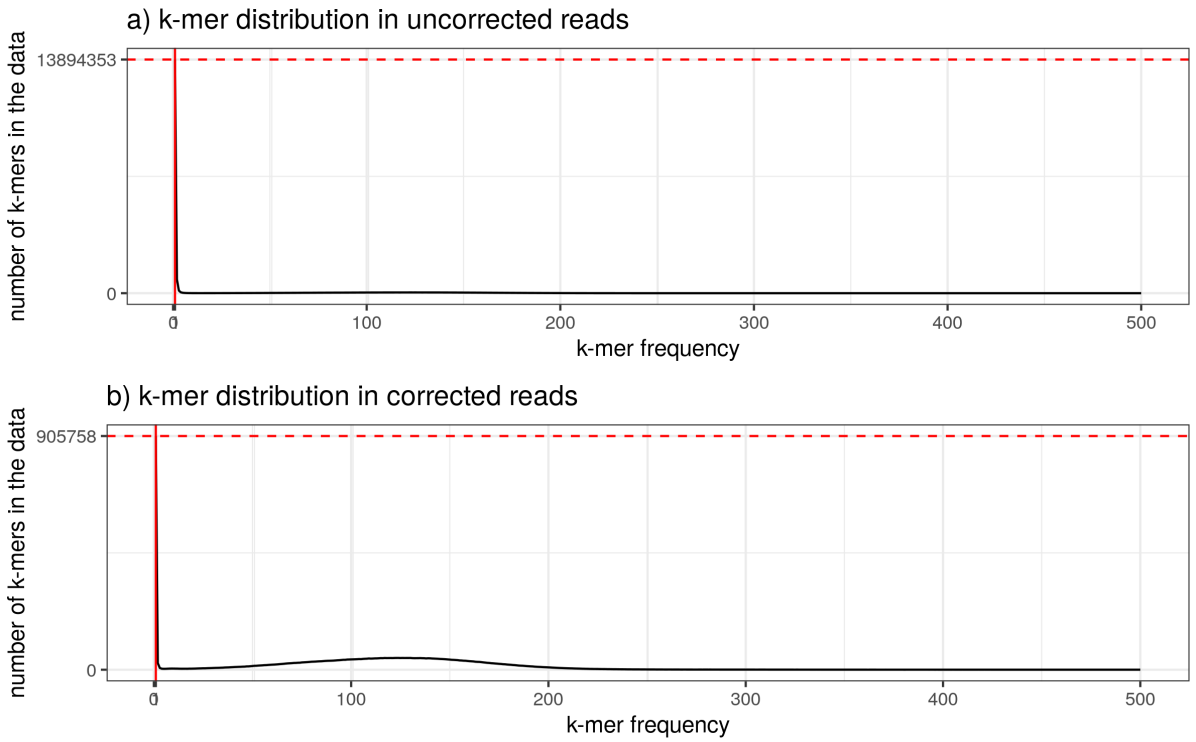
```
setwd("/media/daria/DaryaNika/IB_fall_2020/project3/jellyfish_histo/initial_reads")

theme_set(theme_bw())

require(data.table)
library(ggplot2)

#read data
kmer_histo <- fread("31merhisto_R1_R2")

#skip first rows
kmer_histo <- kmer_histo[c(9:nrow(kmer_histo)), ]

#find a peak and genome size
max_depth <- max(kmer_histo$V2)
peak <- kmer_histo[kmer_histo$V2 == max(kmer_histo$V2), ]$V1

genome_size <- sum(as.numeric(kmer_histo$V1 * kmer_histo$V2))/peak

#draw a plot
my_plot <- ggplot(kmer_histo[c(1:500), ], aes(x = V1, y = V2)) +
  geom_line() +
  geom_vline(xintercept = peak, color = "red") +
  geom_hline(yintercept = max_depth, color = "red", linetype = 2) +
  scale_x_continuous(name = "k-mer size",
                     breaks = c(seq(0, 500, 100), peak)) +
  scale_y_continuous(name = "number of k-mers in the data",
                     breaks = c(seq(0, 75000, 25000), max_depth)) +
  ggtitle("k-mer distribution")
```

```
my_plot
```

#save plot

```
ggsave(filename = "kmer_distribution.png", plot = my_plot, device = "png",
width = 8, height = 3)
```

Figure S1. K-mer distribution for pair-end library (SRR292678) without trimming k-mers with errors



a - uncorrected reads; b - reads corrected by SPAdes

Figure S2. Statistics for *E. coli X* genome assembly

a)

Statistics without reference	contigs	scaffolds
# contigs	210	221
# contigs (>= 0 bp)	386	372
# contigs (>= 1000 bp)	159	158
# contigs (>= 5000 bp)	81	82
# contigs (>= 10000 bp)	67	67
# contigs (>= 25000 bp)	50	50
# contigs (>= 50000 bp)	32	32
Largest contig	300 763	300 763
Total length	5 295 721	5 304 595
Total length (>= 0 bp)	5 334 575	5 336 365
Total length (>= 1000 bp)	5 259 101	5 259 608
Total length (>= 5000 bp)	5 076 685	5 081 904
Total length (>= 10000 bp)	4 977 737	4 977 737
Total length (>= 25000 bp)	4 714 504	4 714 504
Total length (>= 50000 bp)	4 035 821	4 035 821
N50	111 860	111 860
N75	55 279	55 279
L50	14	14
L75	31	31
GC (%)	50.56	50.53
Mismatches		
# N's	0	1790
# N's per 100 kbp	0	33.74

b)

Statistics without reference	contigs	scaffolds
# contigs	105	90
# contigs (>= 0 bp)	369	327
# contigs (>= 1000 bp)	79	54
# contigs (>= 5000 bp)	33	16
# contigs (>= 10000 bp)	30	13
# contigs (>= 25000 bp)	26	10
# contigs (>= 50000 bp)	22	10
Largest contig	698 474	2 815 616
Total length	5 350 156	5 391 554
Total length (>= 0 bp)	5 403 327	5 437 160
Total length (>= 1000 bp)	5 331 230	5 365 719
Total length (>= 5000 bp)	5 202 939	5 258 076
Total length (>= 10000 bp)	5 183 802	5 238 939
Total length (>= 25000 bp)	5 133 691	5 200 270
Total length (>= 50000 bp)	4 975 501	5 200 270
N50	335 515	2 815 616
N75	143 558	391 920
L50	6	1
L75	13	4
GC (%)	50.59	50.57
Mismatches		
# N's	0	33 833
# N's per 100 kbp	0	627.52

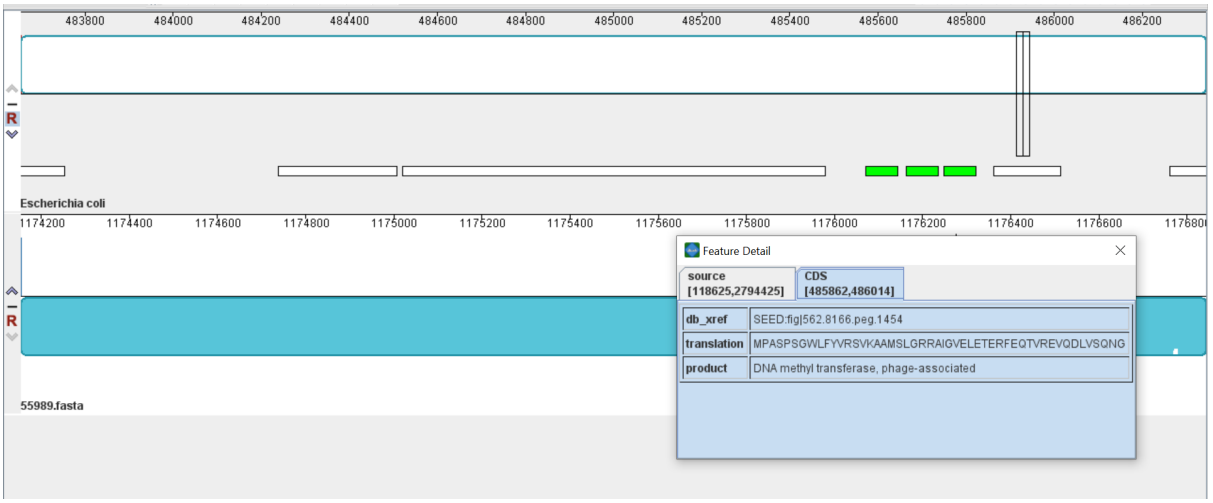
a - only pair-end reads were assembled; b - pair-end and two libraries of mate-pair reads were assembled

Table S1. Identified ribosomal RNA in the *E .coli X*

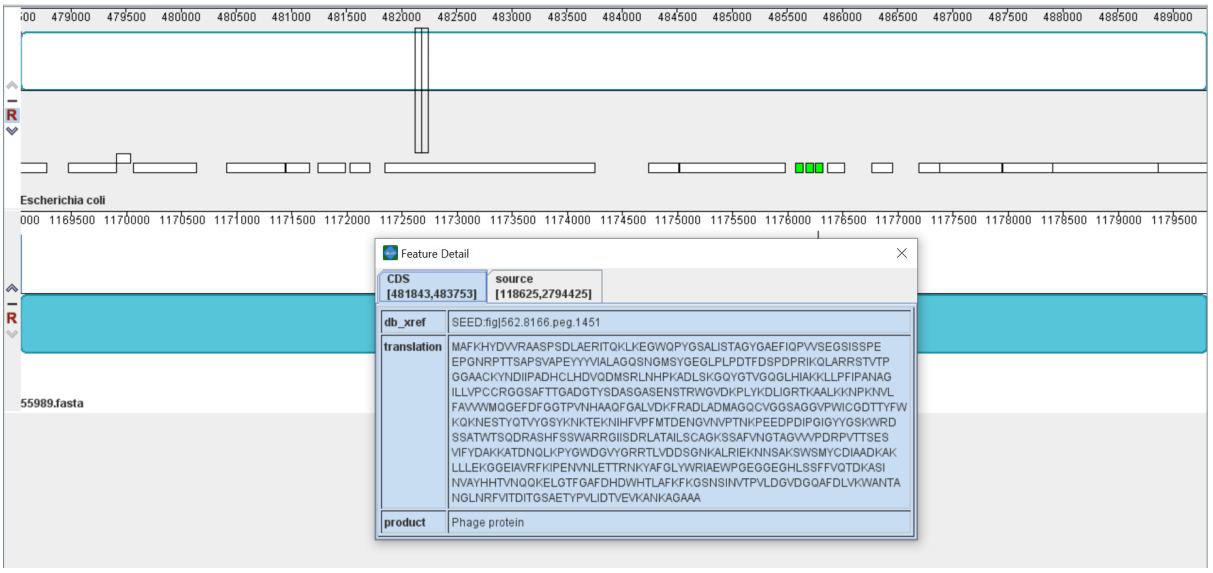
rRNA	start	end	length
16S	353549	355087	1538
16S	185421	186959	1538
16S	764546	766084	1538
16S	313	719	406
23S	5	1023	1018
5S	18	129	111
5S	641221	641332	111

Figure S3. Gene neighborhood of Shiga-toxin related genes in *E.coli X* on the genome-wide alignment with the reference *E. coli* 55989.

a)



b)



Downstream (a) and upstream (b) regions of *stxA* and *stxB* genes are shown. Rast annotation was used to obtain the visualisation in Mauve.