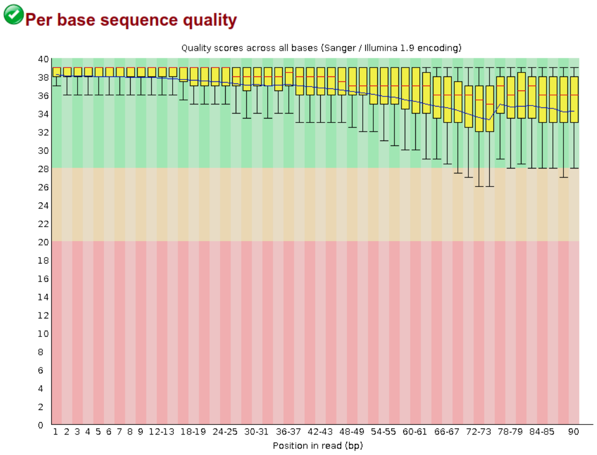
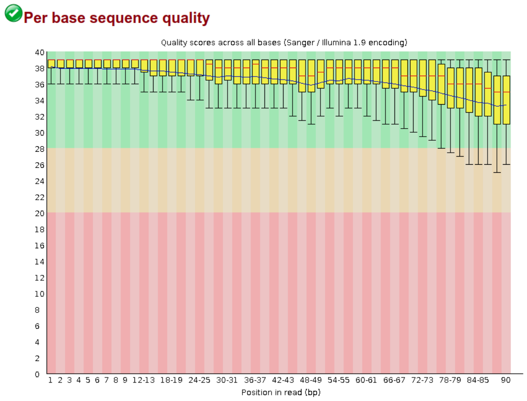
E.coli outbreak - lab notebook

Download datasets & Quality control

78

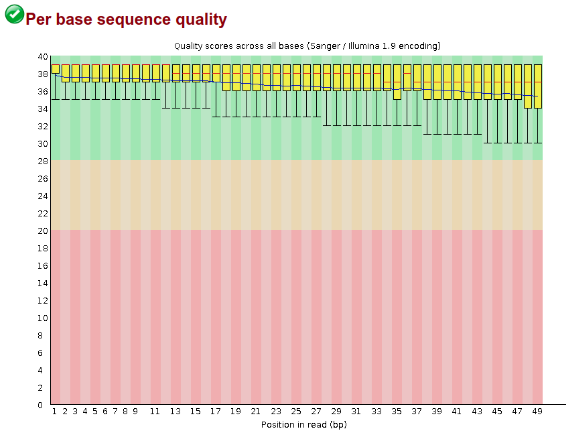
```
wget https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R1_001.fastq.gz

wget https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R2_001.fastq.gz
```

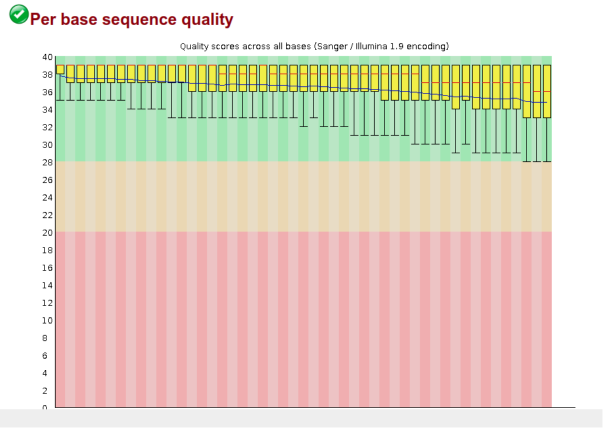




62 - matepair 2

```
wget https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R1_001.fastq.gz

wget https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R2_001.fastq.gz
```

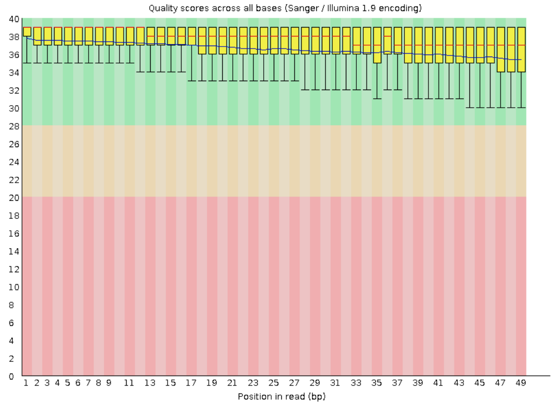




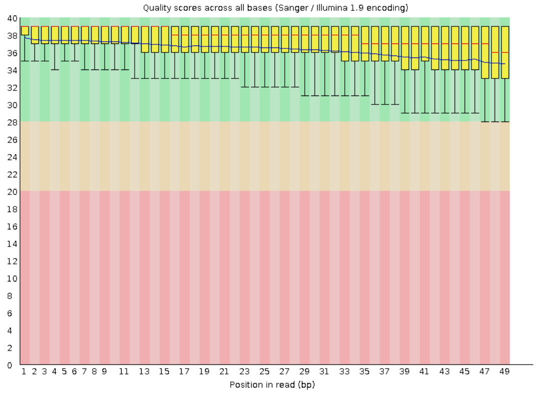
70 - matepair 1

```
wget https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R1_001.fastq.gz
wget https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R2_001.fastq.gz
```

✓ Per base sequence quality



✓ Per base sequence quality

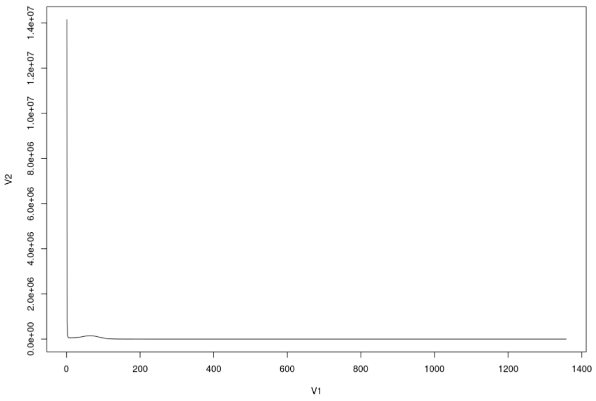


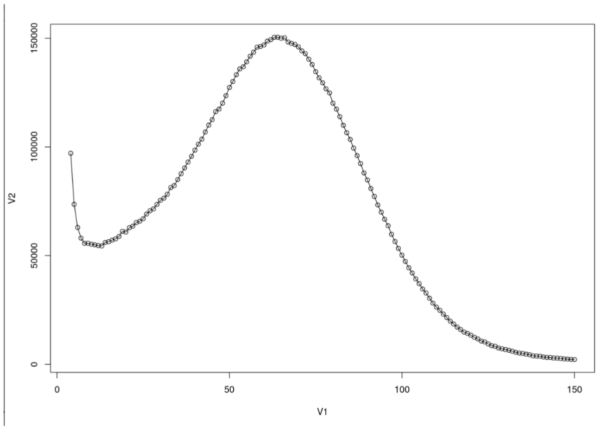
K-mer profile and genome size estimation

Guide

```
(base) anya@anya-laptop:~/Рабочий стол/IB/workshop/project 3$ jellyfish count -m 31 -o output_jelly -s 10000000 '/home/anya/Рабочий стол/IB
```

```
table <- read.csv('output_jelly_histo.txt', header = F)
plot(table[1:948,],type="l")
```





```
table <- read.csv('output_jelly_histo.txt', header = F)
plot(table[4:150,],type="l")
points(table[4:150,])
```

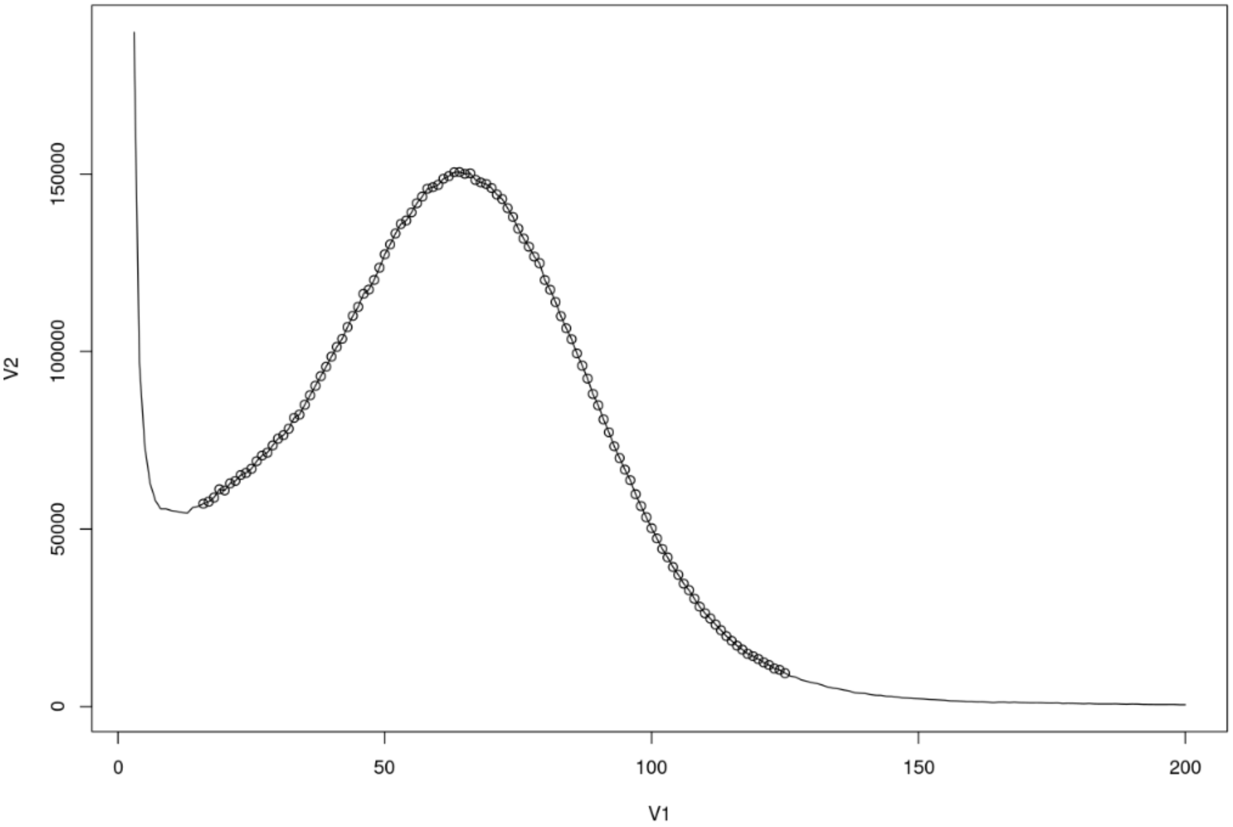
```
sum(as.numeric(table[16:949,1]*table[16:949,2]))
636969672
```

count bases

```
at '/home/anya/Рабочий стол/IB/workshop/project 3/SRR292678sub_S1_L001_R1_001.fastq'  | paste - - - - | cut -f 2 | tr -d '\n' | wc -c
494941140
```

```
table <- read.csv('output_jelly_histo.txt', header = F)
plot(table[3:200,],type="l")
points(table[16:125,])


sum(as.numeric(table[16:949,1]*table[16:949,2]))
636969672

max(table[16:150,2]) # 150522 -> 64

all <- sum(as.numeric(table[16:949,1]*table[16:949,2]))/64
# 952651

single <- sum(as.numeric(table[16:125,1]*table[16:125,2]))/64
# 9309610

(sum(as.numeric(table[16:125,1]*table[16:125,2]))/sum(as.numeric(table[16:949,1]*table[16:949,2])))

# 0.93539

n <- (64*90)/(90-31+1) # depth of coverage
genome_size = 494941140/n
```

read size = 90



```
conda install spades -c bioconda
```

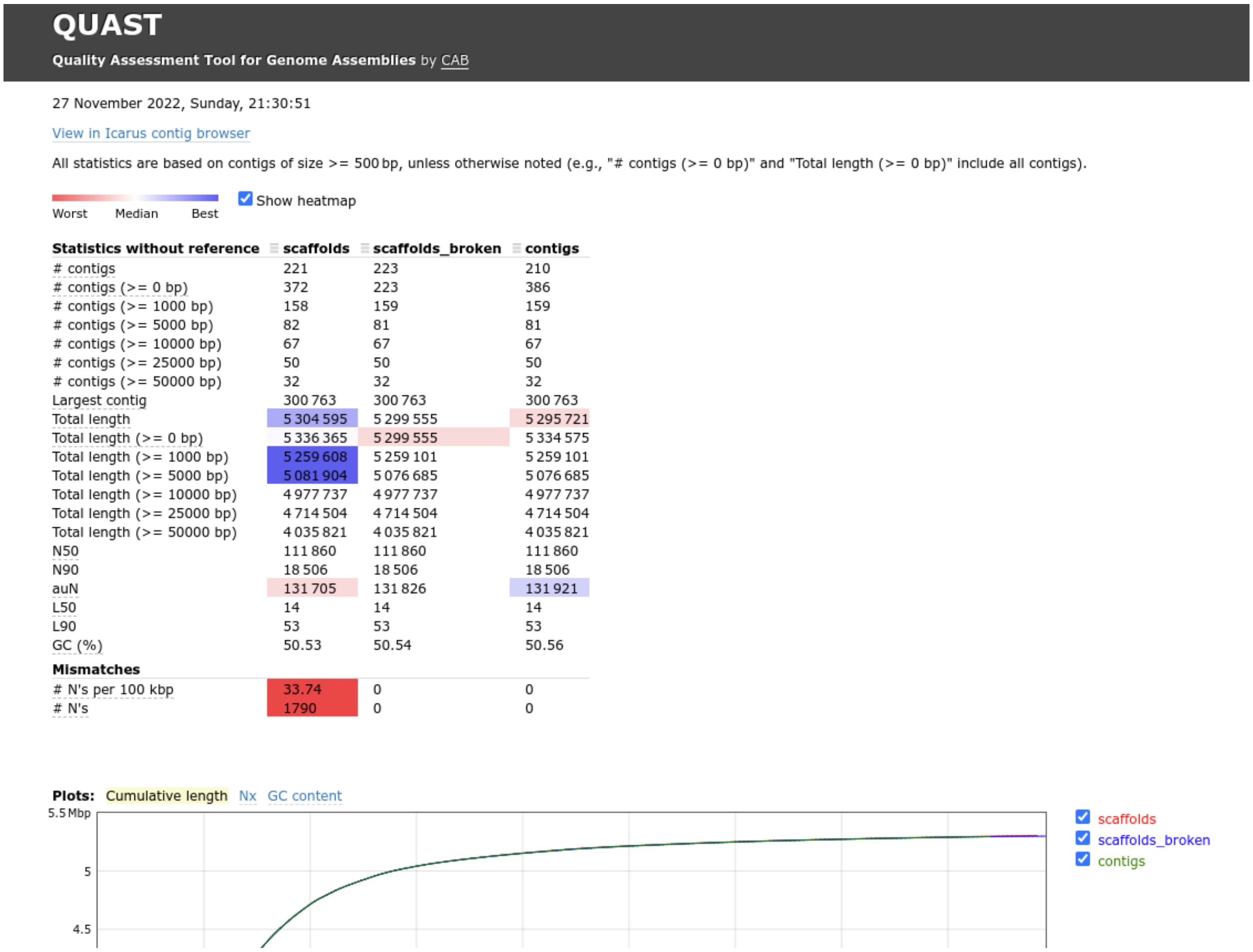
Assembling E. coli X genome from paired reads

SPAdes manuals

```
spades.py -1 'SRR292678sub_S1_L001_R1_001.fastq' -2 'SRR292678sub_S1_L001_R2_001.fastq' -o spades_did_this
```

Assess the quality of the resulting assemble

```
python '/home/anya/miniconda3/bin/quast.py' -s scaffolds.fasta contigs.fasta
```



QUAST

Quality Assessment Tool for Genome Assemblies by CAB

27 November 2022, Sunday, 21:30:51

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

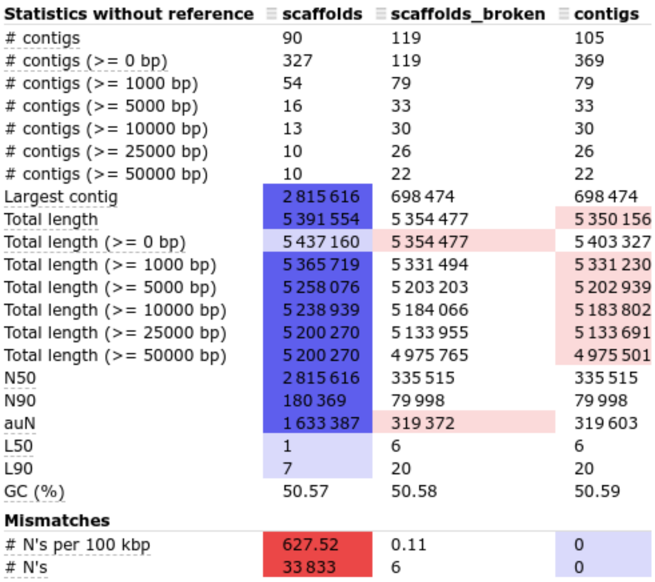
Statistics without reference	scaffolds	scaffolds_broken	contigs
# contigs	221	223	210
# contigs (>= 0 bp)	372	223	386
# contigs (>= 1000 bp)	158	159	159
# contigs (>= 5000 bp)	82	81	81
# contigs (>= 10000 bp)	67	67	67
# contigs (>= 25000 bp)	50	50	50
# contigs (>= 50000 bp)	32	32	32
Largest contig	300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

report.html

Impact of reads with large insert size (Assembling with all three libraries)

```
spades.py --pe1-1 ~/IB/hw3/SRR292678sub_S1_L001_R1_001.fastq.gz --pe1-2 ~/IB/hw3/SRR292678sub_S1_L001_R2_001.fastq.gz --mp1-1~/IB/hw3/SRR29
```

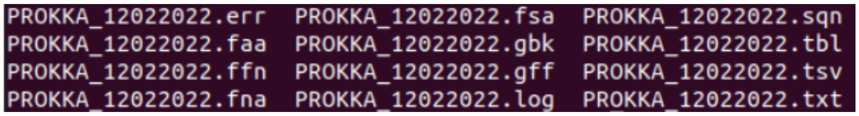
```
python '/home/anya/miniconda3/bin/quast.py' -s scaffolds.fasta contigs.fasta
```



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

Genome Annotation

```
conda activate prokka
prokka --outdir prokka --compliant --centre centre --gcode 11 --evalue 1e-04 --genus Escherichia --species Escherichia_coli --quiet ~/IB/hw
```



```
PROKKA_12022022.err  PROKKA_12022022.fsa  PROKKA_12022022.sqn
PROKKA_12022022.faa  PROKKA_12022022.gbk  PROKKA_12022022.tbl
PROKKA_12022022.ffn  PROKKA_12022022.gff  PROKKA_12022022.tsv
PROKKA_12022022.fna  PROKKA_12022022.log  PROKKA_12022022.txt
```

Finding the closest relative of E. coli X

```
barrnap --quiet -o bout/rrna.fa < scaffolds.fasta > bout/rrna.gff
```

Escherichia coli 55989, NCBI Reference Sequence: NC_011748.1

What is the genetic cause of HUS?

Two shiga toxin were found using Mauve:

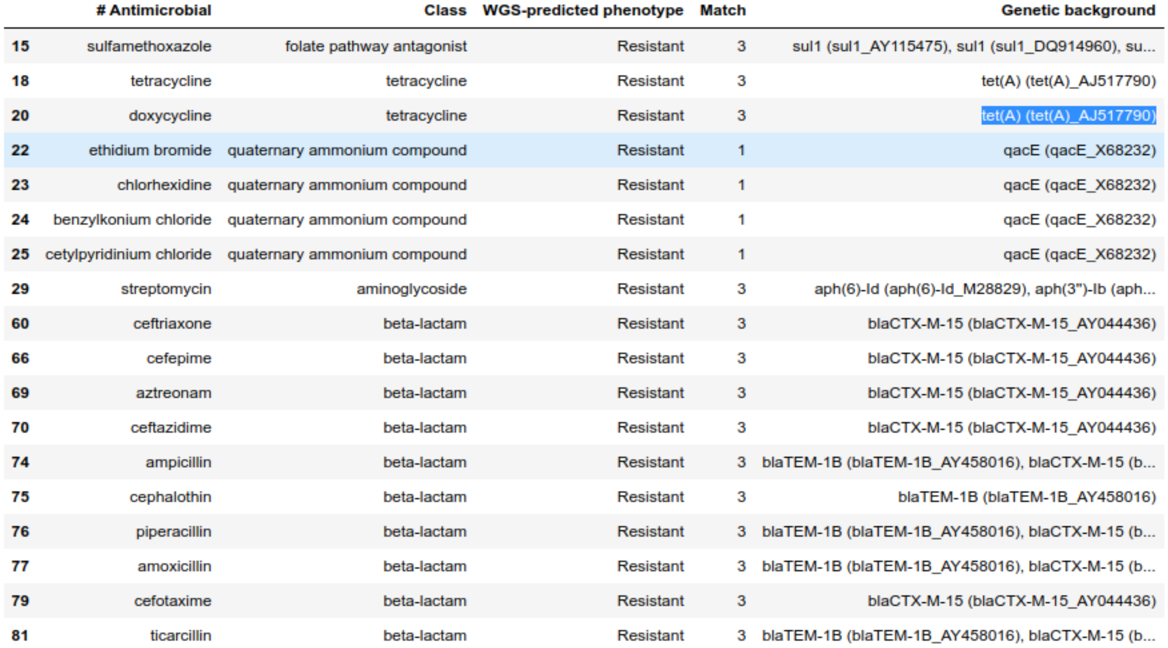
stxA 959 nt locus LHMPOLMA_35 location complement(877..1836)

stxB 269 nt locus LHMPOLMA_35 location complement(596..865)

Tracing the source of toxin genes in E. coli X

The most proteins were phage-related

Antibiotic resistance detection



	# Antimicrobial	Class	WGS-predicted phenotype	Match	Genetic background
15	sulfamethoxazole	folate pathway antagonist	Resistant	3	sul1 (sul1_AY115475), sul1 (sul1_DQ914960), su...
18	tetracycline	tetracycline	Resistant	3	tet(A) (tet(A)_AJ517790)
20	doxycycline	tetracycline	Resistant	3	tet(A) (tet(A)_AJ517790)
22	ethidium bromide	quaternary ammonium compound	Resistant	1	qacE (qacE_X68232)
23	chlorhexidine	quaternary ammonium compound	Resistant	1	qacE (qacE_X68232)
24	benzylkonium chloride	quaternary ammonium compound	Resistant	1	qacE (qacE_X68232)
25	cetylpyridinium chloride	quaternary ammonium compound	Resistant	1	qacE (qacE_X68232)
29	streptomycin	aminoglycoside	Resistant	3	aph(6)-Id (aph(6)-Id_M28829), aph(3")-Ib (aph...
60	ceftriaxone	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
66	cefepime	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
69	aztreonam	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
70	ceftazidime	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
74	ampicillin	beta-lactam	Resistant	3	blaTEM-1B (blaTEM-1B_AY458016), blaCTX-M-15 (b...
75	cephalothin	beta-lactam	Resistant	3	blaTEM-1B (blaTEM-1B_AY458016)
76	piperacillin	beta-lactam	Resistant	3	blaTEM-1B (blaTEM-1B_AY458016), blaCTX-M-15 (b...
77	amoxicillin	beta-lactam	Resistant	3	blaTEM-1B (blaTEM-1B_AY458016), blaCTX-M-15 (b...
79	cefotaxime	beta-lactam	Resistant	3	blaCTX-M-15 (blaCTX-M-15_AY044436)
81	ticarcillin	beta-lactam	Resistant	3	blaTEM-1B (blaTEM-1B_AY458016), blaCTX-M-15 (b...